

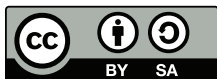
Closing Information Gaps with Need-driven Knowledge Sharing

Zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation
von

Dipl.-Wirtsch.-Inf. Hans-Jörg Happel

Tag der mündlichen Prüfung: 19. Dezember 2017
Referent: Prof. Dr. Rudi Studer
Korreferent: Prof. Dr. Alexander Mädche



This document is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0):
<https://creativecommons.org/licenses/by-sa/4.0/deed.en>

Abstract

Closing Information Gaps with Need-driven Knowledge Sharing

Knowledge management systems for asynchronous knowledge sharing – such as Intranets, Wikis, or file shares – often suffer from a lack of contributions. This is mainly because information providers are *decoupled* from information seekers, and thus have limited awareness about their actual information needs. Therefore, the questions *which knowledge is worth sharing* and *how to motivate people to share knowledge*, are core issues of knowledge management.

To this end, we describe a novel approach called *need-driven knowledge sharing* (NKS), which consists of three elements. The first deals with indicators of information need – especially search queries – which are aggregated in order to derive continuous forecasts of *organizational information needs* (OIN). By comparing with private and shared information spaces, *organizational information gaps* (OIG) are derived to identify *missing information*. These gaps can be made transparent using so called *mediation services* and *mediation spaces*, which help to create awareness for organizational information needs and to guide knowledge sharing. The realization of NKS is illustrated by three tools, which are all based on established knowledge management systems.

Inverse Search is a tool that identifies documents in the private information space of information providers, which may help closing organizational information gaps if moved to a shared information space. *Woogle* extends Wikis with features that identify and prioritize incomplete or missing information based on other users' information needs. Similarly, *Semantic Need* is an extension to Semantic MediaWiki that guides the creation of semantic data by analyzing information needs expressed through structured queries.

The implementation and evaluation of all three tools shows, that need-driven knowledge sharing is technically feasible and can be a helpful extension to knowledge management practices. The concepts of *mediation services* and *mediation spaces* provide a framework to analyze and extend other tools with respect to NKS. Finally, our approach may also spark improvements of Internet-scale services and infrastructures, such as the Wikipedia or the Semantic Web.

Informationslücken schließen durch bedarfsgetriebenen Wissensaustausch

Systeme zum asynchronen Wissensaustausch – wie Intranets, Wikis oder Dateiserver – leiden häufig unter mangelnden Nutzerbeiträgen. Ein Hauptgrund dafür ist, dass Informationsanbieter von Informationssuchenden *entkoppelt*, und deshalb nur wenig über deren Informationsbedarf gewahr sind. Zentrale Fragen des Wissensmanagements sind daher, *welches Wissen besonders wertvoll ist* und *mit welchen Mitteln Wissensträger dazu motiviert werden können, es zu teilen*.

Diese Arbeit entwirft dazu den Ansatz des *bedarfsgetriebenen Wissensaustauschs* (NKS), der aus drei Elementen besteht. Zunächst werden dabei Indikatoren für den Informationsbedarf erhoben – insbesondere Suchanfragen – über deren Aggregation eine fortlaufende Prognose des *organisationalen Informationsbedarfs* (OIN) abgeleitet wird. Durch den Abgleich mit vorhandenen Informationen in persönlichen und geteilten Informationsräumen werden daraus *organisationale Informationslücken* (OIG) ermittelt, die auf *fehlende Informationen* hindeuten. Diese Lücken werden mit Hilfe so genannter *Mediationsdienste* und *Mediationsräume* transparent gemacht. Diese helfen Aufmerksamkeit für organisationale Informationsbedürfnisse zu schaffen und den Wissensaustausch zu steuern. Die konkrete Umsetzung von NKS wird durch drei unterschiedliche Anwendungen illustriert, die allesamt auf bewährten Wissensmanagementsystemen aufbauen.

Bei der *Inversen Suche* handelt es sich um ein Werkzeug das Wissensträgern vorschlägt Dokumente aus ihrem persönlichen Informationsraum zu teilen, um damit organisationale Informationslücken zu schließen. *Woogle* erweitert herkömmliche Wiki-Systeme um Steuerungsinstrumente zur Erkennung und Priorisierung fehlender Informationen, so dass die Weiterentwicklung der Wiki-Inhalte nachfrageorientiert gestaltet werden kann. Auf ähnliche Weise steuert *Semantic Need*, eine Erweiterung für Semantic MediaWiki, die Erfassung von strukturierten, semantischen Daten basierend auf Informationsbedarf der in Form strukturierter Anfragen vorliegt.

Die Umsetzung und Evaluation der drei Werkzeuge zeigt, dass bedarfsgetriebener Wissensaustausch technisch realisierbar ist und eine wichtige Ergänzung für das Wissensmanagement sein kann. Darüber hinaus bietet das Konzept der *Mediationsdienste* und *Mediationsräume* einen Rahmen für die Analyse und Gestaltung von Werkzeugen gemäß der NKS-Prinzipien. Schließlich liefert der hier vorstellte Ansatz auch Impulse für die Weiterentwicklung von Internetdiensten und -Infrastrukturen wie der Wikipedia oder dem Semantic Web.

Acknowledgements

An endeavor of this kind is only possible when standing on the shoulders of giants. Besides the academic community at large, I owe gratitude to many people providing advice, feedback, support, and friendship throughout the time I was writing this thesis.

My first *thank you* goes to Rudi Studer for supervising this thesis and for never giving up hope. I am also grateful to Rudi and Andreas Abecker for providing an inspiring research environment and serving as role models for team leadership. Additional thanks go to Alexander Mädche for serving as second reviewer.

Numerous people provided feedback on individual chapters of this thesis, namely Andreas Abecker, Robin Aly, Mark Hefke, Olaf Grebner, Stephan Grimm, Thomas King, Athanasios Mazarakis, Walid Maalej, Asarnusch Rashid, Tim Romberg, Stefan Seedorf, and Max Völkel. Special thanks to all of you!

Further thanks go to my awesome colleagues at Information Processing (IPE), Software Engineering (SE) and other departments of FZI, as well as to all colleagues at Karlsruhe Institute of Technology. Particular thanks go to my long-term officemates Simone Braun and Stephan Grimm and to the “good soul” of IPE, Heike Döhmer. I am also grateful for the collaboration with friends from other Universities, in particular Walid Maalej and Stefan Seedorf.

Special thanks go to my student assistants throughout the time at FZI – Christoph Bier, Jordan Dukadinov, Paul Hübner, Thomas Hummel, Evgeny Matershev, Ben Romberg, Ingo Steinbauer, Christian Röhr, Sören Schlegel, Marius Treitz, Hristo Valev and Behar Veliqi – who contributed diligently to the implementation of the tools described in this thesis.

I am grateful to all taxpayers for enabling the state of Baden-Württemberg, the German Federal Government, and the European Union to fund large parts of my research. Work presented in this thesis has particularly been supported by the projects CollaBaWue, GlobaliSE, WAVES and TEAM.

Finally, I owe gratitude to my family and friends for their support and for bearing with me throughout this long journey.

“We shape our buildings, and afterwards our buildings shape us.”

—Winston Churchill

Contents

1. Introduction	1
1.1. Scope of the Thesis	1
1.1.1. Problem Statement	2
1.1.2. Research Questions	2
1.2. Content of the Thesis	3
1.2.1. Solution Approach	3
1.2.2. Structure of the Thesis	3
1.2.3. Contributions	4
1.2.4. Publications	5
2. Knowledge Management	7
2.1. Perspectives on Knowledge Management	8
2.1.1. Functional Perspective	8
2.1.2. Structural Perspective	11
2.2. Knowledge	12
2.2.1. Data, Information, and Knowledge	12
2.2.2. Implicit and Explicit Knowledge	14
2.2.3. Individual and Organizational Knowledge	15
2.2.4. Knowledge Maturing	17
2.3. Knowledge Management Strategies	19
2.3.1. Centralized vs. Decentralized	19
2.3.2. Codification vs. Personalization	21
2.3.3. Push vs. Pull	22
2.3.4. Personal vs. Organizational Knowledge Management	22
2.4. Knowledge Management and Organization	23
2.4.1. Knowledge Work	24
2.4.2. Distributed Work	25
2.4.3. Enterprise 2.0	27
2.5. Summary	29
3. Information Seeking and Knowledge Sharing	33
3.1. Information Seeking and Retrieval	33
3.1.1. Information Seeking Process	34
3.1.2. Information Retrieval Systems	39
3.2. Knowledge Sharing	43
3.2.1. Knowledge Sharing Process	44
3.2.2. Knowledge Sharing Systems	49
3.3. Knowledge Sharing as a Communication Process	52
3.3.1. Mediated Knowledge Sharing	52

Contents

3.3.2. Mediated Communication Approaches	53
3.4. Summary	57
4. Need-driven Knowledge Sharing	59
4.1. Definition	59
4.1.1. Goals	60
4.1.2. Scope	61
4.2. Assumptions	63
4.2.1. Organizational Information Needs	63
4.2.2. Organizational Information Gaps	68
4.3. Framework	72
4.3.1. Design Principles	72
4.3.2. Mediation Spaces and Services	73
4.3.3. Information Need Attributes	74
4.4. Keyword-based Instantiation of NKS	76
4.4.1. Storing Information Needs	76
4.4.2. Calculating Organizational Information Gaps	78
4.4.3. Calculating Organizational Information Needs	79
4.4.4. Reference Implementation (TeamWeaver)	80
4.5. Related Work	81
4.5.1. Information Seeking	82
4.5.2. Knowledge Sharing	83
4.5.3. Bridging Information Seeking and Knowledge Sharing	86
4.6. Summary	89
5. Inverse Search: Recommending Users to Share Documents	91
5.1. Problem	92
5.1.1. Information Seeking and Retrieval	92
5.1.2. Knowledge Sharing	93
5.1.3. Motivating Example	94
5.2. Design and Implementation	95
5.2.1. Approach	95
5.2.2. Architecture	97
5.2.3. Design	98
5.2.4. Implementation	100
5.3. Evaluation	101
5.3.1. Organizational Information Gap Analysis	102
5.3.2. Organizational Information Need Survey	104
5.4. Related Work	105
5.4.1. Peer-to-Peer Information Retrieval	105
5.4.2. Prospective Search	106
5.4.3. Enterprise File Sharing	106
5.4.4. User Interfaces for Access Control	107
5.5. Summary	107
6. Woogle: Guiding Contributions to Wikis	109
6.1. Problem	110
6.1.1. Information Seeking and Retrieval in Wikis	110

6.1.2.	Sharing Knowledge in Wikis	112
6.1.3.	Motivating Example	114
6.2.	Design and Implementation	115
6.2.1.	Approach	115
6.2.2.	Architecture	116
6.2.3.	Design	120
6.2.4.	Implementation	122
6.3.	Evaluation	123
6.3.1.	Qualitative Evaluation	124
6.3.2.	Online Field Experiment	125
6.4.	Related Work	128
6.4.1.	Social Search	128
6.4.2.	Collaborative Question Answering	129
6.4.3.	Guiding Contributions in Wikis	130
6.4.4.	Wiki Gardening	131
6.5.	Summary	131
7.	Semantic Need: Guiding Contributions to Semantic Wikis	133
7.1.	Problem	134
7.1.1.	Information Needs in the Semantic Web	134
7.1.2.	Information Provisioning in the Semantic Web	135
7.1.3.	Motivating Example	138
7.2.	Design and Implementation	139
7.2.1.	Approach	139
7.2.2.	Architecture	145
7.2.3.	Design	147
7.2.4.	Implementation	148
7.3.	Evaluation	152
7.3.1.	Public Semantic MediaWiki Analysis	152
7.3.2.	Semantic Need Survey	156
7.4.	Related Work	159
7.4.1.	Semantic Query Log Analysis	159
7.4.2.	Query Sharing and Reuse	160
7.4.3.	Knowledge Extraction from Queries	160
7.4.4.	Guiding and Motivating Semantic Content Creation	161
7.4.5.	Collaborative Knowledge Creation	162
7.4.6.	Sharing Semantic Content	163
7.4.7.	Information Completeness	163
7.4.8.	Why-not Provenance	166
7.4.9.	Valuation of Data	167
7.4.10.	Crowdsourced Information Provisioning	167
7.5.	Summary	168
8.	Summary	171
8.1.	Contributions	171
8.2.	Outlook	172
8.2.1.	Derivative Work	172
8.2.2.	Complementary Work	174

Contents

I. Appendix	177
A. Algorithms	179
B. Woogle Evaluation Participation Dialog	183
C. Semantic Need Public SMW Analysis	185
D. Semantic Need Survey Questionnaire	191
List of Figures	205
List of Tables	207
List of Theorems	209
Nomenclature	211
Bibliography	213
Index	251

In general, new technologies have minimized the technological separation of producer and consumer. It is a shift of some significance that the computer we read on is also the one we write on, whereas the book we read is very different from the manuscript we write.

— Brown and Duguid (1996)

1. Introduction

Documents allow to share knowledge across time and distance. We thus do not just live in an information society, but more precisely in a *document society* (Buckland, 2017), in which the Internet enables universal access to documents.

Search engines have the dominant role in mediating between authors and readers. They are *databases of intentions* (Battelle, 2005) of their users and provide strong incentives for the production of goods and knowledge. Therefore, we do as well live in a *search society*.¹

Knowledge sharing based on documents plays a similarly important role in organizations.² Explicit or *codified* knowledge is even considered as “the most important factor of production in the knowledge economy” (Zack, 1999). In contrast to the demand-driven strategy of *personalization*, which directly engages information seekers with information providers, the *codification* of knowledge traditionally implements a “push”-oriented, centralized distribution of stable and standardized knowledge in organizations.

Approaches for decentralized sharing of codified knowledge based on file shares, intranets, Wikis, or generally knowledge management systems (KMS³), often fail due to a lack of contributions, because information providers have limited resources and not much awareness about the needs of information seekers. Therefore, the questions *which knowledge is worth sharing* and *how to motivate people to share knowledge*, are considered to be major issues in the discipline of knowledge management (King et al., 2002).

1.1. Scope of the Thesis

Literature typically distinguishes between *data* (facts without interpretation), *information* and *knowledge* (“actionable” information) (Davenport and Prusak, 1998, p. 1ff). While knowledge in the strict sense is bound to individuals and cannot be “exchanged” technically, many authors use terms like *knowledge sharing* when they actually mean something like “exchanging information for the purpose of sharing knowledge”.⁴ While our primary emphasis lies on textually codified information, we⁵ will use the expressions *information exchange*,

¹See, e.g., Hillis et al. (2013); König and Rasch (2014); Halavais (2017)

²See, e.g., Hertzum (1999); Hicks et al. (2008)

³See, e.g., Maier (2007)

⁴See also Section 2.2.1 Also note, that in the strict interpretation, the term “KMS” does not make sense at all

⁵We is used throughout the thesis to honor the contribution of prior research and fellow collaborators to this work. Nevertheless, the core contributions of this thesis are original work of its author.

1. Introduction

information sharing and *knowledge sharing* synonymously.

We are furthermore interested in sharing with groups of a certain size, involving project teams, organizations or communities, but not single individuals or the general public (“world”). Our points of focus are summarized in Table 1.1.

Object	Data		<i>Information</i>		Knowledge
Communication mode	<i>Asynchronous</i>		Synchronous		
Subject	Other individuals	<i>(Project) team</i>	<i>Organization</i>	<i>Community</i>	World

Table 1.1.: Categorization of Information Sharing (Scope of the Thesis is *highlighted*)

1.1.1. Problem Statement

While IT provides a technological basis to allow for knowledge sharing, it does not properly solve the problem of deciding *which* knowledge to share. Especially in enterprise settings, where *resources are limited*, users are thus not contributing adequately to knowledge sharing systems. Due to such non-participation of information providers, many KMS projects fail.⁶

On the other hand, several studies indicate that users are actually *willing to share* information with others. However, various barriers such as effort, privacy concerns, and a limited communication bandwidth prohibit this.

Despite of several attempts of knowledge management research, this *knowledge sharing dilemma* is still not solved. Authors have shown that organizational culture has a large impact on knowledge sharing behavior and collaborative tools, allowing for incremental contributions have achieved a broader participation. However, a systematic framework analyzing knowledge sharing processes in asynchronous collaboration settings is yet missing.

1.1.2. Research Questions

Based on the problem description, our general *goal* is to improve and foster the asynchronous sharing of information via documents and content in organizational settings. As means to this end, we consider two important questions which are not yet adequately solved by existing research:

RQ 1 : How can we determine *what* knowledge should be shared in order to *maximize the benefit* for the group or organization?

RQ 2 : How can we *foster the creation and sharing* of this knowledge?

⁶See, e.g., Orlikowski (1992); Kankanhalli et al. (2005)

While RQ 1 has an *analytic and descriptive* emphasis, RQ 2 addresses a *design problem* by aiming to improve existing knowledge management tools and practices. Actual steps to tackle these questions are described in the following.

1.2. Content of the Thesis

This section discusses which scientific methods were applied to answer the research questions, and how the presentation of this thesis is structured. Finally, the major contributions of the thesis will be summarized.

1.2.1. Solution Approach

As explained while presenting the research questions, this thesis consists of an analytic and a constructive part. As for the analytic part, the following methods are applied to derive a detailed view of the current state of practice:

- Extensive literature review concerning KM approaches and practices
- Qualitative user studies
- Quantitative user surveys

As for the design part, further steps were pursued:

- Eliciting user requirements
- Extending KM approaches and models to accommodate for improvements
- Designing and implementing tool and improvements to existing tools
- Qualitative and quantitative evaluation of the implemented tools
- Discussion and reflection of achieved results

1.2.2. Structure of the Thesis

Figure 1.1 illustrates the structure of the thesis. In the following Chapter 2, we describe relevant work from the larger area of *knowledge management*. In Chapter 3, we have a deeper look into the fields of *information seeking and retrieval* (IS&R) as well as *knowledge sharing*, and especially discuss their interrelations. Based on that, the concept of *Need-driven Knowledge Sharing* (NKS) is introduced and described in Chapter 4. We also present the *TeamWeaver* platform as a technical basis for realizing applications based on NKS principles.

The subsequent chapters describe three of such applications, their underlying idea, implementation, and evaluation results. In Chapter 5, *Inverse Search* is presented, which recommends private documents for sharing to users. In the following chapter, *Woogle* will be presented as a tool combining enterprise search with Wiki-based knowledge capturing. Both, *Inverse Search* and *Woogle* deal with classical keyword-based searches and textual documents,

1. Introduction

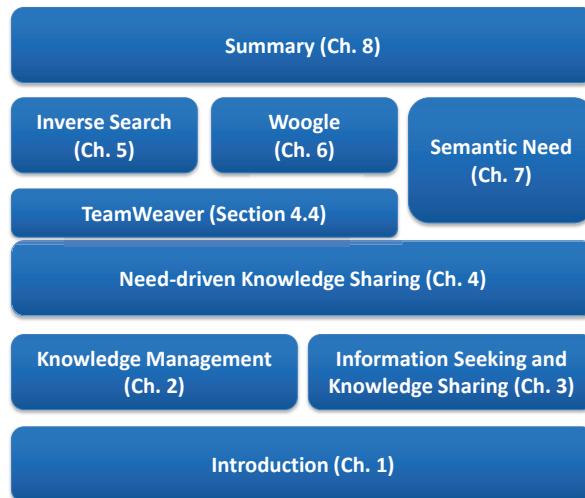


Figure 1.1.: Structure of the Thesis

based on the TeamWeaver platform. The third prototype, presented in Chapter 7, is based on *Semantic Web* technologies. Our tool, called *Semantic Need*, analyzes structured queries in the Semantic MediaWiki software to guide semantic annotations. Chapter 8 finally summarizes this thesis and presents future research opportunities.

1.2.3. Contributions

This thesis contains four major contributions, which also map to its core chapters. First, we introduce the novel concept of *Need-driven Knowledge Sharing* (NKS), which establishes a “missing link” between the so far mostly unconnected areas of information seeking and retrieval (IS&R) and knowledge management (KM). In particular, we propose to analyze and aggregate users’ information needs as, e.g., expressed by queries, and to leverage *gaps* between private and shared information spaces to provide guidance for information providers. The notions of *mediation spaces* and *mediation services* are introduced to structure tool support for knowledge sharing and can also be used to analyze and improve other KMS with respect to NKS.

NKS is implemented and validated by using three different tool implementations, each providing unique functionality within its particular application domain. *Inverse Search* is, to our knowledge, the first tool which systematically guides users in sharing their private documents with others. *Woogle* helps users to recognize which information is missing in a Wiki and to prioritize, how important this information is for the overall group of users. The *Semantic Need* extension transfers this idea to semantic knowledge bases, guiding metadata annotations based on structured queries. It is thus one of the first approaches addressing the “demand side” (Mika et al., 2009) of the Semantic Web.

The thesis contains interdisciplinary and multi-methodological work, bridging several different research areas as well as deriving results with practical relevance from theory and empirical evidence. It is motivated by distributed work settings, as studied in the field of *computer-supported cooperative work* (CSCW), and its intersection with *software engineering* research.⁷ Based on that, we deal with concepts from the areas of *information retrieval*, *knowledge management*, and *semantic technologies*. In each of these areas and communities, we have published scientific papers related to this thesis.

To combine and develop insights from these domains, multiple scientific methods were applied. Starting with an extensive *literature review and analysis*, *exploratory empirical studies* have been carried out. Based on their results and corresponding *user studies*, we designed and realized various *tool implementations*. These were evaluated with *user surveys* and *field experiments*.

While rooted in an interdisciplinary scientific context, an important additional goal of our work was to produce usable tools that can have positive impact in real-world settings. Accordingly, the *TeamWeaver* platform⁸, *Woogle*⁹, and *Semantic Need*¹⁰ have been published as Open Source software, that sparked interest by several organizations. Accordingly, results have been presented at various industry-oriented events throughout the duration of the thesis.

1.2.4. Publications

This section lists core publications which are underlying the work described in corresponding chapters.

Need-driven Knowledge Sharing and Inverse Search (Chapter 4 and 5)

- Happel, H.-J., Stojanovic, L., and Stojanovic, N. Fostering knowledge sharing by inverse search. In K-CAP '07: Proceedings of the 4th international conference on Knowledge capture (2007), ACM, pp. 181-182
- Happel, H.-J., and Stojanovic, L. Analyzing organizational information gaps. In Proceedings of I-KNOW '08: 8th international conference on knowledge management and knowledge technologies (2008), K. Tochtermann, H. Maurer, F. Kappe, and W. Haas, Eds., JUCS, pp. 28-36
- Happel, H.-J. Towards need-driven knowledge sharing in distributed teams. In Proceedings of I-KNOW '09: 9th international conference on knowledge management and knowledge technologies (2009), K. Tochtermann and H. Maurer, Eds., JUCS, pp. 128-139
- Happel, H.-J. Closing information gaps with inverse search. In Proceedings of PAKM 2008: 7th International Conference on Practical Aspects of Knowledge Management (2008), T. Yamaguchi, Ed., Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 74-85

⁷In particular collaborative software development (CSD)

⁸<http://www.teamweaver.org>

⁹<https://www.mediawiki.org/wiki/Extension:Woogle4MediaWiki>

¹⁰https://www.mediawiki.org/wiki/Extension:Semantic_Need

1. Introduction

Woogle (Chapter 6)

- Happel, H.-J. Woogle — on why and how to marry wikis with enterprise search. In *WM2009: 5th Conference on Professional Knowledge Management (2009)*, K. Hinkelmann and H. Wache, Eds., vol. 145 of LNI, GI, pp. 194-205
- Happel, H.-J. Social search and need-driven knowledge sharing in wikis with woogle. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration (2009)*, D. Riehle and A. Bruckman, Eds., ACM, pp. 1-10
- Happel, H.-J., and Mazarakis, A. Considering information providers in social search. In *2nd International Workshop on Collaborative Information Seeking (CIS '10) at CSCW 2010 (2010)*, pp. 1-5

Semantic Need (Chapter 7)

- Happel, H.-J. Growing the semantic web with inverse semantic search. In *1st Workshop on Incentives for the Semantic Web (INSEMTIVE) at ISWC 2008 (2008)*, pp. 1-12
- Happel, H.-J. Semantic need: Guiding metadata annotations by questions people ask. In *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, Eds., vol. 6496 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 321-336
- Happel, H.-J. Semantic need: An approach for guiding users contributing metadata to the semantic web. *Int. J. Knowledge Engineering and Data Mining* 1, 4 (2011), pp. 350-369

2. Knowledge Management

This chapter describes different concepts from the discipline of *knowledge management* (KM). The existing body of work is contrasted with recent organizational and technological developments, such as the emerging theme of *Enterprise 2.0*. Based on this analysis, we derive some major challenges for the next generation of KM tools.

The term *knowledge management* was coined in research literature during the late 1980's¹ and was subsequently adopted in practice.² However, the term and the discipline of KM did not appear out of nowhere, but emerged from related fields of study. Maier (2007, p. 45) identifies organizational learning³ and information management⁴ as the two “historical roots” of KM. Since then, KM has remained an *interdisciplinary* field of research, receiving inputs from – but also providing results for – various other scientific disciplines (see, e.g., Maier, 2007, p. 34f).

Accordingly, there exist various definitions of knowledge management, each with a different perspective or focus (Maier, 2007, p. 52). For the context of our work, we want to adapt the definition by Allan et al. (2004):

Definition 2.1 (Knowledge Management). *Planned and ongoing management of activities and processes for leveraging knowledge to enhance competitiveness through better use and creation of individual and collective knowledge resources.*

This definition stresses some important issues:

- First, KM is considered to be a *consciously planned and designed* effort, as opposed to merely spontaneous and chaotic practices.
- Second, KM efforts are *oriented towards goals* of the organization or group that practices KM.
- Third, KM operates at the *interface between individuals and a collective* in which these individuals collaborate and interact.

¹Maier (2007, p. 22) mentions even earlier works from the 1970's, but notes that KM “emerged again in the mid 80s in the context as it is still used today”.

²See, e.g., Hansen et al. (1999) or Maier (2007, p. 40ff)

³*Organizational learning* studies learning processes in organizations on an individual and organizational level – see, e.g., Argyris and Schön (1978), Duncan and Weiss (1979) or Argote (1999); see also Section 2.2.3.

⁴*Information management* is the discipline of managing information as an organizational resource for preparing decisions and actions – see, e.g., Maier (2007, p. 42f) or Krcmar (2004).

2. Knowledge Management

All three issues might be addressed very differently, depending on the actual setting in an organization. Plans for managing knowledge, organizational goals, and the role of individuals will vary when comparing a manufacturing company to a consultancy. As we will see in the remainder of this chapter, the process of defining goals and the relation between the individual and the collective are also influenced by technological and organizational innovations.

We will now present different *perspectives* on what KM is all about. Afterwards, we discuss the distinction between *data*, *information*, and *knowledge*. This is followed by a description of several KM *strategies* and highlighting the mutual influence between KM and *organizational issues*. At the end of the chapter, we reflect on the current state of KM and identify a number of major *challenges*.

2.1. Perspectives on Knowledge Management

Due to its interdisciplinary roots, KM can be viewed from different angles. We will introduce the well-established *functional perspective* of Probst et al. (2006) and a *structural perspective*, which will also serve as a framework for following subsections.

2.1.1. Functional Perspective

A frequently cited model to describe the discipline of KM are the *building blocks of knowledge management* by Probst et al. (2006).⁵ Its authors stress, that the model should serve as a mere basis for discussion and neither claim to describe reality in organizations, nor that it is intended to serve as a normative blueprint. Nevertheless, the model has become influential in both practical application and research as its number of citations shows. Since other authors describe similar models (e.g., Maier, 2007, p. 207ff), one can also assume a certain level of validity.

Probst et al. (2006) distinguish eight building blocks which are also depicted in Figure 2.1:

Knowledge Goals The definition of goals sets the starting point for all following KM activities (Probst et al., 2006, p. 37). Probst (1998) distinguishes *normative* goals (addressing corporate culture), *strategic* goals (describing a desirable future competence portfolio), and *operative* goals that translate the two prior categories into actions. While this describes a *top-down* process, Probst et al. (2006, p. 59) caution against a “control illusion” when setting KM goals.

Knowledge Identification The next building block seeks to raise transparency about existing knowledge within and outside the organization. According to Probst et al. (2006, p. 61ff), this can help to *make people aware of knowledge* by providing overviews (e.g., knowledge maps) or search

⁵See Probst (1998) and Probst et al. (1999) for English translations

2.1. Perspectives on Knowledge Management

engine access. When contrasted with the *knowledge goals*, transparency can also help to derive *knowledge gaps* of the organization.

Knowledge Acquisition While the term knowledge acquisition typically describes either the learning behavior of individual persons or knowledge-based systems,⁶ Probst et al. (2006, p. 61ff) take an organizational perspective. According to this, organizations can acquire external knowledge by means of cooperations or mergers with other companies, hiring human resources, involving stakeholders, or by purchasing knowledge products such as software or patents.

Knowledge Development If knowledge can not or shall not be acquired externally, the organization has to develop knowledge on its own. This can happen on both, individual and organizational level.⁷

Knowledge Distribution The core concern of knowledge distribution is to provide the existing knowledge *at the right time to the right persons*. The *selection* of the most relevant knowledge is a major challenge, since the *vast amount* of knowledge in an organization requires *efficient allocation* (Probst et al., 2006, p. 147f). Furthermore, individual and organizational *barriers* have to be addressed to let knowledge flow. Especially in distributed organizations (see Section 2.4.2) one can observe a lack of “*natural sharing contexts*” (Probst et al., 2006, p. 144).

Finally, two fundamental modes of knowledge distribution can be distinguished. In the *push* mode, knowledge is distributed top-down from a central position (Probst et al., 2006, p. 151). The *pull* mode instead means that the knowledge user makes a directed knowledge request based on her actual knowledge need.⁸

Knowledge Use The usage and application of knowledge can be considered the *major success criterion*. While usage naturally occurs at the end of the knowledge management cycle, Probst et al. (2006, p. 177) argue that KM should actually work the other way round. Knowledge identification, development and distribution *should take into account the actual needs* of knowledge users, as supported by the *pull* mode of knowledge distribution. The *push* mode instead *decouples* upstream knowledge processes from its usage (Probst et al., 2006, p. 179).

Knowledge Preservation Besides usage in daily operations, knowledge acquired and created needs to be preserved for *future use*. However, this is a costly task, including the *challenge to anticipate* which knowledge might be of value at a later time: “The guiding rule should be to preserve only information that will be usable for a third party in the future” (Probst, 1998).

Knowledge Measurement Finally, proper management of knowledge requires controlling if the previously defined knowledge goals are met. Therefore, organizations need to find suitable measurements. Due to the elusive

⁶See, e.g., Boose and Gaines (1989)

⁷This is sometimes also called *knowledge creation*; see, e.g., Nonaka (1994)

⁸*Pull* and *push* will also be discussed in Section 2.3.3

2. Knowledge Management

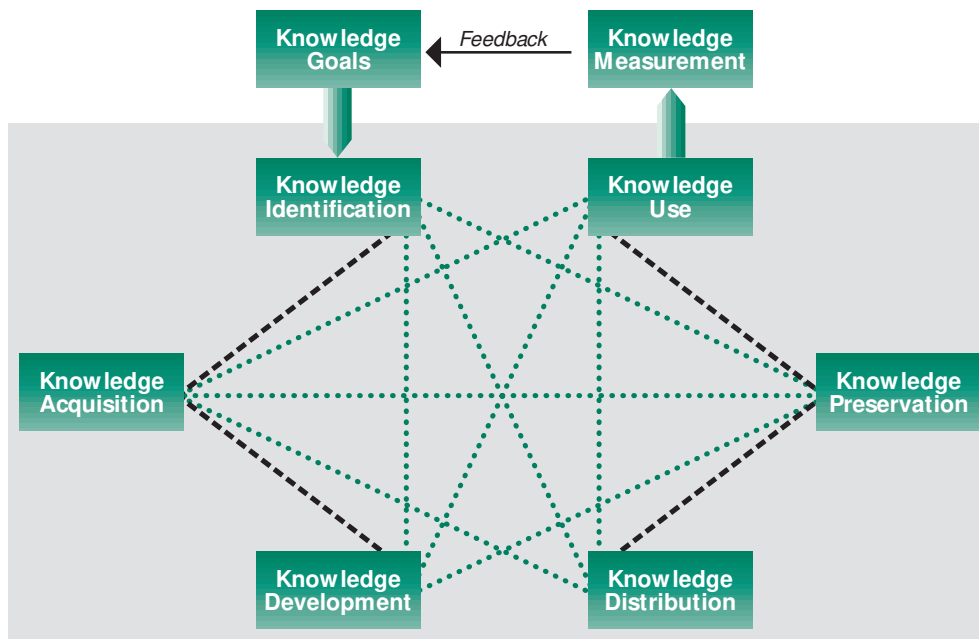


Figure 2.1.: Building Blocks of Knowledge Management (Probst et al., 2006)

character of the subject, this is however a non-trivial effort. Inspired by management accounting, practical measures such as the intellectual capital statement (“Wissensbilanz” in German) have been developed (Edvinsson and Malone, 1997). However, these instruments have a rather high-level and long-term perspective and can not be easily applied to the other building blocks outlined above.

Summarizing, the presented model provides a concise overview of the management tasks involved in KM. The individual building blocks can serve as a good checklist when implementing KM in an organization. However, one has to take into account that the model was primarily developed in cooperation with larger companies (Probst et al., 2006, p. 266). In particular, the book mostly advocates a management-driven *top-down* approach towards KM, although the authors themselves caution against this at some points (e.g., Probst et al., 2006, p. 177).

Furthermore, the model as well as the book have *not* undergone significant changes since its first edition in 1995, although there has been considerable change in the business environment and KM technologies.⁹

Finally, the building blocks have a strong focus on management tasks, independent of the particular environment in which these tasks are executed. Actual subjects and objects of KM, such as the organization, the knowledge

⁹In the recent edition, two isolated pages covering Wikis, Blogs and Microblogging have been added (Probst et al., 2006, p. 243f).

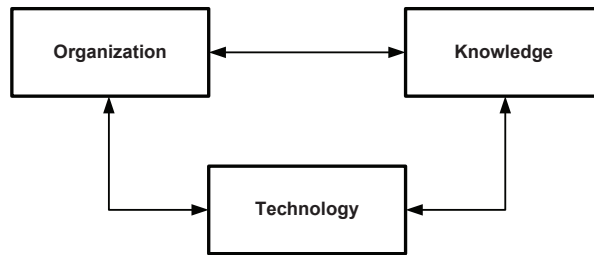


Figure 2.2.: Relationship of Organization, Technology and Knowledge

managed, or actual technologies are often mentioned in an anecdotal fashion only. This makes it difficult to derive actions for concrete practical settings.

2.1.2. Structural Perspective

After discussing the management aspects of KM, we will now take a complementary, structural perspective on KM. Therefore we consider the following elements (as depicted in Figure 2.2):

- The *organization* is the main subject of investigation in KM.
- *Knowledge*, accordingly, is the major object of investigation in KM.
- *Technology* denotes all means employed by the organization to manage knowledge.¹⁰

We propose that there exist mutual relationships between each of these elements:

- The interrelationship between *organization* and *technology* has been well-studied in the field of information systems research.¹¹ *Technological determinism* argues that technology has a strong influence on the organization, while *organizational determinism* claims that organizations adapt and shape technology according to their needs. Both perspectives can also be found in the case of KM. While some authors argue that technology should be chosen according to match the organization (Hansen et al., 1999), others claim that technology enables and drives organizational change (e.g., McAfee, 2006). While either perspective might dominate in a concrete situation, one can conclude that there exists a mutual dependency.¹²
- Similarly, an *organization* shapes the *knowledge* that it produces: "the distribution of knowledge in an organization, or in society as a whole, reflects the social division of labor." (Brown and Duguid, 1998).¹³

¹⁰Note that this not just includes IT systems but also KM processes and methods.

¹¹See, e.g., Markus and Robey (1988); Orlikowski and Robey (1991); Orlikowski (1992, 2000); Majchrzak et al. (2000) or Lehner (2000, p. 47)

¹²This mutual dependency is also called "Emergent perspective" by Markus and Robey (1988)

¹³See also Allen (1984), Henderson and Clark (1990), as well as Section 2.2.3 and 2.4.2

2. Knowledge Management

Nonaka and Konno (1998) describe the example of a company which deliberately designed organizational structures for knowledge creation. On the other hand, knowledge and competencies existing in an organization can have an influence on designing organizational structures, as inherent to the concept of functional organization.¹⁴

- The last relationship indicates that different kinds of *knowledge* require different *technologies*, while technologies might as well influence what kind of knowledge is created and shared.¹⁵

Since organization, technology, and knowledge all offer design choices and are subject to management decisions, their balance influences the overall success of KM. In the following sections, we will discuss these design choices more deeply. We begin by analyzing different properties of *knowledge*, continuing with *technology*, and *organization*.

2.2. Knowledge

This section aims to convey a deeper understanding what *knowledge* actually means in the context of KM. We will therefore discuss the common distinction between data, information, and knowledge and present properties to distinguish different qualities of knowledge.

2.2.1. Data, Information, and Knowledge

Due to its fundamental character, the term *knowledge* has various connotations across different scientific disciplines such as philosophy, sociology or computer science (see, e.g., Maier, 2007, p. 60ff). In the broader field of information systems¹⁶ however, a distinction between *data*, *information*, and *knowledge* is commonly made. These three terms are typically described as:

Data Facts (numbers, symbols) without context and interpretation (Davenport and Prusak, 1998; Allan et al., 2004, p. 2f)

Information Meaningful interpretation of data in a context by the receiver (Davenport and Prusak, 1998, p. 3; Probst et al., 2006, p. 16)

Knowledge A set of data and information [...] which can be used to improve the capacity to act and support decision making (Allan et al., 2004)

Often, these terms are treated as a hierarchy of somehow increasing quality (Machlup and Mansfield, 1983; Case, 2002, p. 64).¹⁷ On the other hand, many

¹⁴See, e.g., Braun and Beckert (1992). Functional organization may occur at large (e.g., at a department level) and on a level of specialized tasks.

¹⁵See Hansen et al. (1999) and in particular Section 2.3.2.

¹⁶This can be considered as a subset of computer science, particular involving scientific communities such as information seeking and retrieval (see Section 3.1) and knowledge management. Different fields of computer science, such as artificial intelligence and knowledge representation, might have different definitions of *knowledge*.

¹⁷Probst et al. (2006, p. 18) argue that knowledge is aggregated from data.

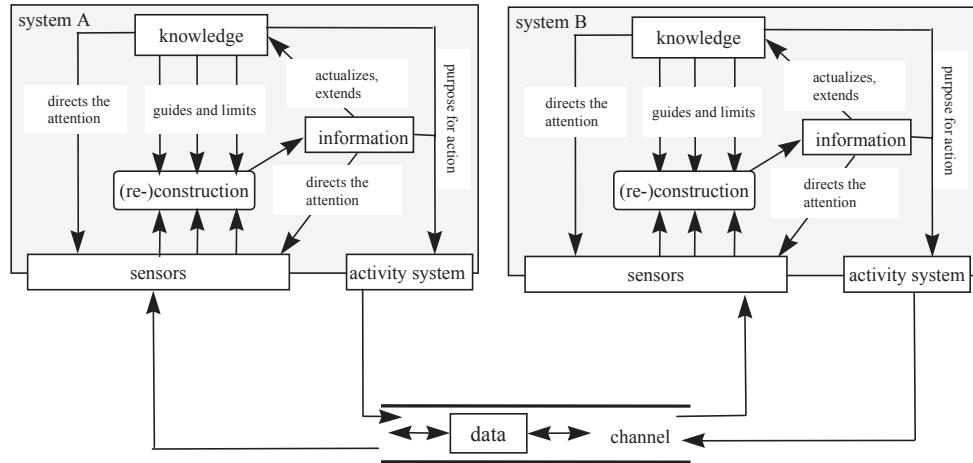


Figure 2.3.: Relationship of Data, Information, and Knowledge (Maier, 2007, p. 71)

studies in the areas of information behavior and knowledge management do not draw clear distinctions between *information* and *knowledge* and often use both terms interchangeably (Nonaka, 1994; Case, 2002, p. 65).

An important distinction is however made concerning the physicality of knowledge and information. As Case (2002, p. 65) puts it, “knowledge [...] is strictly a phenomenon of the human mind, whereas data and information are often represented by tangible, physical objects”. Accordingly, Machlup and Mansfield (1983, p. 644) argue that “information is acquired by being told, whereas knowledge can be acquired by thinking”. Maier (2007, p. 71) even goes a step further, by claiming that only data can be communicated. As depicted in Figure 2.3, data is then interpreted to reconstruct information, which in turn actualizes or extends the knowledge of the “receiving system”.

This perspective also partially explains the common lack of a distinction between information and knowledge, since “Knowledge and information are [...] usually not the same, except that ‘information in the sense of that which is being told may be the same as knowledge in the sense of that which is known, but need not be the same’” (Case, 2002, p. 64).¹⁸ Similarly, Maier (2007, p. 71), explains the notion of “knowledge transfer”, which would normally contradict the non-physical nature of knowledge. He argues however, that the term assumes that “the sender is quite certain that the receiver will interpret the data accordingly, (re-)construct the knowledge and use it to actualize the receiver’s knowledge in a way that the *sender intends*”.

Thus, knowledge and information can be regarded two sides of the same coin: while the “receiver” is seeking information to extend her knowledge, the “sender” shares knowledge which is communicated through information,

¹⁸Case is citing Machlup and Mansfield (1983, p. 644) in the last part of the quote.

2. Knowledge Management

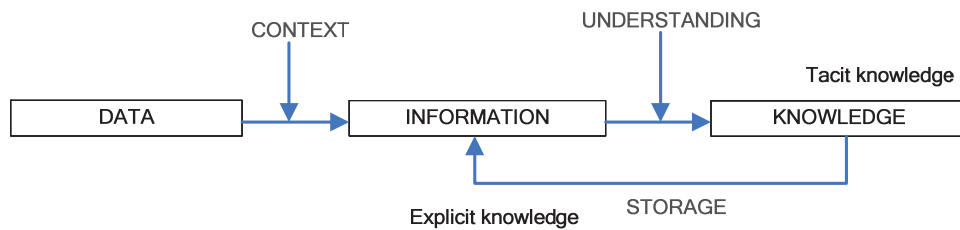


Figure 2.4.: Relationship of Data, Information, and Knowledge (Tang et al., 2006)

assuming that the receiver will be able to re-construct the knowledge.¹⁹ In any case, while the information transmitted is the same for sender and receiver, the “knowledge” derived might differ between what was assumed by the sender and what was re-constructed by the receiver.

2.2.2. Implicit and Explicit Knowledge

Tightly related to the physical nature of information and knowledge, Polanyi (1967) introduces a seminal distinction between implicit (originally referred to as “tacit”) and explicit (also referred to as “codified”) knowledge. *Implicit* knowledge can not easily be expressed and communicated by a person. Typically, it denotes knowledge based on experience or intuition, such as how to ride a bike. *Explicit* knowledge denotes knowledge which can be articulated and communicated between persons, and that can accordingly also be formally captured.

The relation of this distinction to the notions of data, information, and knowledge is depicted in Figure 2.4. Similar to the argumentation in the previous section, *explicit knowledge* is always also *information*. On the other hand, *information* is only *explicit knowledge*, if one can assume that it can be appropriately reconstructed by a receiver.

This “intersubjective” nature of knowledge was adopted by Nonaka (1994), who investigated the dynamic interrelationship between implicit and explicit knowledge more deeply. He regards knowledge creation as a process which starts with individuals. Based on social interaction between the individuals in an organization, knowledge is then “transformed and legitimized”. Nonaka (1994) describes a “spiral model of knowledge creation” which distinguishes four modes of knowledge conversion (see also Figure 2.5):

1. From tacit knowledge to tacit knowledge (socialization) denotes the transfer of knowledge without explicit communication or language. A typical example is learning by observation as performed in craftsmanship.

¹⁹The underlying processes of *information seeking* and *knowledge sharing* will be discussed in more detail in Section 3.1.1 resp. 3.2.1

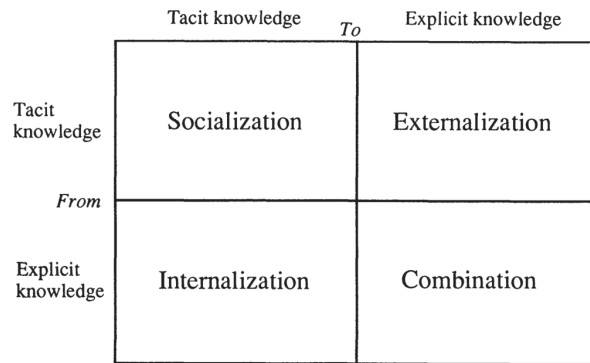


Figure 2.5.: Modes of Knowledge Creation according to Nonaka (1994)

2. From explicit knowledge to explicit knowledge (combination) covers processes to combine different bodies of explicit knowledge (e.g., meetings or phone calls).
3. From tacit knowledge to explicit knowledge (externalization) requires patterns of conversion which iteratively shape explicit knowledge out of tacit knowledge.
4. From explicit knowledge to tacit knowledge (internalization) describes the reverse process of understanding or learning from explicit knowledge.

The distinction between implicit and explicit knowledge stresses the social and intersubjective nature of knowledge. The transformation from information to knowledge does not merely occur in the context of a single, unidirectional act of communication between a sender and a receiver but also involves cycles of conversation which result in the actual construction of *mutual knowledge*. We will now further investigate this mutual or “organizational” nature of knowledge.

2.2.3. Individual and Organizational Knowledge

The spiral model of knowledge implies a distinction between *individual* and *organizational* knowledge. According to Nonaka (1994), knowledge can only be created by individuals and is thus ultimately bound to persons. By means of communication however, knowledge can spread and legitimize among groups of several individuals and thus become what is called *organizational, shared* or *collective* knowledge.²⁰

Wiegand (1998)²¹ introduces the concept of “organizational knowledge communities”²² to describe groups which share knowledge that is “not easily un-

²⁰Maier (2007, p. 530f) describes a similar “knowledge life cycle” which consists of four types of knowledge: individual knowledge, inter-subjective knowledge, institutionalized knowledge and knowledge in use.

²¹See also Lehner (2000, p. 117)

²²Translation by author; original German term: “organisational Wissensgemeinschaften”

2. Knowledge Management

derstood by other members of the organization”. They can foster legitimation and consensus building about knowledge and thus also help to make it explicit and easy to communicate (Lehner, 2000, p. 147). Organizational knowledge communities may emerge in the context of organizational structures, due to proximity, and especially due to specialization of work.²³

From a management perspective, it makes sense that large parts of organizational knowledge are not shared by the whole organization, but only by certain groups. This is mainly due to considerations of an efficient allocation of knowledge which allows to leverage specialization effects due to a division of labor by employing people with specialized and unique knowledge.²⁴ Also, individuals are known to have a *limited information-processing capacity*, which does not allow to store all knowledge relevant for an organization.

However, certain organizational knowledge might become isolated among groups, which is why concepts such as *transactive memory systems* (TMS) provide a theory for *knowledge in organizations* which connects widespread organizational knowledge. A TMS consists of “a set of individual memory systems in combination with the communication that takes place between individuals” (Wegner, 1986). The idea is, that each individual is responsible for a specialized set of knowledge. Thus, knowledge is distributed across individuals, resulting in a “specialized division of (cognitive) labor” (Hollingshead et al., 2002). By means of communication however, all group members have access to the individual memory systems of individuals, forming a TMS.²⁵ In the best case, TMS allow to manage the overall knowledge of a group with minimized burden for the individual. To achieve an optimal knowledge distribution within a group, TMS require very specialized tasks, clear responsibilities, and a low rate of changing knowledge.

Finally, we introduce the concept of the *organizational knowledge base* (also called *organizational memory*; Maier, 2007, p. 25f). According to Probst et al. (2006, p. 23), it consists of “individual and collective knowledge resources which an organization can use to complete its tasks”. The concept is closely related to *organizational learning*, which can be considered the dynamic process of changing and updating the organizational knowledge base (Probst et al., 2006, p. 23). Since the organizational knowledge base is defined to include individual and organizational knowledge, it includes both notions of *organizational knowledge* and *knowledge in organizations* as discussed in the previous paragraphs. In particular Pautzke (1989)²⁶ distinguishes several “horizontal layers” of the organizational knowledge base:

- Knowledge which is shared by all members of an organization
- Individual and collective knowledge which can be accessed by an organization

²³The latter case is similar to the concept of “Communities of Practice” (CoP) (Lave and Wenger, 1991)

²⁴See, e.g., Brown and Duguid (1998), Baldwin and Clark (2000, p. 5ff), Lehner (2000, p. 117) or Probst et al. (2006, p. 147)

²⁵See also Lehner (2000, p. 109)

²⁶See also Lehner (2000, p. 99)

- Individual and collective knowledge which can *not* be accessed by an organization
- Knowledge in the environment of the organization

The important distinction introduced by Pautzke (1989) is the fact that certain knowledge might exist, but is *not accessible* for an organization. While this might include private knowledge of individuals, which is not relevant to the organization, individuals might also have *organizationally relevant knowledge which is not shared* due to “*information or communication pathologies*” (Lehner, 2000, p. 101f). Furthermore, this stresses the *incompleteness* of the organizational knowledge base. Besides the issue of non-accessibility, it implies that knowledge might also *not yet exist*.

As Maier (2007, p. 205) states, organizational knowledge is in the core of KM activities: “KM envisions an organizational memory or organizational knowledge base into which the individual’s knowledge is supposed to be made explicit and which is the basis for (more or less unguided) knowledge transfer”. Lehner (2000, p. 103) notes, that many authors regard increasing the amount of knowledge which is “*explicit and shared by all*” as the main task of KM. However, we have described that it is not useful, and a potentially waste of resources, if *all* knowledge is shared within an organization.

Accordingly, besides fostering the *creation* and *sharing* of organizational knowledge, a core task of KM is to *decide* which knowledge is worth to be shared within an organization. As Probst (1998) puts it: “they should identify core areas of their organizational knowledge base and establish a pragmatic selection process for knowledge to be saved. The guiding rule should be to preserve only what will be usable for a *third party in the future*.”²⁷

2.2.4. Knowledge Maturing

As argued at the beginning of this section, data, information, and knowledge should not be regarded as a strict hierarchy. However, at the level of knowledge, there seems to occur an evolution from individual to organizational forms of knowledge, as we discussed in Section 2.2.3. While the spiral model of Nonaka (1994) provides an abstract model for this, we now describe the tightly related *knowledge maturing* process by Maier and Schmidt (2007).²⁸

Similar to the spiral model, the knowledge maturing process describes an iterative process, during which organizational knowledge slowly *emerges*. This process is separated into five phases, depicted in Figure 2.6, which were derived from case studies, mostly in the e-learning domain. Each phase has a set of typical information artifacts and information systems assigned.

The different phases of the knowledge maturing process are separated by transitions, during which knowledge is transformed to the next level.

²⁷See also “Knowledge Preservation” in Section 2.1.1.

²⁸See also Maier (2007, p. p290ff). The knowledge maturing process was also in the core of the EU integrated project *Mature* (<http://mature-ip.eu/>).

2. Knowledge Management

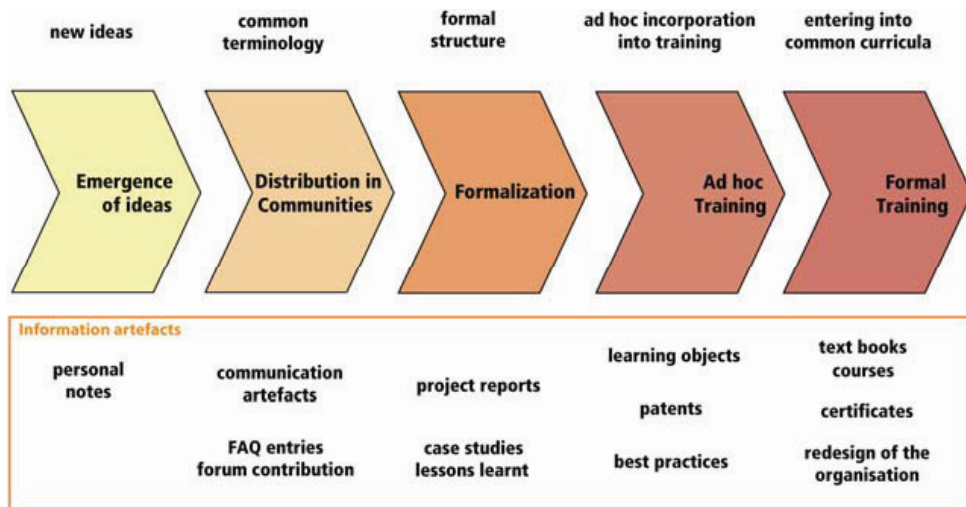


Figure 2.6.: Knowledge Maturing Process (Maier and Schmidt, 2007)

Maier and Schmidt (2007) argue, that *disruptions* between the phases might impede knowledge maturing. These disruptions can be due to different organizational entities or information systems involved at each stage.

The process starts with the transition from the *emergence of ideas*, in which individuals derive and start to discuss new knowledge, to those ideas' *distribution in communities*, in which this knowledge starts to spread within an organization. Barriers during this transitions are also tightly related to the gap between personal and organizational knowledge management.²⁹

Another barrier may exist between the following phases of *distribution in communities* and *formalization*. In the transition process, knowledge is externalized, and thus detached from its originator, as it is supposed to spread within the organization during the following phase. However, individuals might be reluctant to lose control about their initial ideas, which would disrupt the maturing process.³⁰ Maier and Schmidt (2007) suggest that an *integrated tool landscape*, which retains credit to the originator, might help to manage this transition more effectively.

One can also identify a *barrier* between the phases *formalization* and *ad-hoc training*. While learning in the former phase typically happens within organizational units, it is under the responsibility of the HR³¹ or training departments in the latter phase. This is often accompanied by a separation on the IT level, e.g., between document management and special learning management systems. From an individual perspective, learning in the formalization phase is considered an information seeking ("pull") activity, while it is designed as

²⁹See also Section 2.3.4 and in particular Tungare and Perez-Quinones, 2008

³⁰A detailed discussion on individual barriers will be given in Section 3.2.1.

³¹Human Resources

a “push”³² of knowledge in the *ad-hoc training* phase (Maier and Schmidt, 2007).³³

The knowledge maturing process provides a fine-grained model of characterizing the maturity of knowledge and its related information artifacts. It implies that the nature of knowledge is tightly related to its acceptance within the organization. KM measures and tools should be aware of these stages and help to overcome maturing barriers where necessary.

2.3. Knowledge Management Strategies

After discussing several dimensions of *knowledge*, we will now analyze design decisions which organizations can make when implementing KM.

2.3.1. Centralized vs. Decentralized

As depicted in Figure 2.2, KM has tight relations to organizational structures which shape a large amount of communication and thus drive knowledge creation and exchange (Henderson and Clark, 1990). Also, parts of the formal organization, such as guidelines and policies, embody organizational knowledge (Davenport and Prusak, 1998, p. 5).

In organizational design, one can distinguish the archetypes of centralized (or hierarchical) and decentralized (or modular) organizational forms.³⁴ Similarly, scholars note that KM staff and responsibilities can be organized in a centralized or in a decentralized fashion (e.g., Nonaka, 1994).

In the first case, KM is often directed by a separate organizational unit (Maier, 2007, p. 160ff) and KM goals are defined top-down. As Maier (2007, p. 162) stresses, centralized KM often has a decentral component as well: “The role of a centralized unit is only a coordinating and administrating one. Generally, the most important KM-related instruments have to be applied as close to where the knowledge is needed as possible, which is directly in the functional departments or projects.”

A particular risk of centralized KM goals is the “*decoupling of knowledge creation*” (Probst et al., 2006, p. 116). This is the case, if certain KM activities, such as the creation of documentation, have become institutionalized without regard for the actual current needs of knowledge users (see also Maier, 2007, p. 610). Probst et al. (2006, p. 151) argue, that such a centralized dissemination of knowledge along hierarchical structures is inefficient in situations when explicit demand by knowledge seekers is predominant.³⁵

Decentralized organizational approaches towards KM are in turn characterized by bottom-up decisions about KM goals (see, e.g. Maier, 2007, p. 610, p. 614).

³²The distinction of pull and push will be further elaborated in Section 2.3.3.

³³This also reflects in a switch from *descriptive* to *prescriptive* knowledge, as discussed in Happel and Schmidt (2007).

³⁴See, e.g., Bühner (1992); Schilling and Steensma (2001); Picot et al. (2008)

³⁵See also Section 2.3.3

2. Knowledge Management

While such local decisions might consider the needs of knowledge seekers in a better way, they are less efficient from an organizational perspective. This is due to increased overall effort for KM and redundant knowledge creation in different organizational units.³⁶

Besides the *organizational* dimension, centralization and decentralization can as well be found on a technical level (see, e.g., Maier, 2007, p. 318ff). Here, centralized KM tools, which consolidate knowledge at a single place and offer access services, can be considered the dominant paradigm (Maier, 2007, p. 341). However, such a tool infrastructure is expensive to maintain and also takes the control about knowledge away from the individual to a central system.

Accordingly, several authors have devised distributed technical infrastructures for knowledge management. Several approaches were presented around the year 2000, coined *distributed*³⁷, *peer-to-peer*³⁸ or *agent-mediated*³⁹ knowledge management. However, these concepts have not seen widespread adoption in practice yet. This can be partially attributed to the fact, that most approaches are rather technology-oriented and not linked to management practices. Also, organizations might hesitate to introduce a technical infrastructure which they can only partly control.

Similar to the centralization and decentralization of the technical architecture, McAfee (2006) observed central and decentral means of communication. He argues that there exist *platforms*, where content is created and approved by few people but has many readers. Content managed on platforms, such as corporate intranets, is mostly characterized by high *commonality*. On the other hand *channels*, such as email or instant messaging, allow content to be created by anyone, but have low commonality and not much publicity.

Both, the *technical* and the *organizational* dimensions of centralization/decentralization are tightly intertwined. As described in Maier (2007, p. 603ff), a centralized KM organization is often accompanied by a centralized KM tool infrastructure.⁴⁰ While an alignment of technical and organizational structures tends to support information flowing within these structures, it also creates barriers for exchanging information across the structures (Henderson and Clark, 1990). This can reduce the impact of KM, given the need and potential benefits of sharing knowledge across organizational boundaries (Nonaka, 1994; Cummings, 2004).

³⁶See, e.g., Nonaka (1994) or Maier (2007, p. 603ff)

³⁷See, e.g.: Abecker et al. (1998); Bonifacio et al. (2003)

³⁸See, e.g.: Bonifacio et al. (2002); Tsui (2002); Tiwana (2003)

³⁹See, e.g.: Kamei et al. (2003); van Elst et al. (2003); van Elst and Abecker (2004)

⁴⁰Probst et al. (2006, p. 151) even argue, that centralized KM practices should not be combined with a *decentral* technical infrastructure: "Setting up a *non-centralized* information infrastructure [...] is not necessary in the context of a hierarchical top-down approach." (translation by author)

Competitive strategy	Codification	Personalization
Organizational structure	Centralized	Decentralized
Main knowledge type	Explicit	Implicit
KM strategy	“People-to-documents”: Codify, store, disseminate and reuse electronic documents	“Person-to-person”: Channel individual expertise; link people to make them share tacit knowledge
People skills	Knowledge reuse and implementation of solutions	Problem solving; tolerance for ambiguity

Table 2.1.: Codifications vs. Personalization (adapted from Hansen et al., 1999)

2.3.2. Codification vs. Personalization

Another strategic KM decision is roughly based on the distinction between *implicit* and *explicit* knowledge. As discussed in Section 2.2.2, implicit knowledge is difficult (if not impossible) to communicate. Therefore, the only means for sharing implicit knowledge is to bring people together and let them learn from observation and shared practice. Explicit knowledge, on the other hand, can easily be documented and captured. Hansen et al. (1999) introduce the term *codification* for “people-to-document” KM, which mostly relies on explicit knowledge. The KM strategy for implicit knowledge is called *personalization* and denotes “people-to-people” KM practices (see also Table 2.1).

Hansen et al. (1999) argue, that companies in practice rely on both, codification and personalization, but typically stress one or the other based on their company profile.⁴¹ They recommend that companies should balance both strategies at a 80%/20% ratio.⁴² *personalization* should be stressed if a company’s products are customized and innovative and if most knowledge is implicit. *Codification*, on the other hand, should be dominant when dealing with standardized products and mature and explicit knowledge (see Section 2.2.4). The underlying rationale is that standardized processes allow for a *good anticipation* of required knowledge, while more dynamic business practices make it difficult to cast KM decisions a long time beforehand.

⁴¹See also Boh (2007)

⁴²Huysman and de Wit (2003) argue interestingly, that the majority of research contributions does focus on a single aspect only

2. Knowledge Management

2.3.3. Push vs. Pull

Regarding knowledge distribution, one can distinguish the general strategies of *push* and *pull*.⁴³ According to Maier (2007, p. 210) they can roughly be characterized as “the systematic processes of bringing knowledge to the employees who need it (knowledge push) as opposed to knowledge search and retrieval that comprises knowledge being searched for by the employees (knowledge pull).”

Again, both strategies have tight relationships to other strategic dimensions. Probst et al. (2006, p. 151) argue that the *push*-strategy is dominant if there are centralized decisions about knowledge distribution, which is then “pressed” into the organization through well-defined channels.⁴⁴ They also stress that the *choice of the right knowledge* to communicate is crucial to this strategy. Conversely, the *pull* strategy should be applied when it is difficult to choose and *anticipate* the “right” knowledge. As Probst et al. (2006, p. 151) put it: “The pull philosophy in contrast addresses the knowledge user and his needs. In the case of a need, he shall be able to request required knowledge fast. Making directed knowledge requests shall become second nature to him.” (translated by author). However, Probst et al. (2006, p. 151) note, that a pull-strategy requires, that a matching between knowledge sources and knowledge demand can be achieved easily.

Accordingly, Probst et al. (2006, p. 179) stress that “KM measures which focus on concrete knowledge needs of end users (pull) have a [...] higher chance of application than measures which are decoupled from the user (push)” (translated by author). Thus, it seems that a pull strategy is typically more *effective* while a push strategy tends to be more *efficient*, especially in stable environments with standardized processes.

2.3.4. Personal vs. Organizational Knowledge Management

So far, this section has primarily dealt with the *organizational* perspective of KM. While most KM approaches and strategies acknowledge the central role of individuals in KM, these are often treated as a “black box” which produces and consumes knowledge.⁴⁵ In KM literature, one can observe a lack of detailed investigations of individual knowledge behavior and appropriate methodological and tool support. A notable exception is *Distributed KM*, as mentioned in Section 2.3.1, which however focuses on tool aspects, and also lacks a deeper discussion about individual knowledge processes. Only recently, in the context of the *Enterprise 2.0* and *Knowledge Work* discourse (see Section 2.4), the role of individuals in organizational KM has been stressed.

⁴³See also Section 2.1.1

⁴⁴See also Brown and Hagel (2005)

⁴⁵See e.g., Nonaka (1994) or Probst et al. (2006, p. 266). Figure 2.3 might be considered a counterexample to this claim. However, it regards individuals as abstract “information processing systems” without detailed considerations about their particular activities and roles in KM.

2.4. Knowledge Management and Organization

Besides classical organizational KM approaches, there exists a stream of research which concentrates on individual knowledge management. Dating back as long as to the 1940s, this work envisioned to leverage the new possibilities of computation to augment human memory. Seminal examples are the *Memex* system proposed by Bush (1945) and Engelbart (1962), who pioneered many concepts which should become part of the *personal computer*.

Since individual computing has become common, the research field of *Personal Information Management* (PIM) emerged.⁴⁶ While PIM research deals with information such as notes, mails, and calendar entries on a more general level, the subtopic of *Personal Knowledge Management* (PKM) addresses personal KM practices more deeply.⁴⁷

At the interface of personal and organizational knowledge management, the aspect of *keeping* information for future use⁴⁸ is of particular interest (Jones, 2004). In particular, several authors stress that the individual decision to keep information can be influenced by the intention to share that information with others.⁴⁹ While some research started to target the particular interrelation of personal and organizational knowledge management practices, methodological or tool support is still in its infancy.⁵⁰

This overlaps with our observations about gaps in the evolution from personal knowledge to organizational knowledge, as discussed in Section 2.2.4. Traditionally, this gap was often bridged by direct interaction in collocated work groups or *organizational knowledge communities* (see Section 2.2.3). However, with increasing decentralized and distributed work settings (as discussed in the following section), this gap tends to become more challenging to address by methods and tools.

2.4. Knowledge Management and Organization

This section will discuss recent and ongoing external trends that have a major influence on the theory and practice of knowledge management. We describe *knowledge work*, *distributed work* and *Enterprise 2.0*, which define new ways of organization, often tightly related to technical innovations.

We argue that, according to the way organizations change, *knowledge management concepts have to evolve* to adapt to these changes. Conversely, technical innovations in the domain of knowledge management can also foster organizational change.⁵¹

⁴⁶See, e.g., Teevan, Jones and Bederson (2006); Jones and Teevan (2007)

⁴⁷See, e.g., Markus (2001); Wright (2005); Pauleen (2009); Cheong and Tsui (2011); Pauleen and Gorman (2016)

⁴⁸Also called “knowledge preservation” in Section 2.1.1

⁴⁹See, e.g., Erickson (2006); Marshall and Jones (2006); Lutters et al. (2007). This aspect will also be discussed in Section 3.2.1

⁵⁰See, e.g., Bradshaw et al. (2006); Tungare and Perez-Quinones (2008) and Section 5.4.3 and 4.5.2. Gwizdka (2006) argues that “there is a need for a perspective that considers relations between external and internal information spaces”.

⁵¹See also Section 2.1.2

2. Knowledge Management

Criterion	Traditional office work	Knowledge work
Centralization	Central organizational design	Decentral organizational design
Structure	Hierarchy	Network, hypertext organization
Process	Highly structured, deterministic processes; pre-structured workflows	Weakly structured, less foreseeable processes; ad-hoc workflows

Table 2.2.: Traditional Office Work vs. Knowledge Work (excerpt of Maier, 2007, p. 49f)

2.4.1. Knowledge Work

Among the three trends discussed in this section, *knowledge work* is the oldest and probably most general one. With an increasing automation of industrial production, economies in most developed countries observe a continuous increase of a service economy. Authors have coined the term *knowledge society* for this ongoing transformation process.⁵²

In this context, individual workers are no longer an easily exchangeable resource in a well-defined production process, but rather important actors in increasingly complex work settings. A study by Johnson et al. (2005), for instance, reports that the number of jobs in the US which primarily consist of “tacit interactions”⁵³ is significantly increasing, while the number of jobs with a large amount of routine tasks is decreasing.

Accordingly, not only individual skills and knowledge, but also the knowledge in the organization has become a management concern – which also explains the emergence of the KM discipline as such (see, e.g., Probst et al., 2006, p. 3ff). While *knowledge work* is an abstract concept, which might be interpreted quite differently in various contexts, a rough comparison to *traditional* office work, as depicted in Table 2.2, can probably foster a better understanding.

People engaged in knowledge work are called *knowledge workers*. Typical examples for knowledge workers are scientists, engineers, and consultants.⁵⁴ According to the study of Johnson et al. (2005), 70% of new jobs created in the USA between 1998 and 2004 can be considered knowledge workers. This has also sparked additional research interest. Similar to studies on industrial age workers as by Taylor (1911), knowledge workers have become subject to *work studies* which seek to understand how they spend their workday.⁵⁵

⁵²See, e.g., Drucker (1969); Stehr (1994) or Probst et al. (2006, p. 3ff)

⁵³Defined as: “complex interactions requiring a higher level of judgment, involving ambiguity, and drawing on tacit, or experiential, knowledge”

⁵⁴See, e.g., Hansen et al. (1999); North and Guldenberg (2009); Heisig et al. (2010)

⁵⁵See, e.g., Brown et al. (2000); Whittaker and Hirschberg (2001); Bradshaw et al. (2006); Hyldegård (2006); Oren (2006); Singer et al. (2008); Maalej and Happel (2009)

2.4. Knowledge Management and Organization

The advent of knowledge work is also tightly related to the information technology revolution. First, the availability of computers at the workplace has increased the amount of knowledge work as such. On the other hand, computer networks enabled the development of particular tools for knowledge workers (Grudin, 1994). Systems such as groupware, Intranets, Wikis, or file shares provide an infrastructure for seeking information and sharing knowledge.

However, the role of individual users is often rather to consume centrally provided knowledge, or to share knowledge within smaller circles.⁵⁶

2.4.2. Distributed Work

Distributed work describes the situation that several people, who may be physically or organizationally distributed, collaborate on a shared task. People in a distributed work setting are also called a *distributed team*.⁵⁷ The measure for physical distribution is the actual location⁵⁸ of people, while organizational distribution is measured in hierarchical distance.⁵⁹ The concept of distributed work is tightly related to the modern history of production, spanning from the industrial revolution⁶⁰ to the rise of modern information technology.⁶¹

In this section, we sketch the origins and reasons for distributed work. We also highlight the problems and challenges this kind of work setting causes for knowledge management practices.

Challenges of Collaborative Work The development of complex products and services is tightly related to the concepts of *modularity* and *division of labor* which led to a specialization of firms and workers (Baldwin and Clark, 2000). Specialization requires collaboration in order to work on common tasks. The modularization of complex artifacts is thus often mirrored by a modularization of the project teams, which produce these artifacts (Conway, 1968; Herbsleb and Mockus, 2003). For example, distinct parts of an automobile are typically built by separate companies or production teams.

If collaborating workers are separated by organizational or physical distance, this is called *distributed work*. Such work settings have become popular in recent years due to several reasons such as cost reduction, availability of human resources, or intra-organizational collaboration.⁶²

While the number of distributed teams increases, empirical studies show, that

⁵⁶See also the discussion of *platforms* and *channels* in Section 2.3.1

⁵⁷See, e.g., Hinds and McGrath (2006)

⁵⁸Researchers define a mere distance of 30 meters as the starting point for distributed work (Allen, 1984; Olson and Olson, 2000; O'Leary and Cummings, 2007)

⁵⁹Such as, e.g., working in the same team vs. working in different department vs. working in a different company.

⁶⁰See, e.g., Baldwin and Clark (2000); Kind and Frost (2002); O'Leary et al. (2002)

⁶¹See, e.g., Moon and Sproull (2002) and Walsh and Maloney (2002) regarding the impact of the Internet on enabling distributed software engineering resp. research.

⁶²Grinter et al. (1999); Olson and Olson (2000)

2. Knowledge Management

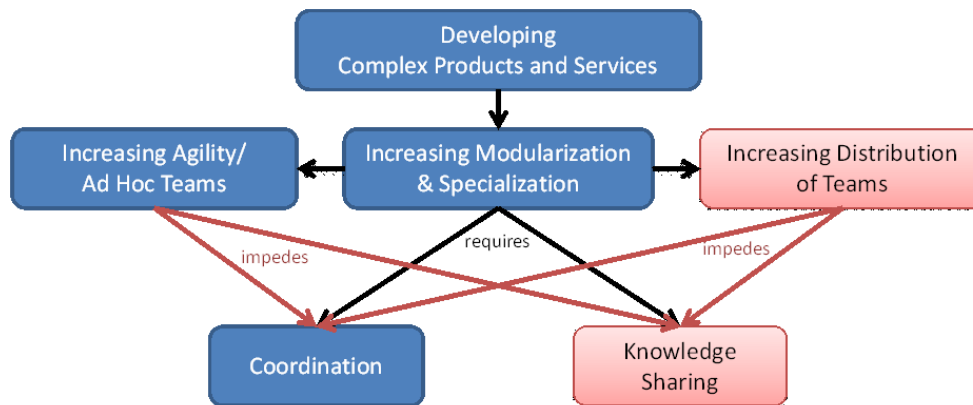


Figure 2.7.: Distributed Work requires, but also impedes Knowledge Sharing

they are typically less efficient than collocated ones.⁶³ Coordination problems are one of the main reasons for this, because the overall coordination capacity in distributed settings is lower, due to a reduced communication bandwidth (Olson and Olson, 2000). Research has also shown, that a distance of 30 meters is already sufficient to significantly reduce *informal communication* (Allen, 1984; Olson and Olson, 2000). Hence, a large fraction of collaboration settings has the characteristics of distributed work.

Collaboration research is focusing on tool-support for creating “virtual 30 meters” (Herbsleb and Mockus, 2003) to alleviate these problems. The main idea is to lower distances in distributed teams by providing information technology which allows for a communication intensity, that is comparable to collocated work settings.

Core challenges in realizing these virtual 30 meters are *coordination* and *knowledge sharing* among the team members. This is stressed by the odd circumstance, that modularization not only *requires* coordination and knowledge sharing for effectiveness, but also *impedes* the efficiency of both. We will explain these relationships, which are also depicted in Figure 2.7, in the following.

Coordination vs. Knowledge Sharing The need for *coordination* stems from the modularity of the artifacts under development. The decomposition of artifacts creates dependencies, which require coordination. Coordination can thus be defined as “the process of managing dependencies between activities” (Malone and Crowston, 1994). Accordingly, most coordination requirements can be traced back to explicit dependencies among the subsystems of artifacts and can be “*objectively*” recognized by the coworkers.⁶⁴ Coordination issues are

⁶³See, e.g., Olson and Olson (2000); Teasley et al. (2000); Olson et al. (2002); Armstrong and Cole (2002); Herbsleb and Mockus (2003)

⁶⁴See also the discussion of *objective information needs* in Section 3.1.1

Collaboration issue	Coordination	Knowledge sharing
Required to manage	Modularity of artifacts	Modularity of the organization
Rooted in	Explicit dependencies of technical artifacts	Different experience and capabilities of the involved persons
Mediated by	Tasks and processes	Organizational communication system
Nature	Objective	Subjective

Table 2.3.: Coordination vs. Knowledge Sharing

central to collaboration research, which, e.g., analyzes suitable coordination mechanisms for different kinds of dependencies (Thompson, 1967) and tools for supporting communication, awareness and workflow management.

In contrast to coordination, organizational *knowledge sharing* is not directly rooted in explicit dependencies of technical artifacts, but in dependencies among organizational entities (see also Table 2.3). The modularity of the organizational subsystems has a deep impact on the communication patterns of an organization, since it forms channels and filters along organizational interfaces to reduce complexity by selecting relevant information (Henderson and Clark, 1990).

However, this has a negative impact on knowledge sharing across organizational units, since the average information flow runs dry with increasing organizational or geographical distance. Also, knowledge sharing needs are highly *subjective*, because they depend on the experience, capabilities and context of the involved persons, which makes them more difficult to address (Cramton, 2001).

Distributed work settings *amplify barriers for knowledge sharing* such as reduced motivation and trust (Cabrera and Cabrera, 2002). The allocation of knowledge can thus be inefficient when collaborating in a distributed organization. This is one reason for our particular interest in the interrelation of knowledge sharing and distribution (i.e., the red boxes in Figure 2.7) in this work.

2.4.3. Enterprise 2.0

The *Enterprise 2.0* concept can be seen as the next step in the development into a knowledge society. Its main driver is the permanent and ubiquitous availability of an Internet connection for a large number of individuals. This has led to the general concept of *Web 2.0*, which can be considered as a mutual effect of several technical, economical, and social innovations.⁶⁵

Major *principles of Web 2.0 applications* can be summarized as follows. First, due to the large amount of users, *collective intelligence* can emerge from

⁶⁵See, e.g., O'Reilly (2005); Musser (2007)

2. Knowledge Management

marginal contributions of many users. The large amount of users can also lead to a better *allocation of expertise*, since users can decide themselves where their contribution is needed. This is often accompanied by self-organized modes of governance (“meritocracy”) in which the influence of individual users depends on their contributions rather than their formal role. The Web 2.0-style of cooperation and production is less formalized and not controlled by a central authority.

As a consequence, artifacts produced in a Web 2.0 environment can be of chaotic structure and lower quality. However, it is part of the philosophy that artifacts are in a state of *perpetual beta*, during which they can continuously improve. The example of Wikipedia shows, that this mode of production can compete with more centrally governed structures.

Indeed, the initial imperfection of certain artifacts can be even considered an advantage. By opening for niche topics, which might receive only few attention, a community is enabled to address the so called *long tail*⁶⁶ of information. For the small number of people interested in such information, also content of (yet) low quality might be of interest – particularly if it is considered a first step of evolution in the perpetual beta cycle.⁶⁷

Finally, the described concepts show, that the Web 2.0 enables individual users to take a more active role in using information systems. Instead of being passive consumers of information produced by a small elite, they can themselves produce and contribute information. This dual role of consumers and producers has accordingly been labeled as *prosumer*.⁶⁸

The adoption of Web 2.0 principles at the workplace has been denoted as *Enterprise 2.0* (McAfee, 2006, 2009). Similar to private Internet connections, broadband access at the workplace has increased in recent years. Studies show that the number of users with Internet access at the workplace has reached 54%.⁶⁹

Enterprise 2.0 is thus the next step in the evolution of the knowledge worker, which is characterized by a high Internet literacy, broadband access at the workplace, and Web 2.0-style collaborative applications which allow to more actively participate in corporate knowledge creation. Studies show, that companies are increasingly adopting such Enterprise 2.0 tools.⁷⁰

However, the open and decentralized nature of Enterprise 2.0 tools also raises some concerns and problems. Enterprise Wikis, for instance, require careful oversight, because a lack of guidance can lead to a “proliferation” of content

⁶⁶The term *long tail* has initially been introduced by Anderson (2004, 2007) to distinguish niche products with low demand from mass production.

⁶⁷This can be considered related to the knowledge maturing process described in Section 2.2.4.

⁶⁸Tapscott and Williams (2006, p. 124ff)

⁶⁹Recent number for Germany according to Statistisches Bundesamt (2016). See also U.S. Bureau of Labor Statistics (2005) and Maier (2007, p. 489f)

⁷⁰See, e.g., Back et al. (2008); Koch and Richter (2009); Andriole (2010); Stocker and Tochtermann (2011); selected properties of Enterprise 2.0 tools will be discussed in Section 3.2.2

which might have a negative impact on tool acceptance (Happel and Treitz, 2008).⁷¹

In summary, Enterprise 2.0 tools and principles seem to be somehow *effective* in improving knowledge sharing in organizations (Andriole, 2010). What remains as open issues, is their level of *efficiency* and a better understanding of how they can be employed to grow organizational knowledge in a *systematic* way.⁷²

2.5. Summary

In the preceding sections, we have given a brief introduction to the field of knowledge management, its underlying notion of *knowledge*, fundamental choices in KM *strategy* and the *organizational* environment which influences KM.

Our main observation is that KM is “trapped” in *two archetypes* of implementation. As Maier (2007, p. 52) argues, there exist a widely accepted distinction between *human-* and *technology-oriented* KM approaches. Both archetypes can be characterized by a set of design choices that often co-occur.⁷³

While there might be a “natural” tendency for certain dimensions to align, we argue that design would allow for different choices, deviating from these archetypes. A good example is the relation between the structure of the organization and the organization of KM. Both dimensions will naturally align due to communication channels established by hierarchy and organizational structures (Henderson and Clark, 1990; Nonaka, 1994).⁷⁴ In contrast to this, several authors claim, that KM is especially useful, if it can help to share knowledge *across* boundaries imposed by formal organization (Nonaka, 1994; Cummings, 2004).

Especially *distributed work* settings and the *Enterprise 2.0* phenomenon challenge the existing archetypes. For instance, distribution makes personalized, face-to-face knowledge exchange much harder. This is due to the loss of a common spatial context, and a limited communication capacity and availability of experts for a special topic. Instead of direct face-to-face interactions, personalized knowledge sharing in distributed settings has increasingly to rely on mediated, asynchronous collaboration.

Authors such as Hansen et al. (1999) have also argued that decentralized organizations, which rely on flexible business processes, do not have much common knowledge to share within the organization. While it is true that different clients require different solutions, we argue that there exists a certain *common core of knowledge* which is reusable across different projects. This *long tail* of common knowledge is however difficult to address under the classic archetypes. A good example for the codification and dissemination of knowledge, which is

⁷¹The case of Enterprise Wikis is particularly addressed and discussed in Chapter 6

⁷²Limitations of Enterprise 2.0 tools will be discussed in Section 3.2.2

⁷³As depicted in Table 2.1

⁷⁴Also refer to the discussion on *organizational knowledge communities* in Section 2.2.3.

2. Knowledge Management

relevant only for a small amount of people, are forums and Q&A sites on the web.⁷⁵

Our particular concern is about the dimensions of *personalization/codification* vs. *individual/organizational knowledge*. The alignment of both dimensions has been useful in classic, centralized organizational forms, where central decisions about codified organizational knowledge were feasible. In more decentralized organizations of knowledge workers however, it is more difficult to predict which knowledge might be useful for larger parts of the organization. Tool-wise, this is reflected by the separation between *platforms* for the centrally controlled dissemination of organizational knowledge, and *channels* for the exchange of personalized, individual knowledge.

When considering an increase of bottom-up style knowledge creation, as it is suggested by the knowledge maturing process, the technological and conceptual *gap* between channels and platforms creates a barrier for knowledge sharing. Benefits from *collective intelligence*, as intended by Enterprise 2.0 concepts, are thus difficult to realize.

We conclude, that many organizations are “trapped” within the existing archetypes of KM implementation. These archetypes combine organizational, technical and strategic choices in a fashion that is problematic under the assumption of distributed knowledge work. Accordingly, Maier (2007, p. 313) calls for “holistic KMS implementations (that) aim at *bridging the gap* between these two architectures”.

Based on the previous discussion, we propose five dimensions which highlight issues on the way to “bridge the gap”:

Anticipation of needs With an increase of knowledge work and “tacit interactions”, information needs of individuals have become difficult to anticipate. Instead of deriving such needs from job roles or business processes,⁷⁶ many information needs *emerge* throughout the organization. The *prediction* of what is valuable organizational knowledge thus becomes more difficult.

Choice The anticipation of information needs serves as a basis to *decide* which organizational knowledge should be created and maintained given limited resources. Again, this choice is increasingly difficult to make on a management level, since there is no global perspective on the emergent information needs of individuals.

Participation The dimension of participation deals with the question *who* is supposed to participate in organizational knowledge creation. As discussed in Section 2.4.3, the *prosumer* role suggests that all individuals in the organization should actively participate in this process. Participation however, is a long-term problem in KM research.⁷⁷

Transition A particular challenge raises due to the emerging role of the *prosumer*. While classical KM approaches focus on a top-down *push* and a

⁷⁵See, e.g., Section 3.3 and 6.4.2

⁷⁶See Section 2.4.2 and 3.1.1

⁷⁷See forthcoming Section 3.2.1

bottom-up *pull* of knowledge, prosumers strive to make bottom-up contributions to organizational knowledge. This results in the gap between PKM and OKM described in Section 2.3.4.

Controlling While it has always been a challenge to measure the success of KM initiatives⁷⁸ this is even more difficult in settings with increased decentralized organizational knowledge creation. Furthermore, the Web 2.0 principles introduced in Section 2.4.3 impose novel requirements for quality control.

We do not consider these dimensions as necessarily complete or mutually exclusive. The list however provides a basis for discussion and will serve as a point of reference throughout this work.

We finally conclude, that there exists a methodological gap as well as a gap of tool support concerning the bottom-up creation of organizational knowledge emerging from individuals. Our focus in the following will thus be on individual *information seeking* and *knowledge sharing* and its transition to an organizational level.

⁷⁸See, e.g., Probst et al. (2006); Jennex et al. (2014)

3. Information Seeking and Knowledge Sharing

While the previous chapter focused on the broader topic of knowledge management, we will now discuss research in the more specific areas of *information seeking and retrieval* and *knowledge sharing*.

When considering knowledge management activities from an individual perspective, one can distinguish the core roles of the *information provider*, who shares her knowledge, and the *information seeker*, who consumes the resulting information.¹ We will review existing research with respect to these roles.

Therefore, we present the established fields of *information seeking and retrieval*, which concentrate on the information seeker, and the more heterogeneous research area dealing with *knowledge sharing*. Afterwards, we discuss the interrelation of both roles by discussing *knowledge sharing as a communication process*.

3.1. Information Seeking and Retrieval

Information seeking (IS) and information retrieval (IR) are two closely related fields of research, which analyze how humans seek and retrieve information. While IS deals with the behavioral aspects of the information seeking process, IR is focused on the technical challenge of matching the information needs of a user with an available set of information.²

We begin by defining core terms such as *information need*, *information seeking*, and *information behavior*. While there exists a similar amount of definitions for these terms, as for *information* and *knowledge*,³ we refer to the basic definitions provided by Case (2002, p. 5):

Information need: An information need is a *recognition that your knowledge is inadequate* to satisfy a goal that you have.⁴

Information seeking: Information seeking is a conscious effort to acquire information in response to a need or gap in your knowledge.

Information behavior: Information behavior encompasses information seeking as well as the totality of other *unintentional* or passive behaviors (such as

¹See Section 2.2.1 for the relationship of *information* and *knowledge* in this context.

²See also Ingwersen and Järvelin (2011).

³See Section 2.2.1

⁴Emphasis added by author of this thesis

3. Information Seeking and Knowledge Sharing

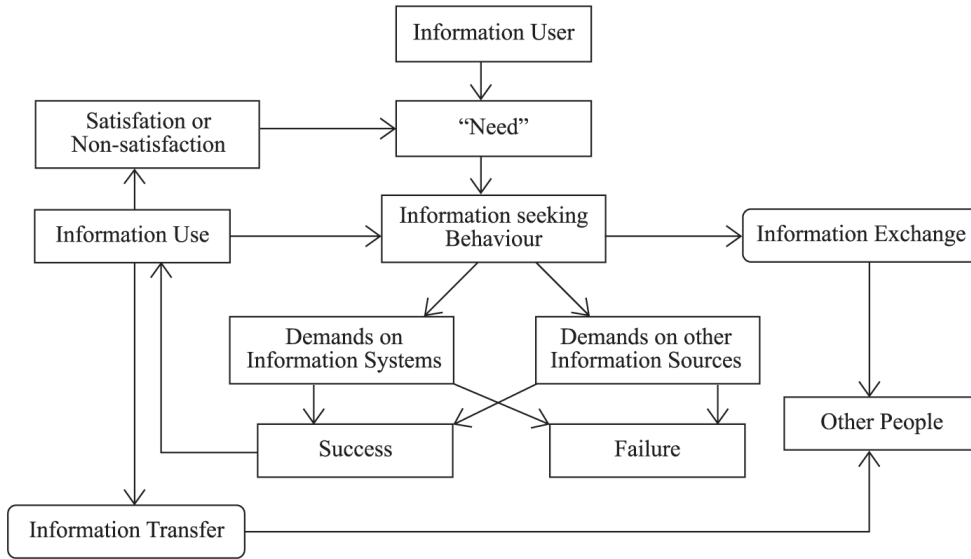


Figure 3.1.: A Model of Information Seeking Behavior (Wilson, 1981)

glimpsing or encountering information), as well as purposive behaviors that do not involve seeking, such as actively avoiding information.

In the following, we will introduce the typical process of information seeking and highlight some critical issues involved. Afterwards, we will discuss information retrieval systems and how they deal with those issues.

3.1.1. Information Seeking Process

In IS research, many models of human information seeking have been devised.⁵ For our purpose, we will focus on presenting one of the the most influential models by Wilson (1981), which is depicted in Figure 3.1. It begins with the information seeker (denoted *information user*) and her *information need*, which triggers *information-seeking behavior*. This behavior results in consulting either *other people*, *information systems*, or *other information sources*. Seeking may result in a *failure* to locate any information or continue with success. In the latter case, *information use* finally reveals if the information can satisfy the information need or not.

We will now elicit several issues related to this process.

Information need vs. information demand One of the most difficult concepts described in the model is the *information need*⁶ which triggers information seeking behavior. According to the seminal distinction by Taylor (1962, 1968), four levels of “question formation” can be distinguished:

⁵See, e.g., Case (2002); Fisher et al. (2005) for an extensive discussion

⁶As the current discussion targets an individual information seeker, we may also denote this more precisely as *personal information need*

Visceral need (Q1) – which denotes an “conscious and *unconscious* need for information”

Conscious need (Q2) – which is a conscious, more concrete state with still some degree of ambiguity

Formalized need (Q3) – which is a concrete statement about “what we like to believe the information system answers”

Compromised need (Q4) – in which “the question is *recast in anticipation* of what the inquirer thinks he will get out of the system”

This differentiation highlights a number of key issues regarding information needs. First, a need is a highly subjective concept which *evolves* in the course of thinking, discussing with others, and interacting with information systems. The expression of the information need might be influenced by the seeker’s interaction with and assumptions about her environment. During this process, a need may be initially unconscious to the information seeker.

According to Case (2002), an information need is a mental concept and can thus not be directly observed. It can only be observed via an explicit formalization, which is called *information demand*. The latter corresponds to the final levels (Q3/Q4) of Taylor’s (1962) model. Case (2002) also discusses if the term of an information “need” is applicable – i.e., if information can be considered a basic human need. Wilson (1981) argues in a similar way and proposes to consider the underlying *physiological*, *affective*, and *cognitive* needs as the “basic” needs underlying a particular information need (see also Figure 3.2).

Another distinction is made in the area of information management, which differentiates subjective and objective information needs. *Subjective* information needs stem from the individual’s “need” and can thus only be derived individually.⁷ *Objective* information needs, on the other hand, are prescribed by organizational processes and tasks.⁸

Context Both, the emergence of a subjective information need, and its evolution during interaction with the environment, are not explicitly shown in Figure 3.1. However, Wilson (1981) also discusses such issues under the notion of *context*. As depicted in Figure 3.2, the “needs” of the information seeker are affected by her current work role and actual tasks performed in a given situation. Both is further influenced by factors such as the work, socio-cultural and physical environment.

Decision process The information seeking process includes several elements of decision which influence its progress. As mentioned, the process is driven by the (information) need of an individual user, who strives to accomplish

⁷(Krcmar, 2004, p. 61) notes that the explicit *information demand* is a subset of the individual information need, which is a simplified view compared to Taylor’s model

⁸See (Krcmar, 2004, p. 60), our discussion about *coordination vs. knowledge sharing* in Section 2.4.2 and the forthcoming discussion of *information requirements analysis* in Section 4.5.1

3. Information Seeking and Knowledge Sharing

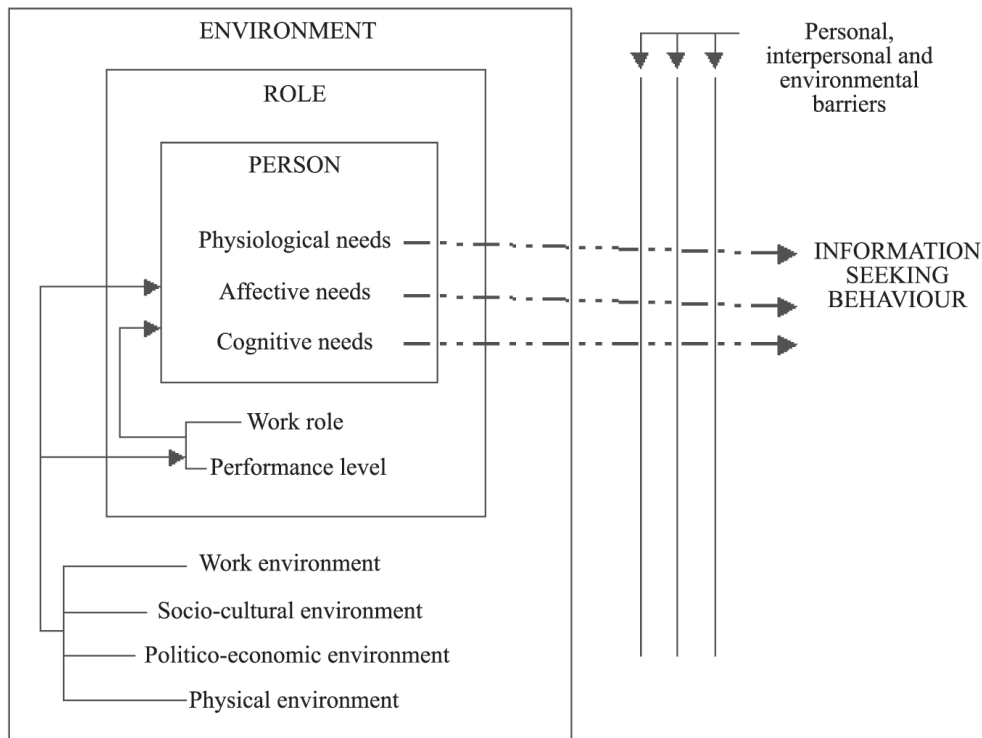


Figure 3.2.: The Context of Information Seeking (Wilson, 1981)

an underlying task. Accordingly, actions are judged by the *expectation* how they can contribute towards completing the task and respectively satisfying the user's needs. The initial decision is thus if to engage in information seeking behavior, once an information need is *recognized*. As (Nicholas, 2003, p. 22) argues, many information needs may remain dormant if the user does not perceive an immediate case for action.

Once a user engages in an information seeking process, its further progress is influenced by what Wilson (1981) calls personal, interpersonal, and environmental *barriers* (see also Figure 3.2). Such barriers may include a *limited awareness* about available information or information sources, or the cost and effort required to engage in information seeking.

Mediation Information seeking can occur in a synchronous or asynchronous fashion. The latter situation is also called *mediated* information seeking, which can refer to a human mediator or an information system mediating between human users. Different forms of such information seeking behavior are shown in Figure 3.3. An information seeker might seek information directly from another actor (a, b, c) or from an information resource (d). In many cases however, she will refer to a human mediator or an information system (f, e).

In the context of mediation, the notion of *anonymity* is of some importance. While synchronous communication is typically not anonymized, asynchronous

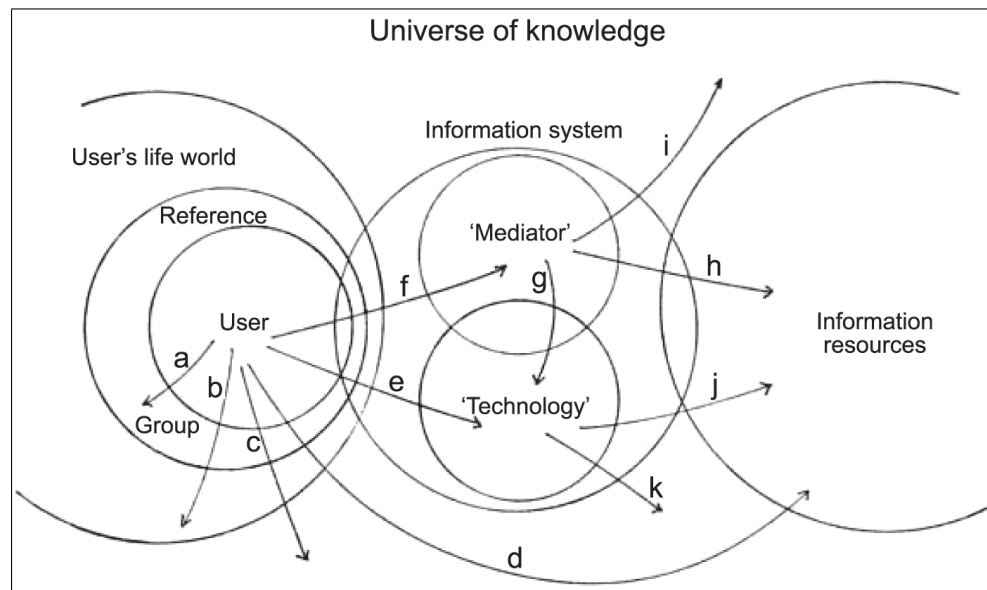


Figure 3.3.: Information Needs and Seeking (Wilson, 1981)

communication is typically anonymized if no additional information about the information seeker or information provider is included.⁹ Given the *affective* nature of certain information needs, handling anonymity is thus an important design choice when using information systems for mediation.

Transience The information seeking process is typically considered with a focus on the information seeker. As depicted in Figure 3.1, activity starts and is driven by the information seeker and her information need. After the process is completed, the major change occurs in the mind of the information seeker. If her need was satisfied, the cognitive system adapts accordingly. If the need was not satisfied, another information seeking activity might be started.

In this conceptualization, the information need itself as well as the subsequent information seeking process do not have any persistent representation. The need and the process are just temporarily represented in the mind of the information seeker, which fades away once the process is completed. They can thus be considered *transient* entities.¹⁰

Solitary Another important aspect is *collaboration*. While Figure 3.1 depicts a solitary user, information systems are used by different information seekers, either subsequently or even in parallel. Most information systems for

⁹This means that asynchronously sent messages by default do not contain information about their originator. As in the case of a postal letter, such information has to be added separately.

¹⁰See, e.g., also Morris and Horvitz (2007)

3. Information Seeking and Knowledge Sharing

information seeking however isolate their users, although various opportunities for leveraging the behavior of other users exist. A well-known example is the technique known as *implicit relevance ranking*¹¹ which tries to improve information suggested to a user based on the information usage behavior of other users.¹²

In recent years, several scholars have proposed additional means for explicit and implicit collaboration throughout the search process.¹³ Information seeking models which explicitly consider multiple information seekers and their interactions are called *collaborative information seeking* or *social search*.¹⁴

Unidirectional Considering the flow of action indicated by the arrows in Figure 3.1, information seeking is an *unidirectional* process. People or other information systems are either able to satisfy an information need or not. In the latter case, the process might be re-iterated. From the perspective of the information seeker, this results in a synchronous interaction with the course of action under her full control. Information is retrieved either instantly or not at all. Accordingly, the information seeking process considers the option of *failure* to locate suitable information, respectively *non-satisfaction* of the initial information need.

A major drawback of this model is, that it does not consider the aspect of *time*. It returns only such information, which is available before or during the information seeking process. In information retrieval research, the model is thus called *retrospective* search (Wyman, 2005; Hearst, 2009, see also Figure 3.4). However, in information environments such as an organization, or the Web, relevant information is created continuously.

This issue is addressed by the paradigm of *prospective* search (Hearst, 2009). Prospective search systems allow users to store their queries permanently. Whenever there is new information which matches these queries, the system sends out a notification to the user.¹⁵

Closed World Assumption The conceptualization of *retrospective search* in IR research is primarily a technical one. Since information seeking behavior occurs at a fixed point in time, information systems can only provide information that has been created beforehand and which is thus available at the time of the request. Consequently, *prospective search* tries to solve the technical problem of informing the information seeker about information that is available at a later point of time. *Prospective* search addresses the information seeker and allows her to search for *information that does not yet exist*.

¹¹See, e.g., Kelly and Teevan (2003); Agichtein et al. (2006)

¹²A typical application of implicit relevance ranking in IR systems (see Section 3.1.2) is the modification of search result rankings based on result clicks of previous users.

¹³See, e.g., Evans and Chi (2008); Golovchinsky et al. (2009); Hearst (2009)

¹⁴See Section 3.1.2 for more information

¹⁵See Sections 3.1.2 and 5.4.2 for more information. See also Section 2.3.3 regarding the related *push* concept in KM.

3.1. Information Seeking and Retrieval

It then checks if newly created information matches the seeker's information need. *How* and *why* this information is created is not considered.

We argue that the notion of retrospective search has a second, more subtle aspect. This is the aspect of interpreting the ontological relationship between an information demand and the actual information to satisfy this demand. Retrospective search considers the actual information as the fixed point and the information demand as a transient (as discussed before) and singular occurrence. This perspective implies that an information demand is depending on the set of available information.¹⁶

In other words, information is considered to precede an information need: if there is no information matching an information need, standard IR systems treat this as if no information would exist. This is tightly related to the so-called *Closed World Assumption* (CWA), which states that “a database includes a representation of every occurrence in the real world environment that it models” (Motro, 1986). While this assumption is often challenged in the database world (see also Section 7.2.1) it's certainly even more questionable in the less structured environment of text-based IR.

In there, the assumption implies a notion of a fixed set of possible information which describes an inter-subjective reality that is shared by an information seeker. This perspective however refutes the possibility that an information demand might *predate* the existence of information. Taking this perspective, an information demand might also spur the creation of information and thus shape the “reality” which is reflected by the information to be created. In other words, information can be arbitrarily created or *designed* in the course of social interaction, without regard to any preexisting common conceptualization.¹⁷

Revisiting Figure 3.1, we can observe that the process of *creating* information is not explicitly considered. While *other people* are conceptualized as sources of information, it appears that this is just for the purpose of *exchanging* information which already exists. We can thus conclude that Wilson (1981) assumes an “objective” nature of information and does not take into account the possibility of information as an artifact that can be designed.

3.1.2. Information Retrieval Systems

IR systems are information systems, which allow users to satisfy their information needs (see Figure 3.1). We shortly introduce *conventional IR systems*, *prospective search systems*, *enterprise search*, and more recent *social search* approaches to give an overview how some issues described in the previous sections are addressed by existing tools.

¹⁶This is also reflected in Taylor's (1968) argument of “recasting information needs in anticipation of available information”, which we introduced in the paragraph on *information need*.

¹⁷See also the discussions on *organizational knowledge* and *knowledge maturing* in Section 2.2.3 resp. 2.2.4

3. Information Seeking and Knowledge Sharing

Conventional IR Systems We call IR systems *conventional*, if they are designed according to what is called the “classic” model of information retrieval (Broder, 2002). Unlike the rather complex models of information seeking discussed in the previous section, models of IR systems have a technical focus and follow an established reference architecture.

Basic elements of this architecture are the information seeker, her information need, the IR system and the corpus of documents, against which a representation of the information need is matched.¹⁸ The main purpose of IR systems is to help users to satisfy their information needs by providing a set of relevant information – which is typically contained in some kind of document.¹⁹

To use an IR system, the user has to express this information need in terms of a query language which can be interpreted by the search system. In most systems, this is a textual, “keyword-based” representation of the information need. As an example, Obama or US president²⁰ could be keywords searching for information about Barack Obama.

Prospective Search From a process perspective, conventional IR systems periodically crawl documents from repositories (such as folders in a file system or the WWW), and analyze them in order to build an index mapping terms to documents. Incoming user queries are matched against this index in order to derive suitable results.

Obviously, this architecture has difficulties to deal with the dynamics of document collections. Since new documents are only added during periodical crawls, users can only retrieve those documents, which were already indexed. As introduced in the previous section, this paradigm is also called *retrospective search* (see Figure 3.4).

This limitation is addressed by a new paradigm called *prospective search* (Wyman, 2005; Irmak et al., 2006, see also Figure 3.4). Instead of querying the status quo of information, prospective search allows users to subscribe to a certain topic and to receive notifications once new information appears. Synonyms terms for this system model are *publish/subscribe* or *continuous querying*.²¹

A popular implementation of such functionality is Google Alerts (Google Inc., 2017a). This feature of the Google search engine allows users to define “alerts” based on keywords and matches them periodically against new documents, which have been discovered during crawling. If there is a match, users are notified about these new results, either by email or by embedding the results in a personalized user interface.²²

¹⁸See, e.g., Broder (2002)

¹⁹IR systems may also target other forms of media such as pictures, videos, or music.

²⁰Barack Obama was still US president at the time of writing this section. Besides that, this can also serve as an example for the inherent ambiguity of information demands expressed by users.

²¹See, e.g., Kukulenz and Ntoulas (2007)

²²Limitations of retrospective and prospective search approaches will be addressed in Chapter 5

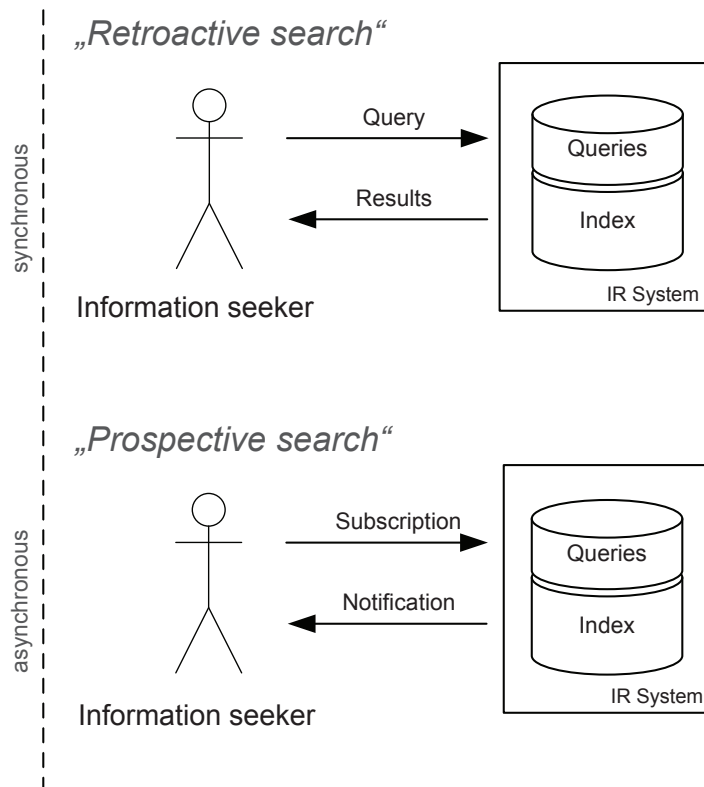


Figure 3.4.: The Dimension of Time in Information Retrieval

Enterprise Search Searching for information has become one of the main activities of knowledge workers to such an extent, that the verb “to google” even made it into contemporary dictionaries.²³ Similar to the vast amount of information on the Web, many enterprises harbor a large set of documents and other information which can be useful for its employees.²⁴

However, search in an enterprise setting suffers from several problems (Fagin et al., 2003). First, the number of cross-links, which is an important building block of popularity-based ranking algorithms, is typically rather low. Secondly, enterprises often have a heterogeneous information systems infrastructure with sophisticated permission schemes that may impede indexing by enterprise search engines.

Enterprise search is particularly cumbersome from a collaboration perspective. Recent research has shown that information seeking is a highly interactive process which can benefit significantly from collaboration (Evans and Chi, 2008; Morris, 2008). However, if, e.g., two employees have a similar information need within a particular timeframe, typical search systems do not allow to interact on this information need, which might be especially helpful if such an

²³See, e.g., <http://www.merriam-webster.com/dictionary/google>

²⁴See also Section 4.2.2

3. Information Seeking and Knowledge Sharing

information need remains unsatisfied.²⁵

Enterprise search systems usually follow the standard IR paradigm, neglecting the dynamic character of information provisioning, and assuming a stable corpus of information which grows only occasionally upon crawling new content.²⁶ Users can typically not directly influence search results, even if there are only few or bad results. This was different during the early days of the WWW, which favored human-maintained “catalogs”. While those did not scale up well with the growing Web (Kobayashi and Takeda, 2000), automated indexing and retrieval approaches, which largely keep human users out of the loop, do not scale down well for rare information needs.

Social Search In recent years, the concept of social and collaborative software has extended the classical model of information seeking and retrieval. Collaboration within the information seeking process has been discussed under the label of “collaborative information retrieval” earlier this decade²⁷ and more recently.²⁸ “Social search” is a related term, which is widely used to label various different approaches. Evans and Chi (2008) define it as:

... an umbrella term used to describe search acts that make use of social interactions with others. These interactions may be explicit or implicit, colocated or remote, synchronous or asynchronous.

While many research approaches in social search have been focusing on *synchronous* collaboration aspects,²⁹ a number of Web 2.0 applications has adopted collaboration features. Notably, even Google adopted features to comment, re-rank, and add search results in its user interface, which had been pioneered in the discontinued Wikia search engine. Collaborative information seeking and social search address important “blind spots” of the classic information retrieval model by considering collaborative activities among fellow information seekers.

While several social search tools are helpful to provide an improved search experience, they only support a limited set of functionality. Recent studies stress however that search is a highly collaborative endeavor, which in particular involves a large amount of information *sharing* activities. These are currently conducted outside of search tools.³⁰ We argue that social search should also develop a more refined understanding of contributions to the social search process, in particular the contribution of content satisfying the information needs of information seekers.

²⁵This aspect will be in the core of Chapter 6

²⁶See Section 3.1.1

²⁷See, e.g., Churchill et al. (1999); Fidel et al. (2000)

²⁸See, e.g., Pickens et al. (2008); Morris (2008)

²⁹See, e.g., Morris (2008)

³⁰See, e.g., Dearman et al. (e.g., 2008); Evans and Chi (e.g., 2008)

3.2. Knowledge Sharing

If knowledge management is considered the overarching activity to manage knowledge in an organization, knowledge sharing rather addresses the behavior of individual members in an organization (Small and Sage, 2006).

Due to the interdisciplinary nature of KM, the term *knowledge sharing* is also extensively used in related areas such as organizational science,³¹ information systems,³² or CSCW.³³ Surprisingly, the term is mostly used in a casual way and rarely defined. In the context of this thesis, we will therefore use our own definition:

Definition 3.1 (Knowledge Sharing). *Knowledge sharing is the conscious act of an individual to translate and communicate own experiences or knowledge, such that it can be understood by other individuals. Knowledge sharing always includes direct or indirect communication, such that other individuals are able to adapt to the shared knowledge.*

The definition can be detailed as follows:

- On an abstract level, knowledge sharing involves a sending and a receiving person. In practice however, *sender and receiver* can each be multiple persons and also roles may switch.
- With indirect communication, the *receiver might not be known* at the time the sender shares information. In such a setting, the sender might *imagine* a concrete individual, an abstract individual, or nobody particular when sharing knowledge. Similarly, the receiver may or may not lack information about the sender.
- Knowledge sharing can be initiated *proactively* by the sender, *or triggered* by an actual demand (e.g., question) of a receiver.
- Shared knowledge is not communicated in its raw form. It is *modified* by the sharing individual, such that she can expect the receiving individual to understand and process the information.³⁴
- Knowledge sharing requires a *decision* of the sender to share knowledge and the willingness of the receiver to acquire knowledge.
- Knowledge sharing between two individuals is a *discrete communication process* with a start and an end. With indirect communication involved, start and end might be decoupled, and the end time might actually date infinitely into the future.

There are various terms which are sometimes used synonymously, or which are closely related to *knowledge sharing*. Without striving for a complete picture, we give a rough overview:

³¹E.g., Cummings (2004)

³²E.g., Alavi and Leidner (2001)

³³E.g., Hollingshead et al. (2002)

³⁴See also Section 2.2.1

3. Information Seeking and Knowledge Sharing

- *Knowledge transfer* denotes that certain knowledge is “transferred” between individuals or groups. While very similar to knowledge sharing, knowledge transfer is often used on a more abstract, organizational level and does not consider a detailed process of individual sharing behavior.³⁵
- *Knowledge or information exchange* is similar to knowledge sharing in a sense that it considers an act of communication between individuals or groups.³⁶
- *Knowledge creation* typically refers to a longer process of generating a new body of knowledge for a given problem. It can thus be considered a precursor of knowledge sharing. In some cases it might occur, that knowledge creation is triggered by information seeking, and that new knowledge is created in the process of sharing knowledge.³⁷

In the following, we will describe the process of knowledge sharing and highlight some critical issues involved. Afterwards, we will introduce typical knowledge sharing systems and how they deal with those issues.

3.2.1. Knowledge Sharing Process

While there exists a decent amount of models describing the information seeking process (see Section 3.1.1), the process of sharing knowledge has received much less attention so far. Furthermore, existing models typically either take an organizational perspective³⁸ or focus on identifying certain variables which influence individual knowledge sharing behavior,³⁹ instead of describing complex interactions in this process (Small and Sage, 2006).

As a basis for our work, we thus decided to draft a more comprehensive model of a knowledge sharing process. This model aims to provide an initial vocabulary to discuss issues in knowledge sharing, and to compare different methodological and technical approaches. It is informed by the nature of the information seeking process and by the existing body of work about knowledge sharing. Similar to early models of the information seeking process, and to the building blocks of KM by Probst et al. (2006), our model does not intend to be *descriptive* in terms of any real world knowledge sharing behavior.

Our model, which is depicted in Figure 3.5, is centered around the role of an individual *information provider*. The process assumes a decision about if to share knowledge in its core, which separates process steps into a *pre-decision*, a *decision* and a *post-decision* phase.

Similar to the role of the information seeker in the information seeking process, an information provider may initially take an active or a passive role. This means that the knowledge sharing process can either be triggered externally

³⁵See, e.g., Argote (1999); Argote et al. (2003)

³⁶See, e.g., Burnett (2000); Wilson (2010)

³⁷See, e.g., Nonaka (1994); Argote (1999)

³⁸See, e.g., Lehner (2003)

³⁹See, e.g., Wasko and Faraj (2005); Kankanhalli et al. (2005); Mooradian et al. (2006); Quigley et al. (2007); Bock et al. (2010)

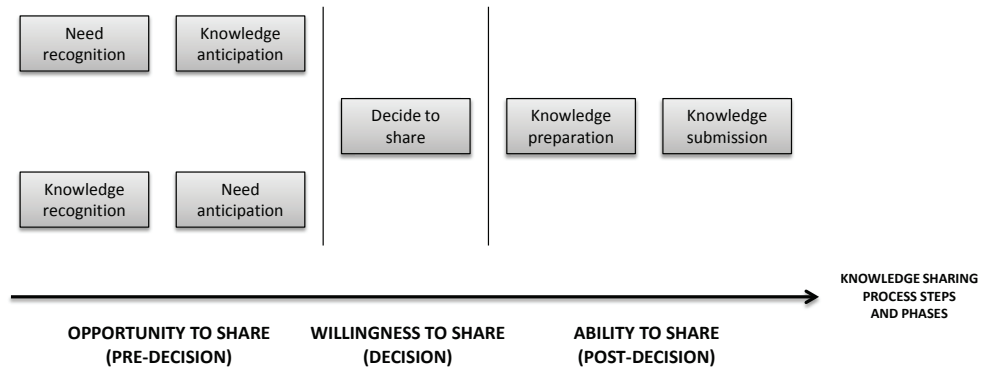


Figure 3.5.: Knowledge Sharing Process

(*pull*) or initiated by the information provider herself (*push*).⁴⁰ Accordingly, two different mental tasks need to be carried out, which are depicted on the left hand side of Figure 3.5.

In the *pull* case, the information provider has to recognize an information need in her environment, upon which she decides to act. The information need can be based on an explicit request of an information seeker, or on the observation, that an implicit need exists for an individual or the organization.⁴¹ Subsequently, the information provider has to decide if she *can* probably satisfy the information need. This step of *knowledge anticipation* involves an estimation, if information available to her can be helpful to satisfy the information need and how it compares to potential alternative information providers.

In the *push* case, an information provider *recognizes knowledge* which is worthwhile to be considered. Recognition might occur due to a serendipitous encounter of knowledge,⁴² or after making a certain special experience. Considering this knowledge leads to an estimation of its *value*, which is derived in *anticipation of information needs*. These needs can be predicted from a personal point of view, or with regard to other people. In the first case, a decision of *keeping* encountered information is cast based on the anticipation of usefulness in the future (Bruce, 2005).⁴³ In the latter case, the consideration process heavily depends on the awareness about others information needs. In both cases, the information provider may decide to *discard* the encountered knowledge or to continue with the knowledge sharing process.

So before casting the actual decision to share knowledge, the information provider has gained an initial understanding of knowledge matching to an

⁴⁰See also Section 2.3.3

⁴¹One can argue that the latter case rather characterizes a *push* behavior. However, in terms of knowledge sharing, the need recognition predates the knowledge recognition. If considered a push behavior, one could coin this case a “need-driven push”.

⁴²See, e.g., Erdelez (2005)

⁴³At a first glance, information kept for personal purposes does not seem to be in the focus for sharing knowledge with others. However, if kept in a private information space, it might be reused in later instantiations of the knowledge sharing process.

3. Information Seeking and Knowledge Sharing

information need. In other words, she has now what we call the *opportunity to share* knowledge.⁴⁴ If this opportunity is taken, depends on the information provider's final *willingness to share*. This is again a complex decision, involving various facets which may influence if the information provider engages in knowledge sharing:

1. The current context or situation
2. A more elaborate estimation of the cost and benefit of sharing⁴⁵
3. The individual motivation, and related cultural norms of the group or organization
4. Affective attitudes concerning the relationship to the information seeker (e.g., trust), or stemming from the personal condition
5. The nature of the information itself (e.g., sensitivity)

As can be seen by the diverse nature of factors, this process step is probably the most complex one in the overall knowledge sharing process and would deserve additional dedication. In fact, the majority of knowledge sharing studies is centering around this focal decision point of reaching the *willingness to share* knowledge. We will discuss some of these findings in the upcoming *decision* section.

After the information provider has decided to share knowledge, the more “practical” or “technical” modalities of the sharing act need to be considered. We call the goal of this phase to reach the *ability to share*. This begins with deciding *what to share* in the step of knowledge preparation. This can involve the encountered knowledge or other existing information. Both might be shared in its raw or in a modified form. Also, the information provider can decide to create a new piece of information or to reuse knowledge from a personal information space. Similar to the *willingness to share*, this decision might be guided by the nature of the information or by information about the intended audience. Activities pursued during this step are also referred to as *packaging* (Markus, 2001) or *audience design* (Rader, 2010). It might also lead to a termination of the sharing process, if the information provider is not able to derive suitable information to share. Furthermore, considering the effort or the nature of the actual information might lead the user to revisit her *willingness* or even the *opportunity to share*.

After the identification of which information to share, the information provider has to decide about suitable means to complete the sharing process (*how to share*). This includes the selection of tools or communication channels to reach the intended audience. This step thus requires various levels of awareness, such as about tools, media and their usage by the intended audience. Also, the information provider needs to consider the actual cost of sharing, which might, for instance, include the cost of setting suitable access rights in a tool.

⁴⁴See also Ipe (2003)

⁴⁵While the initial comparison of cost and value served the purpose of identifying the *opportunity to share*, an elaborate estimation involves if it makes sense to take this opportunity, which includes, e.g., consideration of time, effort and actual benefit for information seekers.

This decision on *how to share* knowledge leads to a final decision to either share or not to share knowledge.⁴⁶

Again, this process is an initial attempt to identify and structure key issues in knowledge sharing. The described steps should thus be considered conceptually and not as a strict linear sequence. Further research is necessary to gain a better and more elaborate understanding of the process. We will now discuss a number of selected issues and problems in more detail.

Context Similar to information seeking, knowledge sharing takes place in a certain context. Several authors have, e.g., stressed the role of organizational culture on individual's willingness to engage in knowledge sharing activities (Orlikowski, 1992). Furthermore, the current work role or task can have an impact. Examples are collaboration on a joint task or project, or mentoring concepts for sharing knowledge with new employees.⁴⁷

A further element of context is the organization of KM. The actual implementation of KM in an organization is called *KM initiative*. As part of its implementation, organizations tend to assign KM tasks either centralized or decentralized (see also Section 2.3.1). In the former case, experts can be identified for certain topics. The important contextual issue is thus, if engagement in knowledge sharing is predefined or open for contributions from all members of an organization.

Awareness Another key concept in each phase of the knowledge sharing process is *awareness*. First of all, users have to be aware about opportunities to share, which breaks down to awareness about their knowledge, and about information needs of information seekers.⁴⁸ The willingness to share can be motivated by awareness about the information seeker (Cuel et al., 2011) and by the availability of means to maintain control about shared information (Voida et al., 2006). Finally, awareness about particular documents that could be shared can influence the ability to share (Whalen et al., 2008b).

Keeping In the information-rich environment of knowledge workers, the identification and preservation of useful information is a key issue. The *usefulness* or value of information is typically measured with respect to its *expected future reuse*.

In the area of PIM/PKM, individual information keeping has received some attention. Rioux (2005) argues, that people build *personal information collections*, which may later reused for sharing with others.

⁴⁶Note that if the channel to share knowledge is personal communication, the information provider might still cancel knowledge sharing even after communication has started.

⁴⁷The concept of *Communities of Practice* has originated from this idea; see also Section 2.2.3

⁴⁸See, e.g., Poltrock et al. (2003); Cataldo et al. (2006); Ko et al. (2007); Herbsleb (2007)

3. Information Seeking and Knowledge Sharing

The aspect of keeping information on behalf of *other people*, i.e., for the purpose of knowledge sharing, is less well understood. Instead of a mere personal decision about the value of a piece of information, a user needs a certain awareness (as mentioned before) regarding others information needs.⁴⁹

Mediation As a communication process, knowledge sharing requires a direct or indirect form of communication (see Section 3.3.1 and also Definition 3.1).

Two aspects are of particular importance in the case of mediation. First, mediated communication may be *synchronous or asynchronous*. As lined out in Section 3.3.2, synchronous communication allows for more direct interaction, while asynchronous communication typically allows for broader reuse. The second aspect is *anonymity*. Since mediation interferes direct communication, it can hide the identity of communication partners, especially in the case of asynchronous communication.

Each of these aspects has an effect on knowledge sharing. Mediation as such *decouples* the direct communication path. Asynchronism leads to a scattering of the communication process over time. Finally, anonymity discards social relations and may reduce trust between communication partners.

Packaging Preparing information for information seekers in the course of knowledge sharing is a multifaceted task. As stated before, knowledge is typically not shared “as such”, but “designed” for an intended audience. Thus, the information provider has to adapt to the information seeker(s) in order to ensure the information can be understood correctly (Rader, 2010).

This is even more difficult in the case of asynchronous knowledge sharing. First, the information provider might not yet know the audience precisely and thus needs to make inferences. Second, the information provider may not know the exact point in time and context, in which information will be reused.

Decision process Finally, the knowledge sharing process can also be regarded as a decision process. This is already reflected in our separation of a *pre-sharing* and *post-sharing* phase in Figure 3.5. The preceding paragraphs have illustrated, that the overall decision *if to share knowledge* can be differentiated in several smaller sub-decisions.

Knowledge sharing research has especially addressed so-called *barriers*, which prevent information providers from sharing, and also devised several categorizations. In the following, we will differentiate between *cognitive*, *social*, *organizational* and *technical* barriers. As various classifications exist, this separation does not claim universal truth, but rather serves as a structure for discussion. Especially, some barriers might be assigned to multiple categories or influence each other.

By *cognitive* barriers we basically refer to a limited *awareness* about the environment in a given knowledge sharing context. As discussed earlier in the

⁴⁹See, e.g., Bernstein et al. (2010)

section, this involves awareness about others information needs, the own capability to contribute, and its relation to the capabilities of other potential contributors.

The category of *social* barriers contains all factors which influence the behavior of individuals by means of social interaction. People may not share, since they do not like to expose their information and expertise to other people.⁵⁰ Furthermore, shared knowledge can typically be considered a *public good*. In order to avoid free-riding (i.e., under-contribution of individual users), people have a sense for *reciprocity*.⁵¹ Related to that, people may experience a lack of personal benefit.⁵² Another factor when sharing knowledge is the loss of control and power.⁵³

Under *organizational* barriers we summarize what is under control of the organization, i.e., parameters which can be actively influenced. One major factor is the general concept of organizational culture.⁵⁴ Furthermore, organizational structure can have severe influence on knowledge sharing by means of physical distance.⁵⁵ A more formal parameter is the time that employees have left for knowledge sharing besides their other work.

Technical barriers mainly concern the knowledge sharing tools. This involves first the mere existence of tools that allow individuals to share their knowledge. Furthermore, these tools need to be usable in order to gain widespread acceptance. Especially, effort for knowledge sharing should be as low as possible. This includes the cost of knowledge capturing, categorization, and setting access rights for documents.⁵⁶

3.2.2. Knowledge Sharing Systems

Tools, especially information and communication technology, are important for mediated knowledge sharing. In a sense, the Web at large can be regarded as an open environment for knowledge sharing. However, while information seeking is broadly supported by tools (see Section 3.1.2), knowledge sharing activities are addressed to a limited extent.

Information systems to support KM activities are typically referred to as *Knowledge Management Systems* (KMS) or *Organizational Memory Systems* (OMS).⁵⁷ Due to the holistic nature of KM, these tools target a broad range of KM activities. Most tools, such as Intranets, file shares, or the Web at large focus on offering *spaces* into which information providers can publish

⁵⁰See, e.g., Ardichvili et al. (2003); Desouza (2003)

⁵¹See, e.g., Ipe (2003)

⁵²See, e.g., Cabrera and Cabrera (2002); Cress and Hesse (2004); Wasko and Faraj (2005)

⁵³See, e.g., Ipe (2003)

⁵⁴See, e.g. Orlikowski (1992); Müller et al. (2005) or (Maier, 2007, p. 223ff)

⁵⁵See Section 2.4.2

⁵⁶See, e.g., Desouza (2003); Desouza and Evaristo (2004); Olson et al. (2005) and Razavi and Iverson (2007)

⁵⁷See, e.g., Lehner (2000); Markus (2001); Alavi and Leidner (2001); Maier (2007)

3. Information Seeking and Knowledge Sharing

information, such that it can be consumed by information seekers. This is often accompanied by certain access control features, which allows information providers to specify the visibility of information. On the other hand, crucial tasks such as *routing*, *keeping* or *audience design* are typically not part of knowledge sharing tools.⁵⁸

In the following, we distinguish between established, “classical” tools and the recent development of social sharing tools inspired by the *Web 2.0* (see also Section 2.4.3).

Classic Knowledge Sharing Tools The basic means to store information on personal computers are electronic files. When personal computers were connected via networks, operating systems allowed to share files with other users, either on their own computer or on a dedicated file server.⁵⁹ Although features for sharing are rather motivated by technical underpinnings of operating systems, file sharing is still very popular for knowledge sharing.

As a response to the limitations of “raw” file sharing, specialized document management systems emerged. Tools such as IBM Notes,⁶⁰ Microsoft SharePoint,⁶¹ or BSCW⁶² offer more usable interfaces and specialized features for file sharing. Examples are easier means of setting access rights (e.g., based on groups), or advanced interaction mechanisms, such as document workflows or change notifications.

Intranets can be considered a further step in the evolution of knowledge sharing tools. Instead of building upon the file metaphor, Intranets are inspired by the Web and its idea of hypertext. Most Intranet tools allow for convenient editing of content directly in the Web browser. However, Intranets are often used as a *push* medium, allowing only a selected set of users to edit content.

Social Sharing In recent years, the paradigms of *Web 2.0* respectively *Enterprise 2.0*⁶³ have sparked tremendous change in the area of knowledge management systems. Various novel genres of tools emerged under the general term “social software”,⁶⁴ including Wikis, Blogs or Social Bookmarking.⁶⁵ All of these tools are stressing the contribution aspect as mandated by the *prosumer* concept of the Web 2.0, as introduced in Section 2.4.3, and adhere to a number of key design principles, which we will shortly discuss.

⁵⁸Exceptions from this rule are *expert finder* (Pipek et al., 2003) or *social tagging* (Millen et al., 2006)

⁵⁹Note that our discussion is focused on file sharing in a fashion also called *enterprise file sharing* (see Section 5.4.3 on Page 106). This has to be distinguished from sharing files for the purpose of exchanging popular media content, sometimes called *illegal file sharing* (Lee, 2003)

⁶⁰See, e.g., Orlikowski (1992)

⁶¹See, e.g., Diffin et al. (2010)

⁶²See, e.g., Bentley et al. (1995)

⁶³See also Section 2.4.3

⁶⁴See Section 2.4.3

⁶⁵See, e.g., Millen et al. (2006); Benz et al. (2010); Dugan et al. (2010)

Openness In contrast to conventional KMS, which consider files or documents authored by an individual as its basic building blocks, social software acknowledges and embraces the evolutionary nature of knowledge and a broad participation by many users. Therefore, social software typically does not implement preemptive access restrictions, but rather allows for the simple contribution of everybody, while malicious edits can be easily reverted as well.

Collaborative authorship The probably most significant paradigm shift concerns the nature of participation and authorship. Instead of few “experts” writing for many knowledge “consumers”, social sharing assumes that expertise is highly scattered. Thus, many people may be able to contribute to a topic. Accordingly, in social sharing systems such as Wikis, content is open to edit for everyone without a primary author acting as a gatekeeper. In the best case, this model of collaboration yields a better allocation of expertise, when many people can contribute to relevant topics.

Incremental improvement Due to the principle of collaborative authorship, social sharing implicitly acknowledges that a certain piece of knowledge is subject to permanent evolution and improvement due to the individual authors’ contributions. Compared to the “old” approach of sharing fully-fledged documents only, this *incremental* way of knowledge sharing has the advantage that intermediary results are visible earlier (and might thus help readers and attract contributors) and improvements might occur continuously. Furthermore, by making this “knowledge maturing”⁶⁶ process more explicit, transparency about the evolution of knowledge is improved.

Many types of contribution Incremental improvement is enabled by the fact, that social sharing platforms allow for many different types of contributions – ranging from contributing actual content to activities such as linking or rating. Thus, the effort required from individual contributors is lowered, while the sum of all contributions can lead to impressive results.

However, social sharing also suffers from a number of drawbacks. The evolutionary character of knowledge, for instance, highlights that the knowledge captured will *never be 100% complete* or perfect – there will always be space for improvements. This in turn makes it difficult for individual contributors to focus and allocate their contributions, such that they provide optimal benefit to the community or organization. Furthermore, the lack of centralized guidance and control may lead to a loss of structure and an increase of outdated content (Happel and Treitz, 2008). In larger communities, this is often compensated by huge organizational efforts. The Wikipedia, for instance, has spent a large amount of resources on discussing and setting up guidelines for the relevance of articles or general quality management.

Summarizing, not only classical KMS systems, but also more recent Enterprise

⁶⁶See Section 2.2.4 and in particular Braun and Schmidt (2007)

3. Information Seeking and Knowledge Sharing

2.0 tools, fall short on supporting many core aspects of knowledge sharing, such as *keeping*, *mediation*, or *packaging*. In particular, information seekers and their particular information needs do not play a major role. The particular problems of file sharing environments and Wikis will be discussed extensively in Section 5.1 resp. 6.1.

3.3. Knowledge Sharing as a Communication Process

The previous two sections described *information seeking* and *knowledge sharing* as two distinct activities – also widely reflected in their corresponding tool suites. However, both are necessarily related due to the mere fact that information seekers and information providers are complementary roles. Depending on the mode of interaction, both roles can tightly intertwine, and a single person can even engage in information seeking and knowledge sharing during one sequence of action.

In this section, we elaborate more deeply on the relationship of information seeking and knowledge sharing processes, and how mediated communication connects both. We will introduce the concepts of *mediation spaces* and *mediation services*, to highlight differences in various mediated communication approaches.

3.3.1. Mediated Knowledge Sharing

On an abstract level, knowledge sharing can be considered as a communication process between information seekers and information providers.⁶⁷ This is evident in most conceptualizations of information seeking (see Figure 3.3) and knowledge sharing (see Section 2.2.1 and in particular Figure 2.3). In this process, information seekers may initiate communication by forming an information demand (or “question”) to satisfy their information needs.⁶⁸ Information providers may then respond with information, based on their interpretation of the demand (“answer”; see also Figure 3.6). Similarly, information providers may “push” information to information seekers in case of an anticipated demand.⁶⁹

In face-to-face communication settings, an information providers’ answer is typically *tailored* towards the request of the information seeker. This is often different in distributed, asynchronous settings, where information providers are *decoupled* from information seekers and their requests. Therefore, *mediation* plays an important role in distributed information seeking and knowledge sharing.

⁶⁷Similarly, one could argue for *information seeking* as a communication process, as both concepts are just different sides of the same coin. However, we prefer to emphasize *knowledge sharing*, as it coincides with a successful information seeking endeavor.

⁶⁸See also the distinction of *information need* and *information demand* discussed in Section 3.1.1

⁶⁹See also Section 2.3.3

3.3. Knowledge Sharing as a Communication Process

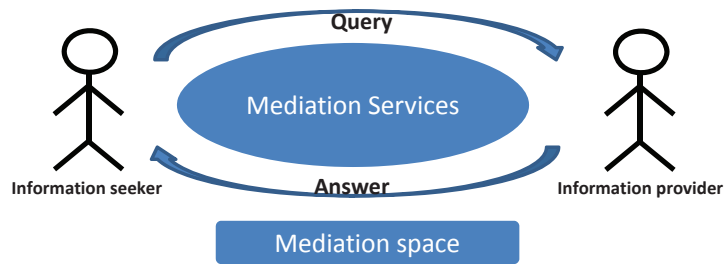


Figure 3.6.: Knowledge Sharing as a Communication Process

In both areas, information seeking and knowledge sharing, this is acknowledged by several authors and approaches, but is not a core concept in many tool implementations, as we have shown previously in this chapter. For the purpose of the further discussion, we introduce a distinction between what we call “mediation services” and “mediation spaces”.

Definition 3.2 (Mediation Service). *A mediation service is a system function which mediates a communication process between two different entities.*

Mediation services can be as simple as a plain communication channel between two actors that seek information resp. share knowledge. They can also involve more complex functionality, such as awareness or notification features.

Definition 3.3 (Mediation Space). *A mediation space is an information space which allows to store information that is primarily used for improving communication processes between different entities.*

Mediation spaces consist of persistent information that is not relevant for satisfying an information need as such, but that captures supplementary information for mediation purposes. Examples for such information are clarification discourses, conceptual mappings, or descriptive metadata like keywords or tags. Examples for mediation spaces are forums or newsgroups, where people can reference and discuss information needs.

Both concepts will be used in the following to discuss different approaches for mediated communication in knowledge sharing.

3.3.2. Mediated Communication Approaches

Communication research typically distinguishes between “face-to-face”-communication and “text-based”⁷⁰ communication approaches (see, e.g.,

⁷⁰Similar to IR (see Section 3.1.2), other media such as audio and video could be mentioned along with text-based communication, although the latter being dominant (see also Dix et al., 2003, p. 495).

3. Information Seeking and Knowledge Sharing

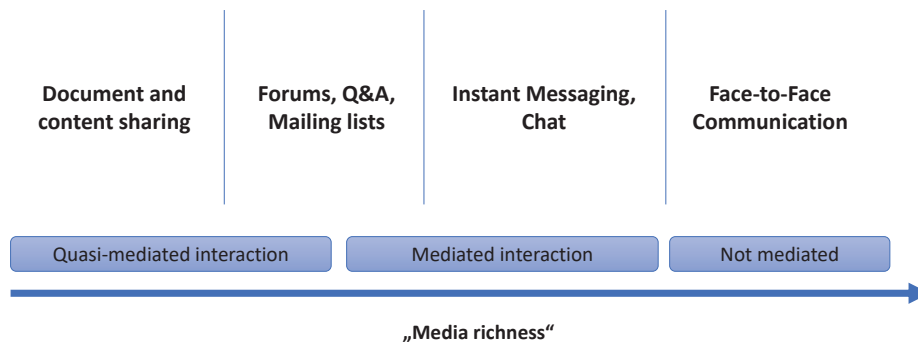


Figure 3.7.: Continuum of Mediated Communication Approaches (partly based on Wouters and Gerbec, 2003)

Dix et al., 2003, p. 476ff). This is related to the distinction of “people-to-people” (personalization) and “people-to-document” (codification) knowledge management, as discussed in Section 2.3.2.⁷¹

For further analysis, we arrange different communication approaches along a continuum, ranging from *text-based* to *face-to-face* communication (see Figure 3.7). Similar to face-to-face communication, *Instant Messaging/Chat* is still a form of synchronous communication, although it is electronically-mediated and text-based. In contrast, *Forums/Q&A/Mailinglists* are asynchronous, while still allowing for conversations between communication partners (see, e.g., Dix et al., 2003, p. 483). This is different for *document and content sharing*, which does not allow for conversation and thus *decouples* communication partners. Thompson (1995) therefore distinguishes *mediated* from *quasi-mediated* interaction, the latter differing in that the audience of the interaction is not known a priori.

In Section 2.3.2 and beyond, we discussed that the choice between personalization and codification is also influenced by organizational context. Similarly, different communication approaches may be considered more or less appropriate for different settings. Media richness theory (Daft and Lengel, 1986) defines a continuum, similar to Figure 3.7, in which face-to-face communication is considered the most “rich” communication mechanism, since it includes multiple means of expression such as voice and body language.⁷² In this model, “richness” is positively associated with communication effectiveness, while efficiency may be affected both positively and negatively.⁷³

⁷¹However, also text-based, electronically-mediated communication – such as Instant Messaging (IM) – can be considered as a tool for personalized knowledge sharing

⁷²Dix et al. (2003, p. 476) argue similarly that “face-to-face is the most sophisticated communication mechanism available” while it is “the most primitive form of communication” (in terms of technology) at the same time.

⁷³Rich communication may be less efficient due to requiring presence or attention while efficiency may benefit from rapid feedback cycles (see also Dennis and Valacich, 1999).

3.3. Knowledge Sharing as a Communication Process

In terms of knowledge sharing, the knowledge maturing framework of Maier and Schmidt (2007)⁷⁴ assigns different communication media to distinct levels of *knowledge maturity*. However, their work is focused on the knowledge creation perspective and does not address any information seeking aspects.⁷⁵

In the following, we want to discuss how different communication approaches compare regarding *mediation services* and *mediation spaces*. Therefore, we distinguish two phases of knowledge sharing, related to Hansen (1999): the phase of searching for knowledge (which we call “pre-transfer”) and the phase of actually transferring knowledge.⁷⁶ For both phases, we discuss mediation services and spaces which address a) the information seeker, b) the information provider and which c) actually mediate between both roles (labeled “matching” resp. “negotiation”).

Face-to-face communication: Face-to-face communication is the primary approach or “gold standard” (Hinds and Kiesler, 2002a) for knowledge sharing in a collocated setting. It is synchronous and usually limited to a small number of people.

Since face-to-face communication happens *in situ*, there is no explicit mediation in the pre-transfer phase. Instead, collocated workers usually have a large shared context,⁷⁷ which allows them to heuristically select persons to ask in case of a certain problem (or to proactively share knowledge with). The effort for provisioning and consuming information in the actual transfer phase is high, since the provider has to invest effort by providing information and the seeker has to pay explicit attention. On the other hand, negotiation is well supported, since both participants can use a broad set of communication means (e.g., voice or gestures) to reach a shared understanding.

Instant Messaging/Chat: Similar to face-to-face communication, electronically-mediated synchronous knowledge sharing using, e.g., instant messaging or chat, is limited to a small number of people. In the pre-transfer phase, there are typically no hints on who could be asked, or what kind of information can be offered. However, complementary electronic tools such as expert finder systems or yellow pages might be used in parallel for mediation purposes. Regarding knowledge transfer, electronically-mediated synchronous communication is less efficient in supporting negotiation, due to a lower communication bandwidth (Olson and Olson, 2000). On the other hand, consuming information is slightly less costly, since reading asynchronous text messages does not require constant attention.

Forums/Q&A/Mailinglists: Forums, question & answer systems, and mailing lists share the characteristics of being asynchronous and being restricted to a small- or medium number of people. Although these media theoretically can support large groups, a specific knowledge sharing incident will

⁷⁴See also Section 2.2.4

⁷⁵While not being the focus of this thesis, combining results from media richness theory, knowledge maturing, and our work could yield a useful framework for guiding the choice of knowledge sharing media in particular settings.

⁷⁶See also Figure 3.5

⁷⁷Which is also an effect of collocation

3. Information Seeking and Knowledge Sharing

get difficult to manage if too many people are involved.⁷⁸ The pre-transfer phase is well supported on the side of the information seeker, since, e.g., forums allow to persistently store demands (and also “offerings” or “announcements”) in a “mediation space” (Burnett, 2000).⁷⁹ This can be leveraged to match suitable information providers. Similarly, answers are persisted in a shared information space.⁸⁰ As for the transfer of information, the situation is similar to Instant Messaging/Chat, while provisioning might be slightly easier – e.g., due to reusing or referencing earlier or similar contributions.⁸¹

Document and content sharing: While (asynchronous) conversations between information seekers and providers are constituent for Forums/Q&A/Mailinglists, document and content (DCS) sharing *decouples* information seekers from direct interaction with information providers. Particular examples for DCS include file shares,⁸² groupware systems, Wikis, or knowledge repositories.⁸³ Transferred knowledge is typically kept within a shared information space and not bound to a specific seeker or contributor. Thus, knowledge is typically less specific for a given case, and covers rather generalized issues. Therefore, document and content sharing scales up very well, since the number of seekers is not limited by the communication bandwidth and/or attention of an information provider.

Document and content sharing barely supports any pre-transfer mediation, as there is typically no way for information seekers to express and persist particular information needs. On the other hand, contributors implicitly offer knowledge when exposing documents or content in a shared information space. Information consumption is solely up to the information seeker, once she has identified suitable information. For information providers, disseminating shared information is quite easy, while there is usually no guidance about what to share. Due to the decoupling of information seekers and providers, bridging terminological problems can be a major issue, since there is no way to establish a shared mental model in joint discourse.⁸⁴

To summarize, the synchronized (personalization-oriented) approaches in general allow for sharing *specific* knowledge with rich negotiation opportunities, while they do not easily scale up to larger numbers of people and do not support the pre-transfer phase. Asynchronous (codification-oriented) approaches in turn, are less suitable for individual needs, but allow for disseminating information to a larger audience more efficiently. *Document and content sharing* scales up particularly well, but is less suited to satisfy more specific needs in a considerable time. *Forums/Q&A systems and mailing lists* are scaling

⁷⁸This argument refers to the actual exchange between an information seeker and information provider. Later-on, when the initial information need is satisfied, many additional information seekers might reuse the information with low additional effort.

⁷⁹As just defined in Section 3.3.1

⁸⁰See also Section 4.2.2

⁸¹See also Section 6.4.2 for a discussion of *Collaborative Question Answering* systems

⁸²See also Sections 3.5 and 5.4.3

⁸³Note that especially groupware systems and knowledge repositories are rather generic concepts that may also include elements of other communication media such as forums.

⁸⁴This is one reason why metadata annotation techniques received lots of attention in the context of search to allow for alternative descriptions of meaning.

up to a medium level. They are unique in providing rich means for demand specification in the pre-transfer phase. Therefore, it is, e.g., easy to compile “Frequently Asked Questions” (FAQs) out of such systems.

In terms of mediation services and mediation spaces, *document and content sharing* falls particularly short. A major mediation service is *search*, although it is not always a core component of a DCS system, but may be an external component.⁸⁵ As for *mediation spaces*, few DCS systems – such as BSCW or SharePoint (see Section 3.2.2) – do support lightweight mediation data like tags or rating information. Accordingly, researchers have identified file and folder names as a core means for mediating between information providers and seekers.⁸⁶

A potential reason for this lack of mediation features might be that both, information providers and seekers, experience DCS tools individually. Communication is not a core design feature, since DCS tools can also “work” as a personal tools for the individual user storing and retrieving files. Ironically, the strong *decoupling* of both roles would make appropriate mediation support even more crucial.⁸⁷

3.4. Summary

In this chapter, we introduced the current state of the art in the areas of *information seeking and retrieval* and *knowledge sharing*. We also highlighted in Section 3.3.1, that both areas address tightly related roles when considering knowledge sharing as a communication process.

A key observation is, that both areas seem to widely neglect their mutual relationship. In *information seeking and retrieval*, the aspect of information provisioning is only a peripheral topic. Similarly, *knowledge sharing* research and tools do not have an elaborated conceptualization of the demand-side of knowledge sharing. To *bridge this conceptual gap*, we point out a number of design aspects that should be addressed by future information seeking and knowledge sharing systems.

Concerning *information seeking and retrieval*, this includes:

Lack of demand treatment: Tools should consider information demands as a resource within the knowledge sharing process. Users should be given additional means to express their information needs.

Lack of awareness: Users need a better awareness about available information and features of the information system.

⁸⁵Like Web search in the case of the WWW, or enterprise search in the case of enterprise file shares

⁸⁶See, e.g., Rader (2010) and the discussion about *packaging* on Page 48

⁸⁷Muller, Shami, Millen and Feinberg (2010) argue in a similar way that “most design decision [in enterprise file sharing systems] are motivated by active roles (uploaders/contributors) even though minorities” and that the work of lurkers and consumption thus in general remains invisible

3. Information Seeking and Knowledge Sharing

Lack of possible actions: Most conventional tools do not allow many activities besides information seeking. However, we have shown that information seeking is not an isolated process, but may also include collaboration with other information seekers, or switching to the role of an information provider.

With respect to *knowledge sharing*, we see the following issues:

Lack of awareness: Tools lack means to make information providers aware of the needs of information seekers and about the capabilities of other information providers.

Lack of guidance: Tools lack functionality that allows information providers to efficiently share knowledge. Support should be given for the selection, authoring, and distribution of information.

Lack of privacy: Most tools require the information provider to disclose information into a shared information space before it can be accessed by information seekers. However, information providers might prefer to keep information private, e.g., if it is sensitive, or if access rights are difficult to set.

In the following chapter, we will introduce a knowledge sharing framework which addresses these issues. The subsequent chapters 5 to 7 will then showcase how particular existing tools are affected from the mentioned issues, and how they can be improved in order to deal with them.

4. Need-driven Knowledge Sharing

In the previous two chapters, we have concluded that the two major roles in knowledge sharing processes are largely *decoupled* from each other. Especially in asynchronous settings, *information providers* often share knowledge without a detailed consideration of actual needs. Conversely, *information seekers* are not capable of communicating their needs adequately.

To overcome these limitations, we will introduce the concept of *need-driven knowledge sharing* (NKS), which is based on the concepts of *mediation services* and *mediation spaces*, which were just introduced in Section 3.3.

Need-driven knowledge sharing refers to the fact, that knowledge shared by someone with another person is tailored towards an actual information need of that person – i.e., ideally satisfies the information need of that person, like it is shown in Figure 3.1. As discussed in Section 3.3.2, this is typically the case in face-to-face conversations, where people interact directly.

While thus “need-driven-ness” appears to be a common case, it turns out that knowledge sharing has become less need-driven with the increased usage of (computer-)mediated communication. In particular, the asynchronous sharing of codified information, which we labeled as “document and content sharing” (DCS), is rarely “need-driven” at all, as we have described in the previous chapter.

Since DCS is an important means for knowledge sharing in distributed teams, we want to explore how the *efficiency* of DCS and the *effectiveness* of need-driven knowledge sharing can be combined. In particular, we want to design *mediation services* and *mediation spaces* which alleviate for the disadvantages of DCS by making it more need-driven.

After a more detailed definition of NKS and its goals, we contrast its underlying assumptions with existing literature. We then introduce the NKS framework and its implementation, which will be used as a basis for subsequent chapters.

Early foundations for the concept of need-driven knowledge sharing haven been described in Happel and Stojanovic (2008) and Happel (2009b).

4.1. Definition

In this section, we will define the underpinnings of NKS. We start with defining the concept and explain the *problem setting* addressed as well as the *scope* and *goals* of the approach in some detail.

4. Need-driven Knowledge Sharing

Definition 4.1 (Need-Driven Knowledge Sharing). *Need-Driven Knowledge Sharing (NKS) is an approach to knowledge sharing in groups, which allows information providers to share knowledge based on the implicit or explicit needs of information seekers.*

This deviates from the Definition 3.1 (*knowledge sharing*) as follows:

- NKS does not primarily target to satisfy single information requests (or *demands*) of individuals, but rather targets abstract *needs* that have relevance for a group.
- NKS is intended for asynchronous and mediated knowledge sharing. It typically involves indirect communication.
- Knowledge sharing can be initiated proactively by the information provider, or triggered by the information need of a group.
- NKS assumes that the information provider has a conceptualization of the information needs and the common knowledge of the intended set of information seekers, such that she can expect the majority of them to understand the information.¹

Note that NKS puts a strong emphasis on the role of the information provider and the kind of knowledge to be shared. Unlike common approaches to information seeking or knowledge sharing (see Chapter 3), NKS includes information seekers and providers in a *single approach*.

Based on this definition, we will now line out the scope and goals of NKS in more detail.

4.1.1. Goals

The basic objectives of NKS are to make the sharing of codified knowledge in groups more effective and efficient. *Effective* means, that the limited resources for knowledge sharing should be allocated to the knowledge which is of *highest value* for the overall group. Assuming that existing knowledge sharing in groups is not 100% effective in the sense that some knowledge produced is never consumed, prioritization can also help to reduce the resources required for sharing valuable knowledge. As a consequence, knowledge sharing gets more *efficient*.

According to Definition 4.1, this shall be achieved by better *mediating* between information seekers and providers. Particular goals of NKS can be derived for different roles and perspectives:

- *Information providers* shall receive better *guidance* about which knowledge is sought. Also, their *motivation* to share knowledge shall be increased.
- *Information seekers* shall receive better *satisfaction* of their information needs. As a supportive measure, they should also receive improved means to *describe* their information needs.

¹See also Section 2.2.1

- From an *organizational or group perspective*, there should be transparency about the relation of information needs and knowledge shared to allow for management interventions. NKS should allow organizations to optimize their flow of knowledge (i.e., to solve the trade-off between limited resources and knowledge sharing).

Note that these goals primarily apply to the conditions described in the following section. By no means, NKS is intended to solve *all* KM problems of an organization or group.

From a technical perspective, NKS should tightly integrate with existing systems or even improve them directly, instead of revolutionizing the IT landscape.

4.1.2. Scope

NKS is not meant as an approach suitable for all possible settings or types of knowledge. This section defines the scope of NKS based on a number of different dimensions. However, this does not generally rule out the application of NKS in situations with different properties.

Organizational NKS is especially intended for large and distributed professional groups. The focus on *large* groups is due to the fact that NKS requires the opportunity for building up organizational knowledge² as well as a certain amount of potential information providers. Thus, we define groups of 20-200 people as an ideal target range. In terms of *distribution*, we expect group members to be separated either physically or organizational to make knowledge sharing barriers (see Section 3.2.1) effective. For larger teams, a certain amount of distribution will be given implicitly anyway.

We target *groups* in the sense that we require a certain amount of *shared context* among the users. Any social entity providing such context – may it be project teams, organizations, or communities – can satisfy this condition. Finally, our focus for NKS is on *professional* groups. The main reason for this is, that knowledge sharing in professional work settings probably accounts for the largest part of people’s overall knowledge sharing. Also, work environments can be assumed to provide a more homogeneous organizational and technical setting when compared to spare-time activities.

Cultural The core idea of NKS is to guide knowledge sharing by collective needs. Also, knowledge sharing is intended to be decentralized – i.e., any group member should generally be able and willing to act as an information provider. Thus, NKS requires an open and trustful organizational culture.

In more competitive or bureaucratic organizational settings, NKS might not be as effective as intended. This is tightly related to the values and principles underlying the Enterprise 2.0 paradigm, as introduced in Section 2.4.3.

²One could similarly argue that smaller teams usually do not need knowledge management in a classical sense. For *organizational knowledge* see also Section 2.2.3f.

4. Need-driven Knowledge Sharing

Informational NKS targets a special subset of information needs, which we call *emergent needs*. Such needs dynamically arise in work situations and can not be easily predicted.³ The individual is expected to actively engage in information seeking, requiring a high level of information literacy: “In the case of a need, [the user] shall be able to request required knowledge fast. Making directed knowledge requests shall become second nature to him.” (Probst et al., 2006, p. 151; translated by author). Accordingly, “pull”-style explicit demand of information is the typical case.⁴

While single information demands are dynamic and individual, NKS assumes that a relevant share of demands concerns information that can be considered relevant for the organization or parts of it. This assumption is based on the fact that users are working on overlapping tasks within a shared context. Thus, different individuals are expected to raise similar information needs. Also, NKS assumes that many information needs occur during a larger period of time, with different individuals having similar information needs at different times.⁵

Expertise When it comes to the satisfaction of information needs, NKS assumes that the expertise, to provide suitable information in response to information demands, is *scattered* across the organization. This implies, that each individual in the organization is relevant as an information provider, and that it is difficult for information seekers to identify a suitable information provider for a particular information need. Due to this “decentralized” setting, we expect that individuals have limited resources for information provisioning, since it is not a dominant part of their job description.

While information providers might share knowledge verbally, we expect that a large fraction of knowledge is shared in a codified fashion due to the distributed nature of the organization. We also expect that the existing organizational knowledge base is inherently *incomplete* (i.e., not able to satisfy all information requests) and constantly evolving.

Technical Technically, we assume that information seekers and providers targeted by NKS have access to a common information system. Users should also be able to consult this information system as information needs occur, i.e., in the context of their daily working activities. We also assume this information system to be centralized, i.e., there should be a single, server-based instance for all users. This assumption is not a conceptual but a pragmatic one, since concepts within this work will be based on a centralized technical architecture. While we acknowledge that distributed technical architectures might even be beneficial for our approach,⁶ suitable realizations need serious investigation and are thus subject to future research.

³Such as *knowledge work* as described in Section 2.4

⁴See also Sections 2.3.1 and 2.3.3

⁵This is, e.g., supported by Heisig et al. (2010), who found that information needs in engineering design show certain patterns of regularity

⁶See also Section 2.3.1

Problem Domains As indicated by the organizational and technical dimensions, we are basically targeting *knowledge workers* that have constant access to information systems. Organizations should stem from domains with a low level of process standardization, and a dynamic work environment which exposes people to frequent emergent information needs (see also Section 2.4.1).

Exemplary domains are software development, any other kind of engineering activities, research, or consulting. We assume that such domains are characterized by a high level of innovation, which makes the maintenance of an organizational knowledge base a particularly difficult and ongoing endeavor.

4.2. Assumptions

In this section, we will elaborate in more detail on some assumptions about information needs that were described in the previous section. In particular, we will discuss evidence from existing research studies and introduce the notions of *organizational information needs* and *organizational information gaps*. Finally, we describe several concrete practical application scenarios for our approach.

For our discussion we will mainly consider *information needs used for querying IR systems*.⁷ This is due to the fact, that searching activities are ubiquitous in modern IT-enabled work environments and accordingly the majority of empirical studies has been conducted in this area. However, as Chapter 7 will showcase, the NKS approach is not limited to a keyword-based IR paradigm.

4.2.1. Organizational Information Needs

We have three main assumptions about information needs that can be made fruitful for knowledge sharing. The first is, that some needs remain *unsatisfied* in the sense that emergent information demands can not be satisfied by existing information. This is a consequence of the assumed incompleteness of the organizational knowledge base. Furthermore, we assume that some needs *recur* over time and are *shared* by different information seekers in an organization. Well will now discuss these three assumptions in more detail.

Unsatisfied Information Needs NKS assumes that a significant amount of information needs (which is expressed as information demands) remains *unsatisfied*. Related to IR systems this means, that the information need of a user can not be resolved based on the results returned by the system. The following definition can be considered a more elaborate specification of the corresponding box in Figure 3.1 (“Satisfaction or Non-Satisfaction”):

⁷See Section 3.1.2

4. Need-driven Knowledge Sharing

Definition 4.2 (Unsatisfied Information Needs). *An information need of a user is considered unsatisfied, as long as the user is not able to retrieve any information to satisfy the need. This may be due to the failure to retrieve any meaningful information, or due to insufficient information contained in material which was retrieved.*

Applied to IR systems, the definition can be illustrated as follows. The relationship between the *documents matched by queries* that were executed against an IR system, the documents *accessible* by that system, and the set of *all documents* is depicted in Figure 4.1.

The intersection of matching and accessible documents (*A*) denotes that these documents have been part of query results.

Accordingly, $C \cup E$ denotes the set of documents, which would have been relevant for a certain set of prior queries Q_{CE} . However, these documents were not part of query result sets, since they were not accessible for the IR system (*C*; e.g., due to being located in a private folder) or did not exist (*E*). Due to those missing documents, we can consider Q_{CE} as *partially satisfied* queries. Each query $q_{CE} \in Q_{CE}$ which does not lead to the satisfaction of an information demand, according to Figure 3.1, can be considered as *unsatisfied* query.

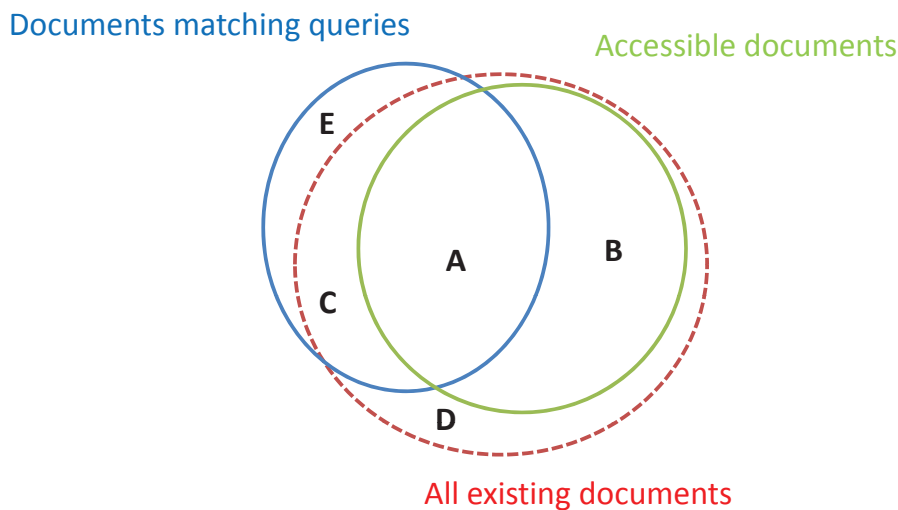


Figure 4.1.: Interrelation of different Sets of Documents for a certain IR system

Table 4.1 describes the different sets of documents and suggests implications from a knowledge sharing perspective for each set. In this context, *annotation* refers to the practice of adding descriptive metadata to a document, which can help matching certain documents which are relevant queries, even if the original document content does not match to those queries (e.g., due different, but synonym terms used in query and document).

As discussed in Section 3.1, IR research tends to address the set of documents (also known as *corpus*) from a technical perspective. It is often treated as

a fixed entity, without any consideration of how and why documents came into being. This is stressed by the fact, that the major evaluation criteria for IR systems, *precision* and *recall*,⁸ focus on the number of relevant documents returned from the *given* corpus ($A \cup B$ in Figure 4.1).

However, at least some researchers have been addressing unsatisfied information needs under the term *failed searches*. Stamou and Efthimiadis (2009) analyzed 908 queries issued by 38 postgraduate students in their favorite Web search engine during one day. The authors put a focus on queries that were not followed by result clicks, which accounted for roughly 13% of the total set.

Out of these, the authors further distinguished those queries which remained without a result click *intentionally* (i.e., because the information need could be satisfied from the result preview snippet) versus those, which remained without a click *unintentionally*, which means that users were not able to satisfy their information need. This latter case applied for 61 queries, which is roughly half of the queries without clicks. Out of these 61 cases (which account for 6.72% of all queries), 9 queries did not have any result at all, 38 did not have any result that seemed *relevant* to the user, and 8 only returned results *already known* to the user.⁹

In a follow-up study, Stamou and Efthimiadis (2010) monitored the queries of six students during one week, resulting in 966 queries. About 27% (261) of these queries were not followed by a result click. Out of these, 81 (8.39%) could not be satisfied.¹⁰

Smyth, Boydell et al. (2005) analyzed the Web searches of a small software company with 50 users over a period of four weeks. Out of the 1572 queries, 32.95% were classified as failed, since they had no follow-up result clicks.

Accordingly, *failed searches*, which we assume to mostly result in *unsatisfied information needs*, can be considered an important problem in Web search. Contrasting the vast amount of documents on the Web with the situation in closed organizations or groups, we assume that this issue is even worse in such closed settings.

Recurring Information Needs Based on the assumption of *unsatisfied* information needs, we add that a certain amount of information needs is *recurring* over time, as elaborated in the following definition:

Definition 4.3 (Recurring Information Needs). *An information need is called recurring, if it actually occurs or can be expected to occur at multiple, separate points of time.*

Recurrence is a crucial property of information needs, since knowledge management puts a strong emphasis on the actual *use* of knowledge which is shared.¹¹

⁸See, e.g., Baeza-Yates and Riberio-Neto (1999, p.75ff)

⁹In the remaining 6 cases, the query session had been interrupted.

¹⁰Due to a different methodology employed, about half of these unsatisfied queries stem from queries with *pre-determined inactivity* – i.e., the information need could not be satisfied from the result snippet *as intended*.

¹¹See also Section 2.1.1

4. Need-driven Knowledge Sharing

Area	Description	Implication
A	Documents that are queried	-
B	Documents not matching any query	Archive or annotate
C	Document not accessible	Share (or create) document
D	Documents not available and not matching any query	Irrelevant private documents
E	No documents matching a query exist	Create documents, annotate existing documents

Table 4.1.: Explanation of the sets depicted in Figure 4.1

Only if some *re-use* can be expected, knowledge sharing will pay-off, as inherent in the definition of organizational knowledge.¹²

There are several causes for recurring information needs:

- The limited ability of users to store and remember information may require to satisfy some information needs over and over again.¹³
- Users may retry to satisfy previously unsatisfied information needs.¹⁴
- Even for some satisfied information needs, users may like to retrieve additional information, that was not available when the initial information seeking took place.¹⁵

The mere existence of features such as Google Suggest,¹⁶ which recommend common queries from past searches, provides initial evidence for recurring information needs. There are also several studies supporting this assumption.

Teevan, Adar, Jones and Potts (2006) analyzed the query repetition patterns in the Yahoo query logs for 114 users over a year. Also syntactically different queries were counted as repetitions, if they led to clicks on the same result. Repetitions within 30 minutes were dismissed from the analysis. In total, 13 060 queries were executed. 40% of the queries led to a click that was repeated in another session. In 71% of these cases, the same query string was used. Conversely, 87% of identical queries led to repeat click. Both facts indicate a large number of information needs which recur over time, both based on the similarity of the query string and the similarity of subsequent result clicks.

Furthermore, the authors distinguish repetitions made to re-find already known results from those made for other purposes (e.g., for finding new results). Out of the 40% of queries with repeat clicks in other sessions, 14% also involved clicking at least one *new* result, while 47% were classified as “navigational” (Broder, 2002) , since only one specific result was clicked during

¹²See Section 2.2.3

¹³Also called “re-finding”; see, e.g., Adar et al. (2008); Teevan et al. (2007); Tyler and Teevan (2010)

¹⁴See also “prospective search” in Section 3.1.2

¹⁵In IR literature this has been coined *standing interest* (Yang and Jeh, 2006)

¹⁶<http://www.google.com/support/websearch/bin/answer.py?hlrm=de&answer=106230>

repetitions. Conversely, 38% of identical queries resulted in clicks on different results, while 25% of the searches involved both, a repeat and a unique click.

Zhang and Lu (2009) performed a similar analysis using 1.9 million queries by 300 000 users from the AOL data set. Measuring the frequency and the recency of a query in history, they found that more frequently and more recently used queries are more likely to recur – i.e.. that “query frequency and recency are indeed highly useful predictors for[...] query recurrence”.

Smyth, Balfe et al. (2005) did an extensive analysis of query repetition using five different data sets with together more than 100 000 queries. While two of the data sets covered “general interest” Web search, three sets covered more special information needs. When considering exact duplicate queries, around 15% of queries in the general interest sets and 55-60% of queries in the specialized data sets occurred at least twice. When considering a similarity metric based on common terms between two queries, 75% of all queries in the general interest sets, and around 90% of queries from the specialized sets, shared at least one term with other queries. In a second study with a small software company (50 users, nine weeks), Smyth, Boydell et al. (2005) found that 60% of all queries shared at least 50% of their terms with other queries.

Summarizing, considerable regularity can be observed when analyzing information needs as expressed by keyword queries. Since most of the reported studies are based on general Web search, one may assume that the amount of recurring information needs in organizational settings is even higher, given the shared working context and the repetitive nature of many organizational processes.

Shared Information Needs Finally, we assume information needs to be *shared*, if they are expressed by different users in an organization or group. Tightly related to the previous discussion of *recurrence*, a certain organizational relevance of information needs is an elementary precondition for any knowledge management activity: the more people that have a similar information need, the more probable it gets, that related information will be (re-)used. Following this argumentation, it would also be desired to rank information needs by the number of people sharing them – especially in group settings with restricted resources for knowledge sharing. Several studies reported in the previous paragraph make statements on the “commonality” of information needs, as we describe in the following.

In the study of Teevan, Adar, Jones and Potts (2006) 18% of all 13 060 queries were repetitions of other users’ queries. Furthermore, 28% of clicks were on URLs clicked more than once by a user, and 7% of URLs were clicked by multiple users.

In their analysis of query recurrence, Zhang and Lu (2009) made a distinction between *individual* and *collective* query recurrence. The latter denotes if any other user issued the same query before. The authors found out that especially query frequency is useful for predicting collective query recurrence, while query

4. Need-driven Knowledge Sharing

recency is useful for predicting individual query recurrence – due to greater drift in individual user’s interests, as the authors assume.

Definition 4.4 (Organizational Information Need). *An organizational information need (OIN) is an information need which is shared by multiple members of an organization or group.*

Based on the concept of (personal) information need introduced in Section 3.1.1, we conceptualize *organizational information need* (OIN)¹⁷ as an aggregate of the personal information needs of members in a group. By *group* we mean the group of users which are able to access a certain information system. Depending on the concrete setup, this can be a team, an organization, or the population of Web users as a whole.¹⁸ The organizational information need thus denotes the overall amount of information, which is required by the members of this group to complete their particular tasks. In terms of keyword searches, an OIN should thus represent the most frequent queries that have been executed by users throughout the organization.

Thus, our approach is similar to the approach used by tools like Google’s Zeitgeist (Google Inc., 2017b), which ranks the queries of all Google users by their popularity. However, while Zeitgeist is rather a descriptive tool to highlight popular queries, we think that the aggregated information need from a limited set of users (such as in an organization) allows for more meaningful interpretations.

This section has shown, that there exist information needs which are *unsatisfied*, *recurring*, and *shared*. Notably, the majority of empirical evidence has been gathered in Web search scenarios with a large document and user base. It can be assumed that organizational environments may even have a higher rate of such needs, due to more homogeneous users and a smaller volume of available information.

What remains unanswered is the question how often the described cases occur. Although some studies show, that recurring queries are often executed by different users, there is no evidence how many of these queries remain unsatisfied.

4.2.2. Organizational Information Gaps

In this section, we elaborate on our main assumptions regarding the *provisioning* (or “contribution”) role in knowledge sharing.

As stated for the aspect of expertise in Section 4.1.2, we assume that information is scattered within the organization, resulting in so called *organizational information gaps*. We furthermore assume that, given a suitable organizational culture, information providers *like to share* knowledge with others. We will now discuss both aspects in more detail.

¹⁷Or *aggregate information need* (AIN)

¹⁸See also our discussion of *scope* in Section 4.1.2

Information Gaps in Teams Information needs can typically be satisfied either from own *private information spaces* or from information available in *shared information spaces*. The private spaces of other users are usually not accessible, although they might contain relevant information.¹⁹

Thus, when considering the overall information space of an organization, this includes the private space of each member plus the public space, which can be accessed by all members of the organization. Note that this is a simplification since there might exist shared information spaces with different access rights. Since a basic definition is suitable for the purpose of this thesis, a more elaborate consideration of information spaces is left to future work.

The following definitions can be derived:

Definition 4.5 (Private Information Space). *The private information space contains the set of information which is accessible solely by its owner.*

Particular examples for personal information spaces can be files stored on a local computer, or emails in ones email account. As discussed in Section 2.3.4, researchers in the domains of Personal Information Management and Personal Knowledge Management pay particular attention to private information.

Definition 4.6 (Shared Information Space). *A shared information space contains information which is not in the private information space of a users, but which can be accessed by her. This includes information shared by other users and information shared within the whole organization (“public information”).*

Examples for shared information spaces are private spaces shared to other users by means of access permissions (e.g., sharing a certain folder), or explicitly shared spaces such as enterprise file shares or Knowledge Management Systems (see also Section 3.2.2).

Together, private information spaces of all users and shared information spaces form a “virtual” organizational information space, containing all information which is theoretically available within an organization:

Definition 4.7 (Organizational Information Space). *The organizational information space consists of the private information spaces of all members of an organization plus the organization’s public information space.*

Regarding the available information which could satisfy a specific personal information need, we expect an unequal distribution of information across an organization. This means, that there are significant differences between information shared by all members, and information that is kept private by the individuals.

¹⁹See Figure 4.1

4. Need-driven Knowledge Sharing

We derive a simple model of information distribution in order to illustrate this. In our model, we distinguish between four general situations as depicted in Figure 4.2:

Information overload denotes the situation when there is enough information concerning a certain information need in the private and the shared space. Thus, if users seek to satisfy an information need, they will be able to retrieve information from multiple sources.

Information shortage characterizes the opposite situation, when there exists only a little amount of information, both in private and shared spaces. From an organizational perspective, this could be a signal to invest in knowledge creation, given there is demand for the information.²⁰

Personal information gap can be identified, if there is plenty of information available in the shared space, while the user has only a little information amount of information in her private space. Thus, the user has to rely on information from the shared space in order to satisfy an information need.

Organizational information gap finally describes the situation, when the user has lots of information regarding a certain information need, while there is only a little information available in shared spaces. This means that other users, searching for that information, might not be able to satisfy their information need, although there is information in the private space of at least one user of the organization.

Definition 4.8 (Organizational Information Gap). *An organizational information gap (OIG) describes the situation when individual people have significantly more information regarding a certain information need within their private information space, compared to information in the shared information space accessible to all members of the organization.*

We are not aware of any empirical studies which have been conducted to analyze the difference between private and shared information spaces.

From an organizational knowledge sharing perspective, the concept of organizational information gaps raises the issue how to determine which information from private spaces should be made available to the organization (“need to share”) and how this diffusion can be achieved, in order to allow satisfying other users’ information needs.²¹

People Like to Share In Section 3.2.1, we argued that the knowledge sharing process contains several *barriers* which prohibit individuals from sharing knowledge. We now take the complementary perspective by discussing means to lower or overcome such barriers.

²⁰See also Section 2.1.1

²¹See also research question RQ 1 in Section 1.1.2

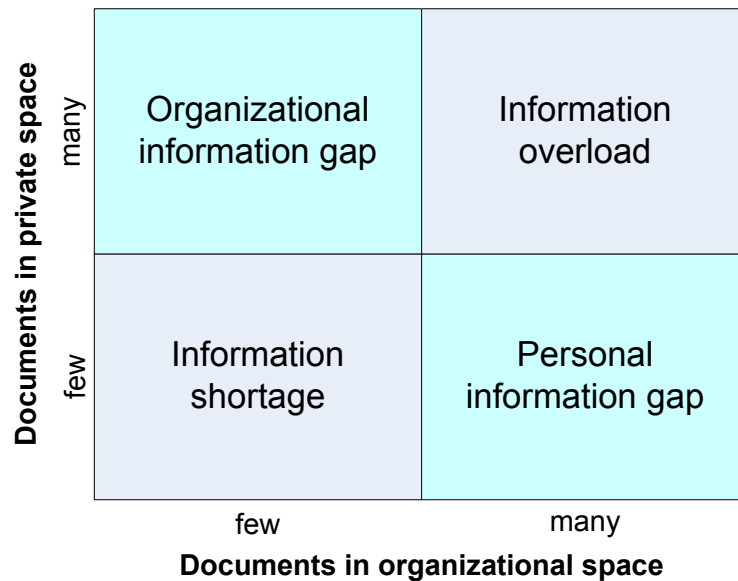


Figure 4.2.: Availability of Information related to some particular Information Need in the Private vs. Organizational Information Space

Research in this area can be separated in descriptive studies and experimental studies, which investigate the effect of certain design interventions. Descriptive studies such as Olson et al. (2005) or Dearman et al. (2008) are often based on diaries or interviews.

Experimental studies, on the other hand, are often based on theories from areas like social psychology (see, e.g., Beenen et al., 2004; Cheshire and Antin, 2008). Authors have argued, that people can be influenced to contribute more actively by *designing* the contribution environment. Accordingly, they conceive experimental setups in which certain parameters of the environment (such as a user interface) are modified. The underlying assumption of these approaches is, that people generally like to contribute, or can at least be motivated to contribute more based on specific design interventions.

According to existing research, contributions can be raised by:

- A simple trigger such as an email (Beenen et al., 2004)
- Offering personal or social advantages to users due to contributions²²
- Setting and communicating contribution goals (Beenen et al., 2004; Jung et al., 2010)
- Giving feedback to contributors²³
- Contribution meets concrete demands²⁴

²²See, e.g., Kustanowitz and Shneiderman (2005); Ames and Naaman (2007)

²³See, e.g., Cheshire and Antin (2008); Jung et al. (2010); Mazarakis and van Dinther (2011)

²⁴See, e.g., Olson et al. (2005); Dearman et al. (2008)

4. *Need-driven Knowledge Sharing*

- Highlighting the relative value of a contribution for all users of subgroups (Rashid et al., 2006)
- Highlighting the uniqueness of a contribution (Beenen et al., 2004)

According to this existing body of knowledge, it seems worthwhile to explore if a notion of other users' information needs as presented in Section 4.2.1, combined with a relative "value" of certain information (e.g., as identified by information gaps), could be fruitful to engage information providers in sharing information.

Summarizing this section, we assume that information is not evenly distributed throughout the information spaces of individual users and their organization. This implies that a situation which we coined *organizational information gap* can occur, in which certain information needs remain unsatisfied, although related information would exist in the private information space of one or more other users.

Besides, we described that research on individual sharing behavior indicates that people like to share information that might satisfy other users' demands. It thus seems worthwhile to explore, if information contained in a user's private information space could help closing *organizational information gaps*.

4.3. Framework

In the previous sections, the *decoupling* of information seekers and information providers has been identified as a major problem for asynchronous, codified knowledge sharing. Chapter 3 has also shown that both, information seeking and knowledge sharing, lack proper conceptualizations of each other – i.e., information seeking does not consider where sought information stems from, while knowledge sharing has a limited understanding of actual information needs.

The idea of NKS is to establish a proper link between information seeking and knowledge sharing to make the latter more *need-driven*. Due to the focus on asynchronicity, *mediation* between both roles plays a major role. Therefore, we propose particular *mediation services* and *mediation spaces* to the knowledge sharing model introduced in Section 3.3.1. We also describe an abstract model of information needs, which serves as a basis for the implementation of NKS in actual systems.

4.3.1. Design Principles

In the following, we introduce several measures to address limitations of existing support for *information seekers* and *information providers* as we identified in Section 3.4.

Information Provider As lined out in Section 3.2, the conceptualization of information providers and the process of knowledge sharing is still in its infancy. We argue that the following design principles can help to improve the knowledge sharing process:

- Make information providers aware of existing information needs and the value of the information they could possibly provide.
- Offer concrete guidance about which information to share and how to share it.
- Overcome knowledge sharing barriers like the desire for privacy or the effort to share.

Information Seeker Section 3.1 has highlighted limitations in the information seeking process. From a knowledge management perspective, especially the consideration of information needs as an artifact should be established:

- Consider particular information needs resp. demands as a means to drive the knowledge sharing process. Therefore, demands need to be stored, analyzed, and aggregated.
- Extend the list of possible actions for information seekers. Besides collaboration with others during the search process (e.g. mediated by representations of the information need), we propose to include means of information provisioning.

Concrete realizations of these abstract principles will be described in the implementation chapters 5 to 7.

4.3.2. Mediation Spaces and Services

In Section 3.3.1, we already introduced the notion of knowledge sharing as a *communication process* between the roles of the information seeker and the information provider. We will now extend this basic model in various dimensions as depicted in Figure 4.3.

For satisfying information needs, people can draw either from their private information space or from a shared information space.²⁵

For the private information space, we distinguish between:

- Explicit information resp. information artifacts (e.g., documents)
- Implicit information, which is not yet formally captured²⁶
- Semi-explicit information, which denotes explicit actions of a user that could be automatically captured and thus easily shared with others (e.g., actions to solve a certain problem)²⁷

²⁵See also Section 4.2.2

²⁶See also Section 2.2.2

²⁷See, e.g., Leshed et al. (2008) and Section 8.2.1

4. Need-driven Knowledge Sharing

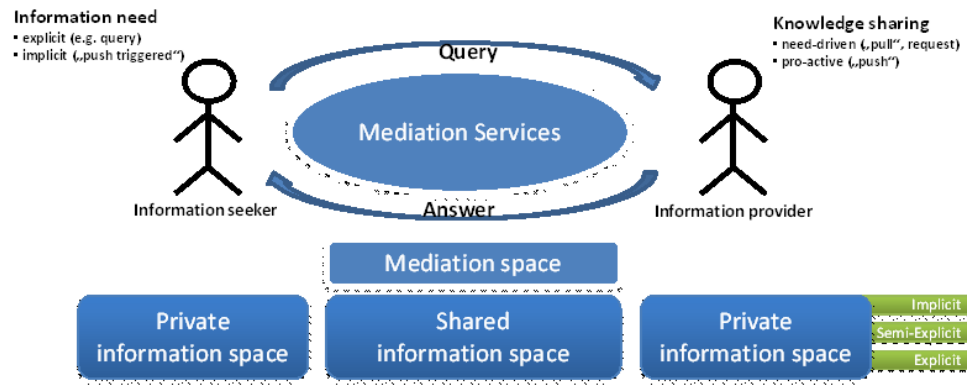


Figure 4.3.: Extended Model of Knowledge Sharing as a Communication Process (adapted from Figure 3.6)

Even though NKS focuses on sharing explicit, codified knowledge, implicit and semi-explicit information is considered relevant, as it might indicate earlier stages of knowledge maturity.²⁸

Both actors can either take a passive or an active role. In most cases, information seekers explicitly formulate their information demand, which is then answered by an information provider. However, information needs might also be implicit – i.e., when a user has a certain problem and is not aware that helpful information exists. Assistance systems are typically used to support users in this scenario. Conversely, information providers can actively “push” certain information, because they think it might be helpful for others.²⁹

Based on these concepts, a number of novel mediation services and spaces can be envisioned. Some examples are depicted in Table 4.2, while others will be presented extensively in chapters 5 to 7. These either focus on the information seeker (such as a better specification of information needs or their persistence in mediation spaces) or the information provider (e.g., displaying most wanted information needs, or providing capturing assistance).

While we are focusing on information seekers and information providers, additional roles such as mediators³⁰ might be useful when considering an organizational perspective of KM. We will leave those aspects to future work.

4.3.3. Information Need Attributes

In previous sections such as 3.1, we have argued that there are different qualities of information needs. In the following, we describe attributes of information needs, which are relevant in the case of NKS. It can serve as a basis for

²⁸See also Section 2.2.4

²⁹See also Section 2.3.3 or Abecker et al. (1998); Maurer and Tochtermann (2002)

³⁰E.g., curators; see also Section 4.5.3

Table 4.2.: Exemplary Mediation Services and Spaces for Need-Driven Knowledge Sharing

Phase	Mediation service/space	
Pre-Transfer	Demanding (Seeker) Matching (How to find suitable sources) Offering (Provider)	Better specification of needs Shared mediation space for information needs/display of most wanted information Sharing recommender system
Transfer	Consumption (Seeker) Negotiation (How to specify what to exchange) Provision/Creation (Provider)	Shared mediation space for information needs Aggregation of most relevant information needs/display of most wanted information Capturing assistant; Semi-automatic, context aware capture

capturing and analyzing information needs when designing systems based on NKS.

Degree of need formalization As described in section 3.1, an information need may translate in a concrete demand in the process of information seeking. Depending on the medium or information system involved, this demand may be specified in a precise, formal language (such as a formal query language) or in plain natural language keywords such as in the conventional IR paradigm.

Need indicators We differentiate various implicit and explicit indicators of an information need. *Explicit* indicators can differ in their degree of formalization – including plain text keywords, elaborate text, controlled language or structured representations. *Implicit* indicators include the general user context, such as location or time, and the working context of a user, such as work artifacts, process steps, or role. Each of these indicators can help to infer the information need of a user.

Type of need For Web searches, Broder (2002) distinguishes navigational and informational searches.³¹ Another distinction could be if the user is seeking for textual information or facts and if the need is of ad-hoc or of long-term interest.

Specificity The specificity of an information need, which might be derived from the length and complexity of queries or from related needs denoted by overlapping queries.

Experience The experience of the information seeker with respect to the need. This can relate to the type of need and its specificity.

Topics Entities that are part of the information need such as persons, places, or artifacts.

Sensitivity An information need might be sensitive in the sense that it exposes information about the information seeker.

Satisfaction The amount of information that is available to satisfy the need – e.g., as indicated by *organizational information gaps*.

³¹See also Rose and Levinson (2004) for a more detailed model

4. Need-driven Knowledge Sharing

Recurrence How frequent the need recurs over time (as discussed in Section 4.2.1).

Organizational value How many different people share particular needs resulting in *organizational information needs* (as introduced in Section 4.2.1).

All these aspects can be considered candidates for deriving information needs in particular implementations. In our own implementation, we'll particularly focus on the last three aspects and leave an extensive consideration of further aspects of information needs to future research.

4.4. Keyword-based Instantiation of NKS

After presenting the abstract conceptual framework for NKS in the previous section, we will now describe its particular application based on plain, keyword-based queries, as they are common for the majority of IR systems, such as Web or enterprise search engines. However, NKS is not strictly limited to this approach. In Chapter 7, we will introduce a second instantiation for structured query languages.

As for the discussion of technical details, we partially use different terms for some concepts that were introduced earlier. In particular, information demand will be denoted by keyword *queries*, which may consist of several natural language *terms*.³² The information exchanged will be denoted as *documents* which contain information. Private and shared information spaces are forming different *corpora* of documents. An *index* is an abstract data structure for the efficient representation of documents in a corpus.

We will begin with introducing our data model for logging resp. storing keyword queries, and describe how organizational information gaps and needs can be computed. We then present the architecture of the *TeamWeaver* software platform, which implements this data model and provides the basis for the reference implementations *Inverse Search* and *Woogle*, which will be presented in Chapter 5 resp. 6.

4.4.1. Storing Information Needs

Keyword queries are typically available in large amounts, but may be too specific to drive knowledge sharing. Therefore we developed the notion of *organizational information need* (OIN) in Section 4.2.1 as an aggregate of the personal information needs of members of a certain group of people.

As for the query log depicted in Table 4.3, we initially aggregate queries based on a normalized³³ representation of the query string. This value, stored in the column “needstring”, is the unique key for the table. Other values, such as

³²E.g., the query “Barack Obama” consists of the two terms *Barack* and *Obama*.

³³I.e., lowercase, with white-spaces trimmed

4.4. Keyword-based Instantiation of NKS

the list of users that have been executing this query (“Queryusers”), will be updated accordingly.

Table 4.3.: Aggregate Query Log

QueryId	Hashed value of the normalized query string (as unique identifier)
Needstring	Normalized query string
Needexec	Number of times the query was executed
Needexecclick	Number of times the query was executed without any result click following
Recency	Average timestamp across all query executions
Firstdate	First time the query was executed
Lastdate	Last time the query was executed
Resulthits	Number of results for the query (at the last time of execution)
Needexecpages	Number of query executions for which at least one further result page was browsed ³⁴
Needexecavg-page	Average amount of result pages users checked for that query
Queryusers	Ids of users executing the query
Sizequeryusers	Amount of users listed in <code>queryusers</code>
Publicusers	Ids of users who agreed to be publicly associated with the query

Aggregating queries provides two major benefits. First, logging individual queries would produce very large amounts of data, and an online processing of individual query instances would be computationally expensive. Second, by aggregating information needs, those shared by a large number of users can be prioritized more easily.

Table 4.4.: Aggregate Query/Click Log

QueryId	QueryId (as in Table 4.3)
ResultId	Id/URI of the clicked result
Clicks	Number of clicks for this result after this query
Position	Position of the result within the result list for the query (at the last time of execution)
Recency	Average timestamp across all clicks
RepoId	Id of the repository (“space”) where the result stems from (to help distinguish different information spaces)
Firstdate	First time the result was clicked for this query
Lastdate	Last time the result was clicked for this query

Besides the *user queries*, we also log *result clicks*, since they allow us to apply heuristics concerning the satisfaction of information needs.³⁵ Our logging scheme is presented in Table 4.3 and 4.4. The scheme and the actual logging data can be persisted within a relational database.

³⁴This assumes that query results are ordered by relevance and that typically 10 result summaries are presented to the user (see, e.g., Baeza-Yates and Riberio-Neto, 1999, p. 289). In order to see additional (less relevant) results, the user has to browse additional result pages (if any).

³⁵I.e., the absence of any result click may hint to *failed queries*

4. Need-driven Knowledge Sharing

4.4.2. Calculating Organizational Information Gaps

After describing how to store *information needs* (resp. information demands), this section deals with *information provisioning*. For the following discussion, we assume a setting with a private information space of each user (e.g., composed of documents on a local computer) and an organization-wide public information space (e.g., a company file share). We further assume that there exists an *index* for all private spaces and the public space, which each user can query using an IR system (e.g., enterprise or Desktop search).

According to our discussion in Section 4.2.2, we are interested to determine organizational information gaps regarding certain information demands by comparing the distribution of related information in private and shared information spaces. Therefore, we calculate two normalized document frequency³⁶ values for a certain term: one based on the documents in the private space of user, and one based on the documents in the public information space.

The normalized document frequency (*NDF*) is computed as follows: Let N be the total number of documents in a corpus and t a term. N_t is then the subset of N which contains t . The document frequency (*DF*) for a given term t can then be defined as the size of N_t . Accordingly, the normalized document frequency (*NDF*) is defined as:

$$NDF(t) = \frac{DF(t)}{N} \quad (4.1)$$

Regarding sharing local information with the organization, we are interested in those terms with a high *NDF* in the private information space, and a low *NDF* in the public information space. The difference of both values denotes the organizational information gap (*OIG*) for a certain term and a certain user:

$$OIG(t, user) = NDF(t)_{private} - NDF(t)_{public} \quad (4.2)$$

To derive a list of organizational information gaps for a particular user, the set of all terms contained in any document in her private information space and the public information space has to be generated.³⁷ Afterwards, Equation 4.2 can be calculated for each term. When listed in descending order, the terms with the highest *organizational information gap* will be shown on top of the list. An empirical evaluation of *OIG* in an organization will be described in Section 5.3.1 on Page 102.

While a high organizational information gap for a term reflects that there is significantly less related information in the public information space, that does not necessarily indicate a problem. From an organizational perspective, action is only required if other users have an information need regarding that term. We will discuss that in following section.

³⁶The document frequency of a term is the amount of documents containing the term.

³⁷Such a list of terms can be easily derived using the *index* of an IR system.

4.4.3. Calculating Organizational Information Needs

In Section 4.2.1, we discussed that information needs which are unsatisfied (i.e., for which no suitable information is available), recurring, and shared by multiple users, are most interesting from an organizational knowledge sharing perspective. As mentioned before, we assume an IR system, which allows users to query both, their private and the public information space. The IR systems maintains a central record of queries and result clicks as, defined before.

Similar to calculating organizational information gaps, we will now describe how to derive an organizational information need for each term. We define the OIN as the weighted sum of individual information needs of the users in the organization, composed out of the following four factors:

Frequency We assume that the OIN regarding a term is the higher, the more often it has been part of a query.

Recency Since the OIN regarding a term is a dynamic value, we also assume that the OIN is higher, if the term has been queried more recently. This allows recent organizational information needs to score a higher value.

Prior information We assume that the OIN is higher, if the querying user has no prior information about the information need. To measure this, we propose the value of $NDF(t)_{user}$, since it tells us, how many documents match the term in the private information space of the querying user. We propose that from an organizational point of view, additional information is less valuable for a user if she already has prior information.

Universality We assume that an OIN is the higher, the more different users queried for a term. The rationale behind this is, that time and resources of information providers are limited and should thus be focused. Accordingly, the more users share a certain information need, the higher the benefit for the organization to provide related information.

In order to formally define the OIN, we combine prior information,³⁸ frequency, and recency into a personal information need (PIN). Q denotes the total number of unique queries, while $Q_{t,user}$ is the subset of queries by a certain $user$ including a certain term t . The index *recent* implies, that the set is constrained to queries that have been executed recently.³⁹ To emphasize recent information needs and to avoid that the according fraction evaluates to zero in case $Q_{t,user,recent}$ is zero, a minimum value of 1 is introduced.

$$PIN(t, user) = \frac{1}{NDF(t)_{user}} \cdot \left(\frac{Q_{t,user}}{Q} \cdot \left(1 + \frac{Q_{t,user,recent}}{Q_{recent}} \right) \right) \quad (4.3)$$

Note that is an initial conceptualization of a personal information need, for the purpose of designing and evaluating specific tool prototypes, as presented in Chapter 5 and 6. A specification including additional factors such as, e.g.,

³⁸Since we want to penalize high experience, we include the inverted value of $NDF(t)$.

³⁹Therefore, the `Lastdate` parameter from Table 4.3 is compared with a certain threshold date (e.g., “now-7 days”).

4. Need-driven Knowledge Sharing

work context, or user profiles, is out of scope here and may be subject to future work.

Accordingly, the OIN is the sum of the values for PIN, normalized by the total amount of querying users ($Users$ denoting the total amount of users and $Users_t$ the subset, which executed a query including term t at least once):

$$OIN(t) = \frac{Users_t}{Users} \cdot \sum_{user_t} PIN(t, user) \quad (4.4)$$

Finally, we can combine the OIN for a certain term t with its OIG for a certain user. The resulting $OVT(t, user)$ denotes the organizational value of the term (OVT) from the subjective perspective of the user.

$$OVT(t, user) = OIG(t, user) \cdot OIN(t) \quad (4.5)$$

The value for OVT for a certain term is the higher, the more frequent this term is in the private information space of the user (when compared with the public information space) and the more users expressed an information need regarding this term (OIN). OVT will in turn be used to help identifying documents in the private information space, that should be shared within the organization, as we will illustrate in the next chapter with our *Inverse Search* tool.

4.4.4. Reference Implementation (TeamWeaver)

In this section, we present TeamWeaver, a tool platform which provides basic features of an IR system (see Section 3.1) and which forms the basis for a number of tools presented later in this thesis.⁴⁰ It can thus be considered a reference implementation for the keyword-based instantiation of the NKS framework. We will now sketch the general architecture of TeamWeaver and describe relevant components in detail.

Architecture

TeamWeaver consists of a backend and a frontend layer. The backend can be configured to crawl various different data sources. It offers an API to search across these sources as well. This is an important feature, since it is more convenient for end users to use a single entry point instead of having to identify which data source to search. Also, a central search interface allows to collect information demands (i.e., user queries) more easily.

The backend includes a QueryManager component, which allows to store and aggregate information needs. An API allows external client applications⁴¹ to retrieve information need data.

⁴⁰TeamWeaver components and source code are available as Open Source. See <http://www.teamweaver.org> for additional information.

⁴¹E.g., the SharingEngine that will be presented in Section 5.2.2

Based on the backend's query service, arbitrary clients can offer search capabilities to end users. By default, the TeamWeaverIS JSF frontend provides a Web-based search interface, like it is provided by many state-of-the-art Web search engines (e.g., Google). Two other frontend user interfaces, *Inverse Search* and *Woogle*, are presented in the forthcoming chapters.

While TeamWeaver's backend and frontend are typically deployed as Web applications in a servlet container,⁴² it can also be deployed as a set of OSGi components within an OSGi framework implementation.⁴³ This allows TeamWeaver services to run within an Eclipse RCP⁴⁴ instance.

Within Eclipse, TeamWeaver search components can interact with the TeamWeaver context component. The latter is a set of sensors and processing components, which allow to observe user interactions in an Eclipse environment, and thus to derive additional information need indicators. The context-capturing tools *MacIntent* and *WinTent* introduced by Maalej (2010) offer similar functionality outside Eclipse, and are also including frontends for integrated search.

The analysis of context information to design assistance systems for software developers has been addressed in the EU research projects TEAM and Fast-Fix.⁴⁵ A particular application of context information for information seeking has been lined out in Happel and Steinbauer (2008).⁴⁶

4.5. Related Work

This section discusses related work with respect to the overall approach of *need-driven knowledge sharing*. It is therefore structured in three parts. The first part presents different approaches in the area of *information seeking* which consider information needs and information demands as entities of their own right.

The second part summarizes approaches in the area of *knowledge sharing*, which aim to motivate and guide information providers contributing to shared information spaces. The last part describes works which particularly try to *bridge information seeking and knowledge sharing* in a holistic way.

⁴²Such as Apache Tomcat (<http://tomcat.apache.org/>)

⁴³OSGi is a specification for distributed, component-based systems standardized by the OSGi Alliance (<http://www.osgi.org>). In particular, the Eclipse project (<http://www.eclipse.org>) is build upon the OSGi framework architecture.

⁴⁴Eclipse RCP is a general framework for Java-based graphical user interfaces. Besides the popular Eclipse IDE for software engineering, many other other applications as for the banking industry have been developed based in Eclipse RCP.

⁴⁵See Happel et al. (2008); Maalej (2009, 2010); Pagano et al. (2012), <http://www.team-project.eu>, and <http://www.fastfixproject.eu/>

⁴⁶See also Happel and Maalej (2008); Robillard et al. (2010)

4. Need-driven Knowledge Sharing

4.5.1. Information Seeking

In this section, we particularly describe areas of work, which address search activity beyond the classic approaches of information seeking and IR as described in Section 3.1. We omit a separate discussion of *social search*, which has already been introduced in Section 3.1.2 and will be revisited in Section 6.4.1.

Information Requirements Analysis

Information requirements analysis (German: *Informationsbedarfsanalyse*; IBA), is a subdiscipline of information management. It mainly aims to analyze and satisfy the *objective information needs* of users, which are mainly shaped by their organizational roles (see also Section 3.1.1).

Within IBA, different user research methodologies can be applied in order to support the planning of information resources, including information systems such as MIS (Koreimann, 1976) or BI solutions (Stroh et al., 2011).⁴⁷ IBA research has also developed a number of analytic models, that contrast information needs with existing information resources, and which are related to the models depicted in Figure 4.1 and Figure 4.2.⁴⁸

Some scholars also discuss the application of information requirements analysis for knowledge management purposes (Krcmar, 2004; Gust von Loh, 2008). They propose to apply different user research methodologies in the design phase of a KMS in order to derive required knowledge resources and structures. This however does not provide a continuous, adaptive perspective on information needs (Stroh et al., 2011), but only a design-time snapshot. However, this could be helpful for bootstrapping need-driven knowledge management.

Query-driven Resource Optimization

On a high level, NKS is an approach to optimize a limited resource (time and cost for knowledge sharing) by means of actual demands (information needs). While this hints to the very general concept of demand-driven optimization,⁴⁹ there exist a number of approaches which are particularly similar to NKS.

Perhaps most related is the area of *search analytics* within the larger discipline of information architecture (Morville and Rosenfeld, 2006). It uses techniques such as search log analysis in order to e.g., drive the creation and maintenance of content on websites (Angiolillo, 2003). In this case, search logs mainly stem from internal “site search” tools.⁵⁰

⁴⁷Therefore, some approaches to information requirements analysis overlap with the requirements engineering (RE) discipline in software engineering (Stroh et al., 2011)

⁴⁸See, e.g., Mujan (2006)

⁴⁹See also Brown and Hagel (2005)

⁵⁰The former Microsoft Encarta encyclopedia was also reported to use such techniques (Wen et al., 2001)

At a larger scope, so-called “content farms” use search logs from public search engines or product marketplaces to create content which is demanded by users, but still scarce on the Web. An extreme example is the Leaf Group company (formerly Demand Media), which automated this process to the extent of crowd-sourcing the creation of content to human editors (Roth, 2009).⁵¹

A more traditional field is the so-called “collection development” in libraries (Fieldhouse and Marshall, 2011). This includes the optimization of acquisition budget and shelf-space. One method used in collection development is *suggestion analysis*, which deals with book requests of library users. Shenton and Johnson (2007) discuss the analysis of user searches in electronic catalogs as a source of suggestions.

Queries as Knowledge

The previously presented approaches acknowledge that queries carry inherently valuable information. However, they are still treated as a teleological *means* to the end of providing or optimizing some other information or process. Another stream of research however argues, that queries can even represent knowledge of their own right – beyond just representing information needs.

Perhaps most famous to this extent is research, which tries to predict trends in the areas of public health or popular culture by aggregating the individual needs of fellow search engine users (Butler, 2008). On a more general level, queries can thus also be considered a means of deriving facts about reality.⁵² Even more, Efron and Winget (2010) argue, based on studying queries in microblogging, that people use queries not only to satisfy their information needs, but also to express themselves.

The interesting point of these discussions is the observation, that there does not always exist a clear distinction between “question” and “answer” (Taylor, 1962). Instead, similar to the fact that information seekers can turn into information providers, questions may turn into answers and vice versa. This again stresses the point, that queries should not be considered transient entities, but deserve representations of their own right in information systems.⁵³

4.5.2. Knowledge Sharing

While many information systems are geared towards assisting information seekers, only few approaches explicitly provide assistance for information providers (as introduced in Section 3.2.1). We summarize work related to

⁵¹In particular, it is considering unsatisfied information needs in the domain of news articles and videos. Based on an analysis of frequent information needs and information offered on the Web, it identifies gaps which are advertised in a marketplace. Media producers can bid on topics to create content filling these gaps. Afterwards the content is fed into content platforms (former Demand Media was, e.g., supposed to be the largest contributor to Youtube according to Roth (2009)) with the goal of attracting users. Revenue is generated based on advertisements.

⁵²See, e.g., Sekine and Suzuki (2007); Richardson (2008); Strohmaier and Kröll (2009)

⁵³See also Section 7.4.2 on Page 160

4. Need-driven Knowledge Sharing

motivating information providers, *guiding* information providers *which* information to contribute, and regarding *group information management*.

Motivating User Contributions

As discussed in Section 4.2.2, people are generally willing to share knowledge with others if circumstances are suitable. We discuss a number of approaches, where “motivational affordances” (Zhang, 2008) in electronically mediated knowledge sharing have been evaluated.

Particularly notable is a series of studies of the *GroupLens* research laboratory, many of them studying the *MovieLens* system for movie rating and recommendation. Beenen et al. (2004) studied the effect of email notifications on contribution behavior. They found out that merely sending an email will raise contributions (also supported by Harper et al., 2007), that variations of email content made a difference, and that setting particular contribution goals led to higher contribution than non-specific ones. Further experiments showed that explaining the uniqueness (Ludford et al., 2004) or the value of a contribution (Rashid et al., 2006) were both effective to motivate users.

Brush et al. (2008) conducted a study using a Desktop application for photo sharing in a family context, which showed a daily reminder to share a *random* photo. They found that most users appreciated those suggestions “without having to think” and that 43% of overall photos shared were those suggested by the system.

While these studies provide insights into means to increase contribution behavior, they do not analyze or address *which* information should be contributed. Therefore, follow-up research on these aspects, discussed in the next paragraph, is more interesting with respect to NKS.

Guided Knowledge Sharing

In this section, we discuss a couple of tools which aim to guide information providers in *which information to contribute* to the respective system.

Cosley et al. (2006) describe an experiment with the *MovieLens* community in which users were asked to contribute based on four strategies: movies a user was predicted to like, movies a user had rated but few others had, movies with the most incomplete data, and random movies. The “rarely rated” strategy clearly outperformed all other strategies, which the authors attribute to the fact, that users certainly knew something about those movies.⁵⁴

FeedMe is an extension⁵⁵ of Google Reader, an online service to consume RSS feeds. It has been developed by Bernstein et al. (2010) to assist Google Reader users in deciding which RSS feed items (i.e., website URLs) to forward to their friends. Based on keyword comparison with prior recommendations,

⁵⁴A related follow-up study using Wikis (Cosley et al., 2007) is discussed in Section 6.4.3

⁵⁵By means of a browser-based GreaseMonkey script

the system recommends friends who might be interested in an article. Besides, the system provides awareness information about if a potential recipient has already seen that page (if known), and how many other recommendations a recipient has received recently. Finally, people can give appreciation feedback to information providers with a single click.

Topika strives to extend email clients by an additional recipient input field, labeled “post to”, which allows to easily send content into “shared spaces” (Mahmud et al., 2011). Suggestions for target spaces are computed by comparing terms of the email with terms derived from existing content in shared spaces. *Topika* is notable, because it not only suggests *what* content to share, but also allows for selecting a suitable destination space, easing the transition from the user’s private information space.⁵⁶

Geyer et al. (2008) describe the *About You* recommender system, which provides users with suggestions to extend their profile page in a social network application based on a number of information inputs. In a study with 2000 users, a significantly larger number of profile entries was created in the group receiving particular suggestions, compared to a control group which was merely asked to update their profile without receiving suggestions.

While the systems described so far make their recommendations mainly based on content, *Blog Muse* was designed by Geyer and Dugan (2010) as a tool that “encourages creation of content that matters in a work context by meeting the needs of information consumers” (Dugan et al., 2010). Therefore, three features were added to an enterprise blogging platform: a simple form to request topics (“Tell others what you would like to read about”), an overview that allows to vote for suggested topics, and an overview showing topics recommendations (“Susan would like to read about *X*”). Clearly, this setup can be considered a *mediation space* for need-driven knowledge sharing, with topics and votes as information need indicators. Dugan et al. (2010) also discuss using search logs for future versions of the system.

Group Information Management

As discussed in Section 2.3.4 and 4.2.2, the tension between private and shared information spaces has been identified a core aspect for knowledge sharing by many authors.⁵⁷ Surprisingly – especially considering the advent of Web 2.0 and Enterprise 2.0 – the amount of relevant research is not very large.

Out of several works providing abstract design guidelines for the interrelation of private and shared spaces,⁵⁸ we want to point out the concept of *information osmosis* described by Tungare and Perez-Quinones (2008). They cite the collaborative music database CDDB, a predecessor of Musicbrainz (Swartz,

⁵⁶Similar does *Mail2Wiki* (Hanrahan et al., 2011), which is discussed in Section 6.4.3 in more detail

⁵⁷See, e.g., Clement and Wagner (1995); Bannon and Bødker (1997); Pipek et al. (2003); Pinelle and Gutwin (2005)

⁵⁸See, e.g., Gwizdka (2006); Erickson (2006); Lutters et al. (2007)

4. Need-driven Knowledge Sharing

2002), as an example for a system that allows information to flow between private and shared spaces.

In terms of tool studies, the *Infotop* system by Maier and Sametinger (2003) considers a central search function and the interrelation of various information spaces. Marshall and Brush (2004) analyze private and shared annotations in a web-based document annotation system. Further tool prototypes are described by Groth and Eklundh (2006), considering Semantic Wikis, and by Iverson et al. (2008), describing a note sharing application. All these systems however do not address how the movement of information between spaces can be improved.

Among the few systems that address this issue are the *Topika* system, described in the previous paragraph, and *Mail2Wiki*, described in Section 6.4.3. Both support the transition between the private information space in email mailboxes and shared information spaces of an organization.⁵⁹

4.5.3. Bridging Information Seeking and Knowledge Sharing

Finally, we describe related work which explicitly considers the duality of *information seeking* and *knowledge sharing*, as discussed in Section 3.3 and refined in this chapter. In the section about *turning readers into contributors*, we summarize work with a focus on different user roles and their interactions. The section about *knowledge curation* finally describes approaches and tools to govern and “curate” the creation and sharing of knowledge.

Turning Readers into Contributors

As discussed in Chapter 3, research on *information seeking* and *knowledge sharing* exists largely independent of each other. In this section, we discuss approaches that provide a holistic perspective of those complementary processes.

Markus (2001) distinguishes the three roles of knowledge producers, knowledge consumers, and knowledge intermediaries. She also stresses, that these roles might be instantiated in different ways – e.g., a single person taking all three roles in a given situation, or some roles being performed by information technology. This fits into our conceptualization of knowledge sharing as a communication process (see Figure 3.6) and in particular the notion of mediation services.

Another stream of research stresses the evolutionary relationship of different roles. Economic perspectives of knowledge sharing often simplify the behavior of information seekers as “lurking” or “free-riding”, leveraging the contributions made by other users. However, concepts such as *legitimate peripheral participation*, which is related to *communities of practice* (Lave and Wenger, 1991, see also Section 2.2.3), stress that reading can be considered a starting

⁵⁹See also Section 5.4.4

behavior which may lead to later contribution. The *reader-to-leader framework* by Preece and Shneiderman (2009) elicits various such contribution roles in online communities.

Such evolutionary processes have, e.g., been described for the Wikipedia (Halfaker et al., 2013).⁶⁰ Antin and Cheshire (2010) argue even further, that reading as such can already been considered a valuable “contribution” to a community.⁶¹ An interesting point here is the distinction of information seeking vs. reading, since we argue in this chapter, that *information seeking* can and should play a pivotal role in knowledge sharing processes.

As we will discuss in more detail in Section 6.4.2, *community question answering* systems are the most comprehensive systems to date, which cover different user roles and their transition. In particular, information seeking easily evolves to publicly shared, elaborate “questions” (Liu et al., 2012), which can be considered a major artifact of their own in CQA communities.

In a similar way, *BlogMuse* (in the area of blogging) and *cattail* (in enterprise file sharing) support information seeking and sharing in an integrated fashion. Both have been discussed above already.

In the area of information behavior, (Rioux, 2004, 2005) developed the concept of *information acquiring-and-sharing* which analyzes how users shift between acquiring and sharing information. Studying information sharing on the Web, he found that users “store” information needs of others and accordingly share information they consider useful for them. He also argues that sharing has increased with the adoption of Internet-based information sources, mainly due to the relative ease of sharing on the Web. Accordingly his studies show that most information that is shared within internet-based environments had also been acquired there.

Several user studies in *knowledge work* domains are consistent with these observations (e.g., Poltrock et al., 2003; Dearman et al., 2008). In the domain of engineering design, the *Through-Life Knowledge Management* approach described by Tang et al. (2006) and Heisig et al. (2010) provides some guidance for knowledge sharing during product lifecycles.

Finally, Bock et al. (2010) take a reverse perspective to the mostly linear evolution from readers to contributors. Complementary to the concept of information overload, they define the concept of *contribution overload* as “a feeling of stress associated with information contribution” due to the need for maintaining ones contribution. Similar to the concept of *leaders* in the work of Preece and Shneiderman (2009), this stresses the importance of mediation and curation, which we will also cover in the following section.

⁶⁰See also Bryant et al. (2005)

⁶¹This is due to the fact that readers contribute attention and can thus easier turn into contributors; see, e.g., Halfaker et al. (2013). Singer et al. (2017) found in a survey, that about half of Wikipedia readers denote themselves as already *familiar* with the topics they are reading about.

4. Need-driven Knowledge Sharing

Knowledge Curation

As discussed in Section 2.3.1, classical knowledge management distinguishes centralized, top-down approaches from decentralized, personalized knowledge sharing. We thus argued in Section 2.5 that *bridging* both paradigms would be required to achieve efficient means for the decentralized creation and sharing of organizational knowledge.

NKS, as presented in this chapter, is our approach to guide knowledge sharing based on organizational information needs. While not elaborated further in this chapter, we argued that an organizational perspective to knowledge management (as inherent in centralized knowledge management) could be complementary to NKS (see Section 4.2.2).

In this direction, a seminal contribution to knowledge sharing is *Answer Garden*, an Organizational Memory System (OMS) developed by Ackerman and Malone (1990); Ackerman and McDonald (1996). It connects information seekers and information providers by triggering potential contributions based on user information needs. However, Answer Garden lacks advanced discussion features (in its initial version), restricts content creation to “experts” and does not incorporate existing document repositories.⁶²

The general concept of Answer Garden has inspired many tools that seek to refine information resources such as forum discussions,⁶³ mailing lists,⁶⁴ or FAQs.⁶⁵ Modern collaborative question answering systems (CQA; see Section 6.4.2) can be regarded ancestors of Answer Garden as well.

Approaches related to FAQs and in the area of CQA do also consider user-defined “valuations” of sought information – e.g., based on the repetition of questions (FAQs) or on votings and comments that can be applied to questions (CQA). A more general model, defining the concept of *Return On Contribution (ROC)* is described by Muller, Freyne, Dugan, Millen and Thom-Santelli (2009). The ROC is defined as the ratio of benefit divided by cost, whereas the “benefit” is the subjectively-defined value of the resource by the accessing person. This is tightly related to our concept of organizational information need as discussed in Section 4.2.1.

Community leadership, as introduced by Preece and Shneiderman (2009), has also been observed and described in several knowledge sharing communities. Studies which have analyzed governance processes in the Wikipedia and in Enterprise Wikis will be described in Section 6.4.4. For the *MovieLens* movie rating system, Cosley et al. (2005) have shown that active curation can provide important signals about the quality of a community to contributors. Muller, Millen and Feinberg (2009) added the notion of *collections* to their enterprise file sharing system, which can be considered a special design feature serving the role of knowledge curators.

⁶²A related, more document-centric concept is *Active Document*, described by Heinrich and Maurer (2000)

⁶³See e.g., *I-DIAG* (Ackerman, Swenson, Cotterill and DeMaagd, 2003), *Arkose* (Nam and Ackerman, 2007), or *Wikum* (Zhang et al., 2017)

⁶⁴See, e.g., Brewer (2000); Hansen et al. (2007)

⁶⁵“Frequently asked questions”; see e.g. Ng’ambi (2002); Ng’ambi and Hardman (2004)

4.6. Summary

In Chapter 3, we described that the *decoupling* of information seekers and information provider is both, a desired property and a problem for document and content sharing (DCS). On the negative side, the asynchronous nature of DCS results in a lack of feedback and motivation which, in turn, make many knowledge management systems suffer from a lack of user contributions.

Accordingly, many information needs can not be satisfied by KMS, resulting in an inefficient allocation of information within an organization. We also highlighted, that users tend to switch between the roles of *information seekers* and *information providers*, and that information needs in the form of keyword queries often *recur* and are *shared* by many users.

We thus introduced the concept of *Need-driven Knowledge Sharing* (NKS), which aims to bridge the separation of information seekers and information providers. This is achieved by considering *information needs* (e.g., in the form of keyword queries) as valuable artifacts, which are aggregated in order to derive continuous forecasts of what we denote *organizational information needs* (OIN). By comparing with private and shared information spaces, *organizational information gaps* (OIG) are derived in order to identify missing information. These gaps can be made transparent using so called *mediation services* and *mediation spaces*, which help to create awareness for organizational information needs and to guide knowledge sharing.

Major contributions of NKS are to provide a conceptual link between the roles of *information seekers* and *information providers*, which are so far considered in isolation in their respective disciplines, as describe in Chapter 3. The notion of *organizational information gaps* helps to identify potentially inefficient allocations of knowledge within what we distinguished as *private* and *shared* information spaces. This provides the foundation to novel mechanisms for moving documents between those information spaces, as we will particularly describe with *Inverse Search* in the next chapter. Finally, the concepts of *mediation services* and *mediation spaces* can serve as a framework to analyze and improve other KMS with respect to NKS.

With this chapter providing an initial outline, NKS also provides various opportunities for future work, which will also be discussed in more detail in Section 8:

- Extending and evolving NKS, in particular the list of information need indicators towards a more holistic *model of information needs*
- Envisioning new approaches and tools to elicit information needs from information seekers
- Developing standards for describing, persisting, and exchanging information needs between applications
- Using NKS as a framework for analyzing knowledge sharing approaches and tools; in particular regarding mediation services and spaces

We finally presented TeamWeaver as a reference implementation of basic NKS features. In particular, TeamWeaver provides a Sharing API, which offers

4. *Need-driven Knowledge Sharing*

aggregated information needs to client applications. In the following chapters, we will present tools – namely *Inverse Search* in Chapter 5, and *Woogle* in Chapter 6 – which realize novel mediation services based on TeamWeaver. In Chapter 7, we present *Semantic Need*, a further tool offering NKS mediation features, which is based on structured information needs and content.

5. Inverse Search: Recommending Users to Share Documents

Current Web- and enterprise search engines allow their users to search for relevant documents among the total set of documents available. Since they can return only those documents, which have been analyzed before the user starts searching, this model is also called *retrospective search* (Hearst, 2009, see also Section 3.1.1).

However, in the Web as well as in the enterprise, new documents are created *continuously*. As a consequence, many users tend to repeat searches in order to find new information (see Section 4.2.1). In order to provide a better user experience, the paradigm of *prospective search* has been conceptualized (Irmak et al., 2006). Prospective search systems, such as Google Alerts (Google Inc., 2017a), allow users to store their search queries for the purpose of receiving notifications once additional results become available.¹

While prospective search acknowledges the “dynamic” nature of document creation, it does not address *how* and *why* new documents become part of the search process. In the end, documents have to be created or shared by individual users, who make them available for others. Studies however show, that a large amount of documents resides within the private information space of users (e.g., their local desktop or private folders²) – hiding them from other users.³

In this chapter, we propose a novel approach called *Inverse Search* which aims to systematically foster the sharing of documents which are “hidden” in private information spaces, but potentially relevant for other users. While both retrospective and prospective search mainly consider an information seeker, her queries and a set of documents, we introduce *information providers* and their *private information space* as additional elements to the search process.

We first analyze issues of existing search and document sharing approaches. Afterwards, we introduce concept, architecture and implementation of Inverse Search. To evaluate our work, we present results from two evaluation studies and a comparison with related approaches. The chapter closes with a summary and an outlook for future work.

This chapter is based on a number of prior publications. The initial idea for Inverse Search has been laid out in Happel et al. (2007). Happel and Stojanovic (2008) describe an exploratory study about the existence of organizational

¹See Section 3.1.2

²See also Section 4.2.2

³See Section 5.1.2

5. Inverse Search: Recommending Users to Share Documents

information gaps. Happel (2008a) describes the architecture and algorithms as well as first evaluation results.

5.1. Problem

This section contains a deeper investigation of the current problems in *retrieving* and *sharing* electronic documents in organizations and distributed teams, based on the issues identified in Section 3.4. At the end, the as-is situation and our envisioned to-be state are illustrated by means of a motivating example.

5.1.1. Information Seeking and Retrieval

Retrieving electronic documents is challenging – especially in an organizational context. In Section 4.3.2 we lined out that electronic documents are either located in *private* or *public* information spaces. Both spaces are typically very fragmented, spanning different document repositories and systems.

While some of these systems come with their own search functionality, the experience of a single search box for all content – such as it exists in Web search – has sparked the desire for similar functionality for the enterprise. So called *enterprise search* applications thus allow users to query many information sources from a single user interface.⁴

However, enterprise search has a focus on public information spaces and typically does not cover private spaces. For the latter, so-called *desktop search* applications allow to search across personal documents, emails or other information. Some of these tools can even be connected to enterprise or Web search engines, thus allowing an almost seamless search experience across these different information spaces.

Lack of demand treatment

Despite of the ubiquity of Web search, the usage dedicated enterprise search solutions in organizations is still far less common (White, 2015). And even if such an infrastructure exists, the landscape of information spaces and search tools is often heterogeneous. This makes it difficult to collect and analyze information needs.

Also, organizational information spaces contain much less information than the public Internet. One can thus assume, that information seekers might often not be able to satisfy information needs from these sources. If captured, this information about *failed searches* could be valuable artifacts to help deriving unsatisfied information needs.⁵

⁴See also Section 3.1.2

⁵See also Section 4.2.1

Lack of awareness

The large amount of information spaces and systems makes it hard for users to maintain awareness about available information.

Additionally, relevant information might exist in private spaces of colleagues (see paragraph “Lack of privacy” below). Typically, neither enterprise nor desktop search applications allow to search information from other users private information spaces. A notable counter-example was the discontinued Google Desktop search⁶, which allowed users to share their desktop search index with certain users. Since the index data is stored on Google servers, this however introduces security issues. Also, users might not want to allow others to search in *all* their private information.

Lack of possible actions

Enterprise search and desktop search tools as mentioned in the previous paragraph are limited to basic search functionality. They typically lack more advanced features such as *prospective* searching and also do not allow information seekers to collaborate.⁷

5.1.2. Knowledge Sharing

This section summarizes problems that exist for the role of information providers in document sharing scenarios. We assume a typical file sharing infrastructure in place, which allows to create folders and assign access rights. This kind of feature set is offered by simple file shares, but also part of larger KMS/groupware systems such as Microsoft SharePoint⁸ or IBM Notes.⁹

Lack of awareness

Many people complain about a lack of awareness concerning the information needs of fellow users (Dearman et al., 2008). This starts with problems of identifying interested receivers in general (Olson et al., 2005) and includes the concern to share only content which is relevant (Bernstein et al., 2010; Dearman et al., 2008) and novel (Bernstein et al., 2010) for them.

Lack of guidance

As electronic information is produced by individuals, a certain infrastructure such as a KMS is required to share information. However, such systems often adhere to certain organizational boundaries. Thus, in many situations users might just lack suitable tools and infrastructure to share information

⁶https://en.wikipedia.org/wiki/Google_Desktop

⁷As described by *social search* approaches; see Section 3.1.2

⁸See, e.g., Diffin et al. (2010)

⁹See, e.g., Orlikowski (1992)

5. Inverse Search: Recommending Users to Share Documents

(Dearman et al., 2008). Also, most KMS offer rather simple sharing features, lacking means to identify receivers (Olson et al., 2005) or keeping control of information after sharing (Whalen et al., 2008a; Volda et al., 2006).

On the other hand, more sophisticated means of sharing might also raise new concerns. For example, studies on the usability of access rights show, that users tend to avoid effort (Dalal et al., 2008; Smetters and Good, 2009; Lau et al., 1999) and often make errors which may lead to “over-sharing” of information (Dalal et al., 2008; Ahern et al., 2007).

Lack of privacy

In Section 4.3.2, we introduced the notion of *private* and *public* information spaces from which information seekers can satisfy their information needs. While it seems to be intuitively clear that a significant amount of information is hidden in private spaces, a number of recent studies stresses this issue (Hicks et al., 2008; Whalen et al., 2008a; Tang et al., 2007). Even in Web 2.0 applications which have open sharing as a default setting, people tend to keep information private. On the Flickr photo share portal for instance, Lam and Churchill (2007) found that 20% of all fotos were nonpublic at the time of their study.

One could argue, that people keep files private for good reasons and are anyway not willing to share them. However, especially in a work context, authors claim that many files are not private due to containing explicitly sensitive content (Schirmer, 2003; Tang et al., 2007). Instead, there are arguments that people tend to keep files private due to a low maturity of content (Olson et al., 2005),¹⁰ *potential* sensitivity (Whalen et al., 2008a), and to avoid being recognized as an expert on a topic (Schirmer, 2003). Also, many files may be kept private due to missing tool support for sharing, as we will discuss in the following paragraph.

5.1.3. Motivating Example

In this section, we illustrate the current *as-is* situation and the *to-be* situation as envisioned in this chapter.

As-is situation Our scenario involves two persons – Alice and Bob – working as colleagues in the same organization. Both of them have private documents stored on their computer which no one else can access. Additionally, there is a shared network drive, containing documents that are available for all members of the organization. Users can access both kinds of documents – private and shared – via a unified search interface. We assume that Alice is interested in information about a new project called “Theseus”. However, there is only one document containing “Theseus” as a keyword, which is in the private information space of Bob (see Figure 5.1). Thus, when Alice submits a query,

¹⁰See also Section 2.2.4

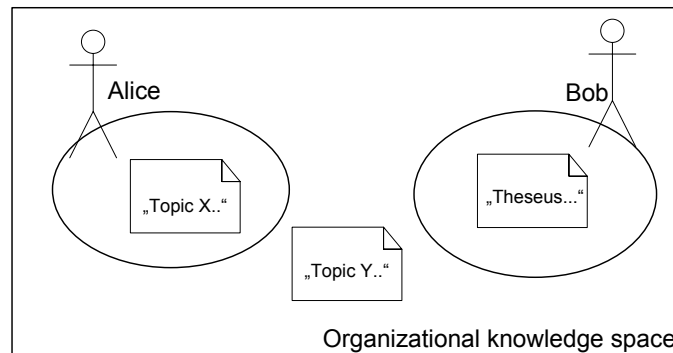


Figure 5.1.: Users and Documents in our motivating example

no results are returned. From a global point-of-view this is inefficient, since there is information in the organization available which could satisfy Alice's information need.

To-be situation In order to improve the sharing of information in the described situation, we propose that Alice's query is not just matched against the corpus of indexed documents (yielding zero results in our example) but also stored in a central query log.¹¹ This information can then be made available to interested clients. Thus, Bob's search application can retrieve a list of sought keywords and automatically compare it to the documents in his private information space. In our example, this would reveal that information from his computer could satisfy Alice's information need. The search application would try to identify documents which contain this information. Afterwards, it presents a small subset of those documents to Bob, indicating that there is an information need that can be satisfied by sharing them. Bob may then choose to move or copy these documents to the public information space. Once Bob shares the information, Alice could be notified about the new results.

5.2. Design and Implementation

In this section, we introduce a novel knowledge sharing mechanism (or *mediation service* in the words of Section 4.3.2) which we call *Inverse Search*. We first present the overall approach, introduce the technical architecture and some algorithms. Finally, we describe the prototype implementation.

5.2.1. Approach

As described in Section 3.1, *conventional* search systems – such as retrospective and prospective search – are focused on information seekers and a public index of documents. Although private information spaces might contain additional

¹¹For privacy reasons, this may also happen in an opt-in or anonymized fashion

5. Inverse Search: Recommending Users to Share Documents

relevant documents, information providers are not part of the standard search model.

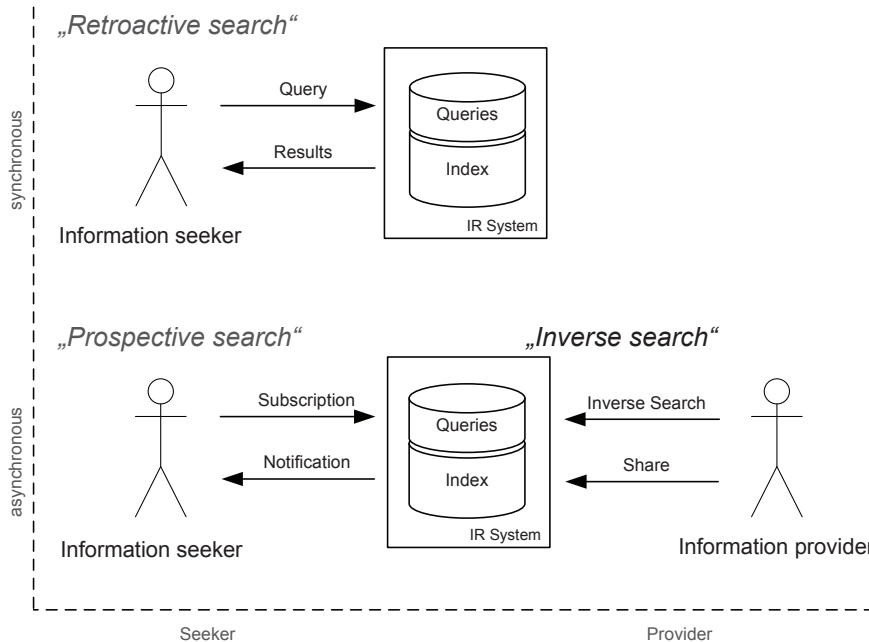


Figure 5.2.: Addressing Information Providers with Inverse Search (extension of Figure 3.4)

However, based on already existing query logs, information providers could find out, if some of their documents should be included into the public index. Thereby, the process of information provisioning to the public collection would no longer have to be a black box, but could be supported and stimulated by existing information needs. As depicted in Figure 5.2, we call this approach *Inverse Search*:

Definition 5.1 (Inverse Search). *Inverse Search is a feature of IR systems targeting information providers, which allows to match documents against a given set of queries – in opposite to conventional search, where information seekers match queries against a given set of documents.*

While users “import” public documents into their private information space in conventional search, Inverse Search helps to move documents from the private information space to the public information space, where they might satisfy the information needs of other users.

With Inverse Search, we aim to conceptualize *how* and *why* documents move from private to public spaces. As depicted in Figure 5.2, this makes our approach orthogonal to the notions of retrospective and prospective search, as they have been described in Section 3.1.2.

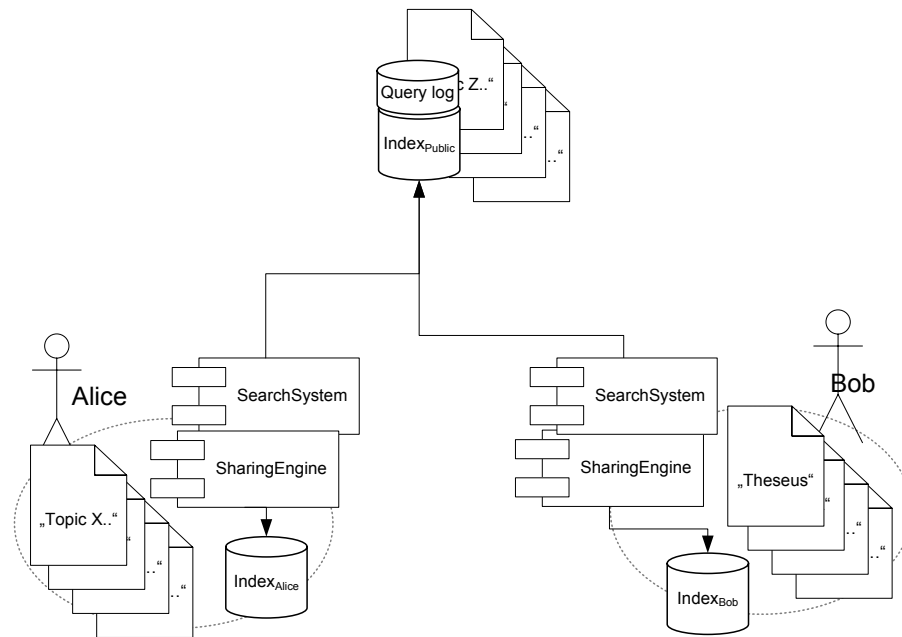


Figure 5.3.: System Architecture

5.2.2. Architecture

We now describe a system architecture that supports the envisioned knowledge sharing process. The whole system is depicted in Figure 5.3.

In order to differentiate between documents in public and private information spaces, the system requires a public index for the documents in the public space ($Index_{Public}$) and a private index of the private documents of each user (e.g., $Index_{Alice}$ and $Index_{Bob}$). This private index is not accessible to any other user.

Queries to the public index are automatically saved to a public *Query log*. Both, the public index and the query logs can be retrieved by any user. In order to retain privacy, queries may be anonymous and do not need to contain information about the querying user. However, if users like to receive automatic notifications when new information arrives, their identity can optionally be stored.

On her local machine, each user runs a *SearchApplication* which allows to query both the local index and the public index. Additionally, each user is running a *SharingEngine*, which periodically compares the local index with the global index and the query log. Thus, the sharing engine can derive an estimation of how useful it would be to share a certain document.

In order to minimize the effort of sharing, several options exist in order to suggest sharing certain documents to the user. This might either happen by enriching existing user interfaces (e.g., by decorating existing document icons with information about its value) or by periodically presenting a ranked list

5. Inverse Search: Recommending Users to Share Documents

of few documents, which the user should share within the organization.

Once information is shared, existing approaches for prospective search¹² can be used to notify information seekers.

5.2.3. Design

Based on this architecture, the sharing engine of a user is able to identify documents in her local space, which can be useful for other users in the organization. This requires two steps, very similar to prospective search:

- Derive unsatisfied information needs from existing queries.
- Select appropriate documents which satisfy these information needs, so that they can be recommended for sharing.

We will now describe these steps in more detail.

Deriving Information Needs

In standard prospective search, the derivation of information needs is trivial, since it is based on “alerts” (corresponding to keyword queries) that are precisely defined by the user. We need to take a different approach, since we are interested in particular kinds of information needs (see Section 4.2.1). First, we want to *automatically* derive information needs, instead of relying on a given set of explicit needs. Second, unlike in prospective search, we do not want to recommend sharing documents related to queries of *single users*. Instead, we need to aggregate the information need of the users in the organization in order to identify the most relevant *organizational* information needs. Therefore, we are interested in needs that are *shared* by a particular amount of people, instead of dealing with individual needs.

For the automatic derivation of needs, Yang and Jeh (2006) define a notion of “standing interest”, which they calculate based on query logs. After identifying a standing interest, a “typical” query for that interest is selected out of the queries of a user. Accordingly, this approach does not require users to manually register “alerts”. On the other hand, it does not accommodate for the aggregation of the *shared* information needs of several users. Especially the notion of “typical queries”, representative for a particular information need, could be problematic in this case.

Thus, instead of using the concrete queries to represent an information need, we use the terms extracted from queries as an approximation for the aggregated organizational information need. Economically spoken, our goal is to determine those terms, which have the highest value for the organization. Candidate documents are selected for recommendation based on this information.

We compute the “value” of a term based on two different concepts:

¹²See Section 3.1.2

- First, those terms are preferred, which appear relatively more often in the private index than in the global index and thus have a large “organizational information gap” (OIG).
- In the second step, information from the query logs is used to compute the aggregate “organizational information need” (OIN).

Formulas for calculating OIG and OIN have already been presented in section 4.4.2 respectively 4.4.3.

Selection and Ranking of Documents

After the identification of suitable organizational information needs, the next step is to identify documents in private information spaces, which could help satisfying these needs. Since Inverse Search does not have a global view on all private information spaces for privacy reasons (see Figure 5.3), the identification will happen decentralized for each user. This can in practice lead to *redundancy* if several users share similar documents. On the other hand, this can turn out beneficial, since this raises the probability that there is suitable information among the documents shared.

To derive documents, prospective search engines typically provide a list of all results matching the registered query, that are new since the previous notification. Documents are thus matched by simply querying the search engine.¹³ Again, we need a different solution, since we do not want to identify *any new documents* which could be shared. Instead we want to limit the number of sharing recommendations to a small set of documents. Furthermore, this small set should probably not just satisfy some particular queries, but a large organizational information need.

Therefore, we need to carry out the following steps:

- Select from the *terms* in the local documents those with the (from this user’s perspective) highest value for the overall organization
- Based on this, select those currently private *documents* with the highest value for the overall organization

After identifying the most valuable terms in the private information space of the user, we need to derive concrete documents that can be recommended to be shared.

A straightforward solution would be to query the local index with the most valuable terms and suggest the top-ranked documents to the user for sharing. However, there are some limitations with this approach, since the top results might cover certain terms redundantly. For example, in the case that the first m top ranked documents cover only one term from the list of n highly-relevant terms, it may happen that a user does not consider sharing documents that are indexed by the other highly relevant terms.

¹³Yang and Jeh (2006) discuss a number of heuristics in order to restrict this to “interesting results”

5. Inverse Search: Recommending Users to Share Documents

Our approach tries to solve this problem. The pseudo code is shown in Listing 1 on Page 179. The algorithm generates a minimal list of documents that “cover” as much as possible of the high-valued terms. Our assumption is, that a document indexed with the more high-valued terms should be ranked high, since it fulfills the highest organizational information need. The goal is to achieve the best representation of high-valued terms in this selection of documents.

5.2.4. Implementation

This section describes the prototypical implementation of the *Inverse Search* architecture and algorithms. The public index, as depicted in Figure 5.3, is realized by means of the TeamWeaver Integrated Search backend.¹⁴ It allows to crawl different public information spaces, which can then be accessed by a single search API. Also, it logs user queries and offers this information via an API.

The private index of each user is maintained by deploying TeamWeaverIS as a desktop search engine. As discussed in Section 4.4.4, the crawling and querying functionality can be bundled into a standalone Eclipse RCP desktop application. This application offers the major features of a desktop search application, i.e., crawling private folders and a search interface. In particular, the search interface can connect to multiple TeamWeaverIS backends, thus providing a seamless search experience, covering multiple private and public information spaces.

Finally, an additional sharing user interface connects to the query log API of the TeamWeaverIS backend. It offers two modes of user interaction. First, users may request a recommendation for documents to share. The system will then present a set of local documents, which could be useful to satisfy information needs of other users. For each document, the user can identify which information needs (i.e., query keywords) are matched.

Figure 5.4 shows a screenshot of our implementation. It lists files in the private information space of the user (“file:/...”) as ranked by the previously described algorithm. For each file, the user can also investigate, which terms contributed to the ranking of each document (“wiki wucherung” etc.). For each term, the UI depicts users which agreed to be associated personally with their query (none in the screenshot) and *OVT* (labeled “results value” in the screenshot).¹⁵ As actions, users can decide to *share* documents – e.g., copy it to a public space, or directly send it to a requesting user – or *capture* the information in a KMS, in case they do not want to share a particular document as-is.¹⁶

Second, users may request a list of information needs with a high *OIN*, based on which they can also capture information. The according part of the user interface is shown in Figure 5.5. It shows a list of terms ranked by their *OIG*.

¹⁴See Section 4.4.4 on Page 80

¹⁵“Clicks value” and “Browsing value” are experimental values which are not relevant for the current discussion

¹⁶The particular KMS, such as a Wiki, can be pre-configured

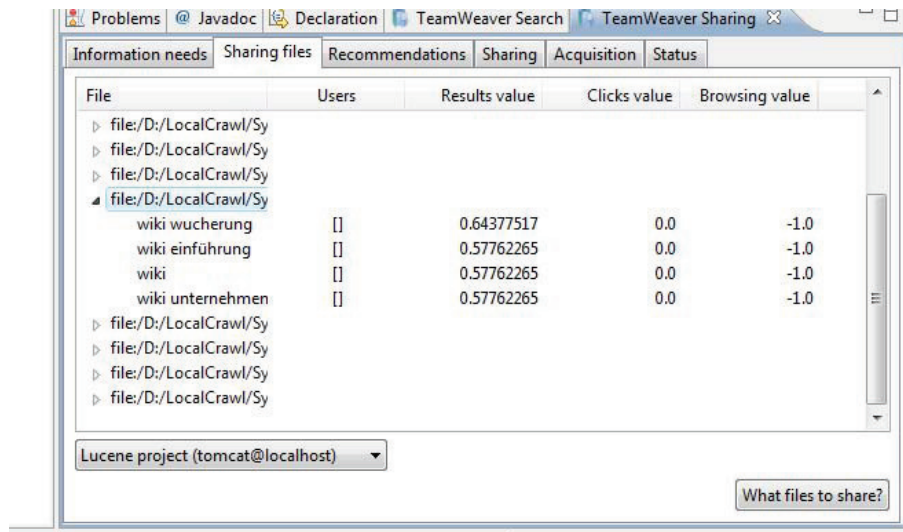


Figure 5.4.: Inverse Search File Sharing UI

In order to understand the context of the information need, overlapping query terms can be explored as well.

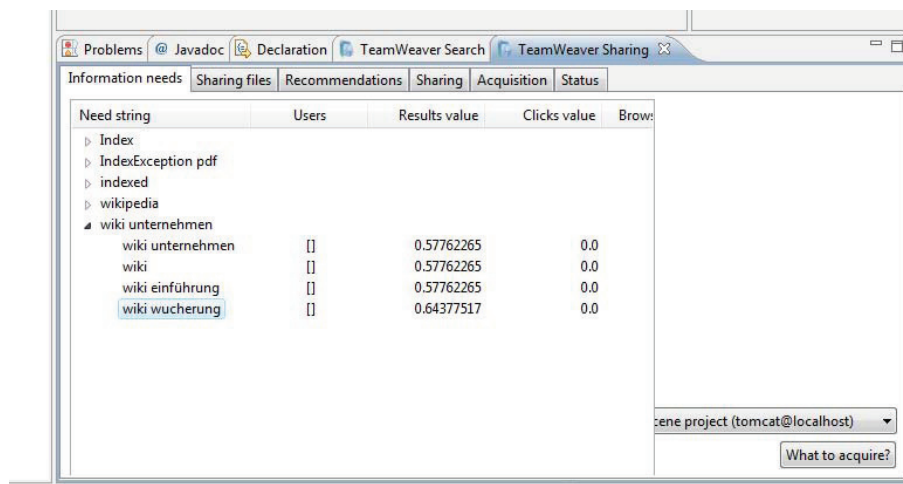


Figure 5.5.: Inverse Search Knowledge Acquisition UI

5.3. Evaluation

The Inverse Search concept is based on three consecutive hypotheses:

- H1** There exist *organizational information gaps* between private and organizational information spaces.

5. Inverse Search: Recommending Users to Share Documents

H2 A significant set of *information needs* can not be satisfied due to information gaps.

H3 Subjects are more *willing to share* information if it meets actual demand.

As Inverse Search seeks to modify existing knowledge sharing practices, the validation of these hypotheses requires data from the field. Hypothesis H3 additionally requires a mature implementation in order to gain considerable usage data. Since such an implementation was not feasible in the context of this work, we will focus on H1 and H2 in the remainder of this chapter. As for H3, studies reported in Section 4.2.2 and data gained from further tool evaluations tends to support the hypothesis. However, a real world investigation remains subject to future work.

5.3.1. Organizational Information Gap Analysis

Design and Process In order to calculate the OIG, we analyzed the private information space (i.e., files on a personal computer) of 13 voluntary users from a research and development organization. Users were allowed to exclude folders containing sensitive or not work-related information. For the public index, we selected the root folder of the fileserver of that organization. It contains shared information about projects, scientific topics, administrative procedures etc. and can be accessed by all users.

We decided to restrict our analysis to Word documents (*.doc, *.docx), because we assume that most of them were created within the organization. We expect that such documents are more often stored in the private information space, in opposite to, e.g., PDF-documents, which might have been downloaded from the Web.

We implemented a program that scans directories for such documents, extracts the terms, and calculates the *OIG* for each term as described in Section 4.4.2 (see Page 78). The term extraction is carried out by using the Lucene library (McCandless et al., 2010) with its default settings. We ran this program to scan for Word documents in the document folders described before (including subdirectories).

Results The basic results are shown in Table 5.1. The fourth column (Unique terms vs. Public) shows the number of terms which only appear in the respective user's index, but not in the public index. The fifth column shows the number of terms with an *OIG* value larger than 0 – i.e., which occur relatively more often in private information space of the respective user, than in the public information space. We excluded terms with a document frequency of 1, which seemed to be mostly typos, and those with a higher document frequency in the public space ($DF(t)_{user} < DF(t)_{public}$). A total number of 6 147 terms has an *OIG* of 0.5 or higher. In average, 13% of user terms had a positive *OIG*.¹⁷

¹⁷Average value for $OIG(t) > 0$ divided by *Terms*

Index	Documents	Terms	Unique terms (vs. Public)	$OIG(t) > 0$
Public	5 330	224 009	n.a.	n.a.
UserA	12	3 624	300	110
UserB	40	13 220	1 795	428
UserC	216	24 059	8 030	3 578
UserD	65	11 754	2 289	637
UserE	710	61 918	25 367	7 901
UserF	35	12 719	1 871	39
UserD	10	3 867	946	930
UserE	1 466	78 251	44 033	19 723
UserF	2 588	114 950	66 321	47 426
UserG	196	19 316	4 450	1 932
UserH	197	21 589	3 802	1 616
UserI	315	21 025	5 215	2 817
UserJ	5	5 274	190	325
$\Sigma_{UserA-J}$	5 855	391 566	164 609	87 462

Table 5.1.: Distribution of Terms in the analyzed Indices

In a second step, we selected only those terms with a minimum document frequency of 5 in one of the 13 private information spaces, in order to ensure a certain relevance. This resulted in a list of 16 714 terms. After removing duplicates from this list, 14 674 terms remained. For each of these terms, we calculated the values $NDF(t)_{public}$ and $NDF(t)_{user-average}$ (NDF_{UA}). The latter is the average $NDF(t)$ value of the respective term across the 13 private information spaces. The overall distribution is depicted in Figure 5.6. As the scales differ, the black line denotes the vector for which $NDF(t)_{public}$ equals NDF_{UA} .

As the plot shows, there are two clusters of outliers at both extremes. Terms with a high value for NDF_{UA} tend to be less frequent in the public information space. On the other hand, terms which occur very frequently in the public information space have no high NDF_{UA} . Accordingly, there seems to be a significant set of terms which has a different distribution across the overall (averaged) private information space and the public information space.

We also sorted the terms of each user’s index by their OIG. A qualitative examination of these lists shows surprisingly little “noise”. Most terms intuitively stem from their owners’ background such as email addresses, zip codes and various terms related to ongoing projects and areas of expertise.

The evaluation shows, that there is a significant number of terms for which information is hidden in the private information spaces of users, which is consistent with previous research (Hicks et al., 2008) and supports our hypothesis H1. Relative term frequencies vary significantly among private and public information spaces in our data set, supporting our initial assumption of scattered information. This indicates that information gaps exist and should be addressed by further research.

5. Inverse Search: Recommending Users to Share Documents

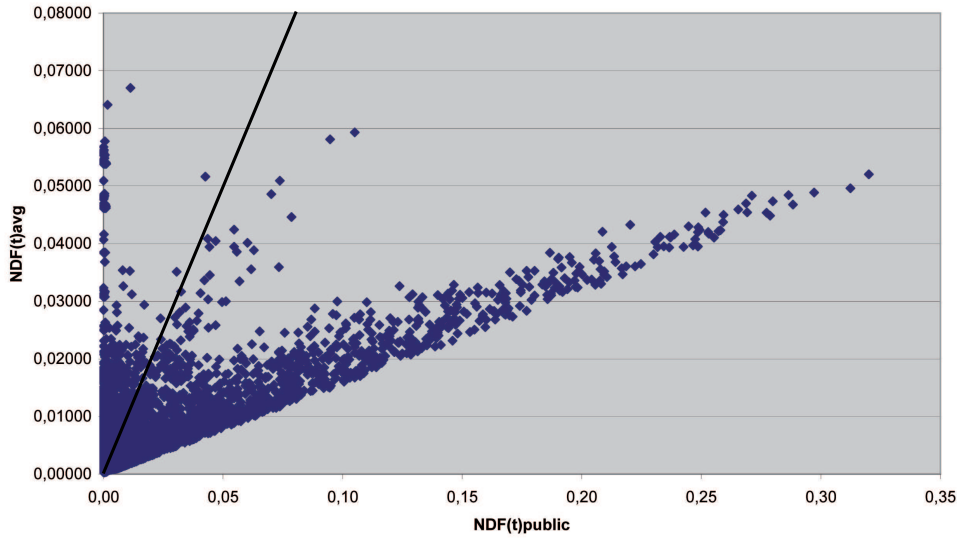


Figure 5.6.: NDF values (*public* and *user – average*) for 14 674 terms

5.3.2. Organizational Information Need Survey

Design and Process As described in Section 5.2.3, our approach for calculating organizational information needs requires query logs from an organization. Since such query logs were not available from the surveyed organization during the time of our study, we decided to estimate the *OIN* for a number of terms by using questionnaires.

Therefore, we collected five meaningful terms from the top-ranked terms (high *OIG*) of the private indices analyzed in the previous section. As “meaningful” we defined terms which are related to entities such as projects, person names, scientific topics, or software tools. Three of the indices were excluded (UserA, UserF and UserJ), since their analysis did not yield a sufficient number of such meaningful terms.

The resulting 50 terms (5 terms from 10 indices) were compiled into a questionnaire, which was then distributed to members of the organization.¹⁸ For each term, the questionnaire provided four options to choose from:

Actively interested: User actively sought information related to the term within the last six months or plans to do so in the future

Passively interested: User is interested in information related to this term, without actively seeking for it

Not interested: User is not interested in the term

¹⁸To avoid any evaluation bias, users for which documents were analyzed before did not receive the questionnaire

Don't know: User does not know/understand the term or is not sure about interest

Results A total number of 12 voluntary users completed the survey. In average, across all 50 terms, 30% of the users selected “don't know”, 37% selected “not interested”, 20% selected “passively interested”, and 13% selected “actively interested”. Out of the 50 terms, 40 terms were marked “actively interested” by at least one user. 10 terms were selected “actively interested” and additional 16 terms were selected “passively interested” by at least 25% of the respondents.

These results show, that users are interested in information related to terms which are more frequent in private information spaces than in the public information space of the organization. While our questionnaire did not ask if the respondents were able to satisfy their active information needs, we assume that they would be interested in results provided by their colleagues.

Clearly, our study suffers from the impreciseness and possible ambiguity of single terms. However, we believe that our selection of entity-related terms allows at least for a rough approximation of real information needs. Also, due to feasibility reasons, our questionnaire was based on a small and subjective selection of 50 terms. While we think that the analysis shows already some promising opportunities for our solution, an *Inverse Search* client installed on the users' computer could discover many additional terms of active interest beyond those 50 terms.

5.4. Related Work

5.4.1. Peer-to-Peer Information Retrieval

There are several approaches that leverage peer-to-peer (P2P) technologies for information retrieval tasks. The main idea is that traditional centralized search engines do not scale naturally, hold centralized control and fail to provide access to the whole web. P2P-based approaches such as Minerva (Bender et al., 2005) or P-Grid (Aberer et al., 2003) assume that peers carry out crawling and indexing tasks, maintain a local index, share parts of it, and provide services for other peers and searchers. Similar to our approach, peers publish statistical metadata (to the peer network). (Podnar et al., 2007) employ query logs to reduce traffic necessary for maintaining the index.

However, all these approaches are based on the assumption of shared information, even if it does not reside on central machines. While users do not have to hand out their results by default, they must provide metadata (i.e. index terms) to have the system working. In contrast to this, our approach does not require information providers to reveal any content they have but information seekers to reveal parts of their information need.

5. Inverse Search: Recommending Users to Share Documents

5.4.2. Prospective Search

There is a number of works dealing with the notification of users when new information with respect to a certain query is available. A popular example for such a system is Google Alerts (Google Inc., 2017a). In literature, this is called *prospective search* (Irmak et al., 2006), *continuous querying* (Kukulenz and Ntoulas, 2007) or *retroactive answering of queries* (Yang and Jeh, 2006).¹⁹ These papers mainly concentrate on selecting relevant documents to existing queries.

Thus, research in this area is orthogonal to our approach, since we strive to diffuse information from private to public spaces. Once this diffusion happened, strategies from prospective search can be employed to notify interested users (see Section 5.1.3).

The work of Yang and Jeh (2006) is particularly related to ours. Similar to our notion of organizational information need, it defines *standing interest* as a measure for queries for which a user would be interested in future results. The authors discuss a number of indicators retrieved from query logs, which partly overlap with our calculation of OIN. However, they focus on deriving standing interest for a specific user, while our notion of OIN relates to a groups of users.

5.4.3. Enterprise File Sharing

Besides descriptive user studies,²⁰ research has not paid much attention to enterprise file sharing in recent years. A notable exception is the *cattail* system, which allows to keep files in private, shared, and confidential spaces, and allows for tagging and annotation, besides the obvious upload and download actions (Muller, Millen and Feinberg, 2009, 2010). Beyond that, *cattail* is also designed as a *social* file sharing system, including social network features with profile pages, recent changes lists, and a “File Page” aggregating activities and comments for each particular file.²¹

Shami et al. (2011) found in an interview study that the those social features of *cattail* significantly raise exploration and also *encouraged contribution* to the system. In our terms, they can be considered as *mediation services* and *mediation spaces*, which could nicely complement *Inverse Search* in future versions. Muller, Shami, Millen and Feinberg (2010), also consider search terms as potential signals of interest that could “help knowledge-creators to serve the needs of their readers”. Thus, *Inverse Search* features could be a natural extension for *cattail* and especially help to move additional files form private to public spaces.

¹⁹See also Section 3.1.2

²⁰See, e.g., Rader (2009); Massey et al. (2014)

²¹Our *Woogle* tool uses similar “Woogle-Pages” for queries (see Section 6.2.2)

5.4.4. User Interfaces for Access Control

Setting access permissions is a pivotal action in the lifecycle of a shared document or artifact, since initial sharing decisions will never be revisited in the majority of cases. Also, access control needs to balance the sensitive “tension between the need for security and the need for access” (Whalen et al., 2008b). Hence, various scholars have turned their attention towards current practice and potential improvements.

Exploratory studies were conducted by a number of researchers, yielding design guidelines based on individual sensitivity for privacy issues (Olson et al., 2005), feature comparisons of sharing tools (Voida et al., 2006), and surveys about the usability of access control (Whalen et al., 2006).

Several researchers also conducted design studies and tool evaluations. Based on studies with a Web URL sharing tool, Lau et al. (1999) argue that users lose control when privacy is merely a property of objects. Instead, they suggest to make access policies objects of their own right. Studying the behavior of users in photo-sharing, Ahern et al. (2007) also discovered that users found it difficult and demanding of attention and time to make good choices about data disclosure. They argue that awareness feedback about the particular audience might be helpful. They also note, that 7% of access settings were changed after the initial upload (2.4% from public to non-public and 4.6% from non-public to public).

This is noteworthy, since most approaches consider sharing as a one-time activity and do not consider the evolution of access policies (defined as “change control” by Razavi and Iverson, 2007). To this end, Dalal et al. (2008) suggest temporary, ad hoc access control policies to accommodate for unplanned sharing activities. Whalen et al. (2008b) extended a file manager user interface which provides special visual indicators to raise awareness about which files are shared. They also provide a *sharing console* which is related to our *File Sharing UI* depicted in Figure 5.4. However, their approach does not include concepts of gaps or needs to signal the potential value of sharing to users.

Finally, Mazurek et al. (2011) present a novel approach called reactive access control, which allows recipients to request access to certain files. While this can be considered as an approach for *retrospective* access control, it is similar to our approach in that way, that file owners are explicitly requested to share certain files. However, the approach raises privacy concerns for information providers, as file names need to be visible to recipients even when files are not accessible for them. Mazurek et al. (2011) found, that providers would tolerate between five and 15 sharing requests per day, and that they were particularly interested, why receivers were requesting access to their files.

5.5. Summary

In this chapter, we presented *Inverse Search* as a novel approach that aims to foster knowledge sharing in organizations. It is designed to stimulate the

5. Inverse Search: Recommending Users to Share Documents

diffusion of relevant documents from private information spaces of particular users to the public information space of an organization. Based on the notion of organizational information need (OIN), Inverse Search recommends people to share private documents containing information relevant for other members of their organization.

As a proof-of-concept, we realized an implementation based on the Eclipse Rich Client platform. The application can be used as a normal desktop search engine for searching within the private information space of the user and the public information space of the organization. However, users might not just *pull* information from the public space, but can use a sharing user interface which provides them with a list of currently private files that might be shared.

The approach contains three major contributions. First, we proposed to extend the scope of information retrieval models by explicitly addressing information providers in order to analyze how and why information should move to public information spaces. Second, we provide a mechanism for identifying documents that should be shared by deriving an organizational information need (OIN) from query logs. Third, our system provides means to foster the diffusion of information by recommending users to share private documents that cover such organizational information needs. To our knowledge, there is currently no comparable approach which systematically guides the diffusion of existing knowledge from private to public information spaces.

To evaluate the approach, we conducted two consecutive evaluations. First, we analyzed the private and public information space of 13 users in a working group, showing that many terms (and hence documents) in their private spaces were unique within the group – i.e., a considerable organizational information gaps exists. In a second step, we checked to which extent these unique terms (and hence corresponding documents) might be helpful for other users in the organization. Out of 50 unique terms selected for evaluation, 10 terms were rated to be significantly related to their past and current information needs by at least four subjects. Given the qualitative nature of this study, these results indicate that a tool-supported analysis of information needs could bring considerable benefits to the organization.

While we have shown the general use and feasibility of Inverse Search, our implementation is not yet optimized for productive use. As discussed in Section 5.2.2, different user interface alternatives should be considered in more detail. Once a productive implementation exists, a long term user study should be conducted to analyze the acceptance of the approach in real settings. Based on those observations, improvements on the user interface and underlying metrics can be made. Interesting follow-up questions would also be notifications to information seekers and statistics about the actual usage of shared documents.

Beyond files and folders considered in this chapter, Inverse Search could also be adapted to other settings which include private and public information spaces. Examples could be social sharing applications such as *social bookmarking*²² or information sharing systems based on mobile phone apps.

²²See, e.g., Millen et al. (2006); Benz et al. (2010)

6. Woogole: Guiding Contributions to Wikis

Wikis can be considered as a category of Web-based groupware systems, which allow for easily capturing and disseminating information in a community or organization. While initial examples, such as the design patterns community Wiki¹ or most notably the Wikipedia,² emerged in the public Internet, there has also been an increasing adoption of corporate Wikis (see, e.g., Bughin and Manyika, 2007; Economist Intelligence Unit, 2007).

Wikis are distinct from conventional groupware systems by stressing the “Web”-aspect (i.e., by requiring a Web-browser and heavily relying on hyperlinks) and favoring the open editing of content instead of access permissions. Maybe most importantly, Wikis enforce conceptual integrity between the URL (e.g., <http://en.wikipedia.org/wiki/Karlsruhe>), title ("Karlsruhe"), and the content of a page. This simple principle plays a key role in making Wikis a powerful tool for fostering the accumulation and maturing³ of information (Happel and Romberg, 2008).

Due to its collaborative editing features, Wikis offer several discussion and awareness mechanisms such as a so-called “recent changes” list, change notifications, and discussion pages. Wikis provide space for discussing, commenting, and linking to other resources outside the Wiki and can thus help to “glue” together distributed information.

In enterprises, Wikis are typically used to collect and refine small pieces of unstandardized, immature information (Braun and Schmidt, 2007). They thus often serve as a common starting point when users are searching for information, similar to the way people start exploring a topic by reading its Wikipedia article. Just as for the Wikipedia, a major challenge for Enterprise Wikis is however to make users actively contribute, as it’s difficult for them to find out *why* they should put *which* information into the Wiki (Majchrzak et al., 2006).

However, while Wikis are a place for both, information seeking and knowledge sharing, these two aspects are only loosely related. Wiki collaboration features are typically not available for search activities and knowledge sharing is not directly guided by the information needs of the users.

We argue that information seeking and knowledge sharing in Wikis are different, but closely related aspects of information processing, which could significantly benefit from each other (see also Section 3.3). To this end, we created

¹<http://c2.com/cgi/wiki?WikiWikiWeb>

²<http://www.wikipedia.org>

³See also Section 2.2.4

6. Woogle: Guiding Contributions to Wikis

Woogle, a tool to improve search with collaboration features (“social search”⁴) and to guide knowledge sharing using actual information needs of the user community (“need-driven knowledge sharing” – see Chapter 4).

In the following, we first discuss the current state of search and knowledge sharing in Wikis. Afterwards, we describe Woogle as a concept to improve both issues by mutually connecting both processes. We line out the general design principles and present our Woogle reference implementation, *Woogle4MediaWiki*. We then describe our evaluation approach and discuss results from two studies. Finally, we summarize related work and give an outlook to future work.

Work described in this chapter is based on a number of prior publications. The dual role of Wikis for knowledge sharing and searching information has been initially discussed in Happel (2009c). Happel (2009a) describes the *Woogle4MediaWiki* tool and first evaluation results. Happel and Mazarakis (2010) elaborate on additional challenges such as collaboration in search and motivational aspects.

6.1. Problem

This section contains a deeper investigation of the current problems in *retrieving* and *sharing* information in Wikis based on the issues identified in Section 3.4. At the end, the as-is situation and our envisioned to-be state are illustrated by means of a motivating example.

Since our reference implementation is based on MediaWiki (Barrett, 2008) and many of our example screenshots stem from the Wikipedia (which is also using MediaWiki), our discussion will mainly draw from this system.⁵ However, most other Wiki implementations are based on similar concepts.

6.1.1. Information Seeking and Retrieval in Wikis

Lack of demand treatment

Wikis offer various means of expressing information needs, which differ with respect to explicitness and expressivity. The most obvious one is the *search* function. Plain MediaWiki does however not maintain query logs, which might be helpful to identify popular searches.

The second, more explicit means is related to linking to other Wiki articles. Wikis are unique in allowing to *reference non-existing pages*. In contrast to normal Web links to non-existing URLs, these references are not rendered with a “404 – not found” error message⁶, but instead, the Wiki provides an empty edit form to write a new article. Within the referencing text, such links

⁴See also Section 3.1.2

⁵Beyond Wikipedia, MediaWiki is also one of the most popular Wiki systems for Wiki-based websites and for enterprise use

⁶See RFC 7231 (<https://tools.ietf.org/html/rfc7231#section-6.5.4>)

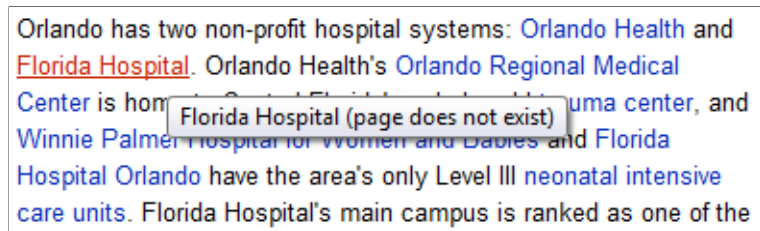


Figure 6.1.: “Red link” in the English Wikipedia

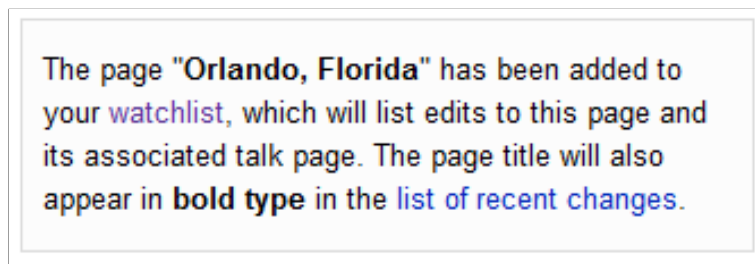


Figure 6.2.: Adding an Article to the Watchlist

are typically rendered in *red color*, which denotes that this page is still to be written and thus implies an information need (see Figure 6.1).

A more personal form of expressing an information need is to *watch* pages in the Wiki (see Figure 6.2). In this case, the user will be notified about changes by email.⁷ While this feature is primarily used for awareness and vandalism protection, some people might use it to learn new information.⁸ This makes sense, since information on a topic might be evolving (Yang and Jeh, 2006), especially in a Wiki system. MediaWiki also allows to watch empty, “red link” articles.

Finally, some communities provide means for explicitly *requesting articles*. For instance, the English Wikipedia community maintains lists with “requested pages”.⁹ Uttering such requests is also suggested from within the Wikipedia search interface, if no results can be found (see Figure 6.3). This is not a built-in feature of the Wiki engine, but rather a community-maintained effort.

Lack of awareness

Wikis have been designed with several features that support awareness for information seekers. First, there is the mentioned *watch* mechanism, which allows users to get notified when existing content changes. Furthermore, most Wiki engines offer a list of *recent changes*¹⁰ which gives a quick overview about

⁷Alternatively, watched pages are highlighted in the list of recently changed articles

⁸Similar to prospective search as discussed in Section 3.1.2

⁹See http://en.wikipedia.org/wiki/Wikipedia:Requested_articles

¹⁰See, e.g., <http://en.wikipedia.org/wiki/Special:RecentChanges>

6. Woogle: Guiding Contributions to Wikis

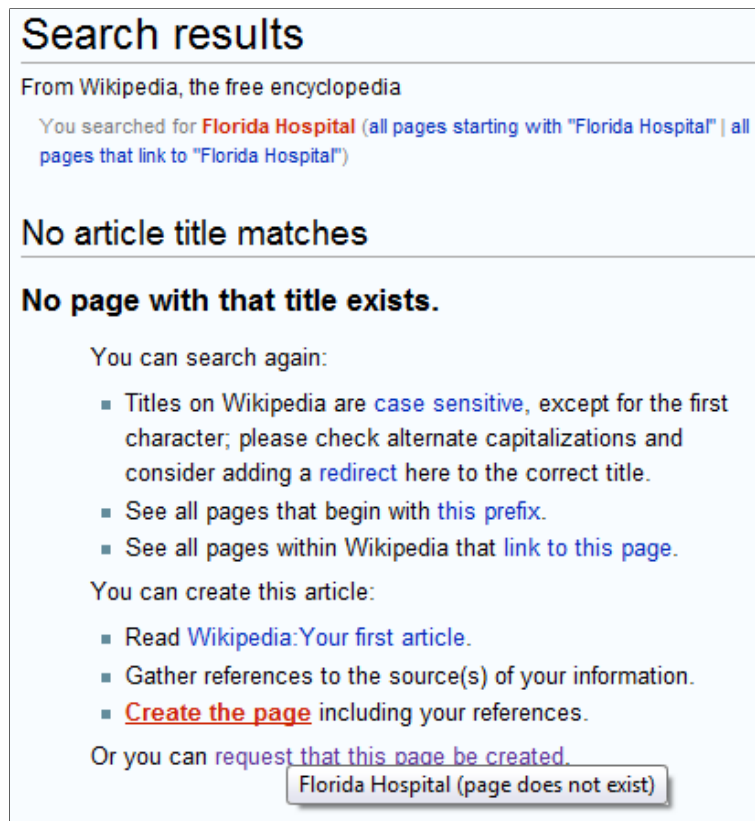


Figure 6.3.: Unsatisfied Search in the English Wikipedia

all modified as well as newly created pages. However, more sophisticated features such as *prospective search* are typically not available.¹¹

Lack of possible actions

Given the absence of restrictive access rights and the ease of contribution, Wikis are all about participation. Surprisingly, the search functionality of common Wiki engines is rather conventional, i.e., there are no explicit means for contribution or collaboration. An exception is the transition from information seeking to contribution, which is supported by offering possible actions if no search results were found (see Figure 6.3).¹²

6.1.2. Sharing Knowledge in Wikis

Wikis are all about knowledge sharing. Thus, the level of tool support is relatively high, as our observations in Section 3.2.2 indicate. However, the

¹¹Besides the simple “watch” mechanism on content pages, which was discussed before

¹²The page has been modified for the English Wikipedia. The same page in plain MediaWiki is much less supportive for potential contributors.

following paragraphs will show that knowledge sharing support in Wikis could be further improved nevertheless.

Lack of awareness

Knowledge sharing in Wikis can occur in various forms. The main approach is to *contribute to an existing article*. A user reading a page can do so by switching into the *edit* mode and enter additional text. However, she typically has to decide on her own if the page needs contributions. Only some larger Wikipedia communities have developed visual markers, which are added manually to indicate missing content.¹³

Red links – i.e., pages without content (see Figure 6.1 in the previous section) – are a second entry point to share knowledge. There are several ways to encounter red links. The obvious one is within an existing article (as in Figure 6.1). Furthermore, MediaWiki provides a page “Special:WantedPages”¹⁴ which lists empty pages ordered by the number of other pages linking to it. While is helpful to identify highly requested pages, it does not allow authors to easily spot missing articles in their personal areas of interest. Larger communities such as the Wikipedia are therefore maintaining separate lists of “requested articles” (see Section 6.1.1).

Although the described measures are effective to raise awareness about missing content, one can argue that they mainly represent the information needs of a small fraction of users. A *red link*, e.g., is typically set by a contributing author. Thus, the existing measures might not be effective to make contributors aware of the information needs of readers who are seeking information.

Lack of guidance

Many of the measures described in the previous paragraph offer some guidance to contributors. *Red links* or lists of *requested articles*, for example, always include the name of the desired content page. The large amount of requested content however makes it difficult to decide where to start contributing. E.g., the number of pages linking to a “red link”-page provides *some* prioritization of demand (considering the number of links an indicator for demand), but this is not necessarily a good measure to estimate the information missed by Wiki readers.¹⁵ Furthermore, scanning any list of wanted pages in order to identify suitable topics to write about to is a tedious task for any contributor.

¹³See, e.g., “stubs” in the English Wikipedia: <http://en.wikipedia.org/wiki/Category:Stubs>

¹⁴See, e.g., <https://en.wikipedia.org/wiki/Special:WantedPages>

¹⁵Also, the ranking of articles on the WantedPages list is often distorted by technical issues such as Wikipedia templates

Lack of privacy

As Wikis do not support sophisticated access rights,¹⁶ there are no spaces in which users can store private or sensitive information. For this reason, a situation that can often be found in practice is to set up multiple Wikis, which are accessible by different groups of people. This might range from personal Wikis for individuals, Wikis for project groups and departments, up to enterprise-wide or public Wikis.

6.1.3. Motivating Example

In this section, we illustrate the current *as-is* situation and the *to-be* situation as envisioned in this chapter.

As-is situation On her first working day, Alice learns about her new working environment. Among other things, she likes to know about the parking facilities and thus seeks for “Parking” in the organization’s Wiki. Since no results are returned, she would like to post a message somewhere, but she is not sure where to post it in the Wiki, and she does not want to send an E-Mail to her department’s mailing list in this case. She decides to ask a colleague over lunch.

Bob is working for the same organization and responsible for the facility administration. When he opens the facility overview page in the Wiki, he sees a couple of links to pages that are not yet filled with content. Among these links, which are highlighted in red color, is a link to the “Parking facilities”.

To-be situation Alice searches for “Parking” in the organizational Wiki. While her search does not return any results, a text on the search result page states, that two other people were searching for “Parking” within the last week. Since there is an “edit” button underneath this text, Alice decides to leave a message, asking for information about parking facilities at the company.

As Bob is watching the facility overview page, he recognizes that the red-colored link to “Parking facilities” is highlighted with a small visual indicator (e.g., “*”). When he hovers over the link with his mouse cursor, a small popup window appears, stating that three users have been searching for related content within the last week. He thus decides to create an initial article, by clicking on the link and editing some basic information.

¹⁶Even though there exist basic access rights for MediaWiki, fine-grained access control is not available, as this is somehow opposed to the overall idea of a Wiki (see also https://www.mediawiki.org/wiki/Manual:Preventing_access#Restrict_viewing_of_certain_specific_pages). Specialized enterprise Wikis however, such as Confluence (Kohler, 2013) do offer more advanced permission schemes.

6.2. Design and Implementation

In this section, we introduce an approach to guide knowledge sharing in Wikis which we call *Woogle*. We present the overall approach, the technical architecture, and its the prototype implementation *Woogle4MediaWiki*.

6.2.1. Approach

The Woogle approach consists of two major facets, which address both roles in the information seeking process (see Figure 4.3). In particular, it helps *information seekers* by providing “social search”¹⁷ features, to allow for mutual interaction and information exchange with fellow searchers, and for contributing to the improvement of the search engine. *Information providers* are targeted by adding “need-driven knowledge sharing” features¹⁸ to motivate and guide their contributions. Our core idea is to bridge the artificial separation of information seeking and information provisioning¹⁹ by seamlessly integrating and mutually improving both processes.

Social Search In particular, we strive to improve information access in terms of better representing desired information, and supporting the elaboration of information needs:

- **Give queries a first order representation** to serve as a common point of reference during information seeking and information provisioning. The idea is, that queries are no longer transient interactions, but get a persistent and referencable representation within the Wiki. In an essence, queries shall become artifacts on their own, which can evolve beyond a search request of a particular user.
- **Provide means for communication and awareness** such as discussion and notification mechanisms. The intended mechanisms should generally be able to cover the whole collaboration process in information seeking which can be separated into the three phases “before search”, “during search” and “after search” (Evans and Chi, 2008).

Need-driven Knowledge Sharing In an enterprise setting – but also in large communities such as Wikipedia – resources for knowledge creation and sharing are limited. Thus, knowledge sharing activities should be prioritized towards knowledge which satisfies the most “popular” information needs. Information provisioning should thus be improved by providing an explicit notion of sought information needs, and by providing means for an easy sharing of information within the information seeking process. In particular, our approach includes:

- **Seamless transitioning** from information seeking to information provisioning, in order to achieve a tight integration and holistic support for

¹⁷See also Section 3.1.2

¹⁸See also Chapter 4

¹⁹See also Section 3.3

6. Woogle: Guiding Contributions to Wikis

user's information behavior. This is intended to ease switching between the tightly intertwined roles of an information seeker and an information provider.²⁰

- **Provide different modes of information provisioning**, such as creating explicit information need descriptions, annotations, and adding new information easily and with low barriers.
- **Provide guidance for content creation** which helps information providers to identify the most important information gaps within the Wiki content.

According to the concept of NKS, information needs derived from search queries are supposed to be a major driver for knowledge sharing activities. However, the search features in most Wikis are rather limited. On the other hand, Wikis often “compete” with *enterprise search* tools for the first option to start an information seeking attempt. Since both tools complement each other quite well, we suggest to resolve this competition by integrating enterprise search features into the Wiki (Happel, 2009c).

This means, that the Wiki search function is supposed to provide not only Wiki pages but also external documents as results. Besides capturing information needs required for NKS, such an approach has also several other benefits. First, it provides a central point of access, as recommended by KM researchers. Second, it can help to raise Wiki acceptance and usage by bootstrapping and complementing information stored within the Wiki.

6.2.2. Architecture

We now describe the architectural considerations underlying our approach. While these are meant to be independent of a particular Wiki technology, some screenshots and examples are taken from MediaWiki, which is used for the reference implementation presented later on. We begin with the *general architecture* of Woogle and continue describing specific features for *social search* and *need-driven knowledge sharing*.

General architecture Woogle is intended as a modification or extension of existing Wiki engines, since most of them have some kind of plug-in mechanism. For its realization, Woogle requires backend features for search and query logging, and modifications to the user interface. Regarding backend functionality, Woogle needs two major features. One is the ability to search for Wiki pages using keyword-based search queries. Furthermore, a database-powered structured query log²¹ is required to store user's search queries and result click information.

Since many Wiki engines offer rather weak search functionality and do not offer programmatically accessible query logs, we decided to allow for two different

²⁰See also Section 3.3

²¹See Section 4.4.1

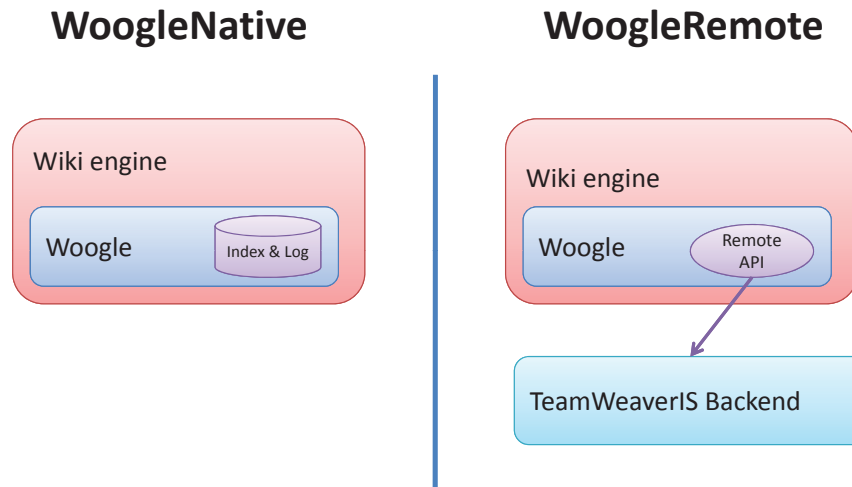


Figure 6.4.: General Architecture of Woogle

ways to instantiate the backend (see Figure 6.4). *WoogleNative* is the standalone implementation of the Woogle concepts, which is completely integrated into an existing Wiki engine.

Alternatively, *WoogleRemote* uses a remote installation of TeamWeaverIS, which is an advanced information retrieval framework that allows to index and search documents from a large set of different repositories and data formats (see also Section 4.4.4). Woogle can access any TeamWeaverIS backend via a Web service interface. Content of the Wiki itself can also be indexed using the remote backend. While requiring a more complex technical setup, WoogleRemote allows realizing enterprise search functionality using the Wiki as a user interface (Happel, 2009c). WoogleNative in turn does not support other content than Wiki pages, but is typically easier to set-up and maintain, as it does not need to connect to another system.

Social Search To give queries a first order representation, we assign a dedicated Wiki page (“Woogle-Page”) for each individual query.²² Once opening such a Woogle-Page – by entering its URL or via the Wiki search feature – a Wiki page which contains a list of search results at its bottom (see Figure 6.5) is shown. For WoogleRemote, the results can also include external documents, which are not part of the Wiki itself.

Besides search results, a “Woogle-Page” presents some immediate search-related “actions” to the user:

- A **freely editable text box** at the top of the page should allow to describe or disambiguate the information need. The text box is restricted in size to prevent the creation of too much “original content” on Woogle

²²The *cattail* enterprise file sharing system uses a similar concept of a “File-Page” (Shami et al., 2011); see also Section 5.4.3

6. Woogle: Guiding Contributions to Wikis

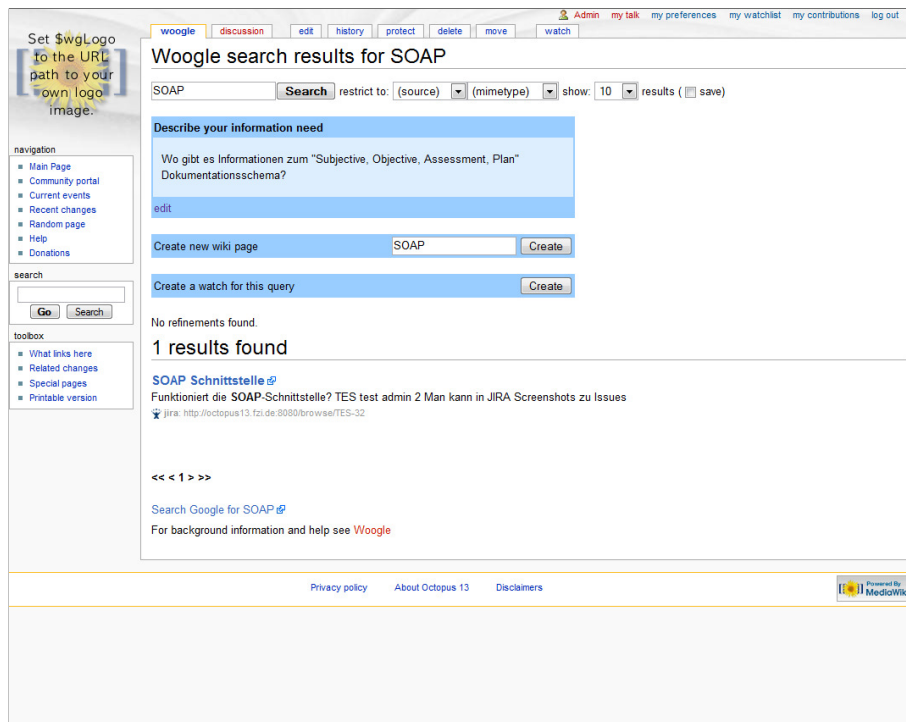


Figure 6.5.: Woogle4MediaWiki Search Results List Screenshot

pages, respectively to force users moving such content to regular Wiki pages.

- A **discussion space** for each query (i.e., “Woogle-Page”).
- The immediate possibility to **create a new Wiki page**, if no suitable search results exist.
- A “watch” feature to receive **notifications** when new results arrive or changes to the “Woogle-Page” occur.

Need-driven Knowledge Sharing For *collecting information needs*, we use “red links”, watches, requests, and queries as described in Section 6.1. In the same section, we identified the following situations as main starting points for knowledge sharing:

1. contributions to existing pages
2. red links
3. search and
4. overview pages

Our current implementation primarily addresses points 2) and 3), while we’ll also discuss the other two points as candidates for future work.

Regarding *red links*, our approach provides Wikis users with direct feedback about how much a certain red linked page is sought-after by others. Such

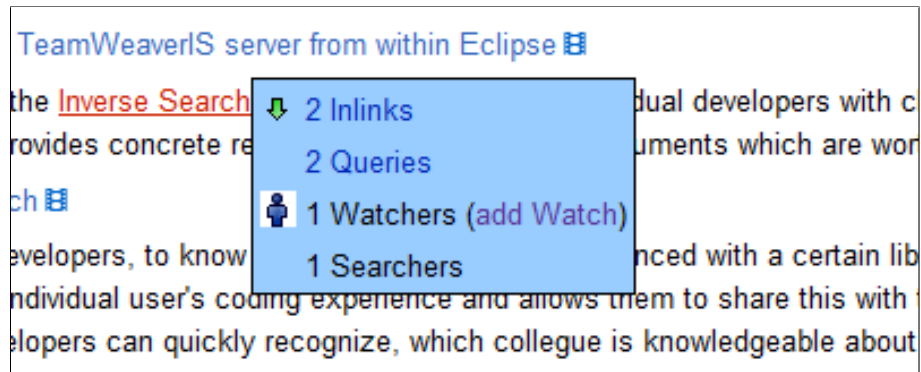


Figure 6.6.: Woogle Mouse-over Display for “Red Links”

feedback can be placed around an red link, or a mouse-over popup, as depicted in Figure 6.6. We aggregate information about “inlinks” of a red link page and queries to a priority value. This is reflected by an “arrow” icon, which symbolizes the relative priority using a three step scale (low, medium and high priority).²³ The number of people *watching* the red link page and the number of users *searching for its title* are aggregated into a “people” icon, denoting the “organizational breadth” of the information need (few people, average, many people).²⁴

Our second focus is the *search* process. As described in Section 3.1.2, search is a highly collaborative activity. Therefore, “social search” features, offering collaboration and knowledge sharing facilities right within the search environment (see Section 6.2.1) are complemented by visual indicators which signal the need for certain information. Some of these indicators are similar to those related to red links. Besides icons for priority and organizational breadth, we add two further signals which characterize the information need. One is a “clock” icon, which denotes how the information need is distributed within time (recent, average, or outdated information need).²⁵ This can help potential contributors to estimate, if there is recent demand for some information. Finally, we add a traffic light icon, which symbolizes the result quality for the given query.²⁶ All these icons are displayed for each query in the search interface as shown in Figure 6.7.

Privacy concerns When discussing social search and knowledge sharing, privacy is a major issue for many users (Burghardt et al., 2008). For the realization of our system, we had to address this in two different ways – concerning information collected for the core operational part, and concerning the evaluation forthcoming in Section 6.3.

²³See the arrow pointing down for low priority in Figure 6.6

²⁴Figure 6.6 shows an icon depicting a single person to denote “few people” are interested in the red link

²⁵In Figure 6.7, the “clock” icon denotes a recent information need, as only few time has passed

²⁶The green traffic light in Figure 6.7 denoting a good result quality

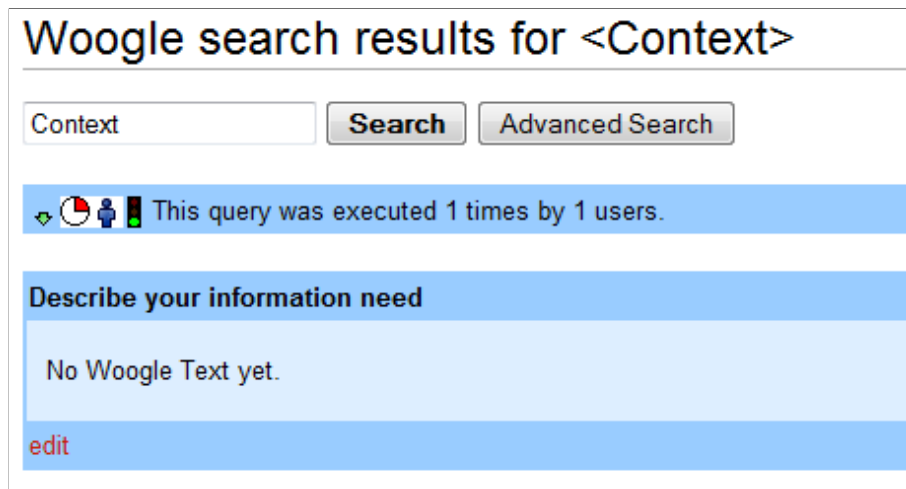


Figure 6.7.: Information Need Indicators in the Woogle4MediaWiki Search UI

The operational part is about the nature of our system design, which is heavily based on user queries and result clicks. Therefore we address privacy issues by introducing a randomized `userId`. This Id, which can also be changed by the user at any time, is only used for query logging and thus keeps data sent to the remote query service anonymized. Users may also completely disable the submission and logging of query and click information.

Scientific evaluation of our system (see Section 6.3) is even more critical from a privacy perspective, since it needs to gather additional data. In order to take user's privacy concerns serious, we implemented a participation dialog²⁷ which explains the data logged by the system in detail, and which asks for explicit agreement of the user. The user may revisit this decision at any time in the preferences, where also the Woogle extension as such can be disabled.

6.2.3. Design

Based on the query log information that is logged,²⁸ the Woogle user interface adapts and presents summarized information about information needs. In particular, this is the case for the so-called “red links” and the main search interface, as described before. We will now elaborate on the computation of the values underlying the visual indicators, which we use to signal meta-information about information needs.

Red Links The goal of applying information need indicators on “red links”, is to help users to better distinguish missing pages based on the underlying demand. Therefore, the popup-display shown in Figure 6.6 lists four pieces of information. Two visual indicators are derived from these values: the “arrow”

²⁷See Figure B.1 on Page 184

²⁸Respectively the TeamWeaverIS backend accessed by Woogle

icon depicting the relative priority, and the “person” icon depicting the so-called “organizational breadth” of the information need.

As for the *relative priority*, the underlying values are the so-called “inlinks” (i.e., the number of Wiki pages referencing the red link), and the number of previous search queries, which match the title of the red link page. The icon to display is derived as follows:

- **Low priority:** If both values (inlinks and queries) are below five
- **Medium priority:** If both values are greater or equal five and smaller or equal ten
- **High priority:** If both values are above ten

Current values are determined based on experience about typical “red links” in smaller Wikis, as we will later target with our evaluation. The validation, respectively improvement, of these values, the scale, and the related icons is subject to future work.

The second icon, depicting the “organizational breadth” is derived from the number of “watchers” (i.e., Wiki users that subscribed to the red link page) and the number of different users underlying the queries matching the red link page title. The actual icon to display is chosen as follows:

- **Few people:** If both values (watchers and searchers) are below five
- **Average number of people:** If both values are greater or equal five and smaller or equal ten
- **Many people:** If both values are above ten

Similar to the icon for *relative priority*, values are based on initial personal experience and should be considered subject to ongoing refinement.

Search UI On the search result page, we show four kinds of icons to indicate different dimensions of the information need. Out of those, the two icons for *relative priority* and *organizational breadth* are similar to those for red links. As additional icons, the result page shows the *recency* of the information need and the *result quality* for the given query.

The *recency* is symbolized using a “clock icon”. The underlying idea is to convey an impression if the query is popular in the recent time, or if it was rather executed some time ago. Based on the *recency* value stored in the query log (see Table 4.3 on Page 77), the icon will display as follows:

- **Recent information need:** Average query timestamp is within the last week (depicted by a quarter-full clock)
- **Average information need:** Average query timestamp is between recent and outdated (depicted by a half-full clock)
- **Outdated information need:** Average query timestamp is more than one year ago (depicted by a full clock)

6. *Woogle: Guiding Contributions to Wikis*

The *result quality* icon intends to show how satisfied querying users were with the results that are displayed for the query. It is symbolized by a “traffic light” icon, which is populated as follows:

- **Green light:** We assume a good result quality, if the number of clicks per query execution is near to one (>0.9 and $<1,1$).
- **Yellow light:** Number of clicks per query execution is >0.8 and <1.2
- **Red light:** A lower result quality is indicated, if it is significantly below one (which means, that there are no result clicks for some queries), above one (which indicates that information may be scattered across many results) or if users tend to browse across result pages (which indicates a low satisfaction with information on the first result page respectively a high information need)

Note again, that the thresholds described need to be explored and refined in future work. In particular, values should dynamically adapt to the size of a Wiki and/or the amount of Wiki users. As our current research interest however is focused on understanding the impact of visual need indicators as such, it should be sufficient if they represent the underlying information needs roughly.

6.2.4. Implementation

The reference implementation of the Woogle concept is realized for the MediaWiki software (Barrett, 2008) and hence called “Woogle4MediaWiki”. Reasons for choosing MediaWiki were a) that it is the most popular Open Source Wiki engine²⁹ and b) that it has a relatively modular architecture, which allows to extend the MediaWiki core using third-party plug-ins.³⁰

MediaWiki defines so-called “Hooks”, which are extension points that external call-back functions can register for. At runtime, each hook will look-up the list of registered functions, and call each of the functions with a defined set of parameters. By means of modifying these parameters, functions can influence the execution of the MediaWiki core. MediaWiki offers more than 400³¹ of such hooks to modify aspects as diverse as user management, parsing, or page layout.

A coherent set of MediaWiki modifications is typically packaged in a so-called “Extension”. MediaWiki extensions provide meaningful features to end users, which are not available from the core implementation. The MediaWiki documentation site alone lists more than 800 extensions rated to be “stable”.³² Woogle4MediaWiki is thus realized as an Extension for MediaWiki.

²⁹Mainly due to the fact that it is used by many Wikimedia projects (see also <http://wiki.c2.com/?TopTenWikiEngines>)

³⁰We have also developed an early prototype of the Woogle for the Atlassian Confluence Wiki engine (Kohler, 2013)

³¹See https://www.mediawiki.org/wiki/Manual:Hooks#Alphabetical_list_of_hooks

³²See https://www.mediawiki.org/wiki/Category:Stable_extensions

Regarding backend functionality, Woogle4MediaWiki introduces two major features. One is a keyword-document index, which is used to improve the search experience.³³ Furthermore, additional database tables are introduced to store query and result click information (see Section 4.4). As described in Section 6.2.2, the backend can be instantiated either as a bundled feature (“WoogleNative”) or using an external search engine (“WoogleRemote”).³⁴

To realize “Woogle-Pages” for queries as depicted in Figure 6.5, Woogle4MediaWiki uses a special “namespace” in MediaWiki. This means, that all pages with an URL-prefix “Woogle:” are processed by our extension. Accordingly, queries are represented as “Woogle:query”, yielding bookmarkable query pages which do not interfere with the regular Wiki content.

Finally, Woogle4MediaWiki contains an instrumentation framework to support scientific evaluation. The framework can ask users for consent to use the extension and can randomly assign them to experimental groups. These groups in turn can be configured to enable or disable certain Woogle4MediaWiki features. The instrumentation framework also generates additional log output to help analyze user behavior retrospectively.

6.3. Evaluation

The main goal of our system is to improve search and knowledge sharing within Wikis by creating a feedback loop between information seeking and knowledge sharing.

Therefore, there are two major claims that need to be validated:

1. Do users understand the system and are they willing to use it?
2. Is the system effective in improving knowledge sharing within a Wiki?

The evaluation of a system like Woogle is challenging for multiple reasons. First, due to the novelty of the proposed approach, it is important to analyze if users are able and willing to use the system. Second, evaluating the underlying functionality requires a large number of users and queries (to allow for meaningful aggregation) and a long time frame (to allow for collecting a broad number of unsatisfied needs).

To address these different issues, we finally chose a mixed-method evaluation approach, consisting of qualitative interviews and an online field experiment. Both evaluations and their results will be presented in the following sections.

³³The built-in search relies on database queries and is thus limited in query features and performance.

³⁴“WoogleNative” is easy to install and particularly suitable for smaller Wikis with a few hundred Wiki pages

6.3.1. Qualitative Evaluation

Design Process

We decided to start with conducting qualitative interviews with a limited numbers of potential users. Goals of these interviews were to:

1. Collect users' opinion on critical issues regarding contribution behavior and privacy awareness,
2. Get feedback on the system as such,
3. Devise design changes from users' feedback.

Therefore, we designed the qualitative interview with two major building blocks. The first one was a semi-structured interview, lasting around 30 minutes, which covered a number of questions regarding information needs, knowledge sharing, and privacy as drawn from existing literature.

The second phase, also lasting around 30 minutes, provided a walkthrough of the current Woogle4MediaWiki system. The system features were explained to the users and they were asked to describe if they understood the system, criticize it, and to optionally provide additional ideas or feature suggestions.

Results

We conducted qualitative interviews with five users. All of these users were experienced in the field of computer science. Three were graduates and two were students of computer science or related disciplines. All users reported to at least occasionally use Wikis.

We will first summarize the initial semi-structured interviews and then describe the feedback the users gave for the Woogle prototype.

Semi-structured interview The interview started with questions related to the *information search infrastructure* and information need behavior. Users primarily relied on searching in email or Desktop search (for their private information space) and Google (for the public information space). “Organizational spaces”, such as Wikis, were far less frequently searched – mostly due to inconvenient search features. Concerning frequency, users search the private sphere approximately five times a day and the web from 10 to 50 times a day.

Users reported that they typically had short term *information needs* – i.e. they were typically only interested in immediate results. As for the internet search, users indicated that they typically continue to search until they find at least a rough solution for their problem. A major problem described for private and internet search was formulating and finding the most appropriate query terms. All users described their search as an exploratory trial- and error process in such situations.

Regarding their *sharing behavior*, the majority of our users followed an “on-demand” approach – i.e., they share information mostly if they are directly

approached by other people. Two users reported to share knowledge in organizational systems occasionally without explicit triggers. Similarly, users differed regarding their treatment of content. Two people shared virtually all of their work-related information openly (e.g., in shared folders), while three people rather kept most of their work-related information private. This was mainly justified with the fear of being attributed for some immature or sloppy drafts of content.

The most interesting results appeared in the *privacy* part of the interview. All five subjects could be rated to be slightly concerned about privacy issues – i.e., they were aware of potential privacy problems but they typically made a trade-off between effort for retaining privacy and benefits they receive in turn. Regarding privacy in the context of search, only one user was really concerned about revealing search queries within the work environment. While the other interview partners were more relaxed, they however preferred a selective revelation of such information (e.g., reciprocally, only if another user searches for the same term) over a general listing of all searches carried out within the organization. They were less concerned, if the query information would not be directly associated with their names.

Evaluation of the Woogle prototype The Woogle4MediaWiki prototype was generally appreciated by our interview partners. Two users explicitly mentioned their impression that the features might motivate people to use the Wiki – although we did not reveal this as an explicit goal of our system.

Critical feedback was raised concerning the current style of representations. Three users were confused by the “information overload” created by the different icons and demanded a more simple representation. One user demanded the same for the social search features attached to the search result page.

Although not explicitly asked by us, four users were interested in getting more information about *who* was seeking certain information. While such features are foreseen in our systems design, we did not yet incorporate them into the live system. However, if feedback remains constantly positive concerning this issue, we might consider it for future versions.

6.3.2. Online Field Experiment

Design & Process

For the second stage of our evaluation, we chose the instrument of an online field experiment. In an online field experiment, groups of users are given different experimental treatments in order to investigate if a statistical difference regarding certain behavior can be observed among these groups. The advantage of this method is that it tries to combine the experimental control of laboratory studies with the field study advantage of operating in real world environments (Konstan and Chen, 2007).

Our major evaluation goal for the online field experiment was to investigate if Woogle can help to motivate users to share knowledge. Therefore we created

6. Woogle: Guiding Contributions to Wikis

three different groups. Group “alpha” was our control group which received the normal, plain MediaWiki search. Group “beta” received an improved Woogle search interface, with the *social search* features depicted in Figure 6.5. However, this group did not receive any automatically derived visual indications about actual information needs. This information, depicted in Figures 6.7 and 6.6 was only shown to group “gamma”. The distinction between “beta” and “gamma” was intended to identify which effects can be accounted to the raw social search interface and which to the information need indicators.

As mentioned in Section 6.2.4, our implementation includes an instrumentation framework which allows to a) collect detailed log data about user’s behavior and b) receive informed consent from users to participate in the experiment. The instrumentation component can be deployed to any Woogle4MediaWiki instance at any time. Once activated, registered users will be presented a participation dialog³⁵ when accessing the Wiki search for the next time. If they agree to participate, they are randomly assigned to one of the experimental groups. Users are free to revoke their participation at any time in the MediaWiki preferences menu.

The online field experiment took place within our institute’s internal Wiki system. The Wiki is used in a daily or weekly fashion for internal organization and knowledge management within our research group. At the beginning of the experiment, it contained 1 967 content pages and had 165 registered users, 77 of them denoted active by the MediaWiki metrics.³⁶

Results

The experiment was running for three months (122 days) in autumn 2009. To participate in the experiment, users were shown an explicit information dialog at the first search activity after our software had been activated. 47 users explicitly agreed to participate, while 30 users declined.

During the study, 924 different queries were executed, among which 263 (=28%) were yielding zero results.³⁷ Another 56 queries had only one search result and 62 queries had two results. 50% of these queries had no follow-up result click (see also Figure 6.8).

Regarding specific Woogle features, three users created seven “Woogle”-pages concerning four topics. All of these pages were used as forwards for synonymous queries.³⁸

Overall, 978 article saves were made by participants during the study. 206 times (i.e., in 21% of the cases), a “navigational search”³⁹ was preceding the edit. This stresses the importance of search as a precursory activity

³⁵See Figure B.1 on Page 184

³⁶I.e., users who have made at least one edit within the last 30 days (see also: <http://en.wikipedia.org/wiki/Special:ActiveUsers>)

³⁷See also Section 4.2.1

³⁸Similar to MediaWiki “redirect”-Pages as discussed in Section 6.4.4

³⁹Broder (2002); see also Section 4.2.1)

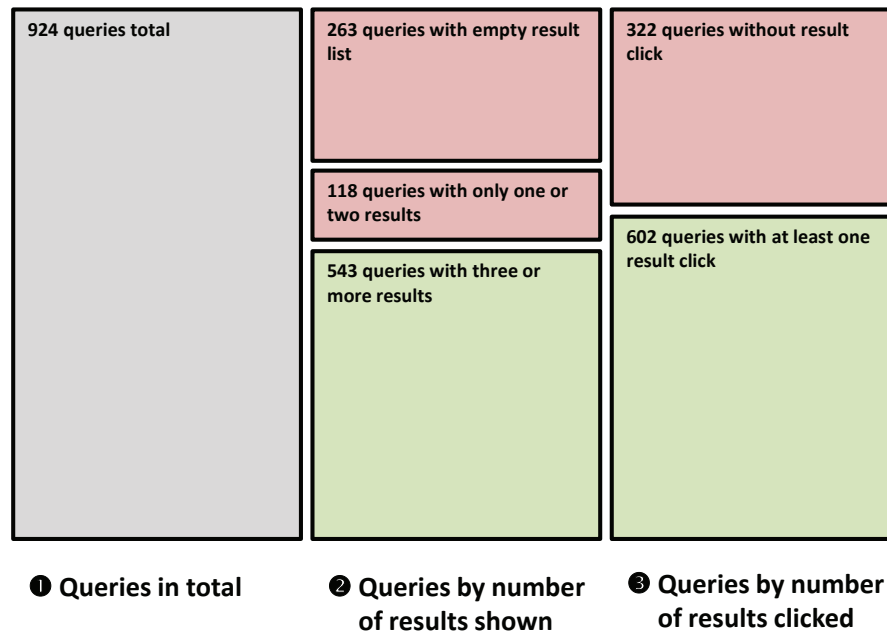


Figure 6.8.: Queries by Number of Search Results and Number of Search Results clicked

of sharing.⁴⁰ In 43 of these 206 cases, a *new* article was created. 25 of the navigational searches failed,⁴¹ which in 16 cases was resulting in the creation of a new article.

Overall, 136 times a “red link” was clicked by users, resulting in the creation of a new page in 60 cases. While 34 of these page creations were done by users which previously had created the “red link” (e.g., a user writing meeting minutes and creating article stubs thereby), 6 of the remaining 20 pages were created by users in the two experimental groups for which the red link popup feature was enabled.

In three of these cases, the popup was indicating no particular demand for that page, while in the other three cases particular demand information was shown. The interesting observation is, that in total 9 418 “red links” with no particular demand were shown to users, while only 1 869 links with demand information were shown, suggesting that the likelihood to contribute was higher in this case. However, due to the low number of pages created, we can not derive any statistical significance from this.

Summarizing, it turned out that the number of pages created due to red links is relatively high (60 out of 924 article edits). However, the subset of these links that benefits from Woogle’s features is too small to derive meaningful statistical results. On the other hand, the analysis of the search and contribution behavior in general indicates several opportunities for Woogle’s features

⁴⁰See also Section 3.3

⁴¹I.e., were yielding zero results.

6. Woogle: Guiding Contributions to Wikis

to increase knowledge sharing – considering, e.g., the high number of failed searches alone (see Figure 6.8).

6.4. Related Work

6.4.1. Social Search

While the general concepts of *Social Search* and *Collaborative Information Seeking* have already been discussed in Section 3.1.2, we will now focus on approaches particularly related to Woogle.

We start by discussing features for collaborating on search results. This may involve a (re-)ranking of results within a search result page as well as commenting, adding, or removing results.⁴² Also major players such as Microsoft (“U-Rank”) and Google (“Google Search Wiki”; GSW) have been experimenting with such features (Hearst, 2009, p. 318). Google Search Wiki has even been a beta feature of the Google search engine for some time. In a demonstration prototype, the creators of Wikiseek made the whole result page for a specific query available as an editable Wiki page, while the initial set of results was bootstrapped with Google results. This idea resembles the idea of human maintained “catalogs”, such as the initial Yahoo site, using Wiki principles. While it seems that such features could not take hold in Web-scale search engines, they might be an interesting complementary add-on to Woogle. To a limited extent, the editable text field on “Woogle-Pages” as described in Section 6.2.2 can be used to add search results.

A central argument of NKS, and Woogle in particular, is to treat queries as artifacts of their own right. While *query log analysis* is an established practice in IR (see, e.g., Section 4.2.1), social collaboration related to queries is less common. Early works by Walkerdine and Rodden (2001) and Morris and Horvitz (2007) introduce systems to manage search sessions including queries and result clicks. While both systems allow for individual use, e.g., to resume prior searches, they also allow to share search information with other users. While such techniques have not been adapted in major search engines so far, the genre of collaborative question answering (CQA) offers broad collaboration possibilities about questions. We will discuss CQA in more detail in the following section.

Finally, there exist approaches which visualize query logs using advanced browsing interfaces (Camp and Ulieru, 2007) or topic maps (Wang et al., 2009). Google Suggest, the query autocompletion feature of Google (Cirasella, 2007), is a further approach to expose query log information in the search user interface.

⁴²See, e.g., Agrahri et al. (2008) or Lüer and Cummins (2009)

6.4.2. Collaborative Question Answering

Collaborative question answering systems (CQA) are both knowledge sharing tools in the ancestry of *Answer Garden*⁴³ and *social search* tools to collaborate about information needs (Hearst, 2009, p. 321). Accordingly, CQA systems can perhaps be considered the most comprehensive implementation of NKS in practice, as we have already argued in Section 4.5.3.

In their core, CQA systems offer the possibility to post questions on the one hand and to provide answers to these questions on the other hand (see, e.g., Mamykina et al., 2011). Similar to our “Woogle-Pages” (see Section 6.2.2), a particular Web page and URL is dedicated to each question. This is helpful for search engine optimization and for cross-referencing similar questions. Answers are listed on question pages, similar to results on search engine result pages. However, CQA systems allow extensive means of interaction, including voting, tagging, or commenting for both questions and answers. *Stack Overflow*,⁴⁴ which is a popular CQA community for technology-oriented topics, even contains a “Meta” information space which allow power users to discuss community rules (Mamykina et al., 2011).

Another – perhaps surprising – feature of most CQA system is keyword-based search, which can be used to search for existing questions and answers. Research has however uncovered interesting connections between *searching* and *asking questions*. Liu et al. (2012) carried out a joint analysis of Yahoo search logs and the Yahoo Answers CQA systems. They found that search engines are a main source for traffic on CQA sites.⁴⁵ In particular, they show that existing questions on CQA pages have a large impact on the transition from “searchers” to “askers”, as 50% of all search sessions leading to questions contained at least one Yahoo Answers link in the list of search results. Similarly, Si et al. (2010) report that 25% of Google’s first page of search results in China contain at least one Q&A link. They also found that “Wh..”-questions and “bad quality” search results are major triggers for asking questions.⁴⁶

These observations stress the role of explicit information need representations for knowledge sharing and in particular for turning readers into contributors. Compared to Woogle, CQA systems help answering individual questions instead of addressing more abstract information needs, although some CQA systems allow to comment and vote on others questions. Furthermore, CQA systems are question-focused in the sense, that they treat content (in the sense of “answers”) as a dependent entity. Content from external information sources is not explicitly considered, besides that answers might contain external references.

Research has also addressed the nature and quality of questions asked in CQA

⁴³Ackerman and Malone (1990); see also Section 4.5.3

⁴⁴<https://stackoverflow.com/>

⁴⁵This is supported by Si et al. (2010), and Mamykina et al. (2011). A similar argument is made for the co-existence of search engines and Wikipedia by McMahan et al. (2017).

⁴⁶From a design perspective, Si et al. (2010) is particularly related to Woogle, as their *Confucius* system explicitly combines web search with CQA

6. Woogle: Guiding Contributions to Wikis

systems.⁴⁷ This is interesting from the perspective of analyzing and classifying information needs (e.g., in terms of urgency) and also concerning *search literacy* as a subtopic of information literacy. Finally, some studies such as Guo et al. (2008) or Liu et al. (2010) deal with routing questions to potential answer providers. Such functionality could be a useful addon to Woogle.

6.4.3. Guiding Contributions in Wikis

This section presents approaches that support Wiki editors in their decision on which information to contribute. More general work on guiding contributions has been discussed in Section 4.5.2.

SuggestBot is a Wikipedia Bot (an autonomous program running on Wikipedia) developed by Cosley et al. (2007). Its purpose is to recommend Wikipedia authors articles which need their help. Articles are recommended from the set of articles that is marked with a so-called “maintenance template” by other Wikipedia authors. Based on similar articles formerly edited by a user, these articles are recommended to edit. In this process, similarity is calculated based on a) formerly edited article titles occurring in the article text, b) links from formerly edited articles to articles and c) articles edited by frequent co-authors. A fixed number of articles is finally derived by concatenating articles from the different maintenance templates. Cosley et al. (2007) found that article recommendations based on these heuristics lead to four times more edits when compared to random recommendations. A similar work is *Intopedia* by Romberg (2010), which leverages various indicators for missing content, including maintenance templates. However, the approach provides a keyword based interface to *search* for contribution opportunities related to a certain keyword, instead of proactively contacting contributors.

Several other approaches highlight contribution opportunities within the Wiki user interface. The *Wikitasks* system (Krieger et al., 2009) shows open tasks (as manually defined) that apply to the current and to related Wiki pages. External content that could be helpful to extend a Wiki page is recommended by *VisualWikiCurator* (Kong et al., 2011), which shows content of emails or special external data sources,⁴⁸ and *IntelWiki* (Nawaz Chowdhury and Bunt, 2014), which shows Google search results matching the current page content. Reversely, the *Mail2Wiki* system by Hanrahan et al. (2011) plugs into an email interface in order to suggest sharing content into a Wiki besides or alternative to writing an email.

A more subtle way of guiding contributions is the *Wiki Scaffolding* approach by Díaz and Puente (2011, 2012). It suggests to initially capture an organization’s core information structures such as glossaries or organigrams using mind maps. Based on special tooling these mind maps can be used to bootstrap Wiki content, which may be helpful for initial contributors.

⁴⁷See, e.g., Ahn et al. (2013); Yang et al. (2014); Baltadzhieva and Chrupa (2015)

⁴⁸Supporting a particular RSS-based interface

6.4.4. Wiki Gardening

Wiki gardening, is a knowledge curation activity (see Section 4.5.3), in which a special group of users restructures Wiki content (Happel and Treitz, 2008) or engages in coordination tasks to sustain the community. These users, called *wiki gardeners* or *shapers* (Majchrzak et al., 2007), do not necessarily contribute original content, but organize contribution work and enable other users to contribute.

Such activities have mostly been analyzed in the context of Wikipedia.⁴⁹ A particularly interesting study has been conducted by Arazy et al. (2016) which do attribute certain shaping activities to artifact maturity levels. On a more technical level, Hill and Shaw (2014) analyzed the usage of “redirect”-Pages in Wikipedia. Although the creation of redirects (e.g., to add synonym page titles) is a rather basic curation activity, it is strongly related to Woogle due to the role of disambiguation in (social) search.⁵⁰ In an Enterprise Wiki context, shaping has been addressed by Yates et al. (2010) and Majchrzak et al. (2013), which however focus on the individual motivation of shapers.

6.5. Summary

In this chapter, we introduced Woogle as a concept to link information seeking and information provisioning within Wiki systems. To this end, we proposed to extend search result pages into Wiki pages, thus turning them into a collaborative artifact. Such search result pages and associated Wiki discussion pages therefore effectively act as *mediation spaces*. Furthermore, search queries are captured in a query log and leveraged in order to extend to notion of missing information in a Wiki and to finally guide its evolution.

As a reference implementation, we presented *Woogle4MediaWiki* (W4M) as an extension of the popular MediaWiki system. W4M has a separate namespace for search queries (“Woogle-Pages”) and introduces visual indicators of information need to the so called “red links”, which denote missing information in MediaWiki. W4M can be configured to search additional information sources besides the Wiki, effectively acting as a user interface for enterprise search.

We claim two major benefits of our approach. First, Woogle eases the transition from *information seekers* to *information providers* and vice versa. It allows for immediately capturing information in a Wiki-style within the enterprise search environment where it is typically accessed. Users can also directly influence and comment on top of the automatically created search index. Second, Woogle can ease the problem of bootstrapping enterprise Wikis, that usually suffer from sparse content, which in turn impedes their adoption by members of the organization. By locating the enterprise search within the Wiki, people are “lured” into the Wiki which raises the chance of contributing information. In reverse, Woogle can also bootstrap incomplete enterprise

⁴⁹See, e.g., Kittur et al. (2009); Arazy et al. (2016)

⁵⁰See also Section 6.3.2

6. Woogle: Guiding Contributions to Wikis

search results by serving as a kind of editable metadata layer on top of a full-text index.

Further steps regarding the implementation of Woogle are possible in several directions. As discussed before, visual indicators and their underlying metrics have to be considered preliminary and should be refined by additional empirical studies. In the direction of *social search*, the introduction of social ranking mechanisms (i.e., allowing users to edit, annotate, or re-rank results) could be an interesting feature extension. The identification of querying users, a mechanism for “requesting” content and leveraging the work of Cosley et al. (2007) could be interesting to target information providers. A “Wiki health report” summarizing missing information and possibly further quality issues could be a useful tool to support Wiki evolution.⁵¹

Finally, the application of Woogle concepts might be worthwhile in other information systems or environments. In this direction, Romberg (2010) describes the implementation of a search engine for contribution opportunities in Wikipedia.

⁵¹See also Happel and Treitz (2008)

7.Semantic Need: Guiding Contributions to Semantic Wikis

The *World Wide Web* has turned out to be a very effective infrastructure for information sharing and retrieval. However, the *efficiency* of using the WWW is limited by the fact that most information is available in textual form only, which makes it difficult to process for algorithms. Thus, Berners-Lee et al. (2001) envisioned the *Semantic Web* as an extension of the WWW, that is populated by machine-understandable metadata based on which agents can reason and act to fulfill tasks for human users. The realization of this Semantic Web largely depends on the availability of such structured semantic metadata for which one can distinguish *users* and *providers* as two different roles.

Semantic Web research has addressed both roles to a considerable extent, although they are typically addressed isolated in research and practical applications. The usage of semantic metadata is supported by various tools, ranging from semantic web service frameworks to ontology-based information retrieval systems.¹ The creation and provisioning of semantic metadata has been studied in terms of manual and (semi-)automatic annotation systems (e.g. Handschuh, 2005) and with respect to exposing existing structured content in the Semantic Web (e.g., Bizer and Cyganiak, 2006).

Although the usefulness of metadata has been claimed for many domains and applications, publicly available metadata in the Semantic Web is still relatively scarce. This is also the case for the growing number of Linked Data sets, for which *incompleteness* has been identified as a core challenge.² This is mainly due to the fact that metadata creation is a costly process which requires upfront effort. Surprisingly, only few research has studied topics such as incentives and methods for guiding the creation of semantic metadata so far, although several authors thus call for better means to “support users in the creation of metadata” (Decker, 2002, p. 148) and “to create incentives for annotations” (Handschuh, 2005, p. 198).

While a number of studies have investigated the forces that drive the creation of metadata by individual users (mostly concerning tagging systems, see also Section 7.1.2), it’s largely unexplored why semantic metadata is created and how it is made available (Thomas and Griffin, 1998). In particular, there is neither a proper notion of *metadata need*, nor a related theory of *guidance which metadata should be created*. Also the Semantic Web vision (Berners-Lee et al., 2001) does not address the *creator* side of metadata, but focuses on

¹See, e.g., Domingue et al. (2011)

²See, e.g., Razniewski, Suchanek and Nutt (2016); Weikum et al. (2016); Abiteboul et al. (2017)

7. Semantic Need: Guiding Contributions to Semantic Wikis

the consumer side and its applications. This is quite similar to the domain of information retrieval, which also neglects the role of information *providers*.³

Within this chapter, we will outline an initial theory about why and how metadata is created and thus how the Semantic Web could be populated. We therefore analyze different aspects of metadata – and in particular the relation of information needs and knowledge bases containing semantic metadata. We then propose to guide metadata provisioning by actual *metadata needs* based on the NKS framework as described in Chapter 4. We argue that this can create motivational incentives to help growing the Semantic Web by creating and sharing metadata and present exploratory and formative evaluation results to support these claims.

The chapter starts with describing the problem setting in more detail, based on an analysis of information needs and information provisioning in the Semantic Web. We then describe how NKS can be applied to the Semantic Web scenario. We introduce heuristics to derive “missing information” and discuss architectural and algorithmic considerations. Based on that, we introduce the Semantic Need extension for Semantic MediaWiki (SMW) as a prototypical system, which uses structured queries to guide users in contributing semantic metadata. We then summarize the results of two evaluation studies – an exploratory analysis of information needs on the Semantic Web, and a formative evaluation of Semantic Need. Finally, we discuss related work and several follow-up research questions.

In Happel (2008b) we presented initial ideas motivating this work. However, that paper has a broader scope – addressing the Semantic Web as a whole – and also discusses the relation of privately/publicly stored metadata (inspired by *Inverse Search* as presented in Chapter 5). The interrelation of metadata creation and consumption on the Semantic Web has been extensively discussed in Happel (2011). Happel (2010) describes the design and initial evaluation of the Semantic Need extension and its underlying heuristics.

7.1. Problem

This section will analyze core problems of metadata provisioning and usage on the Semantic Web. Accordingly, it is structured in two parts – lining out how information needs are expressed on the Semantic Web, and then discussing metadata provisioning.

7.1.1. Information Needs in the Semantic Web

We distinguish two major scenarios that motivate the usefulness of metadata in the Semantic Web. The most prominent one is *resource description for information retrieval*. The need for metadata in this scenario stems either from resources that are not accessible by standard keyword-based search (i.e.,

³See also Section 3.4

photos, videos, or services), or from the fact that resources might not explicitly contain certain keywords by which they might be accessed. Metadata is thus added to provide descriptive information which can incorporate structured classifications or keyword synonyms.

The second case for metadata is rooted in *task automation*. This comprises a whole range from visionary, agent-driven scenarios, which automatically perform actions on behalf of their human owners (e.g., Berners-Lee et al., 2001), to so called “mash-ups”, where data from different sources is joined to provide additional value for users (e.g., Ankolekar et al., 2007).

The default mechanism to express and formalize information needs on the Semantic Web is structured queries. In particular, the SPARQL query language (Prud’Hommeaux and Seaborne, 2008; Valle and Ceri, 2011) can be considered the dominant standard when dealing with RDF data.⁴ Despite of this, an empirical analysis of actual information needs on the Semantic Web is nearly impossible, since there is no standardized way to log and preserve SPARQL queries for later analysis. Finally, formulating structured queries on semantic data is difficult and yet uncommon for most end users, which further restricts the possibilities to gather data for analysis.

7.1.2. Information Provisioning in the Semantic Web

In order to describe the technical underpinnings of the Semantic Web, Decker et al. (2000) coined the metaphor of the *information foodchain*. At its starting point, there is the construction of a metadata schema, which serves as a basis for the annotation of semantic metadata in the following step. Based on a storage system with inferencing capabilities, end user applications, such as semantic portals, can then be used to browse and query this semantic metadata.

Notably, this foodchain separates the creation of the underlying schema (in description logics⁵ terminology referred to as *TBox*) from the creation of actual instance data (*ABox*).⁶ The corresponding engineering processes can be described as *knowledge meta process* and *knowledge process* (Staab et al., 2001). For the latter, we distinguish three different ways of creating metadata: 1) either it comes for free and just needs to be exposed, 2) it can be generated automatically, or 3) has to be created manually.

Exposition is probably the most simple case. If data is already available in some highly structured form – such as in database systems – it can be easily exposed. An example for this could be a cinema which offers metadata about films from its existing booking system. Although supporting

⁴RDF (Resource Description Framework) can perhaps be considered the dominant standard for structured knowledge representation underlying the Semantic Web (Pan, 2009; Gandon et al., 2011)

⁵Description logics (DL) denotes a family of knowledge representation formalisms optimized for knowledge representation and efficient reasoning; see e.g., Baader et al. (2003)

⁶Throughout this chapter, we use the terms “ontology” and “schema” synonymously for *TBox* and “annotations”, “semantic metadata” or “instances” for *ABox*.

7. Semantic Need: Guiding Contributions to Semantic Wikis

tools already exist (e.g., Bizer and Cyganiak, 2006), an initial technical investment is necessary to make such data available for external users.

Automatic creation of metadata generates descriptive metadata using, e.g., machine learning for analyzing documents, pictures, or other content, to automatically assign topics or categories. Such techniques depend on the availability of sophisticated algorithms, suitable input and training data, and suffer from potential imprecision (Kustanowitz and Shneiderman, 2005). Furthermore, they are limited to identify metadata which can be directly derived from artifacts' content. Automatic metadata creation techniques are therefore often used in a semi-automatic fashion to assist human metadata creators.

Manually created metadata is probably the most common way of metadata creation.⁷ Classic approaches for semantic annotation take ontologies as a prerequisite and focus on the *knowledge process* to guide users in creating metadata – e.g., by providing templates based on the ontology structure (Handschuh, 2005, p. 63). However, this is *costly*, since it typically requires skilled users, knowledge engineers, or domain experts to maintain annotations. This kind of metadata creation has thus been common for specific tasks and domains, such as library management, and has also gained popularity in recent years due to the emerging *Web 2.0* phenomenon. Applications like del.icio.us⁸ or Flickr⁹ collect small pieces of metadata from individual users, resulting in large sets of aggregated metadata.

Despite of these diverse options, two main issues can be identified which are impeding a widespread success of metadata (Thomas and Griffin, 1998). First, metadata is additional, descriptive data on top of actual information resources which requires additional effort. Second, the creation of metadata often implies a disparity of providers and beneficiaries (i.e., people using metadata are different from people creating it) and between the time of creation and its use (Ames and Naaman, 2007).¹⁰ The information foodchain supports this view by prescribing an unidirectional process, in which both, ontologies and metadata, have to be created prior to their usage.

From an engineering perspective, we derive three lower-level problems which should be addressed by next-generation Semantic Web tools.

Lack of Guidance

Since the unidirectional “information foodchain” approach is limited when considering the evolving and decentralized nature of the Semantic Web, researchers call for better means to “support users in the creation of metadata”

⁷Although the “Linked Data Cloud” (<http://lod-cloud.net/>) probably exposes most semantic data to date, one might argue that, e.g., DBPedia can only expose Wikipedia data so easily, since it has been manually created by Wikipedia editors beforehand.

⁸<http://del.icio.us>

⁹<http://www.flickr.com>

¹⁰This is a classic problem of collaborative software; see, e.g., Grudin (1994)

(Decker, 2002, p. 148). If cost is high, resources are limited and benefits are unclear, effort should be focused and guided towards creating the most *needed metadata*. Thus, metadata creation should conceptualize and address usage patterns and scenarios.

Lack of Incentives

Even if appropriate guidance is available, this needs to be complemented by incentives for creating metadata in order to overcome the asymmetry of metadata creation and usage. Thomas and Griffin (1998) discuss incentives for metadata sharing on a market scope, identifying advertising and retrieval services as potential contributors. In order to also address the scope of individual contributors, metadata provisioning systems should “create incentives for annotations” (Handschuh, 2005, p. 198). Related issues have been discussed concerning tagging and photo sharing systems in recent years and results highlight the important role of personal and social benefits as functional motivations for individual users.¹¹ Considering Semantic Web applications, Siorpaes and Hepp (2008b) propose to use game-based approaches to leverage users’ contributions.

Hidden Metadata

Even if metadata has been created, it needs to be available for potential consumers. Like any kind of digital resource, metadata can be kept in arbitrary spheres of access – ranging from the private sphere of an individual user to public visibility in the Internet. Also, due to the heterogeneous nature of the Semantic Web, relevant metadata might exist in distributed, but unconnected places.

Private spheres are commonly used because users often hesitate to share data openly. Reasons are low motivation due to a lack of personal benefit,¹² privacy concerns (Ardichvili et al., 2003; Desouza, 2003) and effort for sharing.¹³ Thus, even many open “Web 2.0” applications such as Flickr or del.icio.us allow for storing metadata privately (Lam and Churchill, 2007).

Since users will seldom revisit sharing decisions, most information labeled as “private” will always remain invisible for other users. We thus argue that tools should actively support users in sharing useful but yet “private” information with others.

Figure 7.1 illustrates this situation. It distinguishes the amount of metadata available for a certain information resource in the private information space of a particular user vs. the public information space. Four general situations are

¹¹See, e.g., Kustanowitz and Shneiderman (2005); Marlow et al. (2006); Ames and Naaman (2007)

¹²See, e.g., Cabrera and Cabrera (2002); Cress and Hesse (2004); Wasko and Faraj (2005)

¹³E.g., capturing, categorization and setting access rights (Desouza, 2003; Desouza and Evaristo, 2004)

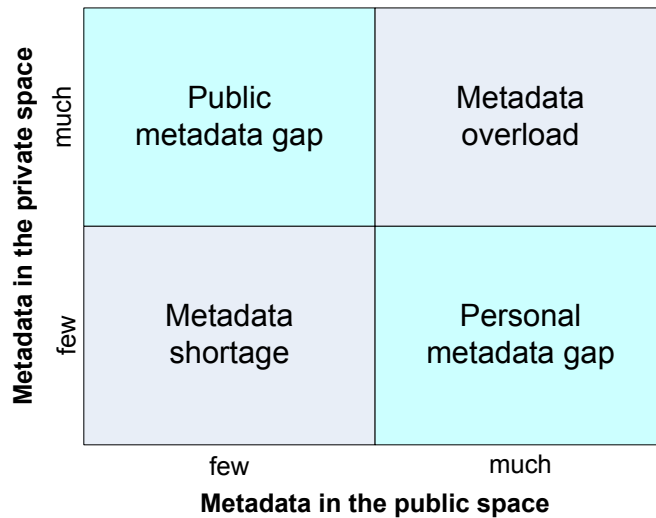


Figure 7.1.: Possible Distributions of Metadata in Private vs. Public Information Spaces (Happel and Stojanovic, 2008, adapted from Figure 4.2)

depicted: in a balanced situation, there either exists few metadata (*Metadata shortage*) or lots of metadata (*Metadata overload*) in both, the private and public information space. If there is more metadata in the public space than in the private space, we call this a *personal metadata gap*. The case of a *public metadata gap* describes that no or only few metadata concerning an information resource exists in the public space, but in the private space of at least one particular user.

When considering the Semantic Web, the situations of a *public metadata gap* and *metadata shortage* are the most problematic ones, since potentially useful metadata is hidden in private spaces or does not exist at all.

7.1.3. Motivating Example

To illustrate the problems discussed in the previous paragraphs, as well as some of our ideas for addressing them, we will now describe a short motivating example.

As-is situation Our scenario involves Chrissy, who wants to buy a birthday present for her boyfriend Dave, who is a movie enthusiast. Chrissy’s initial idea is to book a trip to one of the locations mentioned in Dave’s favorite movie “Casablanca”. Thus, Chrissy queries her favorite Semantic Web search engine for “All locations mentioned in Casablanca”. To her surprise, the application only returns the obvious “Casablanca” as a result – no additional metadata seems to be available in the Web. On the other hand, Dave maintains his own local movie application, in which he keeps data about his favorite films.

His application actually contains “Paris” and “Lisbon” as additional locations mentioned in the movie. However, since this data is within Dave’s private information space, Chrissy cannot retrieve the information. Thus, she finally decides to buy a different birthday present.

Coincidentally, Cosley et al. (2006) motivate their work on *MovieLens* with a closely related anecdote:

Its movie information is incomplete. For most of its life the MovieLens database has been maintained by a single movie guru. When the guru is busy, the database suffers. Sometimes he does not add actors and directors, movies released on DVD are not always updated as “new DVDs,” and so on. About 1/3 of the fields in the database are blank. This has a direct impact on the value of MovieLens, for example, when searches fail to return relevant movies.

To-be situation In order to improve knowledge sharing in the described situation, we propose that Chrissy’s query is not just matched against the available metadata corpus (yielding only one result in our example), but also stored in a central query log. Based on that, Dave’s movie application can retrieve a list of queries and automatically compare it to the metadata in his private space. This would reveal that information from Dave’s private space could help satisfying Chrissy’s information need. The movie application would present a list of metadata items to Dave, indicating that there is an information need that can be satisfied by sharing them. Dave may then choose to contribute this metadata to the public information space. Once Dave shares the information, Chrissy could be notified about the new results for her previous query.

7.2. Design and Implementation

This section presents the *Semantic Need* approach, which is intended to guide and motivate users in contributing metadata to the Semantic Web, respectively specific semantic applications. We also describe the abstract architecture of a technical realization, discuss algorithms and introduce our implementation.

7.2.1. Approach

Before we start explaining Semantic Need, we will dedicate a small section to show how the concept of NKS, as introduced in Chapter 4, can be adapted to Semantic Web settings. We then develop the Semantic Need approach in two parts. First, we analyze which kinds of *information gaps* in semantic knowledge bases exist, and how they could be automatically derived. Afterwards we discuss which kind of *information provisioning* mechanisms can be employed in order to help users filling these gaps.

7. Semantic Need: Guiding Contributions to Semantic Wikis

Paradigm/Issue	Information representation	Information retrieval
Keyword-based information processing	Text in documents	Keyword queries
Logics-based information processing	Logical axioms in a knowledge base	Structured query languages

Table 7.1.: Comparison of Paradigms

Need-driven Knowledge Sharing in the Semantic Web

In Section 4.4, we described the instantiation of NKS in a keyword-based *information retrieval* paradigm. Accordingly, the prototypes described in the two previous chapters were based on *keyword queries* as indicators for information needs, and on textual content and *documents* as information objects that can be shared (see also Table 7.1).

The Semantic Web, in turn, uses logics-based knowledge representation formalisms with well-defined formal semantics (such as description logics; Baader et al. (2003)), to allow users a more precise specification of knowledge. Accordingly, retrieving knowledge from the Semantic Web is based on similarly precise structured query languages. While both, logics-based knowledge representation and retrieval, thus get slightly more complicated than text-based information representation, they allow for more precise processing in turn.¹⁴

Concerning NKS, logics-based information processing however is similar regarding the two separate roles of information providers and seekers. Information needs – as in this case represented by *structured* queries – can generally be used to guide the evolution of semantic knowledge bases. In the following section, we will line out the corresponding technical details.

Information Gaps in Semantic Knowledge Bases

Ultimately, the Semantic Web can be seen as a specialized system for sharing codified information. As when sharing texts and documents, users and providers of information are separated due to the asynchrony of the technology, resulting in reduced motivation and contribution (Happel, 2011). To address this separation, we developed the concept of *Need-driven Knowledge Sharing* (NKS; see also Chapter 4).

It is based on the assumption that information needs re-occur over time and across different information seekers (see also Section 4.2.1), and can thus be used to guide the creation and improvement of information. NKS rejects the understanding of information sharing as a linear process where all information has to be created prior to any request. In turn, it embraces that an information repository is never 100% *complete*, but grows and evolves over time. This perspective acknowledges the real world experience that individual requests

¹⁴See also Table 7.1

might even fail to deliver any appropriate result, if some information is not yet known to the repository (Happel and Mazarakis, 2010).

In a similar fashion, the logical formalisms underlying the Semantic Web share that “information [...] is in general viewed as being incomplete” (Baader and Nutt, 2003, p. 68) and thus make a so-called *Open World Assumption* (OWA).¹⁵ In opposite to “closed world”-systems such as relational databases, facts that cannot be derived are not considered false but (yet) unknown under the OWA. Thus, a semantic knowledge base (*KB*) usually describes only a limited subset of what is considered true in a domain (see Figure 7.2) and might grow over time.

A *KB* can be generally considered as a set of logical statements or axioms.¹⁶ Such axioms might be used to state so-called terminological knowledge, which describes classes and properties of a domain (i.e., “Professor is a subclass of Teacher”) or about named individuals (i.e., “Rudi Studer is a Professor”).¹⁷ If a *KB* cannot answer a request that can – based on full knowledge of the domain – considered to have true results, this can either be due to missing assertions (Baader and Nutt, 2003, p. 68) but also due to an incomplete specification of the terminology.

Although this evolutionary nature of captured knowledge is a fundamental principle underlying the Semantic Web, we are not aware of dedicated methods providing guidance on how a knowledge base should evolve from an information seekers’ perspective – i.e., which axioms should be added to satisfy information needs.¹⁸

We thus propose to use structured queries for this purpose. While there is no universal definition of structured queries, we consider so-called *conjunctive queries*¹⁹ (Hitzler et al., 2009, p. 294), which are composed of conjunctive query atoms. These atoms may contain variables (i.e., “*Professor(x) ∧ worksAt(x, y)*”) which will be assigned concrete instance values from *KB* if suitable results can be derived from the axioms in the *KB*. Formally, a query *q* can be *satisfied* by a knowledge base *KB*, if $\exists \mu : KB \models \mu(q)$. The function μ maps every variable of the query to the name of an individual, ensuring that only known individuals are returned by a query (Hitzler et al., 2009, p. 295).

We choose *QBox* as the set of all structured queries that are formulated against a knowledge base *KB* by its users over time. Due to the inherent incompleteness of *KB*, we expect that there is a set of *unsatisfied* queries *UQ* ($UQ \subseteq QBox$) for which holds: $\neg \exists \mu : \forall q \in UQ : KB \models \mu(q)$. *UQ'* is the subset of *UQ*, for which true results can be assumed, based on full knowl-

¹⁵A related aspect of information needs was discussed in Section 3.1.1

¹⁶In RDF these axioms are called triples (Prud’Hommeaux and Seaborne, 2008)

¹⁷The terminological and assertional part of a *KB* are usually referred to as *TBox*, respectively *ABox* (Baader and Nutt, 2003, p. 46).

¹⁸See Section 7.4.9 for some related work based on data quality measures; in particular Luczak-Rösch (2014). See also Paulheim (2017) for a more general overview on approaches for knowledge base evolution.

¹⁹In particular, we only consider the case of “DL-safe” conjunctive queries (Motik et al., 2004) in this chapter – i.e. we do not allow for non-distinguished variables in query atoms.

7. Semantic Need: Guiding Contributions to Semantic Wikis

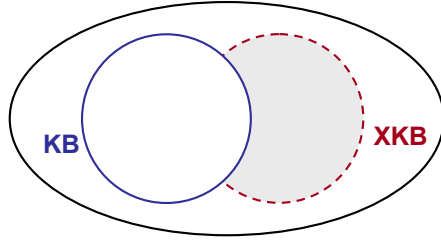


Figure 7.2.: KB denotes the set of all axioms in the knowledge base. XKB denotes the set of all axioms which have to be added to the KB to satisfy all structured queries, that should be satisfiable according to full knowledge of the domain (See also Figure 4.1)

edge of the domain.²⁰ We thus choose a set of logical axioms XKB such that $\exists \mu : \forall q \in UQ' : XKB \cup KB \models \mu(q)$. We assume that KB and $KB \cup XKB$ are *consistent* knowledge bases. Note that XKB thus loosely corresponds to the set of axioms filling the “semantic gap between supply and demand in the Semantic Web” as described by Mika et al. (2009).

Finally, we choose a set of *partially unsatisfied* queries PUQ' ($PUQ' \subseteq UQ'$) by requiring that $\exists \mu : \forall q \in PUQ' : \exists atom \in q : KB \models \mu(atom)$. We consider PUQ' a particularly relevant subset of the $QBox$, since in opposite to queries in $UQ \setminus PUQ'$, queries in PUQ' (“PUQs”) have at least one query atom that can be satisfied from KB . Using PUQs, axioms in XKB , contributing yet missing knowledge, can thus be related to existing KB axioms. We argue that those “PUQs” are a particularly interesting in order to guide contributions to a KB . Accordingly, they will play a central role throughout the next sections.

Information Need Heuristics

In this section, we will line out how the concept of PUQ' can be leveraged for analyzing information needs expressed in Semantic Web environments.

Semantic Web knowledge bases can typically be assumed to have a graph-based data structure. RDF query languages such as SPARQL (Prud'Hommeaux and Seaborne, 2008) mostly consist of two major parts:

- A list of triple patterns (triples containing variables) which should be matched against the knowledge base to constrain the result set (corresponding to a *WHERE* statement in SQL²¹)
- A list of variables for which values should be contained in the result set (corresponding to a *SELECT* statement in SQL)

The *result set* of such a semantic query is thus basically a set of n -tuples, providing bindings for the n variables selected in the query. Result tuples may

²⁰For instance, a query for “All volcanoes in Karlsruhe” would *not* be contained in UQ' since there cannot be any true result (at least if we consider the real world as our domain).

²¹Structured Query Language

either contain knowledge base instances or literals.²²

We thus derive the following two cases to identify gaps stemming from information needs that can not be satisfied by a knowledge base:

Incomplete Result Set denoting the case when expected results do not appear in the result set due to *query constraints*.

Sparse Result Set denoting the case when certain expected values for *query variables* do not appear.²³

By *expected* results and values we denote such instances and literal values which are part of XKB – i.e., which exist based on full knowledge of the domain, but which are not yet formally specified in the KB .

Incomplete Result Set An incomplete result set denotes the situation that an *expected* result is not returned by a structured query. This can either be caused by an incomplete knowledge base or by semantic query constraints, which do not match for a certain result (*WHERE*-part of the query).

In the first case, a result instance might not yet be formally specified in the KB – e.g., a query for all instances of the class *employee* would not yield such employees that are not yet known to the system. Second, instance annotations might be incomplete – e.g., a query for employees with a salary >40.000 would not yield *Employee* instances that lack any information about their salary.

Clearly, it is not obvious to decide if a given result set is incomplete (i.e., if further results are to be *expected*). One option could be to leverage ontological background knowledge such as cardinality statements on properties. However, such statements are not possible in all knowledge representation paradigms. Thus, another option is to heuristically infer missing results. In the following we present one particular heuristic for this purpose.

Near matches As stated before, structured queries often contain multiple conditions to select particular subsets of an ontology class. The previously mentioned query for employees with a certain salary is an example for this. We define *near matches* as instances in the knowledge base which are potentially relevant results for a given query, but which do not appear in its result set due to missing semantic metadata.

To identify such cases, we only consider queries with at least two conditions. Technically, a candidate “near match” has to match at least one condition of a query and must not match at least one other condition, for which it lacks any annotation. This is to avoid considering instances which are properly described (e.g., an employee with a salary of 30 000, who does not match the query by purpose).

²²In the case of SPARQL, also Boolean values or so-called “blank nodes” may be included, which we omit here.

²³Note: in SPARQL, a *sparse result set* is only possible if using the *OPTIONAL* modifier in queries.

7. Semantic Need: Guiding Contributions to Semantic Wikis

Near matches can thus help to indicate missing annotations which “prevent” instances from appearing as a query result. The underlying assumption is, that these instances potentially could match the information need if meta-data would have been properly annotated. Accordingly, we consider them *near matches* and assume that this might offer valuable insights on required metadata for people contributing to a knowledge base.

Sparse Result Set We define *sparse result set* as a case, when variable bindings in a result set remain empty. As stated before, the *SELECT*-part of semantic queries indicates a set of n -tuples presenting variable bindings in the result set. A cell in a column of the result set may remain empty, if there is no appropriate binding for that variable (see Figure 7.3).²⁴ We consider a result set as a *sparse result set*, if at least one of its cells remains empty (i.e., no binding exists for the corresponding variable).

Country	Area	Population	Capital	Currency
Burundi	28,000 km ²	8,700,000	Bujumbura	
Central African Republic		4,400,400		Central African CFA franc
Mauritius	2,040 km ²		Port Louis	Mauritian rupee
Nigeria			Abuja	
Seychelles	455 km ²	87,476		Seychellois rupee
Somalia	637,661 km ²	9,133,000		Somali shilling
South Africa	1,221,037 km ²		Pretoria Bloemfontein Cape Town	Rand
Zimbabwe		12,521,000	Harare	US dollar

Figure 7.3.: Example of a Sparse Result Set

Missing Result Values Such empty cells can be considered an unsatisfied information need, since the semantic query requests a variable binding which can not be satisfied from the knowledge base. Thus, we define *missing result values* as a heuristic to infer missing annotations. Contributors to the knowledge base could be interested in these cases in order to help delivering complete information for queries.

Information Provisioning in Semantic Knowledge Bases

Semantic gaps in the knowledge base, as indicated in the previous section can be addressed at both the schema and the instance level. At the schema

²⁴Note that, e.g., in SPARQL, the default behavior will not show the entire result if at least one variable can not be bound. This default behavior can be changed using the *OPTIONAL* modifier (Prud’Hommeaux and Seaborne, 2008). However, in this case, we would end up with an *incomplete result set* as discussed before.

level, terminological knowledge such as mappings between classes, instances or properties might be defined to resolve incompleteness. If suitable annotations can not be derived by contributing such mappings, annotations (assertional knowledge) have to be explicitly contributed.

Schema-level knowledge and annotations may stem from two primary sources: either it already exists formally in private or distributed information spaces (see Section 7.1.2) and just needs to be *shared*, or it still needs to be *captured* in a formal way.

Capturing is necessary, if knowledge is not yet formalized and thus needs to be added to the knowledge base. This can involve both, schema-level knowledge or data/annotations.

Concerning annotations, “near matches” and “missing result values” can help identifying concrete properties, which are not yet annotated for a knowledge base instance. Thus, users can be provided with an interface denoting all missing properties for a given instance as derived by these heuristics.

Similarly, one can check if “near matches” or “missing result values” are caused by missing schema mappings. This denotes the case when query atoms do not correspond to existing terminological knowledge. This can either imply that parts of the terminological knowledge are missing, or it can be an indicator of synonyms – e.g., if a user queries for `[[Category:Worker]]` instead of `[[Category:Employee]]`. Thus, the system might assist users in finding candidate mappings to improve the terminological knowledge and thus help satisfying information needs.

Sharing knowledge can be done, if information *is* already formally captured, but not available at query time, since it is hidden in a yet unknown or not accessible knowledge base. Information needs might thus be satisfied by either sharing (i.e., copying) semantic information into the queried knowledge base, or by introducing suitable mappings, which allow the query engine to retrieve semantic information from other knowledge bases.

7.2.2. Architecture

In this section, we will line out the architectural components for our approach introduced in the previous section.

The dark blue boxes at the left hand side of Figure 7.4 depict the standard components of a knowledge base. Concerning the backend, this is the persistent knowledge base as such, which can be accessed using a *Query API* and/or a generic *KB API* which allows write access to the knowledge base. While writing into the knowledge base is typically done indirectly – e.g., by third-party tools – there might also exist a corresponding KB manipulation UI.

In order to enable our knowledge capture approach, we introduce a *query log* storage, which is wired to the query API. Besides that, there is a so-called *Need*

7. Semantic Need: Guiding Contributions to Semantic Wikis

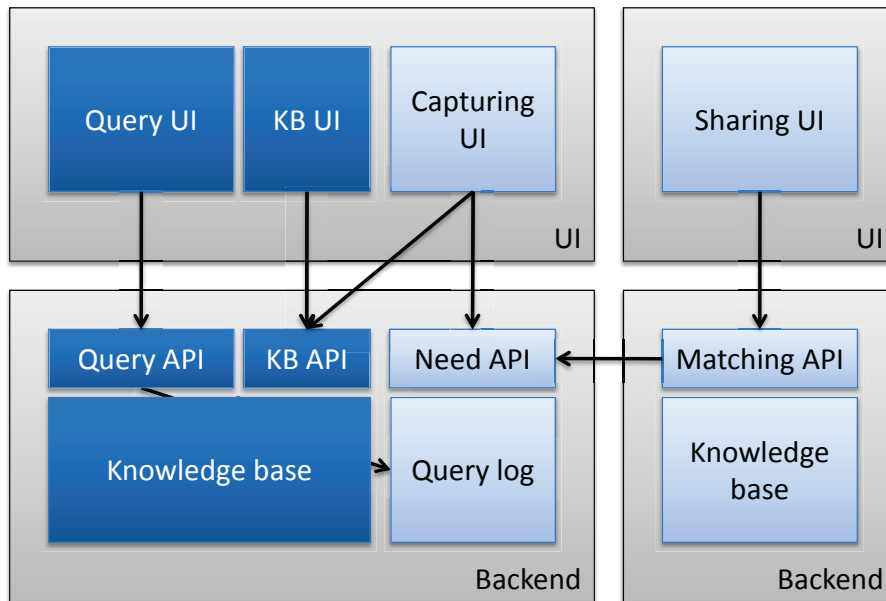


Figure 7.4.: Abstract Architecture of Semantic Need

API which offers metadata need information to outside consumers, which is derived from the query log. Examples for such need information would be “near matches” and “missing result values” as explained previously. Further implementation details will be presented in Section 7.2.3.

Due to exposing query log information, its content is subject to privacy considerations. In contrast to the typical practice of saving one query log entry per request, we maintain an *aggregate query log*. This means that each semantic query is represented by one log entry which also includes how often that query has been executed and by how many different users. Instead of exact execution timestamps, we only store the first, the last, and an average timestamp. Finally, for efficiency reasons, the query log duplicates information about the state of the knowledge base, such as the number of result rows, the number of “near matches” and “missing result values”.

One consumer of such need information would be a *Capturing UI*. This is intended to be a special user interface which, e.g., transforms information needs into a form-based UI, which allows knowledge engineers, domain experts, or end users to contribute potentially missing facts.

Another consumer of need information could be a *Sharing UI*. In contrast to the Capturing UI, it does not allow users to formalize new knowledge, but instead recommends them to share existing knowledge. Such knowledge may, e.g., stem from a different knowledge base, which is only known or accessible to certain users. Technically, this can be realized either by physical or logical separation. Physical separation means, that the private information space is an independent system running, e.g., on the local machine of a user (such

as a Semantic Desktop system²⁵). Logical separation does not require two separate applications, but can be implemented as a feature in a server-based system – e.g. by offering “private” and “public” sharing options. Based on this additional knowledge base – where the knowledge to be shared is formalized – a *Matching API* can select knowledge worth sharing based on information needs derived via a *Need API* and present these suggestions in the *Sharing UI*.

From an *information seeking perspective*,²⁶ the architecture could be complemented by a notification component which helps to alert interested users if new metadata is captured or shared. However, since we focus on *information provisioning* in this chapter, we will leave this topic to future work.

7.2.3. Design

In this section, we describe the services and algorithms proposed by our approach in more detail. As depicted in Figure 7.4, this comprises the *Matching API* and the *Need API* for capturing metadata not yet formalized, which will be in the focus of this section.

We will distinguish two types of services: “global” services, which provide knowledge base-wide need information, and “local” services, which expect an instance identifier (URI) as a parameter and provide need information focused on this particular knowledge base instance. In the following, we shortly describe these services and line out selected implementation details.

Global Services The following services offer information about the state of the knowledge base as such. They can be helpful for both *Capturing UI* and *Sharing UI* implementations. While the first three services address potential missing metadata, the last service is intended to convey an idea about the most popular queries. Depending on the query as such, this could, e.g., help to identify metadata that should be kept up-to-date with priority (i.e., if a query for *phone numbers* is executed often) or to provide additional metadata (e.g., if users regularly seek for images annotated with certain concepts).

getQueriesWithNoResults() – lists all semantic queries that do not yield any result

getQueriesWithMostNearMatches() – lists all semantic queries ranked by the number of *near matches*

getQueriesWithMostMissingResultValues() – lists all semantic queries ranked by the number of *missing result values*

getPopularQueries() – lists semantic queries ranked by popularity

Local Services The “local” services roughly correspond to the “global” ones but provide information focused on a certain knowledge base instance. We

²⁵See also Section 7.4.6

²⁶See also Section 3.1.1

7. Semantic Need: Guiding Contributions to Semantic Wikis

expect these services to be more useful from a knowledge acquisition perspective, since they allow for a very focused application of need information in user interfaces, while general overviews on missing metadata might be less motivating for contributors.

getMissingResultValues(URI) – returns a list of missing metadata for the instance which causes empty result values for some queries (see Algorithm 2)

getNearMatches(URI) – returns a list of missing metadata for the instance which causes this instance to be a *near match* for some queries (see Algorithm 3)

getQueriesMatchingInstance(URI) – returns a list of queries which actually have the instance as a result

While the first two services again primarily address missing information, the last service is intended to inform users about metadata which is actually leveraged by existing queries.

As indicated by Algorithm 2, `getMissingResultValues()` and `getNearMatches()` include a final step to semantically aggregate and rank the result set based on the `aggRank()` algorithm. The idea of this step is first, to semantically aggregate missing metadata (i.e., need on a super-property increases need on its sub-properties) and second, to rank needs based on the amount of demand.

Need aggregation and ranking Need aggregation targets the ranking of queries in terms of identifying those queries with the largest information need. We therefore apply two processing steps to the data in the query log. First, identical queries are aggregated on a per-user basis to calculate a *personal information need*. Second, the different personal information needs concerning a particular query are aggregated into an *aggregate information need*.

7.2.4. Implementation

In this section we present a prototypical realization of the architecture and algorithms described in the previous sections. Therefore we implemented a Semantic Need extension to Semantic MediaWiki (SMW).²⁷ In the following, we thus give a brief introduction to SMW. We then apply the NKS concept to SMW, describing *incomplete* and *sparse result sets* as two heuristics for identifying “PUQs”. We also describe the user interface of the Semantic Need extension.

²⁷<http://www.semantic-mediawiki.org>

Semantic MediaWiki

Semantic MediaWiki is an extension to the widespread MediaWiki engine.²⁸ It allows users to semantically annotate content on Wiki pages, making it easier to export or query data in a structured way.

Wiki pages in the “Category” and “Property” namespace are considered class and property definitions (i.e., terminological knowledge). Within these pages, subclass and subproperty axioms can be expressed by creating a hierarchical structure. Finally, annotations in all other namespaces are interpreted as instance assertions. This allows to define pages to be the *instance of a class* or the *subject of an RDF triple* either relating to other instances or capturing literal value attributes (Bao and Li Ding, 2008). Technically, semantic annotations are parsed after saving pages and stored in a persistent knowledge base, which is by default a set of special tables in the MediaWiki database.²⁹

The major means for satisfying information needs is the SMW query language (“SMW-QL”; see e.g., Bao and Li Ding, 2008). The typical way of using SMW-QL is to embed queries into Wiki pages – so called “inline queries”. Besides that, there is a page `Special:Ask`, which allows to express free form SMW-QL queries. Similar to other structured query languages, SMW-QL consists of two major parts:

- A list of conditions (basically categories and property values, but also named instances) which should be matched against the knowledge base to constrain the result set.
- A list of printout statements from which values should be contained in the result set.

Furthermore, several icons within the Wiki will implicitly create a query when clicked, in order to allow “browsing” the knowledge base. With the help of additional extensions, SMW can also provide an endpoint for submitting SPARQL queries.³⁰

Based on its TBox and ABox content, plain Semantic MediaWiki provides users with a number of cues to analyze their wiki according to what we illustrated in Figure 4.1. First, there are two so called “Special Pages”, `Special:Wantedcategories` and `Special:WantedProperties`, which list schema elements that have been used for stating instance data, but which have not been explicitly specified yet (corresponding to area $A \cup B$ in Figure 4.1). Furthermore, `Special:Unusedcategories` and `Special:UnusedProperties` show schema elements for which no instance data exists yet (Area $G \cup D$).

²⁸<http://www.mediawiki.org>; the concept of MediaWiki extensions is explained in more detail in Section 6.2.4

²⁹Storage alternatives are listed at http://www.semantic-mediawiki.org/wiki/SPARQL_and_RDF_stores_for_SMW

³⁰See http://www.semantic-mediawiki.org/wiki/Help:SPARQL_endpoint

Semantic Need for Semantic MediaWiki

We will now present Semantic Need as an extension for SMW, which realizes some of the features described in the previous paragraphs. In particular, our implementation addresses the *capturing* of semantic *annotations*. While the *sharing* of semantic information and the provisioning of *schema-level knowledge* are foreseen in the system design, these aspects will be left to future work.

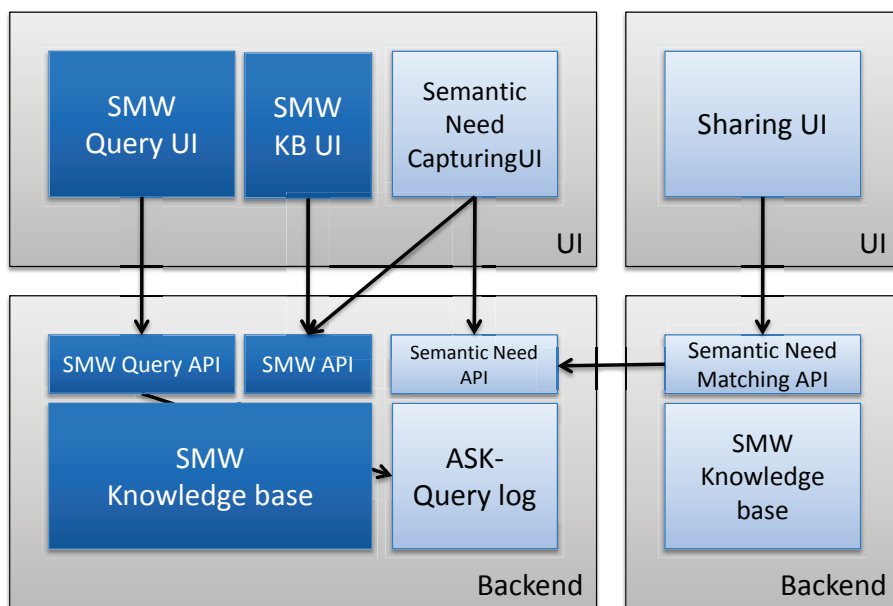
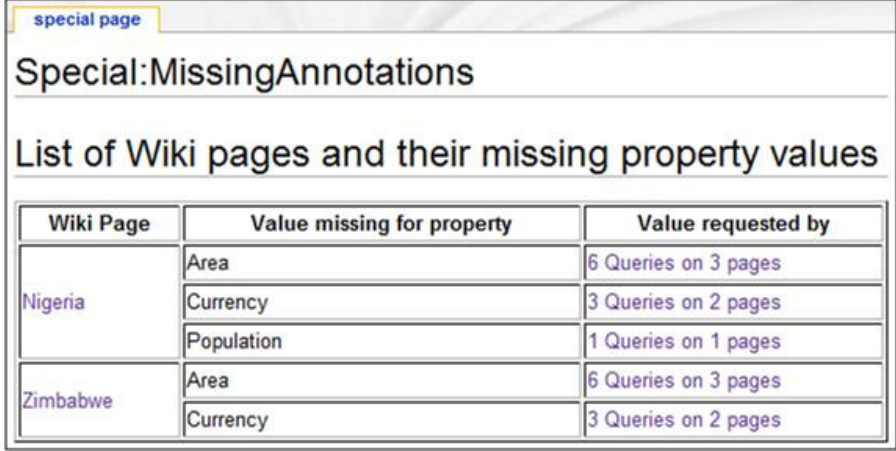


Figure 7.5.: Architecture of Semantic Need for MediaWiki (instantiation of Figure 7.4)

SMW with its annotation and querying functionality provides all core components of a typical semantic knowledge base. The dark blue boxes at the left hand side of Figure 7.5 depict these features, including a query and knowledge base interaction UI and a corresponding backend (see also Krötzsch et al., 2007).

We introduce a *query log* storage, which is wired to the query API. In our implementation, we focus on so-called “inline queries”, which are embedded in Wiki pages. We consider them the most relevant, since many end users might not be able or willing to formulate ad hoc structured queries on their own. Basic information about inline queries is stored in a “semantic query log”, which includes the conditions and printout statements of the query (see Happel, 2008b). Based on the query log, the *Need API* offers *metadata need* information such as “near matches” and “missing result values”.

One consumer of such need information is the *Capturing UI*, a special user interface which allows *knowledge engineers*, *domain experts*, or *end users* to contribute potentially missing facts to the knowledge base. We realized two different types of implementation so far. First, we provide “global” overview



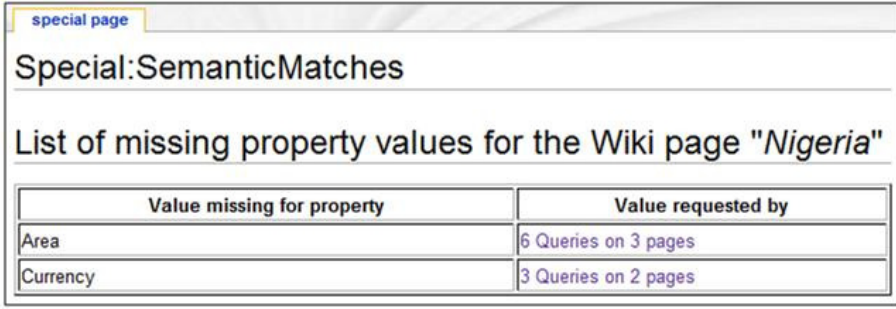
The screenshot shows a web interface for 'Special:MissingAnnotations'. It features a title bar 'special page' and a main heading 'Special:MissingAnnotations'. Below this is a subtitle 'List of Wiki pages and their missing property values'. The main content is a table with three columns: 'Wiki Page', 'Value missing for property', and 'Value requested by'. The table lists missing annotations for 'Nigeria' and 'Zimbabwe'.

Wiki Page	Value missing for property	Value requested by
Nigeria	Area	6 Queries on 3 pages
	Currency	3 Queries on 2 pages
	Population	1 Queries on 1 pages
Zimbabwe	Area	6 Queries on 3 pages
	Currency	3 Queries on 2 pages

Figure 7.6.: Wiki-wide Overview of Pages and their Missing Annotations

pages for the semantic query log, which list all semantic queries – in particular those without results – and a Wiki-wide overview of pages and their missing annotations (see Figure 7.6).

Second, the same feature is applied to individual pages, resulting in an overview of missing annotations for a specific Wiki page (see Figure 7.7. This can be considered a semantic counterpart to the MediaWiki page `Special:WhatLinksHere`, which helps users to find out how a Wiki page is *syntactically* interlinked with other Wiki pages.



The screenshot shows a web interface for 'Special:SemanticMatches'. It features a title bar 'special page' and a main heading 'Special:SemanticMatches'. Below this is a subtitle 'List of missing property values for the Wiki page "Nigeria"'. The main content is a table with two columns: 'Value missing for property' and 'Value requested by'. The table lists missing annotations for 'Nigeria'.

Value missing for property	Value requested by
Area	6 Queries on 3 pages
Currency	3 Queries on 2 pages

Figure 7.7.: Missing Annotations for a Specific Wiki Page (i.e., specialization of Figure 7.6)

Although usability issues are not in the core focus of our work, we also thought about how to address actual end users who might contribute to the Wiki more directly (see Figure 7.8). Besides this, several other ways to inform users about contribution possibilities can be imagined – including integration in Java-Script based annotation UIs,³¹ game-based interfaces (e.g., Siorpaes and Hepp, 2008b), or identifying and approaching potential contributors directly (e.g., by email). We consider these usability aspects an important issue for

³¹Such as described by Pfisterer et al. (2008)

7. Semantic Need: Guiding Contributions to Semantic Wikis

further research.



Figure 7.8.: In-page Display of Missing Annotations

Another consumer of need information could be a *Sharing UI*. In contrast to the Capturing UI, it does not allow users to formalize new knowledge, but instead recommends them to share existing knowledge. Such knowledge may, e.g., stem from another SMW instance, which is only known or accessible to certain users.

7.3. Evaluation

So far, we argued that semantic knowledge base tend to be incomplete and developed *Semantic Need* as an approach to help information providers to contribute metadata annotations that satisfy information needs of information seekers. This results in two claims which we want to evaluate in the following:

1. There exist *semantic gaps* between information needs and existing metadata in semantic knowledge bases.
2. *Semantic Need* with its underlying heuristics can be helpful to fill these gaps.

As for the first issue, we present an exploratory empirical study on SMW instances running in the public Internet. We extracted persistent semantic queries from eight of these Wikis for analysis. In a second study, we asked 30 experienced SMW administrators to provide feedback on our general concept and on *Semantic Need* in particular.

7.3.1. Public Semantic MediaWiki Analysis

To make our case for the existence of *semantic gaps*, we wanted to investigate the cases of “missing result values” and “near matches” (see Section 7.2.1)

in more detail. Since semantic query data is not widely available for research purposes, we decided to rely on data extracted from SMW installations running in the Web. Since SMW is a popular MediaWiki extension, there exists a large number of publicly accessible installations which we could use for this purpose. In the following, we will first describe the design of our study and afterwards discuss its results.

Design

Since we are not aware of any previous studies of semantic query data (especially concerning SMW queries),³² our study will primarily have an exploratory and descriptive focus. We follow the basic research interest how many *missing result values*, respectively *near matches*, exist for real world structured queries. In terms of information need indicators, we will rely on the analysis of inline queries, since these are the only information needs in SMW which currently have a persistent representation which is accessible for analysis.

To select public SMW instances, we derived an initial list by consulting overview pages and search engines. By dismissing Wikis that did not contain much SMW-data (less than three queries and 250 annotations), we cut down our list from around 200 Wikis to 100. We then ruled out Wikis which were not accessible via a public API, or which could not be accessed due to connection problems during the test runs of our evaluation tooling. Due to the massive amount of data, we decided to carry out deeper investigations on eight randomly selected Wikis described in Table 7.2.

Sitename	Pages	ANN^{33}	PG_{ANN}^{34}	IQ^{35}	IQ_{EC}	IQ_{ECPO}	IQ_{ECCJ}
CS Wiki (CS)	195	1 591	67	5	5	5	4
Eroge Wiki (ER)	340	1 853	182	3	1	0	0
HAR2009 (HA)	2 892	3 468	940	38	0	0	0
Historiographus (HI)	998	2 724	390	19	14	10	8
Mount Wiki (MN)	2 662	1 422	833	199	0	0	0
Protege Wiki (PR)	1 545	253	367	11	10	6	4
Sharing Buttons (SH)	122	590	18	7	0	0	0
territoile (TR)	1 801	3 135	502	3	1	1	1
Σ	10 564	15 036	3 299	285	31	22	17

Table 7.2.: Overview of Surveyed SMW Installations

³²Mika et al. (2009) have argued for semantic gaps based on keyword queries from Yahoo search logs

³³Overall number of semantic annotations

³⁴Number of pages containing at least one semantic annotation

³⁵Number of inline queries. Further columns indicate subsets of IQ constrained by evaluation conditions as described in the text.

Process

In order to retrieve data for our analysis, we wrote a crawler³⁶ which accesses the MediaWiki API.³⁷ It extracts all semantic annotations and structured queries from the pages and stores them into a database.

After retrieving the data, we applied further processing in order to restrict the number of queries for analysis. First, we chose an evaluation condition (“EC”) which selects queries that a) are “ask”-queries (ruling out “show” queries) and that b) have either “table” (=default) or “broadtable” as output format (ruling out, e.g., RSS exports of query results). The number of queries satisfying the evaluation condition is shown in Table 7.2 as IQ_{EC} . For compatibility with the goals of our analysis, we applied a final selection step. For the analysis of *missing result values*, we selected those queries that actually contain printout statements (IQ_{ECPO}). Accordingly, we selected only conjunctive queries for the analysis of *near matches* (IQ_{ECCJ}).

Overall, this processing resulted in 22 queries satisfying IQ_{ECPO} and 17 queries satisfying IQ_{ECCJ} . Due to overlaps of both sets (see Table 7.3 and 7.4) this results in 25 distinct queries.³⁸ As a first step of analysis, we derived the number of results for each query. Since many queries were located on template pages,³⁹ the corresponding fields in Table 7.3 and 7.4 denote “n.a.”, as in their case, the number of results depends on the page embedding the template. Instead, we computed the number of instances for the `[[Category:]]` part of the query ($Results_{CAT}$) as an approximation.⁴⁰

Results

Missing result values Table 7.3 summarizes the analysis of the IQ_{ECPO} query set. We computed the number of missing result values (e.g., empty cells in a result table as depicted in Figure 7.3) in the result set. For queries on normal pages, this is the actual number of empty cells visible. For queries on template pages, we counted missing result values for each instance contained in the “virtual” set of all instances ($Results_{CAT}$). A single missing annotation of a value on a Wiki page would thus only count once, even if the missing result value would appear on multiple Wiki pages embedding the query.

As it can be seen from the results, all queries on normal pages (i.e., not template pages) provide a complete result set without any empty cells. However, for queries on template pages, up to 63% of cells in the query result set were empty. To estimate if these empty cells were really due to missing information

³⁶ Available at <http://www.teamweaver.org/wiki/index.php/MediaWikiTools>

³⁷ <http://www.mediawiki.org/wiki/API>

³⁸ See <http://www.teamweaver.org/downloads/data/sneed/sneed-smw-queries.pdf> or Section C.1

³⁹ A MediaWiki template page can be transparently included in other pages in order to reuse content across several pages in a Wiki

⁴⁰ This approximation assumes, that each instance of the `[[Category:]]` part of the query occurs at least once in a query result among all the Wiki pages embedding the query.

(and not consciously omitted), we manually investigated three empty cells for each of five randomly selected queries. It turned out, that only two of those 15 empty cells checked could *not* be considered missing information. Accordingly, those randomly selected queries lack result values to a considerable extent. In average, 16% of cells remained empty across all queries surveyed.

ID	Results	<i>Results_{CAT}</i> ⁴¹	Empty cells	Printout requests	% Empty cells
CS1	n.a.	8	19	4	59%
CS2	n.a.	7	0	3	0%
CS3	n.a.	1	0	1	0%
CS4	n.a.	16	0	2	0%
CS5	7	7	0	4	0%
HI1	1	18	0	3	0%
HI2	28	65	0	2	0%
HI4	n.a.	18	27	3	50%
HI5	n.a.	65	22	2	17%
HI7	n.a.	24	60	4	63%
HI8	n.a.	4	1	3	8%
HI9	n.a.	35	6	4	4%
HI10	n.a.	15	3	4	5%
HI11	n.a.	14	2	4	4%
HI12	n.a.	15	9	4	15%
PR1	72	80	0	1	0%
PR2	n.a.	80	13	1	16%
PR3	n.a.	91	1	1	1%
PR4	n.a.	91	57	1	63%
PR5	n.a.	91	75	2	41%
PR6	n.a.	91	1	1	1%
TR1	70	102	0	1	0%
		Σ 938	Σ 296	Ø2,5	Ø16%

Table 7.3.: Missing Result Values for the IQ_{ECPO} Queries

Near matches For the conjunctive queries, we first observed that all 17 queries under consideration consisted of exactly two conjunctions. In most cases, this is a category statement combined with a restriction on one property (e.g. PR2: `[[Category:Plugin]] [[For Application::PAGENAME]]`). In order to derive near matches, we computed the number of instances which completely lack the annotation of the restricted property. The rationale behind this is, that these instances might qualify to appear in the query result set, once a correct value for the property is annotated.

As shown in the last column of Table 7.4, up to 94% of instances lacked the annotation on the selection property in extreme cases. Again, we performed a deeper analysis of three *near matches* for five randomly selected queries. Out of these 15 cases, five turned out be “false positives” – i.e., were lacking annotations by purpose – and ten did correctly indicate a missing result in a therefore incomplete result set. While near matches might thus not be a strict indicator for “missing” annotations, they are nevertheless a strong hint. On

⁴¹Number of instances for the `[[Category:]]` part of the query

7. Semantic Need: Guiding Contributions to Semantic Wikis

average, across all surveyed queries, up to 22% of results might not show up in result sets due to missing annotations.

ID	Results	$Results_{CAT}$ ⁴¹	Missing selection property	% Missing selection property
CS1	n.a.	8	6	75%
CS2	n.a.	7	0	0%
CS3	n.a.	1	0	0%
CS4	n.a.	16	4	25%
HI1	1	18	17	94%
HI2	28	65	10	15%
HI3	1	3	1	33%
HI4	n.a.	18	17	94%
HI5	n.a.	65	10	15%
HI6	n.a.	3	0	0%
HI7	n.a.	24	13	54%
HI8	n.a.	4	2	50%
PR1	72	80	8	10%
PR2	n.a.	80	9	11%
PR3	n.a.	91	0	0%
PR4	n.a.	91	18	20%
TR1	70	102	32	31%
		Σ 676	Σ 147	\emptyset 22%

Table 7.4.: Near Matches for the IQ_{ECCJ} Queries

Although our analysis is based on a rather small set of queries, this selection can already help to identify up to 296 missing printout statements and up to 147 missing selection properties within the surveyed Wikis. Given the fact that we only analyzed around 9% of the overall inline queries (due to our evaluation conditions), this stresses the potential for using “missing result values” and “near matches” as heuristics for guiding semantic annotations.

7.3.2. Semantic Need Survey

The prototype implementation of the MediaWiki extension for Semantic Need was evaluated by a survey among experienced SMW administrators.

Design and Process

The main goal of the survey was to gather feedback on our concept and its realization. We thus decided to include a small example scenario with screenshots of SMW and our extension. Since this requires a) prior knowledge of SMW, b) a holistic view of an existing SMW installation and its usage, and c) results in a rather large questionnaire, our main target group consists of experienced SMW administrators, rather than end users.

The questionnaire⁴² consists of five major components. Two parts address the problems of a *sparse* and *incomplete result set*, asking respondents about the frequency and severity of these issues. Another part deals with semantic

⁴²See <http://www.teamweaver.org/downloads/data/sneed/sneed-survey.pdf> or Section D.1

annotation practices. People are asked how they find out missing annotations in a standard SMW. Afterwards, screenshots of Semantic Need are shown (including Figure 7.6 and 7.8) and people are asked if they agree that Semantic Need might be effective to a) generally help maintaining annotations, b) focus annotation effort, and c) motivate users to provide contributions.

Two other parts of the survey address the usage context of SMW. We asked about the knowledge domain captured in the Wiki, the structure and content of the knowledge base, about the users, and about the nature of the information needs which users typically seek to satisfy in the Wiki.

The final questionnaire has 34 questions and was pre-tested by five persons resulting in some minor modifications and clarifications. To gather participants for the survey, we followed two strategies. Since we were interested in frequent SMW users, we advertised our survey on the official SMW user and developer mailinglists. Furthermore, we directly contacted 15 persons which are known to drive own SMW projects. The survey was launched at the end of June 2010 and remained open for interested participants for two weeks.

Results

We received 30 complete answers. A majority of 15 answers came from Germany and seven from the US. The remaining participants were scattered across eight different (mostly European) countries. Concerning their experience with SMW, 15 respondents describe themselves as “intermediate”, eleven as “expert” and four as “novice”. On average, they are using SMW for 2.3 years.

The knowledge domain captured in SMW is characterized as “fixed/standardized” in eight cases, as a “generally open domain without many predetermined entities and properties” in six cases and as a mix of both options in 15 cases. The semantic data model is largely prescribed by Semantic Forms⁴³/Templates in 19 cases. Only seven SMWs have an equal level of prescribed and ad hoc structure and another four rely mostly on free-form annotations. None of the Wikis surveyed do *not* use Semantic Forms/Templates at all. Twelve people answered that no particular methodologies, practices, or tools are used to maintain the semantic data, while five people claim to follow simple informal practices, and seven people implement changes based on more advanced measures such as scripts, documentation, and team decisions. In seven cases, the data stored in SMW is driven by the structure of external data and systems.

The problem of *sparse result set* was observed “often” or “sometimes” by 18 people, while twelve indicated “rarely” or “never”. 15 people rate the issue as “not problematic” while twelve answered “somehow problematic”. No one rated query result sparseness as “very problematic”. In their free text justification, people made the point that the question if query result sparseness is an actual problem depends on the application context (four answers) and the nature of the data itself (five).

⁴³Semantic Forms, also known as “Page Forms” (https://www.mediawiki.org/wiki/Extension:Page_Forms), is a MediaWiki extension which makes it more user-friendly to enter semantic annotations

7. Semantic Need: Guiding Contributions to Semantic Wikis

For *incomplete result sets*, 19 people answered to have observed the issue “often” or “sometimes” while nine observed it “rarely” or “never”. Furthermore, only five people consider the issue “not problematic”, while 18 answered “somehow problematic” and five even “very problematic”. This is stressed by the free text justifications in which 16 respondents repeated that query result incompleteness is a problematic issue. Key aspects are the “invisibility” of the issue – which makes it a larger issue than sparse query results – and which increases if the dataset grows large: *“due to the nature of our wiki (IT company) it is hard to know when a query is incomplete. For example, there are hundreds of pages on servers so impossible to know when one or several are missing.”*

We also asked how people would deal with finding out missing annotations for a particular Wiki page and clustered the free-form answers in four main categories. Six answers suggest to make a comparison with annotations on similar Wiki pages. Related to that, seven people would check the schema (i.e., properties) and forms related to that page. Another four people would do an analysis of the page text to identify additional content that could be formalized. Finally, ten answers suggested to create specific ask-queries for this purpose. It turns out that *decisions* are a core part of this process – as one answer puts it: *“Write down a list of all the quantifiable data on the page. - Then decide if any of these are excessive in depth for most users. - In this case I would add part of Africa, size, population, and currency.”*

The global overview about Wiki pages and their missing annotations is generally appreciated in the survey. On a 5-point scale ranging from “strongly disagree” to “strongly agree”, most respondents agree that this feature can be effective to maintain semantic annotations in SMW (8/18/2/2/0⁴⁴). The agreement is slightly less strong about if it can help to guide annotation efforts towards most crucial information needs (5/18/6/1/0), and about if it can motivate users to provide missing annotations (9/13/5/2/1). The page-specific features of Semantic Need are even more appreciated. 15 respondents strongly agree that it can be effective to maintain semantic annotations in SMW (15/11/2/2/0). Concerning annotation guidance and user motivation, 26 respondents at least chose “agree” in both cases (12/14/3/1/0). Finally, 20 participants (66%) are interested in using the Semantic Need extension on their own Wiki.

Summarizing, we can observe that SMW usage differs largely among participants: ranging from prescribed data structures to more open, Semantic Web-inspired scenarios. While the first group argues that data quality and completeness is crucial in their case and thus considers missing annotations a serious problem, others stress the evolving nature of Semantic Web applications: *“I don’t see this is a ‘problem’ - it’s the way things are, always in flux, always perfecting, always coming to stasis. Law of Thermodynamics.”* Semantic Need however, was considered helpful by both groups – either to help raising data quality or to provide guidance in less predefined settings.

⁴⁴Amount of answers stating: strongly agree/agree/neutral/disagree/strongly disagree

7.4. Related Work

Due to the permanent growth of research in the area of data processing, there exists a reasonable body of related work. As usual, we roughly distinguish approaches which are primarily targeting *information seekers* and *information providers* or which try to *bridge or combine* both of these perspectives.

Concerning *information seeking*, we discuss the topics of *Semantic Query Log Analysis*, *Query Sharing and Reuse*, and *Knowledge Extraction from Queries*.

Related to *information provisioning*, we cover *Guiding and Motivating Semantic Content Creation*, *Collaborative Knowledge Creation*, and *Sharing Semantic Content*.

Finally, we describe *Incomplete Information*, *Why-not Provenance*, *Valuation of Data*, and *Crowdsourced Information Provisioning* which take a combined or holistic perspective on both information seeking and provisioning.

7.4.1. Semantic Query Log Analysis

Query log analysis is a well-established method to study users' information needs in the area of information retrieval (Jansen, 2006). Accordingly, also researchers from the Semantic Web community have pursued similar studies.

Due to the fact that issuing structured queries is far less common among end users, several researchers took existing keyword queries. The probably earliest study by Mika et al. (2009) used keyword queries from the Yahoo search engine in order to find out how much semantic data was returned for such queries. Resubmitting 7 081 queries with at least one result to the Yahoo BOSS search engine (which also returns metadata), 59% of the results actually contained metadata. Accordingly, structured data was missing for 41% of information needs.

Halpin (2009) did a similar study using query logs from Microsoft Live.com.⁴⁵ He extracted 509 659 queries denoting entities (here: people or places) and 6 698 queries related to abstract concepts (such as “weather”) using WordNet. Both sets were massively reduced by excluding queries that did not occur at least ten times in the log. The remaining queries were re-issued to the *FALCON-S* Semantic Web Search engine (Cheng et al., 2008). For all concept-related queries, at least 10 results (URIs) were returned. For the entity-related questions, 12% did return less than ten results. 30% out of these did not return any result, 12% one, and 10% two results. Thus, even though infrequent, *long tail* queries were excluded from the analysis, a significant number of entity-related queries returned only few results from the Semantic Web.

An early study of structured (SPARQL) queries was presented by Möller et al. (2010), who obtained access logs of four popular LOD datasets, including DBpedia. Analyzing request mechanisms, they concluded, that a majority of queries were submitted by means such as bots, crawlers, or other tools.

⁴⁵Meanwhile known as *Bing*

7. Semantic Need: Guiding Contributions to Semantic Wikis

In terms of query structure, out of roughly twelve million queries from the DBpedia set, 42% contained more than one triple pattern. This is noteworthy since our approach, as discussed in Section 7.2.1, requires at least two triple patterns as part of a query.

7.4.2. Query Sharing and Reuse

Similar to considerations about *social search* (see Section 3.1.2), researchers in data processing have argued that structured queries are valuable artifacts which deserve consideration especially for reuse and collaboration (Khossainova et al., 2009).

In terms of reuse, approaches for query autocompletion have been developed for both database (Khossainova et al., 2010) and Semantic Web technologies.⁴⁶ Related to that, a number of approaches allow for recommendation and interaction with prior queries of other users.⁴⁷ Additional application scenarios of query logs are discussed by Sellam and Kersten (2017).

Further approaches allow for explicit collaboration and sharing of structured queries. Wahl et al. (2017) present a collaborative approach for knowledge sharing and data source integration for data scientists. Demartini, Trushkowsky, Kraska and Franklin (2013) use crowdsourcing to understand the structure of keyword queries for refining lists of structured query templates. Similarly, Collis and Frommholz (2017) describe a crowdsourced approach for labeling, saving and voting on linked data query templates. Finally, Varga et al. (2017) developed an ontology to describe multidimensional queries in an OLAP setting. Patrick (2003) even proposes to store questions in a dedicated part of a knowledge-base denoted “QBox”.

Notably, the presented approaches do not address the underlying data sets, but address knowledge sharing at the level of how to best explore those datasets. Some of the approaches thus open a *mediation space* concerning the discussion and potential prioritization of information needs. Additionally, refined formal models of information needs could be helpful for future extensions of *Semantic Need*.

7.4.3. Knowledge Extraction from Queries

Besides treating structured queries as a valuable source of information about the exploration of data, some works treat them as a source of primary knowledge (see also Section 4.5.1 discussing “Queries as Knowledge”). Most of these approaches are concerned with information extraction from keyword-based queries (Paşca et al., 2007), respectively the interrelation of keyword-based queries and structured data, such is the case of Google Knowledge Graph.⁴⁸

⁴⁶See, e.g., Kramer et al. (2013); Campinas (2014); Rafeş et al. (2017)

⁴⁷See, e.g., Sellam and Kersten (2013); Eirinaki et al. (2014); Sellam and Kersten (2016)

⁴⁸See, e.g., Gupta et al. (2014); see (Uyar and Aliyu, 2015) relating search engine knowledge graphs more generally

One can distinguish approaches targeting the extraction of terminological (TBox) and assertional knowledge (ABox). As for terminological knowledge, there exist works targeting the extraction of unknown entity attributes⁴⁹ or for deriving entity types (Zhang et al., 2015).

Concerning assertional knowledge, there exist approaches for entity extraction⁵⁰ or for class attribute values (Pasca, 2014). The EU-Funded OPTIQUE project⁵¹ investigates methods for query-driven ontology extensions in ontology-based data access scenarios.⁵² Li et al. (2017) present a novel holistic approach which combines text and query logs for the verification of extracted information.

While not directly related to the current realization of *Semantic Need*, knowledge extraction from queries points out two promising kinds of opportunities. First, it looks worthwhile to consider the extraction of knowledge from *structured* queries. This is particularly interesting in the context of Semantic MediaWiki, as it already maintains a notion of desired categories and attributes, based on the concept of “red links”. Second, techniques concerning entity attributes could be used to improve the derivation of Semantic Need – e.g., to prioritize certain missing values.

7.4.4. Guiding and Motivating Semantic Content Creation

This section describes work which is geared towards guiding and motivating users to create semantic content.

A number of approaches is related to so called “mixed initiative information extraction” (Hoffmann et al., 2009), which gathers user feedback in order to confirm algorithmically extracted knowledge. Singh et al. (2016) describe how entities extracted from text are presented to users by means of short text snippets with some surrounding context. Kondreddi et al. (2014) describe a game-based approach in which users need to confirm relations between entities. There exist a number of further approaches using gamification techniques⁵³ as this is identified a promising means of motivating users (Cuel et al., 2011).

Another set of works is addressing users annotating structured data using forms. Chen et al. (2010) line out a dynamic approach to form generation, which takes into account data quality aspects, e.g., by re-asking for certain information. Eberius et al. (2013) target the alignment of open data schemas while users are uploading that data to a portal.

Most related to *Semantic Need* is the so-called “Collaborative Adaptive Data Sharing platform (CADS)” described by Ruiz et al. (2014). It describes an attribute/value-based annotation system for text documents. Values can be

⁴⁹See Pasca (2012); Gupta et al. (2014); Halevy et al. (2016)

⁵⁰See, e.g., Jain and Pennacchiotti (2011); Alasiry et al. (2014); Zhai et al. (2016)

⁵¹<http://www.optique-project.eu>

⁵²See Grau et al. (2013)

⁵³See, e.g., Siorpaes and Hepp (2008c); Thaler, Simperl and Siorpaes (2011); Thaler, Siorpaes, Mear, Simperl and Goodman (2011) and Siorpaes and Hepp (2008a) for a general discussion

7. Semantic Need: Guiding Contributions to Semantic Wikis

entered using a form, which is composed based on a “querying value (QV)” and a “content value (CV)” of each attribute. While CV denotes the relevance of an attribute based on keywords contained in the text, QV denotes the relevance of an attribute based on prior queries submitted to the system. Both values are combined and ranked in order to decide which attributes to show in an input form, which basically relates to our notion of an *information gap*.

7.4.5. Collaborative Knowledge Creation

This section describes tools and communities dealing with the collaborative creation of knowledge. On the tool side, this particularly includes *Semantic MediaWiki*, which is also the basis for *Semantic Need*. Community-wise, we primarily focus on studies about Wikipedia-related knowledge creation communities such as DBpedia and Wikidata (Vrandečić and Krötzsch, 2014).

Semantic MediaWiki communities have been studied in a series of studies by Gil et al.⁵⁴ The core focus of these studies was the schema of the Semantic MediaWiki instances, namely categories, concepts, and properties. Property assertions are only considered in Gil et al. (2013). Accordingly, there is no deeper analysis on the quality of semantic annotations. Also, semantic queries are not addressed. In a related study, Walk and Strohmaier (2014) studied the evolution and dynamics of 79 Semantic MediaWiki instances. They found that the first 6-12 months are crucial to determine if the Wiki later follows a path of positive or negative growth. They furthermore found that a smaller group of focused contributors seems more beneficial to a positive long-term evolution than a large set of contributors.

There are also a number of attempts towards better tool support for semantic annotations. Pfisterer et al. (2008) describe an improved annotation user interface including autocompletion, a toolbar, and inline annotations. A similar design for inline annotations, but informed by prior information extraction is presented by Hoffmann et al. (2009). Filipiak and (2014) propose a bootstrapping approach based on the import of existing domain ontologies. Finally, Vrandečić (2009) discusses the usage of integrity constraints in order to prevent or detect data quality issues.

Further work has been addressing structured data from Wikipedia, mostly in the form of DBpedia and Wikidata systems. The completeness of so-called article “infoboxes” was analyzed by Lewoniewski (2017) across seven different language versions of Wikipedia. A detailed study of data quality metrics comparing five different public knowledge graphs was conducted by Färber et al. (2017). Both studies found significant evidence of data quality issues, in particular missing information.

Besides such descriptive studies, a number of tools have been developed to support data quality assurance. *COOL-WD*⁵⁵ is a Wikidata extension which allows to assert completeness on attributes in Wikidata. Färber and Hepp

⁵⁴See Gil et al. (2013); Gil and Ratnakar (2013); Gil et al. (2015)

⁵⁵See Darari et al. (2016); Prasojo et al. (2016); Darari et al. (2017)

(2010)⁵⁶ and Kontokostas et al. (2014) present approaches using pre-defined SPARQL queries to identify common data quality issues in linked data. Among these are also queries to detect incomplete or missing property values. However, all those checks need to be defined and executed by a data curator in contrast to *Semantic Need*, which derives missing values based on user queries.

7.4.6. Sharing Semantic Content

Another facet of provisioning structured knowledge is *sharing* knowledge from a private information space into a shared information space.

The notion of private knowledge spaces is a primary concern of the vision of the *Semantic Desktop*, which enables users to structure their personal data such as files, emails, or address books. Interestingly, although denoted “social”, the actual exchange of structured data is not in the core focus of Semantic Desktop systems (Sauermann et al., 2005; Groza et al., 2007). Even if explicitly considered, data flow between private and public spaces is typically unidirectional – “enriching” information on the local Desktop with public data (Drăgan et al., 2011). Notable exceptions include Drăgan et al. (2010) and David et al. (2010), describing the sharing of personal notes respectively lecture notes as linked data. However, both approaches do not address the process of sharing as such (as discussed in Section 3.2.1).

There also exist some work on access control (Ioannou et al., 2007) and information disclosure in data exchange (Miklau and Suciu, 2004; Benedikt et al., 2016), which is however rather technical and not focused on assisting information providers.

Besides Semantic Desktop applications, also cloud-based platforms often distinguish private and shared information spaces. Popular examples are photo sharing platforms, such as Flickr, or bibliography tools, such as Bibsonomy (Benz et al., 2010). Similar to Semantic Desktop tools however, we are not aware of any such system that allows to systematically revisit sharing decisions about content. Initial ideas in this direction have been discussed in prior work of ours (Happel, 2008b). In terms of *Semantic Need*, sharing semantic content could be relevant across federated Semantic MediaWiki instances.

7.4.7. Information Completeness

A major logical assumption about query processing in relational databases is the so-called “closed world assumption” (CWA; see also Section 7.2.1). Under CWA, information in a database is supposed to describe a *complete* set of facts modeling the intended part of reality, which yields the conclusion that any fact not explicitly stated is considered *false*. The opposite is typically the case for *knowledge bases*, which mostly follow an “open world assumption” (OWA), which considers its set of facts about the world it models as incomplete and hence evaluates questions for unknown information as *unknown*.

⁵⁶See also Fürber (2015)

7. Semantic Need: Guiding Contributions to Semantic Wikis

Information completeness is typically considered a part of overall *data quality*, which also deals with other issues such as accuracy, currency, or consistency (Wang and Strong, 1996). Procedures of ensuring data quality are called “data cleaning” (Rahm and Do, 2000). Data quality is crucial for the effectiveness of IT systems and accordingly, poor data quality can have negative impact on business performance.⁵⁷ Such impact is often attributed to other quality aspects than completeness, which is neglected by many data cleaning approaches (Fan, 2015) – perhaps also due to the CWA.

Recent trends in data processing, such as information extraction, information integration, or big data have however increased interest in this topic (Fan, 2015). Given that most available data cleaning tools offer poor support for detecting incomplete data,⁵⁸ this is considered a major open issue for data quality research and practice (Chu et al., 2016; Abedjan et al., 2016).

General concepts used in the area of *information completeness* are tightly related to concepts used for *Semantic Need*. First, information completeness is mostly defined in contrast to a comparison database which is considered to be complete, similar to Figure 7.2.⁵⁹ Furthermore, authors typically distinguish complete tuples (or rows) missing in query results and missing property values⁶⁰ – similar our notion of *missing result values*.

Nevertheless, information completeness has been in discussion since the early days of relational databases. In terms of attribute values, e.g., Codd (1986) proposed alternate means to model “NULL” values, which he semantically distinguished into “missing and applicable” (i.e., can be entered later) and “missing and inapplicable” (i.e., there exists no meaningful value). Instead of a “NULL” value for empty predicates, he thus suggested so-called “A-marks”⁶¹ and “I-marks”.⁶² This suggestion however was never adopted by the SQL standard.

Later approaches to address information completeness have suggested to explicitly assert the completeness of queries (views; Motro, 1989) or database table contents (asserting so-called “local-completeness”; Levy, 1996). Razniewski and Nutt (2011) combine and extend the approaches of Motro (1989) and Levy (1996).⁶³ Other work by Deng et al. (2016) derives completeness from existing master data tables which are assumed to be complete. Finally, Darari et al. (2013) and Galárraga, Hose and Razniewski (2017) proposed approaches for modeling information completeness in Semantic Web standards such as RDF and SPARQL.

While information completeness approaches discussed so far require *explicit assertions* concerning the information completeness of certain sets of data,

⁵⁷See, e.g., Redman (1998); Haug et al. (2011)

⁵⁸See, e.g., Müller and Freytag (2003); Barateiro and Galhardas (2005)

⁵⁹See, e.g. also Levy (1996) or Nutt et al. (2012)

⁶⁰See, e.g., Grahne (2009); Nutt et al. (2012)

⁶¹short for “missing-but-applicable-value mark”

⁶²or “inapplicable-value mark”

⁶³Further, more formal notions for modeling completeness information on a logical level have been proposed in the form of so-called “c-tables” (Imieliński and Lipski, 1984) and recently “m-tables” (Sundarmurthy et al., 2017)

Galárraga, Razniewski, Amarilli and Suchanek (2017) recently presented a rule mining approach for inferring completeness and incompleteness of parts of a knowledge base.⁶⁴ To learn completeness or incompleteness, certain oracles (heuristics) are discussed, such as the popularity or change rate of an entity over time.⁶⁵ While that work is still preliminary and focuses largely on *completeness* rules, the authors also provide one particular example for an *incompleteness* rule derived: “a person who has a date of death, but no place of death, is incomplete for the place of death.” Certainly, rules of this kind could be a beneficial extension of *Semantic Need*’s derivation of missing facts.

In general, work on information completeness is highly related, but mostly complementary to *Semantic Need*. A common ground is the overall setting, addressing the problem of missing tuples and empty result values. The goal of *information completeness* research however, is to provide users with insights and guarantees about the completeness of query results (based on explicit completeness statements), whereas *Semantic Need* aims to derive insights about *incompleteness* (based on user queries). Notably, both notions are not strictly inverse, as missing completeness information would not necessarily imply incompleteness. Reversely however, completeness information could be beneficial to *Semantic Need*, as it can help ruling out potentially near matches or missing values, for which is known, that they are missing by purpose. Particular tool support in that direction, even in the context of the Wikidata project (which is a sister project of Semantic MediaWiki) is described by the COOL-WD approach, described in the forthcoming section about Collaborative Knowledge Creation.

There also exists some initial work though, which focuses on the *incompleteness* of query results rather than their completeness. Razniewski, Savkovic and Nutt (2016) discuss a notion of “completeness-as-default” (CAD) in opposite to the assumption of “incompleteness-of-default” (IAD), as proposed by *information completeness* approaches. They also present logical means that allow to explicitly assert a potential *incompleteness* of parts of a database. Related to this, Nikolaou and Koubarakis (2016) propose an extension to RDF (called *RDFi*) containing so-called “e-literals” that “can be used to represent values of properties that exist but are unknown or partially known”. Such incompleteness statements could complement calculations of missing result values and empty cells in *Semantic Need*. Furthermore, there exist approaches dealing with explanations about why certain tuples do not appear as results for a certain queries. This problem, denoted *why-not provenance*, is discussed in the following.

⁶⁴There exists also recent work by Pellissier Tanon et al. (2017) on using *completeness* assertions for the evaluation and ranking of mining completeness rules. Other related work by Mirza et al. (2017) aims to derive completeness-related cardinalities by information extraction from text. A general discussion of algorithmic means to increase completeness of knowledge graphs (such as e.g., type inference; Paulheim and Bizer, 2013) is presented in Paulheim (2017).

⁶⁵I.e., high popularity (measured in the relative number of facts and Wikipedia article metrics) and a slow change rate (no new relations added since prior version of KB) are considered signals for *completeness*

7.4.8. Why-not Provenance

Data provenance is a topic closely related to information completeness. In the context of queries, it is typically defined as the means to explain “the derivation of a piece of data that is in a query result” (Cheney et al., 2009). Such information is typically sought by querying users, which want to understand *why* certain results are part of a query.

In the past years, researchers have also started to consider the opposite question of why certain data is *not* part of a particular query result. The first works in this direction were presented by Huang et al. (2008) and Chapman and Jagadish (2009), who coined the term *why-not* provenance. Particular approaches allow users to ask a system why certain results are not shown in order to “debug” query evaluation and either recommend query modifications to incorporate a certain result or data modifications in the underlying database (Gao et al., 2015).

We are especially interested in the latter case, as it is closely linked to the idea of incomplete information in a database or knowledge base. For this, the work of Huang et al. (2008) is particularly interesting. They describe the scenario of a database built from information extraction which is hence considered partially incomplete. For a given query, they denote all tuples “non-answers” which are not returned as a result. Those non-answers are further distinguished in “potential answers”, that might become part of the query result when inserting tuples or attribute values, and “never-answers” for which this is not the case.⁶⁶ Note that the notion of “potential answers” is closely related to our concept of a *near match*, introduced in Section 7.2.1.

Work presented by ten Cate et al. (2015) suggest to leverage domain ontologies in order to provide insights to users why certain results were not returned for a query. So instead of suggesting particular data modifications (or query modifications), an ontology-based, conceptual explanation is provided. Drosou and Pitoura (2013) describe a way to return additional results that are *similar* to actual results of a query. Although not limited to missing values, the approach could be interesting in terms of addressing the missing results problem in *Semantic Need*.⁶⁷

Finally, there has recently been some work on why-not provenance in the area of Semantic Web technologies. While Yao et al. (2015) and Vasilyeva et al. (2016)⁶⁸ primarily target to derive (SPARQL) query refinements, Bienvenu et al. (2016) target knowledge base modifications in ontology-based data access (OBDA) scenarios. In particular, they assume relaxed and potentially incorrect query answering semantics over incomplete data to provide “possible” or “almost sure” results.⁶⁹ In particular, their system automatically derives

⁶⁶The distinction is drawn based on integrity constraints and a notion of *trusted* tables that is similar to information completeness statements discussed before

⁶⁷Slightly related work in a Semantic Web context has been presented by Hurtado et al. (2006) and Troumpoukis et al. (2017)

⁶⁸More detailed in Vasilyeva (2017)

⁶⁹This is slightly related to the work of Drosou and Pitoura (2013)

so-called “repair plans” to add or delete certain ABox statements in order to “fix” the knowledge base.

7.4.9. Valuation of Data

While the notion of *data quality* takes a data-centric perspective, the related concept of *data governance* adds a business perspective and “aims at maximizing the value of data assets in enterprises” (Otto, 2011). Due to the fact that a “perfect” level of data quality is not possible or economically infeasible, various approaches have been conceived to complement data quality with notions of business value.

In particular, Even et al. (2007) and Even and Shankaranarayanan (2007) describe means for combining the cost respectively the utility (e.g., based on customer order value) of ensuring data quality in a business context. Even et al. (2010) describe a case study about maintaining different levels of information quality for different people in an alumni CRM system, based on their donation behavior.

Issues concerning the value of maintaining semantic open data are theoretically discussed by Brennan (2017). Alborzi et al. (2015) introduce a “Data Readiness Level (DRL)” as a quantitative measure of the value of a piece of structured data. Closely related to *Semantic Need* is the work by Luczak-Rösch (2014). While generally taking a broad perspective on data quality in DBpedia, particular experiments were carried out analyzing “failed queries” (defined as yielding no results) derived from DBpedia SPARQL query logs. While that analysis stresses the case for Semantic Need, it does not explicitly target individual information providers, but rather conceptualizes data quality in a larger scope of an ontology engineering lifecycle. Also, the particular notion of failed queries is much more restrictive than notions of near matches and missing result values in Semantic Need.

Summarizing, especially business-driven means of data value could be interesting complements to our derivation of semantic need. However, in a knowledge management context, query information probably is still a viable approximation of actual business needs. Nevertheless, one could, e.g., consider to differentiate queries by their particular relevance or business value. Related to this, the database community has recently discussed a number of approaches towards query-based pricing for data marketplaces⁷⁰ which might be worthwhile to explore in this direction.

7.4.10. Crowdsourced Information Provisioning

An approach which tightly combines information seeking and provisioning is crowdsourced query processing. Initially conceived in the database community, early approaches basically extend SQL syntax (Franklin et al., 2011) or make use of so-called “user-defined functions” in SQL (Marcus et al., 2011) in order

⁷⁰See, e.g., Koutris et al. (2013, 2015); Deep and Koutris (2017)

7. Semantic Need: Guiding Contributions to Semantic Wikis

to allow information seekers to explicitly specify to *crowdsource* parts of a SQL query. A crowdsourcing engine implements the distribution of crowdsourcing tasks and combines results into a common SQL query response.

There also exist multiple advanced approaches to crowdsourcing which explicitly target missing or incomplete information. Fan et al. (2015) analyze the WHERE clause of SQL queries and crowdsources predicate values which are NULL for a given tuple. Park and Widom (2014) use a slightly different approach for user input, showing a complete table with missing values to crowdworkers. Other approaches like Bergman et al. (2015) focus on crowdsourcing the removal of incorrect respectively the contribution of missing tuples (i.e., rows). Selke et al. (2012) finally even support the usage of attributes, that are not yet contained in the database schema, within queries.

Similar strategies have also been adopted in a Semantic Web context. The *HARE* system by Acosta et al. (2017) uses a RDF-based completeness model in order to determine missing values in SPARQL queries, which are subsequently crowdsourced. Other works have been addressing data integration (Sarasua et al., 2012; Demartini, Difallah and Cudré-Mauroux, 2013) and query answering (Lin et al., 2010; Simperl et al., 2012) on the Semantic Web. A broader discussion of crowdsourcing opportunities is provided by Sarasua et al. (2015).

Overall, crowdsourced information provisioning approaches can be considered means for demand-driven information provisioning. This involves a tight coupling of the query and the information provisioning process – at query-time, and for each individual query. *Semantic Need* in contrast aggregates demand across potentially multiple queries, and does not enforce the completion of missing information, but raises awareness by potential information providers. Also, missing information is openly visible while browsing and editing the semantic database, allowing for a broader timeframe and audience for information provisioning.

7.5. Summary

In its core, the Semantic Web is about the creation, collection and interlinking of metadata on which agents can perform tasks for human users. While many tools and approaches support either the creation or usage of semantic metadata, there is neither a proper notion of *metadata need*, nor a related theory of guidance which metadata should be created.

This chapter has described three major contributions. First, we have argued for considering information needs – in particular structured queries – as drivers for the process of creating semantic metadata. To this end we introduced the *Semantic Need* approach, which guides contributors to create metadata which has the most value for other users in the Semantic Web. Second, we described an extension for SMW, which guides contributors to create metadata which has most value for other users, as a proof-of-concept realization of this approach.

Third, we conducted two empirical studies to validate our approach. An exploratory analysis of public SMW installations shows, that the current ap-

plication areas of Semantic Need – *missing result values* and *near matches* – occur in the surveyed dataset to a considerable extent and are thus of practical relevance. This is also stressed by the result of an expert survey among 30 experienced SMW administrators. Their feedback provides initial evidence that Semantic Need can be an effective tool to support the guided growth of semantic knowledge bases.

On a more general level, Semantic Need shows, that NKS is not only beneficial for improving knowledge sharing in text-based information processing, but can also be applied in structured data environments.

While this stresses the general feasibility of our ideas, we think that a realization within a larger Semantic Web scope is possible as well (see also Happel, 2008b). Besides, there are many additional directions for future work.

First of all, a live evaluation of Semantic Need could be helpful for gaining insights about user acceptance and potential improvements. For this purpose, the existing instrumentation framework built in the context of Woogle (see Section 6.2.4) could be leveraged in order to conduct online field experiments.

In a technical direction, the calculation of actual *near matches* and *missing result values* could be refined by means such as schema-level constraints (see, e.g., Vrandečić, 2009) or completeness rules (Galárraga, Razniewski, Amarilli and Suchanek, 2017). While Semantic Need currently focuses on instance-level (ABox) knowledge, structured queries in Semantic MediaWiki might also be used in order to evolve terminological knowledge (i.e., SMW categories or properties) if missing.⁷¹

Finally, Semantic Need might be extended – either on a SMW-level or generally – to incorporate further knowledge bases as additional information spaces. Challenges in that direction could be modeling and exchanging information needs *as knowledge* on the Semantic Web (i.e., not as mere SPARQL queries) and adapting concept such as *Inverse Search* (see Chapter 5) for *sharing semantic content* across different information spaces (see also Section 7.4.6).

⁷¹See also Selke et al. (2012); Grau et al. (2013)

8. Summary

We conclude this thesis by summarizing contributions and providing an outlook on future work.

8.1. Contributions

We have provided an extensive analysis of sharing explicit, codified knowledge in organizations. We first lined out, that existing *codification* strategies in knowledge management are focused on the centralized, “push”-style dissemination of knowledge in an organization. They fall short in supporting decentralized knowledge sharing, which is a major problem for the success of knowledge management systems (KMS).

KMS, such as file shares, Intranets, or Wikis, often fail due to a lack of contributions, because information providers have limited resources and limited awareness about the needs of information seekers. We carved out, that this *decoupling* of information seekers and information providers can be considered a root cause for some central questions of knowledge management, particularly *which knowledge is worth sharing* (RQ 1) and *how to foster sharing of such knowledge* (RQ 2).

To this end, we described a novel approach called *need-driven knowledge sharing* (NKS), which consists of three elements. The first part deals with indicators of information need, which are aggregated in order to derive continuous forecasts about *organizational information needs* (OIN). By comparing with private and shared information spaces, an *organizational information gap* (OIG) is derived. This meta information can be made transparent using so called *mediation services* and *mediation spaces* in order to create awareness for organizational information needs, and to guide the creation of *knowledge that is worth sharing* (RQ 1) The realization of these elements is illustrated by three tools, which are all based on established knowledge management systems, and which provide examples for *how to foster sharing of such knowledge* (RQ 2).

Inverse Search is a tool which helps information providers to identify documents in their private information space, which may help closing organizational information gaps. This is, to our knowledge, the first approach which can derive such recommendations automatically based on existing information collections.

Content in Wikis – such as Wikipedia or enterprise Wikis – evolves incrementally, based on small contributions of their users. *Woogle* extends Wikis with features that help to identify and prioritize missing information. Woogle thus

8. Summary

provides guidance for the evolution of Wiki content based on actual information needs of Wiki users.

The Semantic Web, which combines contributions of many users to a large knowledge base, suffers from similar problems as Wikis. Taking Semantic MediaWiki as an example, we show how *Semantic Need* can be used to guide the creation of structured semantic data.

The implementation and evaluation of all three tools shows, that need-driven knowledge sharing is technically feasible and can be an important extension for knowledge management practices.

8.2. Outlook

With respect to describing future work, we differentiate two distinct dimensions: *derivative* work, which directly extends and deepens the concepts described in this thesis, and *complementary* work, which refers to ideas that are orthogonal to our current work.

Besides these two dimensions, the chapters on NKS, Inverse Search, Woogle, and Semantic Need contain particular future work related to each individual approach.

8.2.1. Derivative Work

Web-scale Deployment of NKS

While the tools in this thesis primarily target enterprise knowledge sharing scenarios, our concepts can in part also be applied to larger communities. A particular example which is *Intopedia* (Romberg, 2010), which is a search engine that does not search existing articles in Wikipedia, but articles that do not yet exist. Thus, it helps potential Wikipedia contributors to find articles they can write or improve.

Furthermore, as discussed in Section 7.1.2, many *Linked Data* or *Open Data* initiatives have an insufficient awareness about their users' information needs. Applying NKS principles could be helpful in order to focus data quality efforts on information which is most valuable for information seekers.

Extending the Scope of Information Needs

Our list of information need attributes, as described in Section 4.3.3 on Page 74, is certainly preliminary due to our focus on queries (either keyword or structured). While queries can be obtained relatively easy and are probably the most natural representation of information needs, there exist opportunities for an extension towards a more holistic model of information needs. In particular, we have been conducting initial investigations on *working context* as a means to derive information needs. Since software developers are an important

group of knowledge workers, we derived context in the development environment, which might help to provide more extensive models of information need (Happel and Steinbauer, 2008; Maalej, 2010). Another interesting aspect in terms of knowledge sharing could be the urgency of information needs (Liu et al., 2009).

Extending Mediation Spaces and Mediation Services

The concepts of *mediation spaces* and *mediation services* have been useful for structuring interactions in information seeking and information provisioning during the development of NKS and related tools. A refinement of both concepts could provide a helpful framework for describing and comparing properties of knowledge sharing approaches and tools such as enterprise file sharing (Section 5.4.3) or CQA (Section 6.4.2).

Extending the Notion of Information Spaces

Our distinction of private and public information spaces has proven helpful to consider the diffusion of information. However, this private/public duality falls short to capture the full complexity of information spaces in real world applications. Once considering a multitude of information spaces, novel description techniques and algorithms would be required in order to apply NKS principles in more decentralized settings without a central query log or detailed information about available information spaces.¹

User Interfaces for Guiding and Motivating Contributions

A clear follow-up concerns analyzing the design of user interfaces and its influence on the contribution behavior of users. In Happel and Mazarakis (2010) we suggested that especially *personalized* approaches to motivate users – based on their personal psychological traits – would be an interesting and novel topic of research.

Besides the display of *motivating* factors, also the user experience of *guiding* contributions – i.e., conveying information needs of other users – could be more systematically analyzed. This could involve novel approaches to *audience design* (Rader, 2009) making contributors more precisely aware of their intended audience, or *capturing assistance* systems, which semi-automatically record user activities, which might be leveraged to capture new knowledge.²

Conducting Online Field Experiments

Conducting field experiments, as designed in Section 6.3.2 is a complex and long-term task. Especially real world evaluations in corporate settings or on-

¹E.g., considering the “Hidden” or “Deep Web” (He et al., 2007)

²See, e.g., Linton (2003); Leshed et al. (2008)

8. Summary

line communities require thorough preparation. However, to achieve statistically significant results, there do not seem to be many alternative ways. The *instrumentation framework* built into Woogle (see Section 6.2.4) can provide a good basis to conduct further experiments on different aspects of NKS and its tool implementations.

8.2.2. Complementary Work

Application to other areas of Knowledge Management

Finally, our approach also enables the investigation of further boxes in Figure 4.2. While this thesis concentrated on deriving which information needs to be shared in a situation of an *organizational information gap*, the situation of a *personal information gap* might be analyzed in terms of relevance for individuals in the organization (“need-to-know”). Also, an analysis of the further boxes could be interesting from an organizational perspective. While *information shortage* might be tackled by incentivizing the creation of content, *information overload* might be addressed by sophisticated retrieval techniques.

In combination, Figure 4.2 could be extended towards some kind of “organizational information dashboard”, providing reports and guidance for decision making about the status of information and knowledge sharing within an organization.

Beyond Knowledge Management

Unsatisfied needs can not just occur related to information, but as well related to physical goods. Similar to information search engines, product search engines do typically not allow to search for “future products” or “desired features”.³ In an age of mass-customization and user innovation toolkits (von Hippel, 2001) this suggests a convergence of classical market research and the analysis of users’ product needs towards what might be coined “prescriptive search”. Content farms, as discussed in Section 4.5 can be considered examples of this idea.

Also electronic communication might benefit from NKS principles. Email, for example, is still the most popular online activity besides search (Purcell, 2011) and also has a tight relation to knowledge sharing (Whittaker et al., 2006). Besides considering this overlap more deeply (like Mahmud et al., 2011; Hanrahan et al., 2011), it could be interesting to consider the information needs of email recipients during message composition.⁴

Private and Corporate Databases of Intentions

Google and other major search engines have achieved huge impact by establishing their *databases of intentions* (Battelle, 2005). Furthermore, this thesis

³See also Hasan et al. (2011)

⁴In this direction, see also Malone et al. (1986); Kraut et al. (2002)

has shown several examples concerning the usefulness of query information. Thus, it is even more surprising that there is not more consideration about these artifacts for usage on a personal (Amanda and Jansen, 2007; Jansen, 2007) or organizational level.

Especially in the case of companies, one could argue that capturing both, internally-rooted information demand (e.g., employees using a Web search engine) and incoming information demands by customers, could be far more useful beyond existing practices such as *search analytics* (as discussed in Section 4.5).

Part I.
Appendix

A. Algorithms

Algorithm 1 SelectDocuments(T , ID , OD , $Threshold$)

Require: T - set of terms, ID - initial set of documents, OD - ordered list of the recommended documents, $Threshold$ - the number of the documents that should be returned

```
1: while not empty  $T$  or  $Threshold > 0$  do
2:   /* Calculate the importance of a document for  $T$  */
3:   for all  $d \in ID$  do
4:      $Importance(d) = 0$ 
5:     for all  $t \in T$  do
6:       if  $Indexed(d, t)$  then
7:          $Importance(d) = Importance(d) + RelativeFrequency(d, t)$ 
8:       end if
9:     end for
10:  end for
11:  /* Rank documents from  $ID$  based on their importance by creating a
12:  new ranked list  $RD$  */
13:   $RankDocuments(RD, ID, Importance)$ 
14:  /* Select the most valuable document as a first in the ranked list  $RD$ 
15:  and put it as a leaf in the list of recommended documents  $OD$  */
16:   $OD = ADD(OD, RD(1))$ 
17:  /* Change the term set by removing all terms that index  $RD(1)$  */
18:  for all  $t \in T$  do
19:    if  $Indexed(RD(1), t)$  then
20:       $T = diff(T, t)$ 
21:    end if
22:  end for
23:  /* Remove the document  $RD(1)$  from the document set */
24:   $ID = diff(ID, RD(1))$ 
25:  /* Decrement threshold */
26:   $Threshold --$ 
27:  /* Repeat the procedure */
28:   $SelectDocuments(T, ID, OD, Threshold)$ 
29: end while
```

A. Algorithms

Algorithm 2 GetWantedResultValues(I, Q) - get all properties of an instance which have no value but are requested in queries

Require: I - knowledge base instance uri, Q - set of all queries with empty result values

```
 $R = resultTuple(property, queries, ain, marked)$ 
{Iterate all conjunctive queries with empty result values}
for all  $q \in Q$  do
   $T = extractAllTriples(q)$ 
5: {Iterate all triples in query}
  for all  $t \in T$  do
    if matches( $i, t$ ) then
      {Check if triple matches  $i$  as a subject}
       $o = object(t)$ 
10:  $S = selectVariables(q)$ 
      for all  $s \in S$  do
        if partOf( $o, s$ ) then
          {Check if triple object is an output variable}
           $p = extractProperty(t)$ 
15: if NOT isSet( $i, p$ ) then
             $R \leftarrow (p, q)$ 
          end if
        end if
      end for
    end if
  end for
20: end if
  end for
end for
{Semantically aggregate and properties in  $R$ }
 $aggRank(R)$ 
25: RETURN  $R$ 
```

Algorithm 3 GetNearMatches(I, Q) - get all properties of queries which nearly match an instance

Require: I - knowledge base instance uri, Q - set of all conjunctive queries with near matches

```

1:  $R = resultTuple(property, queries, ain, marked)$ 
2: /* Iterate all conjunctive queries with near matches */
3: for all  $q \in Q$  do
4:    $T = setOfAllTriplesIn(q)$ 
5:    $numNotSet \leftarrow 0$ 
6:    $numSatisfied \leftarrow 0$ 
7:   for all  $t \in T$  do
8:     if matches(i, t) then
9:       /* Check if triple matches i as a subject */
10:       $p = propertyOf(c)$ 
11:      if isSet(i, p) then
12:        /* Check if value for property p is set for i */
13:         $V = valuesOf(p, i)$ 
14:        for all  $v \in V$  do
15:          if satisfies(i, d, v) then
16:            {Check if one value for p satisfies d - as soon as one property
              is set consciously - drop conjunction}
17:             $numSatisfied ++$ 
18:            break
19:          end if
20:        end for
21:      else
22:        /* No value for property p is set for i */
23:         $numNotSet ++$ 
24:         $R \leftarrow p, q$ 
25:        /* Add p and q to result object */
26:      end if
27:    end if
28:  end for
29:  /* Only store missing properties if other properties satisfy query con-
30:   constraints */
31:  if ( $numNotSet + numSatisfied$ ) ==  $size(C)$  then
32:     $R \leftarrow T$ 
33:  end if
34: end for
35: /* Calculate ain for all properties of i causing near matches */
36: for all  $p \in R$  do
37:    $q = queriesFrom(R)concerning(p)$ 
38:    $a = ain(p, q)$ 
39:    $R \leftarrow (p, a, q)$ 
40: end for
41: /* Semantically aggregate and properties in R */
42:  $aggRank(R)$ 
43: RETURN  $R$ 

```

B.Woogle Evaluation Participation Dialog

special page

[Studyuser3](#)
[my/wiki](#)
[my/preferences](#)
[my/watchlist](#)
[my/contributions](#)
[log out](#)

Would you like to participate in the evaluation of Woogle?

By clicking "Participate"

- you agree to evaluate the "Woogle" search extension for MediaWiki
- you receive an improved search experience in this Wiki
- for the time of the study, your data (actions you pursue in the wiki such as page creation, searches but not reading pages) is logged for scientific purpose
- this data is not more (actually even less), than what is typically tracked by Webserver log files
- logs will be saved employing an anonymized userID which you can change at any time in [MediaWiki:Preferences](#)
- your data will not be used for anything but scientific evaluation and improvement of the Woogle extension for MediaWiki
- you may receive an online-questionnaire at the end of the study
- your participation is voluntary
- you may withdraw your participation at any time in [MediaWiki:Preferences](#)

By clicking "No, thank you.:"

- you will not take part in the evaluation of Woogle
- you will not receive new functionality - i.e. you can use this Wiki as usual
- the only data logged will be
 - a) your queries and clicks in an anonymous fashion. This data will only be employed for statistical aggregations
 - b) your actions (e.g. creating a page) in an completely anonymized fashion
- note that this is far less data than anyway collected by typically Webserver log files
- however, you may revoke both issues by checking the preference "log queries" in [MediaWiki:Preferences](#)
- you can revisit your decision about participation at any time

General remarks

We would be glad if you **support our scientific study** by clicking "Participate". If you have any questions, do not hesitate to contact [Hans-Joerg Happel \(happel@rz.tu.de\)](mailto:Hans-Joerg.Happel@rz.tu-dresden.de). Additional information about Woogle can be obtained at <http://www.teamweaver.org/wiki/Woogle#>. You can find the information of this page at [Woogle Study](#) at any time.

Participate
No, thank you

Figure B.1.: Screenshot of Participation Dialog for the Woogle Evaluation (see Section 6.3.2)

C.Semantic Need Public SMW Analysis

Id	Sitename	URL	Comment
CS	CS Wiki	http://cswiki.nudgenudge.eu/wiki/	Offline
ER	Eroge Wiki	http://eroge.wikia.com/wiki/	Still active
HA	HAR2009	https://wiki.har2009.org	Content has been replaced since analysis
HI	Historiographus	http://www.historiographus.org/wiki/	Offline; dump available from Internet Archive (https://archive.org/details/wiki-historiographusorg_wiki)
MN	Mount Wiki	http://mountwiki.com/wiki/view/Main_Page	Still active
PR	Protege Wiki	https://protegewiki.stanford.edu/wiki/Main_Page	Still active
SH	Sharing Buttons	http://www.sharingbuttons.org	Offline
TR	territoile	http://territoile.org/index.php	Offline

Table C.1.: List of Semantic MediaWiki Instances analyzed in Section 7.3.1

Queries analyzed for ISWC 2010 Submission #169 ("Semantic Need: Guiding Metadata Annotations by Questions People Ask")

All queries were extracted from Public Semantic Media/Wiki installations (see query URLs below)
 Crawl tooling available at <https://waves.fzi.de/svn/waves/trunk/MediaWikiTools/> (login: anonymous/anonymous)

QueryID	Results	CAT	Missing	selection	property%	Missing	NUM_CONJ	Empty	printouts	NumPrintOuts	Empty	cells
CS1	n.a.	8	6	75%	2	19	4	59%				
CS2	n.a.	7	0	0%	2	0	3	0%				
CS3	n.a.	1	0	0%	2	0	1	0%				
CS4	n.a.	16	4	25%	2	0	2	0%				
CS5	7	7	0	0%	4	0	4	0%				
HI1	1	18	17	94%	2	0	3	0%				
HI2	28	65	10	15%	2	0	2	0%				
HI3	1	3	1	33%	2	0	2	0%				
HI4	n.a.	18	17	94%	2	27	3	50%				
HI5	n.a.	65	10	15%	2	22	2	17%				
HI6	n.a.	3	0	0%	2	60	4	63%				
HI7	n.a.	24	13	54%	2	1	3	8%				
HI8	n.a.	4	2	50%	2	6	4	4%				
HI9	n.a.	35	0	0%	2	3	4	5%				
HI10	n.a.	15	0	0%	2	2	4	4%				
HI11	n.a.	14	0	0%	2	9	4	15%				
HI12	n.a.	15	0	0%	2	0	1	0%				
PR1	72	80	8	10%	2	13	1	16%				
PR2	n.a.	80	9	11%	2	1	1	1%				
PR3	n.a.	91	0	0%	2	57	1	63%				
PR4	n.a.	91	18	20%	2	75	2	41%				
PR5	n.a.	91	0	0%	2	1	1	1%				
PR6	n.a.	91	0	0%	2	0	1	0%				
TR1	70	102	32	31%	2	0	1	0%				

blue fields = used for near matches analysis

red fields = used for missing result values analysis

Figure C.1.: Details about Queries used in the Evaluation described in Section 7.3.1

QueryID	URL
CS1	http://cswiki.nudgenudge.eu/wiki/Template:Country
CS2	http://cswiki.nudgenudge.eu/wiki/Template:Country
CS3	http://cswiki.nudgenudge.eu/wiki/Template:Country
CS4	http://cswiki.nudgenudge.eu/wiki/Template:Publisher
CS5	http://cswiki.nudgenudge.eu/wiki/Research_Groups
HI1	http://www.historiographus.org/wiki/index.php?title=Historiographus:Sandbox
HI2	http://www.historiographus.org/wiki/index.php?title=Historiographus:Sandbox
HI3	http://www.historiographus.org/wiki/index.php?title=Historiographus:Sandbox
HI4	http://www.historiographus.org/wiki/index.php?title=Template:Fields_of_study
HI5	http://www.historiographus.org/wiki/index.php?title=Template:Fields_of_study
HI6	http://www.historiographus.org/wiki/index.php?title=Template:Fields_of_study
HI7	http://www.historiographus.org/wiki/index.php?title=Template:Persons
HI8	http://www.historiographus.org/wiki/index.php?title=Template:Persons
HI9	http://www.historiographus.org/wiki/index.php?title=Template:Persons
HI10	http://www.historiographus.org/wiki/index.php?title=Template:Persons
HI11	http://www.historiographus.org/wiki/index.php?title=Template:Published_works
HI12	http://www.historiographus.org/wiki/index.php?title=Template:Published_works
PR1	http://protegewiki.stanford.edu/index.php/Protege_Plugin_Library
PR2	http://protegewiki.stanford.edu/index.php/Template:Application
PR3	http://protegewiki.stanford.edu/index.php/Template:Application
PR4	http://protegewiki.stanford.edu/index.php/Template:VersionOfApplication
PR5	http://protegewiki.stanford.edu/index.php/Template:Plugin
PR6	http://protegewiki.stanford.edu/index.php/Template:Plugin
TR1	http://territoile.org/index.php?title=territoile:Villes

QueryID Query string

CS1 {{#ask: [[Category:Conference]] [[Country:{{PAGENAME}}]] | ?Subject | ?Organization | ?City | ?ConferenceDate | sort=ConferenceDate | default=Currently, the wiki doesn't contain information about conferences in {{PAGENAME}}. }}
CS2 {{#ask: [[Category:Research Group]] [[Country:{{PAGENAME}}]] | ?Subject | ?Organization | ?City | default=Currently, the wiki doesn't contain information about research groups from {{PAGENAME}}. }}
CS3 {{#ask: [[Category:Publisher]] [[Country:{{PAGENAME}}]] | ?Homepage | default=Currently, the wiki doesn't contain information about publishers from {{PAGENAME}}. }}
CS4 {{#ask: [[Category:Book]] [[Publisher:{{PAGENAME}}]] | ?Creator=Author | ?Subject | default=Currently there are no books listed in this wiki published by {{PAGENAME}}. }}
CS5 {{#ask: [[Category:Research Group]] | ?Subject | ?Organization | ?Country | ?Homepage }}
H11 {{#ask|format=table |intro="Events" in the field of study of "Botany": [[Category:Events|Event]] [[Has field of study:Botany]] | ?Category | ?Starting year=Start | ?Ending year=End }}
H12 {{#ask|format=table |intro="Persons" in the field of study of "Botany": [[Category:Persons|Person]] [[Has field of study:Botany]] | ?Born year=b. | ?Deceased year=d. }}
H13 {{#ask|format=table |intro="Scientific views" existing in the area of "Botany": [[Category:Scientific views|Scientific view]] [[Exists in area:Botany]] }}
H14 {{#ask|format=table |intro="Events" in the field of study of " {{PAGENAME}} ": [[Category:Events|Event]] [[Has field of study: {{PAGENAME}}]] | ?Category | ?Starting year=Start | ?Ending year=End }}
H15 {{#ask|format=table |intro="Persons" in the field of study of " {{PAGENAME}} ": [[Category:Persons|Person]] [[Has field of study: {{PAGENAME}}]] | ?Born year=b. | ?Deceased year=d. }}
H16 {{#ask|format=table |intro="Scientific views" existing in the area of " {{PAGENAME}} ": [[Category:Scientific views|Scientific view]] [[Exists in area: {{PAGENAME}}]] }}
H17 {{#ask: [[Category:Published works]] [[Has creator:{{PAGENAME}}]] | format=broadtable | intro="Published works": | ?Title | ?Document sub title=Subtitle | ?Starting date=1st ed | ?Has citation=Citation | sort=Starting date }}
H18 {{#ask: [[Category:Online publications]] [[Is about:{{PAGENAME}}]] | format=broadtable | intro="Online resources": | ?Resource type=Type | ?In language=Lang | ?URL }}
H19 {{#ask: [[Has authority:{{PAGENAME}}]] | format=broadtable | intro="Citations": | ?Title | ?Address | ?Publisher | ?Year | sort=Year }}
H110 {{#ask: [[Is about:{{PAGENAME}}]] | format=broadtable | ?Author | ?Title | ?Publisher | ?Year | sort=Author }}
H111 {{#ask: [[Is citation of:{{PAGENAME}}]] | ?Title | ?Address | ?Publisher | ?Year | sort=Year }}
H112 {{#ask: [[Is about:{{PAGENAME}}]] | ?Title | ?Address | ?Publisher | ?Year | sort=Year }}
PR1 {{#ask: [[Category:Plugin]] [[Last update:{{}}] | ?Last update=Updated | mainlabel=Plugin | limit=20 | sort=Last update | order=desc | format=broadtable | default=No news yet. }}
PR2 {{#ask: [[Category:Plugin]] [[For Application:{{PAGENAME}}]] | ?Has topics=Associated topics | mainlabel=Plugin | default=No plugins available. | format=broadtable }}
PR3 {{#ask: [[Category:Version]] [[Version of:{{PAGENAME}}]] | ?ChangeLog | mainlabel=Version | default=No version information available. | format=broadtable | sort=Version number | order=desc }}
PR4 {{#ask: [[Category:Version]] [[Compatible with:{{PAGENAME}}]] | ?Depends on | mainlabel=Plugin | format=broadtable | default=No plugin versions are compatible with this application version. }}
PR5 {{#ask: [[Category:Version]] [[Version of:{{PAGENAME}}]] | ?Compatible with | ?Depends on=Dependencies | mainlabel=Version | format=broadtable | sort=Version number | order=desc | default=No version information available. }}
PR6 {{#ask: [[Category:Version]] [[Version of:{{PAGENAME}}]] | ?ChangeLog=Changes in this version | mainlabel=Version | format=broadtable | sort=Version number | order=desc | default=No version information available. }}
TR1 {{#ask: [[Category:Plasticidens]] [[city:{{}}] | intro=artistes plasticidens par ville : | limit=200 | ?city | sort=city }}

D.Semantic Need Survey Questionnaire

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

<http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...>

Semantic Annotations in Semantic MediaWiki

This is a questionnaire about annotations to Wiki pages in Semantic MediaWiki (SMW) created by Hans-Jörg Happel (FZI Karlsruhe). Filling out the questionnaire **requires that you are familiar with Semantic MediaWiki (SMW)** and work with at least one instance of this system on a regular basis.

Thanks in advance for filling out the survey. Your support is very much appreciated! Filling out this questionnaire will **help us to develop and improve extensions that assist you in working with Semantic MediaWiki**.

Filling out the questionnaire will take approximately 15-20 minutes. Your responses will be processed anonymously.

Answers submitted until 2010-07-12 will be part of the final analysis. However, we will carry out an intermediate analysis on 2010-06-26 and thus kindly ask you to participate before this date already.

Note: Some questions ask for a particular SMW instance you work with. If you work with several SMW instances, please answer regarding the instance you consider the most relevant one. If the Wikis you work with differ significantly with respect to our questions, you may also elaborate on this in the final open input field!

There are 34 questions in this survey

Query result sparseness

By "query result sparseness" we denote the situation that one or more cells in an ASK-query result are empty (as depicted in Figure 1).

Country	Area	Population	Capital	Currency
Burundi	28,000 km ²	8,700,000	Bujumbura	
Central African Republic		4,400,400		Central African CFA franc
Mauritius	2,040 km ²		Port Louis	Mauritian rupee
Nigeria			Abuja	
Seychelles	455 km ²	87,476		Seychellois rupee
Somalia	637,661 km ²	9,133,000		Somali shilling
South Africa	1,221,037 km ²		Pretoria Bloemfontein Cape Town	Rand
Zimbabwe		12,521,000	Harare	US dollar

Figure 1: Sparse ASK-query result (some empty cells)

1 How often did you observe the problem of "query result sparseness" in any SMW instance? *

Please choose **only one** of the following:

- Don't know
- Never
- Rarely
- Sometimes
- Often

1 von 14

24.06.2010 16:07

Figure D.1.: Questionnaire used for the Survey described in Section 7.3.2

2 To which extent do you generally consider "query result sparseness" problematic if it occurs in a SMW instance? *

Please choose **only one** of the following:

- Don't know
- Not problematic at all
- Somehow problematic
- Very problematic

3 Please justify shortly why you consider "query result sparseness" problematic (or not)! *

Please write your answer here:

4 Please note if you have any additional remarks to the phenomenon of query result sparseness!

Please write your answer here:

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...

Query result incompleteness

By "query result incompleteness" we denote the situation that one or more instances do not appear in ASK-query results (although they should) due to missing annotations which are conditions of the ASK-Queries.

In the example depicted below, the "Nigeria" page lacks a proper semantic annotation of its population (see Figure 2) and is thus not listed as a result in a query it should actually appear in (Figure 3).

Nigeria

Nigeria (pronounced /naiˈdʒɪəriə/), officially the Federal Republic of Nigeria, is a country is located in West Africa. Its size is just under 923,768 km2 with an estimated population of almost 154.729.000. Its capital is Abuja. The currency used is Naira.

Category: Country

Facts about Nigeria ⓘ RDF feed 🔗

- HasCapital Abuja + 🔍
- OfContinent Africa + 🔍

Figure 2: Wiki page with some annotations

How many countries (defined in this wiki) have at least a population of 10,000,000?

<input checked="" type="checkbox"/> Country	<input checked="" type="checkbox"/> Population	<input checked="" type="checkbox"/> Continent
Zimbabwe	12,521,000	Africa
Germany	81,800,000	Europe

Figure 3: Incomplete ASK-query result → „Nigeria“ (Fig. 2) does not appear due to missing semantic annotation of its population)

5 How often did you observe the problem of "query result incompleteness" in any SMW instance? *

Please choose **only one** of the following:

- Don't know
- Never
- Rarely
- Sometimes
- Often

6 To which extent do you generally consider "query result incompleteness"

3 von 14

24.06.2010 16:07

problematic if it occurs in a SMW instance?

*

Please choose **only one** of the following:

- Don't know
- Not problematic at all
- Somehow problematic
- Very problematic

7 Please justify shortly why you consider "query result incompleteness" problematic (or not)! *

Please write your answer here:

8 Please note if you have any additional remarks to the phenomenon of query result incompleteness!

Please write your answer here:

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...

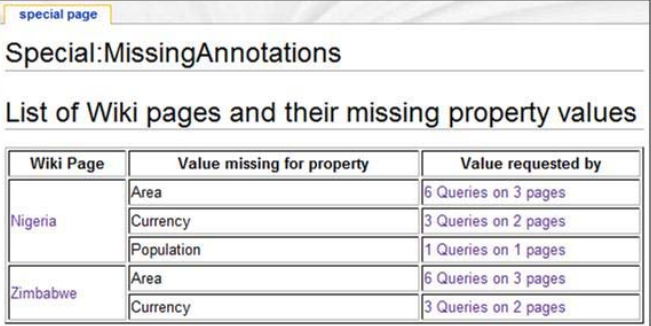
Semantic Annotation

9 Imagine a semantically annotated Wiki page such as "Nigeria" depicted in Figure 2.

Please describe how you would find out which additional semantic annotations should be added to that page in order to satisfy the information needs of Wiki users! (please describe short actions/process steps using bullet points)! *

Please write your answer here:

10 The Semantic Need extension for SMW provides a special page which lists Wiki pages and missing annotations as derived from Inline-ASK-Queries within the Wiki (see Figure 4 below).



The screenshot shows a web interface for a special page. At the top, there is a tab labeled 'special page'. Below it, the title is 'Special:MissingAnnotations'. Underneath, the heading reads 'List of Wiki pages and their missing property values'. A table follows with three columns: 'Wiki Page', 'Value missing for property', and 'Value requested by'. The table contains data for 'Nigeria' and 'Zimbabwe'.

Wiki Page	Value missing for property	Value requested by
Nigeria	Area	6 Queries on 3 pages
	Currency	3 Queries on 2 pages
	Population	1 Queries on 1 pages
Zimbabwe	Area	6 Queries on 3 pages
	Currency	3 Queries on 2 pages

Figure 4: Special page listing Wiki pages with missing annotations

Do you agree that this feature can be effective to help maintaining semantic annotations? *

Please choose **only one** of the following:

- Strongly disagree
 Disagree

5 von 14

24.06.2010 16:07

- Neutral
- Agree
- Strongly agree

11 Do you agree that this feature can help to guide annotation effort towards satisfying the most crucial information needs? *

Please choose **only one** of the following:

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

12 Do you agree that this feature can help motivating Wiki users to provide missing annotations? *

Please choose **only one** of the following:

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

13 Besides a Wiki-wide overview, the Semantic Need extension also provides a special page "Special:SemanticMatches/PAGENAME" which lists missing annotations for a particular page (as depicted in Figure 5 below).

special page	
Special:SemanticMatches	
List of missing property values for the Wiki page "Nigeria"	
Value missing for property	Value requested by
Area	6 Queries on 3 pages
Currency	3 Queries on 2 pages

Figure 5: Special page listing missing annotations for a particular Wiki page

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...

Furthermore, this information can also be shown within the article itself (see the modified "Nigeria" article screenshot in Figure 6 below).

The screenshot shows a Wikipedia article for Nigeria. The text reads: "Nigeria (pronounced /naiˈdʒɪəriə/), officially the Federal Republic of Nigeria, is a country is located in West Africa. Its size is just under 923,768 km2 with an estimated population of almost 154,729,000. Its capital is Abuja. The currency used is Naira." Below the text, there is a "Hint" box with an orange warning icon and the text: "The population of Nigeria is not annotated but requested on 4 pages. Please add the population of Nigeria." To the left of this box is a black arrow pointing to it with the word "Hint" in white. Below the hint box is a "Category: Country" field and a "Facts about Nigeria" section with "HasCapital Abuja" and "OfContinent Africa".

Figure 6: Hint to missing annotations within the Wiki page body

Do you agree that this feature can be effective to help maintaining semantic annotations? *

Please choose **only one** of the following:

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

14 Do you agree that this feature can help to guide annotation effort towards satisfying the most crucial information needs? *

Please choose **only one** of the following:

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

15 Do you agree that this feature can help motivating Wikis users to provide missing annotations? *

7 von 14

24.06.2010 16:07

Please choose **only one** of the following:

- Strongly disagree
- Disagree
- Neutral
- Agree
- Strongly agree

16 Please note if you have any additional remarks on this topic?

Please write your answer here:

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

<http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...>

Knowledge Engineering

17 With how many SMW installations do you work on a regular basis? *

Please write your answer here:

18 What are the rough general characteristics of these Wikis (#Pages, #Inline ASK-Queries, #Users, Purpose) (If you work with more than three Wikis regularly, just describe the three most important ones)? *

Please write your answer here:

You may retrieve information about pages and users from Special:Statistics. For inline queries you might differentiate if queries occur on regular or template pages. **It is fine if you just give some rough estimation for these numbers!**

19 What characterizes the knowledge domain covered by your SMW instance best? *

Please choose **only one** of the following:

- Fixed set of entities and properties from a standardized or well-understood domain (e.g. bibliographic data, project management)
- Fixed core of well-understood entities and properties but additional entities and properties might emerge
- Open domain without many predetermined entities and properties
- Other

20 How would you characterize the semantic data model of your SMW instance? *Please choose **only one** of the following:

- Mostly prescribed in Templates/Semantic Forms
- Mostly prescribed in Templates/Semantic Forms but also some free-from annotations
- A roughly equal mix of Templates/Semantic Forms and free-from semantic annotations
- Mostly free-from semantic annotations but also few Templates/Semantic Forms
- Mostly free-from semantic annotations within the Wiki text
- Other

21 How often do changes to your semantic data model occur (e.g. new categories or properties are added)? *Please choose **only one** of the following:

- Don't know
- Less or never
- At least once per month
- At least once per week
- At least once per day

22 How often do changes to your semantic annotation data occur (e.g. new Wiki pages or annotation values are added)? *Please choose **only one** of the following:

- Don't know
- Less or never
- At least once per month
- At least once per week
- At least once per day

23 How would you characterize the user base of your SMW instance? *Please choose **only one** of the following:

- Fixed set of users with no or very few new users joining over time
- Fixed set of users with some new users joining over time
- Open user base (new users may constantly access/join the Wiki)
- Other

24 How would you characterize the information needs that the users typically

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

<http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...>

seek to satisfy in your SMW instance? *

Please choose **only one** of the following:

Don't know

Information needs of users are mostly predictable

Information needs of users are somehow predictable, but users might also have unanticipated information needs

Information needs of users can hardly be predicted

Other

25 What are your personal roles in the SMW instance? *

Please choose **all** that apply:

Wiki technical administrator (technical administration/user management)

Wiki gardener (organizing content)

Wiki contributor (contributing content)

Wiki reader (reading content)

Other:

26 Do you employ any specific methodology, practices or tools to guide the evolution of the semantic data model and/or content of your SMW instances? Please shortly describe what and how you use it! *

Please write your answer here:

Final Questions

27 How experienced are you in working with SMW? *

Please choose **only one** of the following:

- Novice
 Intermediate
 Expert

28 Since when are you working with SMW? *

Please write your answer here:

29 In which country are you working? *

Please write your answer here:

30 What is the primary usage context of the SMW instances you work with? *

Please choose **only one** of the following:

- Industry
 Academia
 Non-profit (Open Source Projects)
 Non-profit (Culture/Arts)
 Non-profit (Other)
 Other

31 Would you be interested in using a stable version of Semantic Need (as depicted in Figures 4-6) in one of your SMW installations?

Please choose **only one** of the following:

- Yes
 No
 Don't know

32 Would you be willing to participate in scientific evaluations (e.g. small user studies) of Semantic Need in your Wiki environment?

Please choose **only one** of the following:

D. Semantic Need Survey Questionnaire

LimeSurvey - Semantic Annotations in Semantic MediaWiki

<http://amazonas.fzi.de/limesurvey/admin/admin.php?action=showprinta...>

- Yes
- No
- Don't know

33 Please enter your E-Mail address here (will be used for contact/clarification purposes on this topic only)!

Please write your answer here:

34 Thanks a lot for your participation! Any final thoughts you wish to share?

Please write your answer here:

List of Figures

1.1. Structure of the Thesis	4
2.1. Building Blocks of Knowledge Management (Probst et al., 2006)	10
2.2. Relationship of Organization, Technology and Knowledge . . .	11
2.3. Relationship of Data, Information, and Knowledge (Maier, 2007, p. 71)	13
2.4. Relationship of Data, Information, and Knowledge (Tang et al., 2006)	14
2.5. Modes of Knowledge Creation according to Nonaka (1994) . . .	15
2.6. Knowledge Maturing Process (Maier and Schmidt, 2007)	18
2.7. Distributed Work requires, but also impedes Knowledge Sharing	26
3.1. A Model of Information Seeking Behavior (Wilson, 1981) . . .	34
3.2. The Context of Information Seeking (Wilson, 1981)	36
3.3. Information Needs and Seeking (Wilson, 1981)	37
3.4. The Dimension of Time in Information Retrieval	41
3.5. Knowledge Sharing Process	45
3.6. Knowledge Sharing as a Communication Process	53
3.7. Continuum of Mediated Communication Approaches (partly based on Wouters and Gerbec, 2003)	54
4.1. Interrelation of different Sets of Documents for a certain IR system	64
4.2. Availability of Information related to some particular Information Need in the Private vs. Organizational Information Space	71
4.3. Extended Model of Knowledge Sharing as a Communication Process (adapted from Figure 3.6)	74
5.1. Users and Documents in our motivating example	95
5.2. Addressing Information Providers with Inverse Search (extension of Figure 3.4)	96
5.3. System Architecture	97
5.4. Inverse Search File Sharing UI	101

5.5.	Inverse Search Knowledge Acquisition UI	101
5.6.	NDF values (<i>public</i> and <i>user – average</i>) for 14 674 terms	104
6.1.	“Red link” in the English Wikipedia	111
6.2.	Adding an Article to the Watchlist	111
6.3.	Unsatisfied Search in the English Wikipedia	112
6.4.	General Architecture of Woogle	117
6.5.	Woogle4MediaWiki Search Results List Screenshot	118
6.6.	Woogle Mouse-over Display for “Red Links”	119
6.7.	Information Need Indicators in the Woogle4MediaWiki Search UI	120
6.8.	Queries by Number of Search Results and Number of Search Results clicked	127
7.1.	Possible Distributions of Metadata in Private vs. Public Information Spaces (Happel and Stojanovic, 2008, adapted from Figure 4.2)	138
7.2.	KB denotes the set of all axioms in the knowledge base. XKB denotes the set of all axioms which have to be added to the KB to satisfy all structured queries, that should be satisfiable according to full knowledge of the domain (See also Figure 4.1)	142
7.3.	Example of a Sparse Result Set	144
7.4.	Abstract Architecture of Semantic Need	146
7.5.	Architecture of Semantic Need for MediaWiki (instantiation of Figure 7.4)	150
7.6.	Wiki-wide Overview of Pages and their Missing Annotations	151
7.7.	Missing Annotations for a Specific Wiki Page (i.e., specialization of Figure 7.6)	151
7.8.	In-page Display of Missing Annotations	152
B.1.	Screenshot of Participation Dialog for the Woogle Evaluation (see Section 6.3.2)	184
C.1.	Details about Queries used in the Evaluation described in Section 7.3.1	187
D.1.	Questionnaire used for the Survey described in Section 7.3.2	192

List of Tables

1.1. Categorization of Information Sharing (Scope of the Thesis is <i>highlighted</i>)	2
2.1. Codifications vs. Personalization (adapted from Hansen et al., 1999)	21
2.2. Traditional Office Work vs. Knowledge Work (excerpt of Maier, 2007, p. 49f)	24
2.3. Coordination vs. Knowledge Sharing	27
4.1. Explanation of the sets depicted in Figure 4.1	66
4.2. Exemplary Mediation Services and Spaces for Need-Driven Knowledge Sharing	75
4.3. Aggregate Query Log	77
4.4. Aggregate Query/Click Log	77
5.1. Distribution of Terms in the analyzed Indices	103
7.1. Comparison of Paradigms	140
7.2. Overview of Surveyed SMW Installations	153
7.3. Missing Result Values for the <i>IQ_{ECPO}</i> Queries	155
7.4. Near Matches for the <i>IQ_{ECCJ}</i> Queries	156
C.1. List of Semantic MediaWiki Instances analyzed in Section 7.3.1	186

LIST OF TABLES

List of Theorems

2.1. Definition (Knowledge Management)	7
3.1. Definition (Knowledge Sharing)	43
3.2. Definition (Mediation Service)	53
3.3. Definition (Mediation Space)	53
4.1. Definition (Need-Driven Knowledge Sharing)	60
4.2. Definition (Unsatisfied Information Needs)	64
4.3. Definition (Recurring Information Needs)	65
4.4. Definition (Organizational Information Need)	68
4.5. Definition (Private Information Space)	69
4.6. Definition (Shared Information Space)	69
4.7. Definition (Organizational Information Space)	69
4.8. Definition (Organizational Information Gap)	70
5.1. Definition (Inverse Search)	96

LIST OF THEOREMS

Nomenclature

AIN	<i>Aggregate Information Need</i>
API	<i>Application Programming Interface</i>
CIR	<i>Collaborative Information Retrieval</i>
CIS	<i>Collaborative Information Seeking</i>
CQA	<i>Collaborative Question Answering</i>
CSCW	<i>Computer-Supported Cooperative Work</i>
CSD	<i>Collaborative Software Development</i>
DCS	<i>Document and Content Sharing</i>
DL	<i>Description Logics</i>
FAQ	<i>Frequently Asked Questions</i>
HR	<i>Human Ressources</i>
IM	<i>Instant Messaging</i>
IR	<i>Information Retrieval</i>
IS	<i>Information Seeking</i>
IT	<i>Information Technology</i>
KB	<i>Knowledge Base</i>
KMS	<i>Knowledge Management Systems</i>
KR	<i>Knowledge Representation</i>
MW	<i>MediaWiki</i>
NKS	<i>Need-driven Knowledge Sharing</i>
OIG	<i>Organizational Information Gap</i>
OIN	<i>Organizational Information Need</i>
OMIS	<i>Organizational Memory Information Systems</i>
OSGi	<i>Open Services Gateway interface</i>

LIST OF THEOREMS

OWL *Web Ontology Language*

P2P *Peer-to-peer*

Q&A *Question & Answer*

RCP *(Eclipse) Rich Client Platform*

RDF *Ressource Description Framework*

SMW *Semantic MediaWiki*

SPARQL *SPARQL Protocol And RDF Query Language*

TMS *Transactive Memory System*

UI *User Interface*

URI *Uniform Ressource Identifyer*

URL *Uniform Ressource Locator*

WWW *World Wide Web*

Bibliography

- Abecker, A., Bernardi, A., Hinkelmann, K., Kuhn, O. and Sintek, M. (1998). Toward a technology for organizational memories, *Intelligent Systems and their Applications, IEEE* **13**(3): 40–48. 20, 74
- Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M. and Tang, N. (2016). Detecting data errors: Where are we and what needs to be done?, *Proc. VLDB Endow.* **9**(12): 993–1004. 164
- Aberer, K., Cudré-Mauroux, P., Datta, A., Despotovic, Z., Hauswirth, M., Puceva, M. and Schmidt, R. (2003). P-grid: A self-organizing structured p2p system, *SIGMOD Rec.* **32**(3): 29–33. 105
- Abiteboul, S., Arenas, M., Barceló, P., Bienvenu, M., Calvanese, D., David, C., Hull, R., Hüllermeier, E., Kimelfeld, B., Libkin, L., Martens, W., Milo, T., Murlak, F., Neven, F., Ortiz, M., Schwentick, T., Stoyanovich, J., Su, J., Suciú, D., Vianu, V. and Yi, K. (2017). Research directions for principles of data management (abridged), *SIGMOD Rec.* **45**(4): 5–17. 133
- Ackerman, M. S. and Malone, T. W. (1990). Answer garden: a tool for growing organizational memory, *Proceedings of the ACM SIGOIS and IEEE CS TC-OA conference on Office information systems*, ACM, New York, NY, USA, pp. 31–39. 88, 129
- Ackerman, M. S. and McDonald, D. W. (1996). Answer garden 2: merging organizational memory with collaborative help, *CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 97–105. 88
- Ackerman, M. S., Pipek, V. and Wulf, V. (eds) (2003). *Sharing Expertise: Beyond Knowledge Management*, MIT Press. 230, 233, 241
- Ackerman, M. S., Swenson, A., Cotterill, S. and DeMaagd, K. (2003). I-diag: From community discussion to knowledge distillation, in M. Huysman, E. Wenger and V. Wulf (eds), *Communities and Technologies*, Kluwer, pp. 307–325. 88
- Acosta, M., Simperl, E., Flöck, F. and Vidal, M.-E. (2017). Enhancing answer completeness of sparql queries via crowdsourcing, *Web Semantics: Science, Services and Agents on the World Wide Web* **45**(Supplement C): 41–62. 168
- Adar, E., Teevan, J. and Dumais, S. T. (2008). Large scale analysis of web revisitation patterns, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08*, ACM, New York, NY, USA, pp. 1197–1206. 66

BIBLIOGRAPHY

- Agichtein, E., Brill, E. and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information, *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 19–26. 38
- Agrahri, A. K., Manickam, D. A. T. and Riedl, J. (2008). Can people collaborate to improve the relevance of search results?, *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, ACM, New York, NY, USA, pp. 283–286. 128
- Ahern, S., Eckles, D., Good, N. S., King, S., Naaman, M. and Nair, R. (2007). Over-exposed?: privacy patterns and considerations in online and mobile photo sharing, *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, ACM, New York, NY, USA, pp. 357–366. 94, 107
- Ahn, J., Butler, B. S., Weng, C. and Webster, S. (2013). Learning to be a better q'er in social q&a sites: Social norms and information artifacts, *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, American Society for Information Science, Silver Springs, MD, USA, pp. 4:1–4:10. 130
- Alasiry, A., Levene, M. and Poulouvasilis, A. (2014). Mining named entities from search engine query logs, *Proceedings of the 18th International Database Engineering & Applications Symposium*, IDEAS '14, ACM, New York, NY, USA, pp. 46–56. 161
- Alavi, M. and Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues, *MIS Quarterly* **25**(1): 107–136. 43, 49
- Alborzi, F., Chirkova, R., Doyle, J. and Fathi, Y. (2015). Determining query readiness for structured data, *Big Data Analytics and Knowledge Discovery: 17th International Conference, DaWaK 2015, Valencia, Spain, September 1-4, 2015, Proceedings*, Springer, pp. 3–14. 167
- Allan, N., Heisig, P., Iske, P., Kelleher, D., Mekhilef, M., Oertel, R., Olesen, A. J. and Leeuwen, M. V. (2004). European guide to good practice in knowledge management: Part 5: Km terminology, *Technical report*, CEN/ISSS.
URL: <ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/KM/CWA14924-05-2004-Mar.pdf> 7, 12
- Allen, T. J. (1984). *Managing the Flow of Technology*, MIT Press. 11, 25, 26
- Amanda, S. and Jansen, B. J. (2007). People's Query Logs: Personal Information Management, in E. Amitay, C. G. Murray and J. Teevan (eds), *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*. 175
- Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media, *CHI '07: Proceedings of the SIGCHI conference on*

- Human factors in computing systems*, ACM, New York, NY, USA, pp. 971–980. 71, 136, 137
- Anderson, C. (2004). The long tail, *Wired* **12**(10). 28
- Anderson, C. (2007). *The Long Tail: How Endless Choice is Creating Unlimited Demand*, Random House Business. 28
- Andriole, S. J. (2010). Business impact of web 2.0 technologies, *Commun. ACM* **53**(12): 67–79. 28, 29
- Angiolillo, J. (2003). Search log analysis as a usability engineering tool, *CHI'03 Workshop on Best practices and future visions for search user interfaces*. 82
- Ankolekar, A., Krötzsch, M., Tran, T. and Vrandečić, D. (2007). The two cultures: mashing up web 2.0 and the semantic web, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 825–834. 135
- Antin, J. and Cheshire, C. (2010). Readers are not free-riders: Reading as a form of participation on wikipedia, *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10*, ACM, New York, NY, USA, pp. 127–130. 87
- Arazy, O., Daxenberger, J., Lifshitz-Assaf, H., Nov, O. and Gurevych, I. (2016). Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction, *Information Systems Research* **27**(4): 792–812. 131
- Ardichvili, A., Page, V. and Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice, *Journal of Knowledge Management* **7**(1): 64–77. 49, 137
- Argote, L. (1999). *Organizational Learning: Creating, Retaining, and Transferring Knowledge*, 1st edn, Kluwer Academic Publishers, Norwell, MA, USA. 7, 44
- Argote, L., McEvily, B. and Reagans, R. (2003). Introduction to the special issue on managing knowledge in organizations: Creating, retaining, and transferring knowledge, *Management Science* **49**(4): v–viii. 44
- Argyris, C. and Schön, D. (1978). *Organizational Learning: A Theory of Action Perspective*, Addison-Wesley Company. 7
- Armstrong, D. J. and Cole, P. (2002). *Managing Distances and Differences in Geographically Distributed Work Groups*, in Hinds and Kiesler (2002b), chapter 7, pp. 167–186. 26
- Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D. and Patel-Schneider, P. F. (eds) (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press. 135, 140, 215
- Baader, F. and Nutt, W. (2003). Basic description logics, in Baader et al. (2003), pp. 43–95. 141
- Back, A., Gronau, N. and Tochtermann, K. (2008). *Web 2.0 in der Unternehmenspraxis: Grundlagen, Fallstudien und Trends zum Einsatz von Social Software*, Oldenbourg. 28

BIBLIOGRAPHY

- Baeza-Yates, R. and Riberio-Neto, B. (1999). *Modern Information Retrieval*, ACM Press. 65, 77
- Baldwin, C. Y. and Clark, K. B. (2000). *Design Rules: The Power of Modularity Volume 1*, MIT Press. 16, 25
- Baltadzhieva, A. and Chrupa, G. (2015). Question quality in community question answering forums: A survey, *SIGKDD Explor. Newsl.* **17**(1): 8–13. 130
- Bannon, L. and Bødker, S. (1997). Constructing common information spaces, *Proceedings of the Fifth Conference on European Conference on Computer-Supported Cooperative Work, ECSCW'97*, Kluwer Academic Publishers, Norwell, MA, USA, pp. 81–96. 85
- Bao, J. and Li Ding, J. A. H. (2008). Knowledge representation and query in semantic mediawiki: A formal study, *Tetherless World Constellation (RPI) Technical Report*, pp. TW–2008–42. 149
- Barateiro, J. and Galhardas, H. (2005). A survey of data quality tools, *Datenbank-Spektrum* **14**: 15–21. 164
- Barrett, D. J. (2008). *MediaWiki*, 1 edn, O'Reilly Media, Inc. 110, 122
- Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Portfolio Hardcover. 1, 174
- Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P. and Kraut, R. E. (2004). Using social psychology to motivate contributions to online communities, *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 212–221. 71, 72, 84
- Bender, M., Michel, S., Zimmer, C. and Weikum, G. (2005). The minerva project: Towards collaborative search in digital libraries using peer-to-peer technology, *Proceedings of the 6th Thematic Conference on Peer-to-Peer, Grid, and Service-Oriented in Digital Library Architectures, DELOS'04*, Springer, Berlin, Heidelberg, pp. 80–95. 105
- Benedikt, M., Bourhis, P., ten Cate, B. and Puppis, G. (2016). Querying visible and invisible information, *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '16*, ACM, New York, NY, USA, pp. 297–306. 163
- Bentley, R., Horstmann, T., Sikkil, K. and Trevor, J. (1995). Supporting collaborative information sharing with the world wide web: The bscw shared workspace system, *WWW '95: Proceedings of the 4th international conference on World Wide Web*, O'Reilly Associates, pp. 63–74. 50
- Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C. and Stumme, G. (2010). The social bookmark and publication management system bibsonomy, *The VLDB Journal* **19**(6): 849–875. 50, 108, 163
- Bergman, M., Milo, T., Novgorodov, S. and Tan, W.-C. (2015). Query-oriented data cleaning with oracles, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15*, ACM, New York, NY, USA, pp. 1199–1214. 168

- Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic Web, *Scientific American* **284**(5): 34–43. 133, 135
- Bernstein, M. S., Marcus, A., Karger, D. R. and Miller, R. C. (2010). Enhancing directed content sharing on the web, *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, ACM, New York, NY, USA, pp. 971–980. 48, 84, 93
- Bienvenu, M., Bourgaux, C. and Goasdoué, F. (2016). Query-driven repairing of inconsistent dl-lite knowledge bases, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, AAAI Press, pp. 957–964. 166
- Bizer, C. and Cyganiak, R. (2006). D2r server-publishing relational databases on the semantic web (poster), *The Semantic Web - ISWC 2006: 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006. Proceedings*, Springer, Berlin, Heidelberg. 133, 136
- Bock, G.-W., Mahmood, M., Sharma, S. and Kang, Y. J. (2010). The impact of information overload and contribution overload on continued usage of electronic knowledge repositories, *Journal of Organizational Computing and Electronic Commerce* **20**(3): 257–278. 44, 87
- Boh, W. F. (2007). Mechanisms for sharing knowledge in project-based organizations, *Information and Organization* **17**(1): 27 – 58. 21
- Bonifacio, M., Bouquet, P., Mameli, G. and Nori, M. (2003). Peer-mediated distributed knowledge management, in *AMKM van Elst and Abecker (2004)*, pp. 31–47. 20
- Bonifacio, M., Cuel, R., Mameli, G. and Nori, M. (2002). A peer-to-peer architecture for distributed knowledge management, *In Proceedings of 3rd International Symposium on Multi-Agent Systems, Large Complex Systems, and E-Businesses (MALCEB2002)*, pp. 8–10. 20
- Boose, J. H. and Gaines, B. R. (1989). Knowledge acquisition for knowledge-based systems: Notes on the state-of-the-art, *Machine Learning* **4**(3): 377–394. 9
- Bradshaw, S., Light, M. and Eichmann, D. (2006). (Bee)Dancing on the Boundary between PIM and GIM, *Proceedings of the 2nd Invitational Workshop on Personal Information Management at SIGIR 2006*. 23, 24
- Braun, G. E. and Beckert, J. (1992). *Funktionalorganisation*, 3 edn, Schäffer-Poeschel, Stuttgart, pp. 640–652. 12
- Braun, S. and Schmidt, A. (2007). Wikis as a technology fostering knowledge maturing: What we can learn from wikipedia, *Proceedings of the 7th International Conference on Knowledge Management (I-KNOW 2007)*, JUCS. 51, 109
- Brennan, R. (2017). Challenges for value-driven semantic data quality management, *Proceedings of the 19th International Conference on Enterprise Information Systems*, pp. 385–392. 167

BIBLIOGRAPHY

- Brewer, R. S. (2000). Improving problem-oriented mailing list archives with mcs, *Proceedings of the 2000 International Conference on Software Engineering*, pp. 95–104. 88
- Broder, A. (2002). A taxonomy of web search, *SIGIR Forum* **36**(2): 3–10. 40, 66, 75, 126
- Brown, B. A. T., Sellen, A. J. and O’Hara, K. P. (2000). A diary study of information capture in working life, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’00, ACM, New York, NY, USA, pp. 438–445. 24
- Brown, J. S. and Duguid, P. (1996). The social life of documents, *First Monday* **1**(1). 1
- Brown, J. S. and Duguid, P. (1998). Organizing knowledge, *California Management Review* **40**(3): 90–111. 11, 16
- Brown, J. S. and Hagel, III, J. (2005). From push to pull: The next frontier of innovation, *McKinsey Quarterly* (3): 82–91. 22, 82
- Bruce, H. (2005). The pain hypothesis, in Fisher et al. (2005), chapter 46, pp. 270–274. 45
- Brush, A. B., Inkpen, K. M. and Tee, K. (2008). Spares: Exploring sharing suggestions to enhance family connectedness, *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW ’08, ACM, New York, NY, USA, pp. 629–638. 84
- Bryant, S. L., Forte, A. and Bruckman, A. (2005). Becoming wikipedia: Transformation of participation in a collaborative online encyclopedia, *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP ’05, ACM, New York, NY, USA, pp. 1–10. 87
- Buckland, M. (2017). *Information and Society*, MIT Press. 1
- Bughin, J. and Manyika, J. (2007). How businesses are using web 2.0, *McKinsey Quarterly* . 109
- Burghardt, T., Buchmann, E., Böhm, K. and Clifton, C. (2008). Collaborative search and user privacy: How can they be reconciled?, *CollaborateCom*, pp. 85–99. 119
- Burnett, G. (2000). Information exchange in virtual communities: a typology, *Information Research* **5**(4): 1–1. 44, 56
- Bush, V. (1945). As We May Think, *Atlantic Monthly* **176**(1): 641–649. 23
- Butler, D. (2008). Web data predict flu., *Nature* **456** **7220**: 287–8. 83
- Bühner, R. (1992). *Spartenorganisation*, 3 edn, Schäffer-Poeschel, Stuttgart, pp. 2274–2287. 19
- Cabrera, A. and Cabrera, E. F. (2002). Knowledge-sharing dilemmas, *Organization Studies* **23**: 687–710. 27, 49, 137

- Camp, G. and Ulieru, M. (2007). In-order: Enhancing google via stigmergic query refinement, *International Journal of Computer Systems Science and Engineering* **22**(4): 128
- Campinas, S. (2014). Live sparql auto-completion, *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, CEUR-WS.org, Aachen, Germany, pp. 477–480. 160
- Case, D. O. (2002). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior (Library and Information Science)*, 1st edn, Academic Press. 12, 13, 33, 34, 35
- Cataldo, M., Wagstrom, P. A., Herbsleb, J. D. and Carley, K. M. (2006). Identification of coordination requirements: Implications for the design of collaboration and awareness tools, *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW '06*, ACM, New York, NY, USA, pp. 353–362. 47
- Chapman, A. and Jagadish, H. V. (2009). Why not?, *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, ACM, New York, NY, USA, pp. 523–534. 166
- Chen, K., Chen, H., Conway, N., Hellerstein, J. M. and Parikh, T. S. (2010). Usher: Improving data quality with dynamic forms, *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 321–332. 161
- Cheney, J., Chiticariu, L. and Tan, W.-C. (2009). Provenance in databases: Why, how, and where, *Found. Trends databases* **1**(4): 379–474. 166
- Cheng, G., Ge, W. and Qu, Y. (2008). Falcons: Searching and browsing entities on the semantic web, *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, ACM, New York, NY, USA, pp. 1101–1102. 159
- Cheong, R. K. F. and Tsui, E. (2011). From skills and competencies to outcome-based collaborative work: Tracking a decade's development of personal knowledge management (pkm) models, *Knowledge and Process Management* **18**(3): 175–193. 23
- Cheshire, C. and Antin, J. (2008). The social psychological effects of feedback on the production of internet information pools, *Journal of Computer-Mediated Communication* **13**(3): 705–727. 71
- Chu, X., Ilyas, I. F., Krishnan, S. and Wang, J. (2016). Data cleaning: Overview and emerging challenges, *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, ACM, New York, NY, USA, pp. 2201–2206. 164
- Churchill, E. F., Sullivan, J. W., Golovchinsky, G. and Snowdon, D. (1999). Collaborative and co-operative information seeking: Cscw'98 workshop report, *SIGGROUP Bull.* **20**(1): 56–59. 42
- Cirasella, J. (2007). Google sets, google suggest, and google search history: Three more tools for the reference librarian's bag of tricks, *The Reference Librarian* **48**(1): 57–65. 128

BIBLIOGRAPHY

- Clement, A. and Wagner, I. (1995). Fragmented exchange: Disarticulation and the need for regionalized communication spaces, *Proceedings of the Fourth Conference on European Conference on Computer-Supported Cooperative Work*, ECSCW'95, Kluwer Academic Publishers, Norwell, MA, USA, pp. 33–49. 85
- Codd, E. F. (1986). Missing information (applicable and inapplicable) in relational databases, *SIGMOD Rec.* **15**(4): 53–53. 164
- Collis, N. and Frommholz, I. (2017). Aquacold - a crowdsourced query understanding and query construction tool for the linked data web, *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017)*, Thessaloniki, Greece, pp. 74–86. 160
- Conway, M. E. (1968). How do committees invent?, *Datamation* **14**(4): 28–31. 25
- Cosley, D., Frankowski, D., Kiesler, S., Terveen, L. and Riedl, J. (2005). How oversight improves member-maintained communities, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, ACM, New York, NY, USA, pp. 11–20. 88
- Cosley, D., Frankowski, D., Terveen, L. and Riedl, J. (2006). Using intelligent task routing and contribution review to help communities build artifacts of lasting value, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, ACM, New York, NY, USA, pp. 1037–1046. 84, 139
- Cosley, D., Frankowski, D., Terveen, L. and Riedl, J. (2007). Suggestbot: using intelligent task routing to help people find work in wikipedia, *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, ACM, New York, NY, USA, pp. 32–41. 84, 130, 132
- Cramton, C. D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration, *Organization Science* **12**(3): 346–371. 27
- Cress, U. and Hesse, F.-W. (2004). Knowledge sharing in groups: experimental findings of how to overcome a social dilemma, *ICLS '04: Proceedings of the 6th international conference on Learning sciences*, International Society of the Learning Sciences, pp. 150–157. 49, 137
- Cuel, R., Morozova, O., Rohde, M., Simperl, E., Siorpaes, K., Tokarchuk, O., Wiedenhofer, T., Yetim, F. and Zamarian, M. (2011). Motivation mechanisms for participation in human-driven semantic content creation, *IJKEDM* **1**(4): 331–349. 47, 161
- Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing, *Management Science* **50**(3): 352–364. 20, 29, 43
- Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design, *Manage. Sci.* **32**(5): 554–571. 54
- Dalal, B., Nelson, L., Smetters, D., Good, N. and Elliot, A. (2008). Ad-hoc guesting: when exceptions are the rule, *Proceedings of the 1st Conference*

- on Usability, Psychology, and Security*, USENIX Association, Berkeley, CA, USA, pp. 9:1–9:5. 94, 107
- Darari, F., Nutt, W., Pirrò, G. and Razniewski, S. (2013). Completeness statements about rdf data sources and their use for query answering, *The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, Springer, Berlin, Heidelberg, pp. 66–83. 164
- Darari, F., Prasojo, R. E., Razniewski, S. and Nutt, W. (2017). COOL-WD: A completeness tool for wikidata, *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, 2017*. 162
- Darari, F., Razniewski, S., Prasojo, R. E. and Nutt, W. (2016). Enabling fine-grained rdf data completeness assessment, *Web Engineering: 16th International Conference, ICWE 2016, Lugano, Switzerland, June 6-9, 2016. Proceedings*, Springer, pp. 170–187. 162
- Davenport, T. H. and Prusak, L. (1998). *Working Knowledge*, Harvard Business School Press. 1, 12, 19
- David, C., Kohlhase, M., Lange, C., Rabe, F., Zhiltsov, N. and Zholudev, V. (2010). Publishing math lecture notes as linked data, *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part II, ESWC’10*, Springer, Berlin, Heidelberg, pp. 370–375. 163
- Dearman, D., Kellar, M. and Truong, K. N. (2008). An examination of daily information needs and sharing opportunities, *CSCW ’08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 679–688. 42, 71, 87, 93, 94
- Decker, S. (2002). *Semantic web methods for knowledge management*, PhD thesis, Universität Karlsruhe. 133, 137
- Decker, S., Jannink, J., Melnik, S., Mitra, P., Staab, S., Studer, R. and Wiederhold, G. (2000). An information food chain for advanced applications on the www, in J. L. Borbinha and T. Baker (eds), *ECDL*, Vol. 1923 of *Lecture Notes in Computer Science*, Springer, pp. 490–493. 135
- Deep, S. and Koutris, P. (2017). Qirana: A framework for scalable query pricing, *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD ’17, ACM, New York, NY, USA, pp. 699–713. 167
- Demartini, G., Difallah, D. E. and Cudré-Mauroux, P. (2013). Large-scale linked data integration using probabilistic reasoning and crowdsourcing, *The VLDB Journal* **22**(5): 665–687. 168
- Demartini, G., Trushkowsky, B., Kraska, T. and Franklin, M. J. (2013). Crowdq: Crowdsourced query understanding, *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*. 160
- Deng, T., Fan, W. and Geerts, F. (2016). Capturing missing tuples and missing values, *ACM Trans. Database Syst.* **41**(2): 10:1–10:47. 164

BIBLIOGRAPHY

- Dennis, A. R. and Valacich, J. S. (1999). Rethinking media richness: Towards a theory of media synchronicity, *HICSS '99: Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 1*, IEEE Computer Society, Washington, DC, USA, p. 1017. 54
- Desouza, K. C. (2003). Barriers to effective use of knowledge management systems in software engineering, *Commun. ACM* **46**(1): 99–101. 49, 137
- Desouza, K. C. and Evaristo, J. R. (2004). Managing knowledge in distributed projects, *Commun. ACM* **47**(4): 87–91. 49, 137
- Díaz, O. and Puente, G. (2011). Wiki scaffolding: Helping organizations to set up wikis, *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, ACM, New York, NY, USA, pp. 154–162. 130
- Díaz, O. and Puente, G. (2012). Wiki scaffolding: Aligning wikis with the corporate strategy, *Inf. Syst.* **37**(8): 737–752. 130
- Diffin, J., Chirombo, F., Nangle, D. and de Jong, M. (2010). A point to share: Streamlining access services workflow through online collaboration, communication, and storage with microsoft sharepoint, *Journal of Web Librarianship* **4**(2-3): 225–237. 50, 93
- Dix, A., Finlay, J. E., Abowd, G. D. and Beale, R. (2003). *Human-Computer Interaction (3rd Edition)*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 53, 54
- Domingue, J., Fensel, D. and Hendler, J. A. (eds) (2011). *Handbook of Semantic Web Technologies*, 1st edn, Springer. 133, 224, 247
- Drăgan, L., Delbru, R., Groza, T., Handschuh, S. and Decker, S. (2011). Linking semantic desktop data to the web of data, *The Semantic Web – ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part II*, Springer, Berlin, Heidelberg, pp. 33–48. 163
- Drăgan, L., Passant, A., Groza, T. and Handschuh, S. (2010). Publishing semantic personal notes as linked data, *Proceedings of the EKAW2010 Poster and Demo Track, Lisbon, Portugal, October 11 - 15, 2010*. 163
- Drosou, M. and Pitoura, E. (2013). Ymaldb: exploring relational databases via result-driven recommendations, *The VLDB Journal* **22**(6): 849–874. 166
- Drucker, P. (1969). *The age of discontinuity: guidelines to our changing society*, Harper & Row. 24
- Dugan, C., Geyer, W. and Millen, D. R. (2010). Lessons learned from blog muse: Audience-based inspiration for bloggers, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM, New York, NY, USA, pp. 1965–1974. 50, 85
- Duncan, R. and Weiss, A. (1979). Organizational learning - implications for organizational design, in B. Staw (ed.), *Research in Organizational Behaviour*, number 1, JAI Press. 7

- Eberius, J., Damme, P., Braunschweig, K., Thiele, M. and Lehner, W. (2013). Publish-time data integration for open data platforms, *Proceedings of the 2Nd International Workshop on Open Data*, WOD '13, ACM, New York, NY, USA, pp. 1:1–1:6. 161
- Economist Intelligence Unit (2007). Serious business: Web 2.0 goes corporate. **URL:** http://socialmediagroup.ca/wp-content/uploads/2007/06/smg_eiu_web20.pdf 109
- Edvinsson, L. and Malone, M. S. (1997). *Intellectual Capital: Realizing Your Company's True Value by Finding Its Hidden Brainpower*, Collins. 10
- Efron, M. and Winget, M. (2010). Questions are content: A taxonomy of questions in a microblogging environment, *Proceedings of the American Society for Information Science and Technology* **47**(1): 1–10. 83
- Eirinaki, M., Abraham, S., Polyzotis, N. and Shaikh, N. (2014). Querie: Collaborative database exploration, *IEEE Transactions on Knowledge and Data Engineering* **26**(7): 1778–1790. 160
- Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework, *Technical report*, Air Force Office of Scientific Research. 23
- Erdelez, S. (2005). Information encountering, in Fisher et al. (2005), chapter 29, pp. 179–185. 45
- Erickson, T. (2006). From pim to gim: Personal information management in group contexts, *Commun. ACM* **49**(1): 74–75. 23, 85
- Evans, B. M. and Chi, E. H. (2008). Towards a model of understanding social search, *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 485–494. 38, 41, 42, 115
- Even, A. and Shankaranarayanan, G. (2007). Utility-driven assessment of data quality, *SIGMIS Database* **38**(2): 75–93. 167
- Even, A., Shankaranarayanan, G. and Berger, P. D. (2007). Economics-driven data management: An application to the design of tabular data sets, *IEEE Transactions on Knowledge and Data Engineering* **19**(6): 818–831. 167
- Even, A., Shankaranarayanan, G. and Berger, P. D. (2010). Inequality in the utility of customer data: Implications for data management and usage, *Journal of Database Marketing & Customer Strategy Management* **17**(1): 19–35. 167
- Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A. and Williamson, D. P. (2003). Searching the workplace web, *WWW '03: Proceedings of the 12th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 366–375. 41
- Fan, J., Zhang, M., Kok, S., Lu, M. and Ooi, B. C. (2015). Crowdop: Query optimization for declarative crowdsourcing systems, *IEEE Transactions on Knowledge and Data Engineering* **27**(8): 2078–2092. 168
- Fan, W. (2015). Data quality: From theory to practice, *SIGMOD Rec.* **44**(3): 7–18. 164

BIBLIOGRAPHY

- Färber, M., Bartscherer, F., Menne, C. and Rettinger, A. (2017). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO, *Semantic Web Journal* pp. 1–53. 162
- Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S., Grudin, J. and Poltrock, S. (2000). Collaborative information retrieval (cir), *New Rev. Inf. Behav. Res.* **1**(January): 235–247. 42
- Fieldhouse, M. and Marshall, A. (2011). *Collection Development in the Digital Age*, Facet Publishing. 83
- Filipiak, D. and , A. (2014). Generating semantic media wiki content from domain ontologies, *Proceedings of the Third International Conference on Semantic Web Collaborative Spaces - Volume 1275*, SWCS'14, CEUR-WS.org, Aachen, Germany, pp. 68–76. 162
- Fisher, K., Erdelez, S. and McKechnie, L. (eds) (2005). *Theories of Information Behavior*, ASIST monograph series, American Society for Information Science and Technology. 34, 218, 223, 243
- Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S. and Xin, R. (2011). Crowddb: Answering queries with crowdsourcing, *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, ACM, New York, NY, USA, pp. 61–72. 167
- Fürber, C. (2015). *Data Quality Management with Semantic Technologies*, Springer. 163
- Fürber, C. and Hepp, M. (2010). Using semantic web resources for data quality management, *Knowledge Engineering and Management by the Masses: 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11-15, 2010. Proceedings*, Springer, Berlin, Heidelberg, pp. 211–225. 162
- Galárraga, L., Hose, K. and Razniewski, S. (2017). Enabling completeness-aware querying in sparql, *Proceedings of the 20th International Workshop on the Web and Databases*, WebDB'17, ACM, New York, NY, USA, pp. 19–22. 164
- Galárraga, L., Razniewski, S., Amarilli, A. and Suchanek, F. M. (2017). Predicting completeness in knowledge bases, *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, ACM, New York, NY, USA, pp. 375–383. 165, 169
- Gandon, F. L., Krummenacher, R., Han, S.-K. and Toma, I. (2011). *Semantic Annotation and Retrieval: RDF*, in Domingue et al. (2011), 1st edn, pp. 117–155. 135
- Gao, Y., Liu, Q., Chen, G., Zheng, B. and Zhou, L. (2015). Answering why-not questions on reverse top-k queries, *Proc. VLDB Endow.* **8**(7): 738–749. 166
- Geyer, W. and Dugan, C. (2010). Inspired by the audience: A topic suggestion system for blog writers and readers, *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, ACM, New York, NY, USA, pp. 237–240. 85

- Geyer, W., Dugan, C., Millen, D. R., Muller, M. and Freyne, J. (2008). Recommending topics for self-descriptions in online user profiles, *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, ACM, New York, NY, USA, pp. 59–66. 85
- Gil, Y., Kapoor, D., Markham, R. and Ratnakar, V. (2015). The provenance bee wiki: Tracking the growth of semantic wiki communities, *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, ACM, New York, NY, USA, pp. 10:1–10:8. 162
- Gil, Y., Knight, A., Zhang, K., Zhang, L., Ratnakar, V. and Sethi, R. (2013). The democratization of semantic properties: An analysis of semantic wikis, *2013 IEEE Seventh International Conference on Semantic Computing*, pp. 206–211. 162
- Gil, Y. and Ratnakar, V. (2013). Knowledge capture in the wild: A perspective from semantic wiki communities, *Proceedings of the Seventh International Conference on Knowledge Capture*, K-CAP '13, ACM, New York, NY, USA, pp. 49–56. 162
- Golovchinsky, G., Qvarfordt, P. and Pickens, J. (2009). Collaborative information seeking, *Computer* **42**: 47–51. 38
- Google Inc. (2017a). Google alerts.
URL: <http://www.google.com/alerts> 40, 91, 106
- Google Inc. (2017b). Google zeitgeist.
URL: <http://www.google.com/press/zeitgeist.html> 68
- Grahne, G. (2009). *Incomplete Information*, Springer US, Boston, MA, pp. 1405–1410. 164
- Grau, B. C., Giese, M., Horrocks, I., Hubauer, T., Jiménez-Ruiz, E., Kharlamov, E., Schmidt, M., Soylu, A. and Zheleznyakov, D. (2013). Towards Query Formulation and Query-Driven Ontology Extensions in OBDA Systems, *OWL Experiences and Directions Workshop (OWLED)*. 161, 169
- Grinter, R. E., Herbsleb, J. D. and Perry, D. E. (1999). The geography of coordination: Dealing with distance in R&D work, *GROUP'99: International Conference on Supporting Group Work*, pp. 306–315. 25
- Groth, K. and Eklundh, K. S. (2006). Combining personal and organizational information, *Proceedings of the SIGIR Workshop: Personal Information Management*, ACM Press, New York, NY, USA. 86
- Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G. and Gudjónsdóttir, R. (2007). The nepomuk project - on the way to the social semantic desktop, in T. Pellegrini and S. Schaffert (eds), *Proceedings of I-Semantics' 07*, JUCS, pp. 201–211. 163
- Grudin, J. (1994). Groupware and social dynamics: Eight challenges for developers, *Communications of the ACM* **37**(1): 93–105. 25, 136
- Guo, J., Xu, S., Bao, S. and Yu, Y. (2008). Tapping on the potential of q&a community by recommending answer providers, *Proceedings of the 17th*

BIBLIOGRAPHY

- ACM Conference on Information and Knowledge Management, CIKM '08*, ACM, New York, NY, USA, pp. 921–930. 130
- Gupta, R., Halevy, A., Wang, X., Whang, S. E. and Wu, F. (2014). Biperpedia: An ontology for search applications, *Proc. VLDB Endow.* **7**(7): 505–516. 160, 161
- Gust von Loh, S. (2008). Wissensmanagement und informationsbedarfsanalyse in kleinen und mittleren unternehmen. teil 1: Grundlagen des wissensmanagement., *Information - Wissenschaft und Praxis* **59**. 82
- Gwizdka, J. (2006). Finding to keep and organize: Personal information collections as context, *Proceedings of the SIGIR Workshop: Personal Information Management*, ACM Press, New York, NY, USA. 23, 85
- Halavais, A. (2017). *Search Engine Society*, Wiley. 1
- Halevy, A., Noy, N., Sarawagi, S., Whang, S. E. and Yu, X. (2016). Discovering structure in the universe of attribute names, *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 939–949. 161
- Halfaker, A., Keyes, O. and Taraborelli, D. (2013). Making peripheral participation legitimate: Reader engagement experiments in wikipedia, *Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13*, ACM, New York, NY, USA, pp. 849–860. 87
- Halpin, H. (2009). A query-driven characterization of linked data, *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009*. 159
- Handschuh, S. (2005). *Creating ontology-based metadata by annotation for the semantic web*, PhD thesis, Universität Karlsruhe. 133, 136, 137
- Hanrahan, B., Bouchard, G., Convertino, G., Weksteen, T., Kong, N., Archambeau, C. and Chi, E. H. (2011). Mail2wiki: Low-cost sharing and early curation from email to wikis, *Proceedings of the 5th International Conference on Communities and Technologies, C&T '11*, ACM, New York, NY, USA, pp. 98–107. 85, 130, 174
- Hansen, D. L., Ackerman, M. S., Resnick, P. J. and Munson, S. (2007). Virtual community maintenance with a collaborative repository, *Proceedings of the American Society for Information Science and Technology* **44**(1): 1–20. 88
- Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits, *Administrative Science Quarterly* **44**: 82–111. 55
- Hansen, M. T., Nohria, N. and Tierney, T. (1999). What's your strategy for managing knowledge, *Harvard Business Review* **77**(2): 106 – 116. 7, 11, 12, 21, 24, 29, 207
- Happel, H.-J. (2008a). Closing information gaps with inverse search, in T. Yamaguchi (ed.), *Proceedings of PAKM 2008: 7th International Conference on*

- Practical Aspects of Knowledge Management*, Lecture Notes in Computer Science, Springer, pp. 74–85. 92
- Happel, H.-J. (2008b). Growing the semantic web with inverse semantic search, in K. Siorpaes, E. Simperl and D. Vrandečić (eds), *1st Workshop on Incentives for the Semantic Web (INSEMTIVE) at ISWC 2008*, pp. 1–12. 134, 150, 163, 169
- Happel, H.-J. (2009a). Social search and need-driven knowledge sharing in wikis with woogle, in D. Riehle and A. Bruckman (eds), *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, ACM, pp. 1–10. 110
- Happel, H.-J. (2009b). Towards need-driven knowledge sharing in distributed teams, in K. Tochtermann and H. Maurer (eds), *Proceedings of I-KNOW '09: 9th international conference on knowledge management and knowledge technologies*, JUCS, pp. 128–139. 59
- Happel, H.-J. (2009c). Woogle - on why and how to marry wikis with enterprise search, in K. Hinkelmann and H. Wache (eds), *WM2009: 5th Conference on Professional Knowledge Management*, pp. 194–205. 110, 116, 117
- Happel, H.-J. (2010). Semantic need: Guiding metadata annotations by questions people #ask, *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, Springer, Berlin, Heidelberg, pp. 321–336. 134
- Happel, H.-J. (2011). Semantic need: An approach for guiding users contributing metadata to the semantic web, *Int. J. Knowledge Engineering and Data Mining* 1(4): 350–369. 134, 140
- Happel, H.-J. and Maalej, W. (2008). Potentials and challenges of recommendation systems for software development, *Proceedings of the 2008 International Workshop on Recommendation Systems for Software Engineering, RSSE '08*, ACM, New York, NY, USA, pp. 11–15. 81
- Happel, H.-J., Maalej, W. and Stojanovic, L. (2008). Team: towards a software engineering semantic web, *Proceedings of the 2008 international workshop on Cooperative and human aspects of software engineering, CHASE '08*, ACM, New York, NY, USA, pp. 57–60. 81
- Happel, H.-J. and Mazarakis, A. (2010). Considering information providers in social search, *2nd International Workshop on Collaborative Information Seeking (CIS '10) at CSCW 2010*, pp. 1–5. 110, 141, 173
- Happel, H.-J. and Romberg, T. (2008). Wikis - die wissensmanagementlösung für agile unternehmen?, in K. Haasis and N. Zaboura (eds), *A digital lifestyle: Leben und Arbeiten mit Social Software*, MFG Innovationsagentur, Stuttgart, Germany, pp. 19–30. 109
- Happel, H.-J. and Schmidt, A. (2007). Knowledge maturing as a process model for describing software reuse, *4th Conference Professional Knowledge Management - Experiences and Visions (WM 2007), Potsdam. Volume*, pp. 155–164. 19

BIBLIOGRAPHY

- Happel, H.-J. and Steinbauer, I. (2008). Learning how developers search for source code with implicit relevance feedback, *in* R. L. Feldmann and M. Wessner (eds), *Proceedings of the 10th International Workshop on Learning Software Organizations (LSO 2008)*, pp. 0–0. 81, 173
- Happel, H.-J. and Stojanovic, L. (2008). Analyzing organizational information gaps, *in* K. Tochtermann, H. Maurer, F. Kappe and W. Haas (eds), *Proceedings of I-KNOW '08: 8th international conference on knowledge management and knowledge technologies*, JUCS, pp. 28–36. 59, 91, 138, 206
- Happel, H.-J., Stojanovic, L. and Stojanovic, N. (2007). Fostering knowledge sharing by inverse search, *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, ACM, pp. 181–182. 91
- Happel, H.-J. and Treitz, M. (2008). Proliferation in enterprise wikis, *in* P. Hassanaly, A. Ramrajsingh, D. Randall, P. Salembier and M. Tixier (eds), *Proceedings of 8th International Conference on the Design of Cooperative Systems (COOP'08)*, pp. 123–129. 29, 51, 131, 132
- Harper, F. M., Frankowski, D., Drenner, S., Ren, Y., Kiesler, S., Terveen, L., Kraut, R. and Riedl, J. (2007). Talk amongst yourselves: Inviting users to participate in online conversations, *Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI '07*, ACM, New York, NY, USA, pp. 62–71. 84
- Hasan, M. A., Parikh, N., Singh, G. and Sundaresan, N. (2011). Query suggestion for e-commerce sites, *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, ACM, New York, NY, USA, pp. 765–774. 174
- Haug, A., Zachariassen, F. and van Liempd, D. (2011). The costs of poor data quality, *Journal of Industrial Engineering and Management* **4**(2): 168–193. 164
- He, B., Patel, M., Zhang, Z. and Chang, K. C.-C. (2007). Accessing the deep web, *Commun. ACM* **50**(5): 94–101. 173
- Hearst, M. A. (2009). *Search User Interfaces*, Cambridge University Press. 38, 91, 128, 129
- Heinrich, E. and Maurer, H. A. (2000). Active documents: Concept, implementation and applications, *Journal of Universal Computer Science* **6**(12): 1197–1202. 88
- Heisig, P., Caldwell, N. H., Grebici, K. and Clarkson, P. J. (2010). Exploring knowledge and information needs in engineering from the past and for the future - results from a survey, *Design Studies* **31**(5): 499 – 532. 24, 62, 87
- Henderson, R. M. and Clark, K. B. (1990). Architectural innovation: the reconfiguration of existing product technologies and the failure of established firms, *Administrative Science Quarterly* **35**: 9–30. 11, 19, 20, 27, 29
- Herbsleb, J. D. (2007). Global software engineering: The future of socio-technical coordination, *2007 Future of Software Engineering, FOSE '07*, IEEE Computer Society, Washington, DC, USA, pp. 188–198. 47

- Herbsleb, J. D. and Mockus, A. (2003). An empirical study of speed and communication in globally-distributed software development, *IEEE Transactions on Software Engineering* **29**(6): 481–494. 25, 26
- Hertzum, M. (1999). Six roles of documents in professionals' work, *Proceedings of the Sixth Conference on European Conference on Computer Supported Cooperative Work, ECSCW'99*, Kluwer Academic Publishers, Norwell, MA, USA, pp. 41–60. 1
- Hicks, B. J., Dong, A., Palmer, R. and Mcalpine, H. C. (2008). Organizing and managing personal electronic files: A mechanical engineer's perspective, *ACM Trans. Inf. Syst.* **26**: 23:1–23:40. 1, 94, 103
- Hill, B. M. and Shaw, A. (2014). Consider the redirect: A missing dimension of wikipedia research, *Proceedings of The International Symposium on Open Collaboration, OpenSym '14*, ACM, New York, NY, USA, pp. 28:1–28:4. 131
- Hillis, K., Petit, M. and Jarrett, K. (2013). *Google and the Culture of Search*, Routledge. 1
- Hinds, P. J. and Kiesler, S. (2002a). *Managing Distance over Time: The Evolution of Technologies of Dis/Ambiguation*, in Hinds and Kiesler (2002b), chapter 4, pp. 83–111. 55
- Hinds, P. J. and Kiesler, S. (eds) (2002b). *Distributed Work*, MIT Press, Cambridge, MA, USA. 215, 229, 231, 237, 239, 248
- Hinds, P. and McGrath, C. (2006). Structures that work: Social structure, work structure and coordination ease in geographically distributed teams, *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW '06*, ACM, New York, NY, USA, pp. 343–352. 25
- Hitzler, P., Krötzsch, M. and Rudolph, S. (2009). *Foundations of Semantic Web Technologies*, Chapman & Hall/CRC. 141
- Hoffmann, R., Amershi, S., Patel, K., Wu, F., Fogarty, J. and Weld, D. S. (2009). Amplifying community content creation with mixed initiative information extraction, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, ACM, New York, NY, USA, pp. 1849–1858. 161, 162
- Hollingshead, A. B., Fulk, J. and Monge, P. (2002). *Fostering Intranet Knowledge Sharing: An Integration of Transactive Memory and Public Goods Approaches*, in Hinds and Kiesler (2002b), chapter 14, pp. 335–356. 16, 43
- Huang, J., Chen, T., Doan, A. and Naughton, J. F. (2008). On the provenance of non-answers to queries over extracted data, *Proc. VLDB Endow.* **1**(1): 736–747. 166
- Hurtado, C. A., Poulouvasilis, A. and Wood, P. T. (2006). A relaxed approach to rdf querying, *The Semantic Web - ISWC 2006: 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006. Proceedings*, Springer, Berlin, Heidelberg, pp. 314–328. 166

BIBLIOGRAPHY

- Huysman, M. and de Wit, D. (2003). *A critical evaluation of knowledge management practices*, in Ackerman, Pipek and Wulf (2003), pp. 27–55. 21
- Hyldegård, J. (2006). Using diaries in group based information behavior research: A methodological study, *Proceedings of the 1st International Conference on Information Interaction in Context*, IiiX, ACM, New York, NY, USA, pp. 153–161. 24
- Imieliński, T. and Lipski, Jr., W. (1984). Incomplete information in relational databases, *J. ACM* **31**(4): 761–791. 164
- Ingwersen, P. and Järvelin, K. (2011). *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer. 33
- Ioannou, E., Coi, J. L. D., Koesling, A. W., Olmedilla, D. and Nejdil, W. (2007). Access control for sharing semantic data across desktops, *Proceedings of the ISWC'07 Workshop on Privacy Enforcement and Accountability with Semantics (PEAS 2007)*, Busan, Korea, Nov. 12, 2007. 163
- Ipe, M. (2003). Knowledge sharing in organizations: A conceptual framework, *Human Resource Development Review* **2**(4): 337–359. 46, 49
- Irmak, U., Mihaylov, S., Suel, T., Ganguly, S. and Izmailov, R. (2006). Efficient query subscription processing for prospective search engines, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 1037–1038. 40, 91, 106
- Iverson, L., Razavi, M. N. and Mirzaee, V. (2008). Personal and social information management with OPNTAG, in J. Cordeiro and J. Filipe (eds), *ICEIS 2008 - Proceedings of the Tenth International Conference on Enterprise Information Systems, Volume HCI, Barcelona, Spain, June 12-16, 2008*, pp. 195–203. 86
- Jain, A. and Pennacchiotti, M. (2011). Domain-independent entity extraction from web search query logs, *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, ACM, New York, NY, USA, pp. 63–64. 161
- Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it, *Library Information Science Research* **28**(3): 407 – 432. 159
- Jansen, B. J. (2007). Preserving the Collective Expressions of the Human Consciousness, in E. Amitay, C. G. Murray and J. Teevan (eds), *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*. 175
- Jennex, M. E., Smolnik, S. and Croasdell, D. (2014). Knowledge management success in practice, *2014 47th Hawaii International Conference on System Sciences*, pp. 3615–3624. 31
- Johnson, B. C., Manyika, J. M. and Yee, L. A. (2005). The next revolution in interactions, *McKinsey Quarterly* (4): 20–33. 24
- Jones, W. (2004). Finders, keepers? the present and future perfect in support of personal information management, *First Monday* **9**(3). 23

- Jones, W. and Teevan, J. (2007). *Personal Information Management*, University of Washington Press. 23, 234
- Jung, J. H., Schneider, C. and Valacich, J. (2010). Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments, *Manage. Sci.* **56**(4): 724–742. 71
- Kamei, K., Yoshida, S., Kuwabara, K., Akahani, J.-i. and Satoh, T. (2003). An agent framework for inter-personal information sharing with an rdf-based repository, in D. Fensel, K. Sycara and J. Mylopoulos (eds), *The Semantic Web - ISWC 2003*, Vol. 2870 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 438–452. 20
- Kankanhalli, A., Tan, B. C. Y. and Wei, K.-K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation, *MIS Quarterly* **29**(1): 113–143. 2, 44
- Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography, *SIGIR Forum* **37**(2): 18–28. 38
- Khoussainova, N., Balazinska, M., Gatterbauer, W., Kwon, Y. and Suciu, D. (2009). A case for A collaborative query management system, *CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*. 160
- Khoussainova, N., Kwon, Y., Balazinska, M. and Suciu, D. (2010). Snipsuggest: Context-aware autocompletion for sql, *Proc. VLDB Endow.* **4**(1): 22–33. 160
- Kind, J. L. and Frost, R. L. (2002). *Managing Distance over Time: The Evolution of Technologies of Dis/Ambiguation*, in Hinds and Kiesler (2002b), chapter 1, pp. 3–26. 25
- King, W. R., Peter V. Marks, J. and McCoy, S. (2002). The most important issues in knowledge management, *Commun. ACM* **45**(9): 93–97. 1
- Kittur, A., Pendleton, B. and Kraut, R. E. (2009). Herding the cats: The influence of groups in coordinating peer production, *Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WikiSym '09*, ACM, New York, NY, USA, pp. 7:1–7:9. 131
- Ko, A. J., DeLine, R. and Venolia, G. (2007). Information needs in collocated software development teams, *Proceedings of the 29th International Conference on Software Engineering, ICSE '07*, IEEE Computer Society, Washington, DC, USA, pp. 344–353. 47
- Kobayashi, M. and Takeda, K. (2000). Information retrieval on the web, *ACM Comput. Surv.* **32**(2): 144–173. 42
- Koch, M. and Richter, A. (2009). *Enterprise 2.0: Planung, Einführung und erfolgreicher Einsatz von Social Software in Unternehmen*, Oldenbourg. 28
- Kohler, S. (2013). *Atlassian Confluence 5 Essentials*, Packt Publishing, Limited. 114, 122

BIBLIOGRAPHY

- Kondreddi, S. K., Triantafillou, P. and Weikum, G. (2014). Combining information extraction and human computing for crowdsourced knowledge acquisition, *2014 IEEE 30th International Conference on Data Engineering*, pp. 988–999. 161
- Kong, N., Hanrahan, B., Weksteen, T., Convertino, G. and Chi, E. H. (2011). Visualwikicurator: Human and machine intelligence for organizing wiki content, *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, ACM, New York, NY, USA, pp. 367–370. 130
- Konstan, J. A. and Chen, Y. (2007). Online field experiments: Lessons from communitylab, *Proceedings of e-Social Science 2007*. 125
- Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R. and Zaveri, A. (2014). Test-driven evaluation of linked data quality, *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, ACM, New York, NY, USA, pp. 747–758. 163
- Koreimann, D. (1976). *Methoden der Informationsbedarfsanalyse*, de Gruyter. 82
- Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B. and Suciu, D. (2013). Toward practical query pricing with querymarket, *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, ACM, New York, NY, USA, pp. 613–624. 167
- Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B. and Suciu, D. (2015). Query-based data pricing, *J. ACM* **62**(5): 43:1–43:44. 167
- Kramer, K., Dividino, R. and Gröner, G. (2013). Space: Sparql index for efficient autocompletion, *Proceedings of the 12th International Semantic Web Conference (Posters & Demonstrations Track) - Volume 1035*, ISWC-PD '13, CEUR-WS.org, Aachen, Germany, pp. 157–160. 160
- Kraut, R. E., Morris, J., Telang, R., Filer, D., Cronin, M. and Sunder, S. (2002). Markets for attention: Will postage for email help?, *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, ACM, New York, NY, USA, pp. 206–215. 174
- Krcmar, H. (2004). *Informationsmanagement*, Springer, Berlin; Heidelberg. 7, 35, 82
- Krieger, M., Stark, E. M. and Klemmer, S. R. (2009). Coordinating tasks on the commons: Designing for personal goals, expertise and serendipity, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM, New York, NY, USA, pp. 1485–1494. 130
- Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H. and Studer, R. (2007). Semantic Wikipedia, *Journal of Web Semantics*, *5*, pp. 251–261 . 150
- Kukulenz, D. and Ntoulas, A. (2007). Answering bounded continuous search queries in the world wide web, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 551–560. 40, 106

- Kustanowitz, J. and Shneiderman, B. (2005). Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives, *Technical Report HCIL-2004-18*, University of Maryland, College Park, MD, USA. 71, 136, 137
- König, R. and Rasch, M. (eds) (2014). *Society of the Query Reader: Reflections on Web Search*, Institute of Network Cultures, Amsterdam. 1
- Lam, S. T. K. and Churchill, E. (2007). The social web: global village or private cliques?, *Proceedings of the 2007 conference on Designing for User eXperiences*, DUX '07, ACM, New York, NY, USA, pp. 16:1–16:7. 94, 137
- Lau, T., Etzioni, O. and Weld, D. S. (1999). Privacy interfaces for information management, *Commun. ACM* **42**: 88–94. 94, 107
- Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press. 16, 86
- Lee, J. (2003). An end-user perspective on file-sharing systems, *Commun. ACM* **46**(2): 49–53. 50
- Lehner, F. (2000). *Organisational Memory - Konzepte und Systeme für das organisatorische Lernen und das Wissensmanagement*, C. Hanser, München. 11, 15, 16, 17, 49
- Lehner, F. (2003). Information sharing und wissensaustausch in unternehmen., in A. Geyer-Schulz and A. Taudes (eds), *Informationswirtschaft: Ein Sektor mit Zukunft*, Vol. 33 of *LNI*, GI, pp. 301–319. 44
- Leshed, G., Haber, E. M., Matthews, T. and Lau, T. (2008). Coscripiter: Automating & sharing how-to knowledge in the enterprise, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, ACM, New York, NY, USA, pp. 1719–1728. 73, 173
- Levy, A. Y. (1996). Obtaining complete answers from incomplete databases, *Proceedings of the 22th International Conference on Very Large Data Bases*, VLDB '96, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 402–412. 164
- Lewoniewski, W. (2017). Completeness and reliability of wikipedia infoboxes in various languages, *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28-30, 2017, Revised Papers*, Springer, pp. 295–305. 162
- Li, F., Dong, X. L., Langen, A. and Li, Y. (2017). Knowledge verification for long-tail verticals, *Proc. VLDB Endow.* **10**(11): 1370–1381. 161
- Lin, H., Davis, J. and Zhou, Y. (2010). Ontological services using crowdsourcing. 168
- Linton, F. (2003). *OWL: A system for the automated sharing of expertise. Sharing expertise*, in Ackerman, Pipek and Wulf (2003), pp. 383–401. 173
- Liu, Q., Agichtein, E., Dror, G., Maarek, Y. and Szpektor, I. (2012). When web search fails, searchers become askers: Understanding the transition, *Proceedings of the 35th International ACM SIGIR Conference on Research*

BIBLIOGRAPHY

- and Development in Information Retrieval*, SIGIR '12, ACM, New York, NY, USA, pp. 801–810. 87, 129
- Liu, Q., Liu, Y. and Agichtein, E. (2010). Exploring web browsing context for collaborative question answering, *Proceedings of the Third Symposium on Information Interaction in Context*, IiX '10, ACM, New York, NY, USA, pp. 305–310. 130
- Liu, Y., Narasimhan, N., Vasudevan, V. and Agichtein, E. (2009). Is this urgent?: Exploring time-sensitive information needs in collaborative question answering, *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, ACM, New York, NY, USA, pp. 712–713. 173
- Luczak-Rösch, M. (2014). *Usage-dependent maintenance of structured Web data sets*, PhD thesis, Freie Universität Berlin. 141, 167
- Ludford, P. J., Cosley, D., Frankowski, D. and Terveen, L. (2004). Think different: Increasing online community participation using uniqueness and group dissimilarity, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, ACM, New York, NY, USA, pp. 631–638. 84
- Lüer, C. and Cummins, J. (2009). Collaborative web search with wikilinks, *2009 Sixth International Conference on Information Technology: New Generations*, pp. 1691–1692. 128
- Lutters, W. G., Ackerman, M. S. and Zhou, X. (2007). Group information management, in *Personal Information Management: Challenges and Opportunities*. Jones and Teevan (2007), chapter 7, pp. 236–248. 23, 85
- Maalej, W. (2009). Task-first or context-first? tool integration revisited, *Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering*, ASE '09, IEEE Computer Society, Washington, DC, USA, pp. 344–355. 81
- Maalej, W. (2010). *Intention-Based Integration of Software Engineering Tools*, 1 edn, Verlag Dr. Hut, München, Germany. 81, 173
- Maalej, W. and Happel, H.-J. (2009). From work to word: How do software developers describe their work?, *Proceedings of the 6th International Working Conference on Mining Software Repositories, MSR 2009 (Co-located with ICSE), Vancouver, BC, Canada, May 16-17, 2009, Proceedings*, pp. 121–130. 24
- Machlup, F. and Mansfield, U. (1983). *The Study of Information : Interdisciplinary Messages*, Wiley, New York. 12, 13
- Mahmud, J., Matthews, T., Whittaker, S., Moran, T. and Lau, T. (2011). Topika: Integrating collaborative sharing with email, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM, New York, NY, USA, pp. 3161–3164. 85, 174
- Maier, R. (2007). *Knowledge Management Systems*, Springer. 1, 7, 8, 12, 13, 15, 16, 17, 19, 20, 22, 24, 28, 29, 30, 49, 205, 207

- Maier, R. and Sametinger, J. (2003). Infotop - A shared-context information workspace, *Proceedings of the Fifteenth International Conference on Software Engineering & Knowledge Engineering (SEKE'2003), San Francisco Bay, CA, USA, July 1-3, 2003*, pp. 534–541. 86
- Maier, R. and Schmidt, A. (2007). Characterizing knowledge maturing: A conceptual process model for integrating e-learning and knowledge management, *4th Conference Professional Knowledge Management - Experiences and Visions (WM 2007), Potsdam. Volume*, pp. 325–334. 17, 18, 19, 55, 205
- Majchrzak, A., Rice, R. E., Malhotra, A., King, N. and Ba, S. (2000). Technology adaptation: The case of a computer-supported inter-organizational virtual team, *MIS Quarterly* **24**(4): 569–600. 11
- Majchrzak, A., Wagner, C., Riehle, D., Thoeny, P., Shah, S. and Cunningham, W. (2007). The role of shapers in knowledge-sharing, *Virtuality and Virtualization: Proceedings of the International Federation of Information Processing Working Groups 8.2 on Information Systems and Organizations and 9.5 on Virtuality and Society, July 29–31, 2007, Portland, Oregon, USA*, Springer US, Boston, MA, pp. 383–386. 131
- Majchrzak, A., Wagner, C. and Yates, D. (2006). Corporate wiki users: results of a survey, *Proceedings of the 2nd International Symposium on Wikis, WikiSym '06, ACM, New York, NY, USA*, pp. 99–104. 109
- Majchrzak, A., Wagner, C. and Yates, D. (2013). The impact of shaping on knowledge reuse for organizational improvement with wikis, *MIS Q.* **37**(2): 455–470. 131
- Malone, T. W. and Crowston, K. (1994). The interdisciplinary study of coordination, *ACM Comput. Surv.* **26**(1): 87–119. 26
- Malone, T. W., Grant, K. R. and Turbak, F. A. (1986). The information lens: An intelligent system for information sharing in organizations, *SIGCHI Bull.* **17**(4): 1–8. 174
- Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. and Hartmann, B. (2011). Design lessons from the fastest q&a site in the west, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM, New York, NY, USA*, pp. 2857–2866. 129
- Marcus, A., Wu, E., Madden, S. and Miller, R. C. (2011). Crowdsourced databases: Query processing with people, *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*, pp. 211–214. 167
- Markus, M. L. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success, *J. Manage. Inf. Syst.* **18**(1): 57–93. 23, 46, 49, 86
- Markus, M. L. and Robey, D. (1988). Information technology and organizational change: Causal structure in theory and research, *Management Science* **34**(5): 583–598. 11

BIBLIOGRAPHY

- Marlow, C., Naaman, M., Boyd, D. and Davis, M. (2006). Ht06, tagging paper, taxonomy, flickr, academic article, to read, *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, ACM, New York, NY, USA, pp. 31–40. 137
- Marshall, C. C. and Brush, A. J. B. (2004). Exploring the relationship between personal and public annotations, *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '04, ACM, New York, NY, USA, pp. 349–357. 86
- Marshall, C. C. and Jones, W. (2006). Keeping encountered information, *Commun. ACM* **49**(1): 66–67. 23
- Massey, C., Lennig, T. and Whittaker, S. (2014). Cloudy forecast: An exploration of the factors underlying shared repository use, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM, New York, NY, USA, pp. 2461–2470. 106
- Maurer, H. and Tochtermann, K. (2002). On a new powerful model for knowledge management and its applications, *Journal of Universal Computer Science* **8**(1): 85–96. 74
- Mazarakis, A. and van Dinther, C. (2011). Feedback mechanisms and their impact on motivation to contribute to wikis in higher education, *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, ACM, pp. 215–216. 71
- Mazurek, M. L., Klemperer, P. F., Shay, R., Takabi, H., Bauer, L. and Cranor, L. F. (2011). Exploring reactive access control, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM, New York, NY, USA, pp. 2085–2094. 107
- McAfee, A. (2009). *Enterprise 2.0: How to Manage Social Technologies to Transform Your Organization*, Harvard Business Review Press. 28
- McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration, *MIT Sloan Management Review* **47**(3): 21–28. 11, 20, 28
- McCandless, M., Hatcher, E. and Gospodnetic, O. (2010). *Lucene in Action, Second Edition*, Manning Publications Co., Greenwich, CT, USA. 102
- McMahon, C., Johnson, I. and Hecht, B. (2017). The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies, *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, AAAI Press, pp. 142–151. 129
- Mika, P., Meij, E. and Zaragoza, H. (2009). Investigating the semantic gap through query log analysis, *The Semantic Web - ISWC 2009: 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, Springer, Berlin, Heidelberg, pp. 441–455. 4, 142, 153, 159
- Miklau, G. and Suci, D. (2004). A formal analysis of information disclosure in data exchange, *Proceedings of the 2004 ACM SIGMOD International*

- Conference on Management of Data*, SIGMOD '04, ACM, New York, NY, USA, pp. 575–586. 163
- Millen, D. R., Feinberg, J. and Kerr, B. (2006). Dogear: Social bookmarking in the enterprise, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, ACM, New York, NY, USA, pp. 111–120. 50, 108
- Mirza, P., Razniewski, S., Darari, F. and Weikum, G. (2017). Cardinal virtues: Extracting relation cardinalities from text, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pp. 347–351. 165
- Moon, J. Y. and Sproull, L. (2002). *Essense of Distributed Work: The Case of the Linux Kernel*, in Hinds and Kiesler (2002b), chapter 16, pp. 381–404. 25
- Mooradian, T., Renzl, B. and Matzler, K. (2006). Who trusts? personality, trust and knowledge sharing, *Management Learning* **37**(4): 523–540. 44
- Morris, M. R. (2008). A survey of collaborative web search practices, *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, pp. 1657–1660. 41, 42
- Morris, M. R. and Horvitz, E. (2007). S3: Storable, shareable search, *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction*, INTERACT'07, Springer, Berlin, Heidelberg, pp. 120–123. 37, 128
- Morville, P. and Rosenfeld, L. (2006). *Information Architecture for the World Wide Web*, O'Reilly Media, Inc. 82
- Motik, B., Sattler, U. and Studer, R. (2004). Query answering for owl-dl with rules, *The Semantic Web – ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004. Proceedings*, Springer, Berlin, Heidelberg, pp. 549–563. 141
- Motro, A. (1986). Completeness information and its application to query processing, *Proceedings of the 12th International Conference on Very Large Data Bases*, VLDB '86, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 170–178. 39
- Motro, A. (1989). Integrity = validity + completeness, *ACM Trans. Database Syst.* **14**(4): 480–502. 164
- Mujan, D. (2006). *Informationsmanagement in Lernenden Organisationen: Erzeugung von Informationsbedarf durch Informationsangebot - Was Organisationen aus der Informationsbedarfsanalyse lernen können*, Logos. 82
- Müller, H. and Freytag, J.-C. (2003). Problems, Methods and Challenges in Comprehensive Data Cleansing, *Technical Report HUB-IB-164*, Humboldt-Universität zu Berlin, Institut für Informatik. 164

BIBLIOGRAPHY

- Muller, M. J., Freyne, J., Dugan, C., Millen, D. R. and Thom-Santelli, J. (2009). Return on contribution (roc): A metric for enterprise social software, *ECSCW 2009*, Springer London, London, pp. 143–150. 88
- Muller, M. J., Millen, D. R. and Feinberg, J. (2009). Information curators in an enterprise file-sharing service, *ECSCW 2009*, Springer, pp. 403–410. 88, 106
- Muller, M., Millen, D. R. and Feinberg, J. (2010). Patterns of usage in an enterprise file-sharing service: Publicizing, discovering, and telling the news, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM, New York, NY, USA, pp. 763–766. 106
- Muller, M., Shami, N. S., Millen, D. R. and Feinberg, J. (2010). We are all lurkers: Consuming behaviors among authors and readers in an enterprise file-sharing service, *Proceedings of the 16th ACM International Conference on Supporting Group Work*, GROUP '10, ACM, New York, NY, USA, pp. 201–210. 57, 106
- Müller, R. M., Spiliopoulou, M. and Lenz, H. J. (2005). The influence of incentives and culture on knowledge sharing, *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 247b–247b. 49
- Musser, j. (2007). *Web 2.0. Principles and Best Practices*, O'Reilly Radar. 27
- Möller, K., Hausenblas, M., Cyganiak, R. and Grimnes, G. A. (2010). Learning from linked open data usage: Patterns & metrics, *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. 159
- Nam, K. K. and Ackerman, M. S. (2007). Arkose: Reusing informal information from online discussions, *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, ACM, New York, NY, USA, pp. 137–146. 88
- Nawaz Chowdhury, M. N. and Bunt, A. (2014). Intelwiki: Recommending resources to help users contribute to wikipedia, *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings*, Springer, pp. 393–404. 130
- Ng'ambi, D. (2002). Dynamic "intelligent handler" of frequently asked questions, *Proceedings of the 7th International Conference on Intelligent User Interfaces*, IUI '02, ACM, New York, NY, USA, pp. 210–211. 88
- Ng'ambi, D. and Hardman, J. (2004). Towards a knowledge-sharing scaffolding environment based on learners' questions, *British Journal of Educational Technology* **35**(2): 187–196. 88
- Nicholas, D. (2003). *Assessing Information Needs: Tools, Techniques and Concepts for the Internet Age*, Taylor & Francis. 36
- Nikolaou, C. and Koubarakis, M. (2016). Querying incomplete information in rdf with sparql, *Artif. Intell.* **237**(C): 138–171. 165
- Nonaka, I. (1994). A dynamic theory of organizational knowledge creation, *Organization Science* **5**: 14–37. 9, 13, 14, 15, 17, 19, 20, 22, 29, 44, 205

- Nonaka, I. and Konno, N. (1998). The concept of "ba": Building a foundation for knowledge creation, *California Management Review* **40**(3): 40–54. 11
- North, K. and Guldenberg, S. (2009). *Produktive Wissensarbeit(er): Antworten auf die Management-Herausforderung des 21. Jahrhunderts Mit vielen Fallbeispielen Performance messen Produktivität steigern Wissensarbeiter entwickeln*, Gabler. 24
- Nutt, W., Razniewski, S. and Vegliach, G. (2012). Incomplete databases: Missing records and missing values, *Database Systems for Advanced Applications: 17th International Conference, DASFAA 2012, International Workshops: FlashDB, ITEMS, SNSM, SIM3, DQDI, Busan, South Korea, April 15-19, 2012. Proceedings*, Springer, Berlin, Heidelberg, pp. 298–310. 164
- O’Leary, M. B. and Cummings, J. N. (2007). The spatial, temporal, and configurational characteristics of geographic dispersion in teams, *MIS Q.* **31**(3): 433–452. 25
- O’Leary, M., Orlikowski, W. and Yates, J. (2002). *Distributed Work over the Centuries: Trust and Control in the Hudson’s Bay Company, 1670-1826*, in Hinds and Kiesler (2002b), chapter 2, pp. 27–54. 25
- Olson, G. M. and Olson, J. S. (2000). Distance matters, *Human-Computer Interaction* **15**(2/3): 139–178. 25, 26, 55
- Olson, J. S., Grudin, J. and Horvitz, E. (2005). A study of preferences for sharing and privacy, *CHI ’05: CHI ’05 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, pp. 1985–1988. 49, 71, 93, 94, 107
- Olson, J. S., Teasley, S., Covi, L. and Olson, G. (2002). *The (Currently) Unique Advantages of Collocated Work*, in Hinds and Kiesler (2002b), chapter 5, pp. 113–135. 26
- O’Reilly, T. (2005). What is web 2.0? design patterns and business models for the next generation of software.
URL: <http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-2.0.html> 27
- Oren, E. (2006). An overview of information management and knowledge work studies: Lessons for the semantic desktop, *Proceedings of the 5th International Conference on Semantic Desktop and Social Semantic Collaboration - Volume 202*, SemDesk’06, CEUR-WS.org, Aachen, Germany, pp. 14–24. 24
- Orlikowski, W. J. (1992). Learning from notes: organizational issues in groupware implementation, *CSCW ’92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, ACM Press, New York, NY, pp. 362–369. 2, 11, 47, 49, 50, 93
- Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations, *Organization Science* **11**(4): 404–428. 11

BIBLIOGRAPHY

- Orlikowski, W. J. and Robey, D. (1991). Information technology and the structuring of organizations, *Information Systems Research* **2**(2): 143–169. 11
- Otto, B. (2011). Data governance, *Business & Information Systems Engineering* **3**(4): 241–244. 167
- Paşca, M., Van Durme, B. and Garera, N. (2007). The role of documents vs. queries in extracting class attributes from text, *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, ACM, New York, NY, USA, pp. 485–494. 160
- Pagano, D., Juan, M. A., Bagnato, A., Roehm, T., Brüggel, B. and Maalej, W. (2012). Fastfix: monitoring control for remote software maintenance, *Proceedings of the 2012 International Conference on Software Engineering*, ICSE 2012, IEEE Press, Piscataway, NJ, USA, pp. 1437–1438. 81
- Pan, J. (2009). Resource description framework, in S. Staab and R. Studer (eds), *Handbook on Ontologies*, 2nd edn, Springer, Berlin, Heidelberg, pp. 71–90. 135
- Park, H. and Widom, J. (2014). Crowdfill: Collecting structured data from the crowd, *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, ACM, New York, NY, USA, pp. 577–588. 168
- Pasca, M. (2012). Attribute extraction from conjectural queries, *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 2177–2190. 161
- Pasca, M. (2014). Acquisition of noncontiguous class attributes from web search queries, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pp. 386–394. 161
- Patrick, T. B. (2003). Using description logic to manage question corpora, *AMIA 2003, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 8-12, 2003*. 160
- Pauleen, D. (2009). Personal knowledge management: Putting the "person" back into the knowledge equation, *Online Information Review* **33**(2): 221–224. 23
- Pauleen, D. and Gorman, G. (2016). *Personal Knowledge Management: Individual, Organizational and Social Perspectives*, Taylor & Francis. 23
- Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web* **8**(3): 489–508. 141, 165
- Paulheim, H. and Bizer, C. (2013). Type inference on noisy rdf data, *The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, Springer, Berlin, Heidelberg, pp. 510–525. 165
- Pautzke, G. (1989). *Die Evolution der organisatorischen Wissensbasis: Bausteine zu einer Theorie des organisatorischen Lernens*, Vol. 58 of *Münchener Schriften zur angewandten Führungslehre*, Kirsch, Herrsching. 16, 17

- Pellissier Tanon, T., Stepanova, D., Razniewski, S., Mirza, P. and Weikum, G. (2017). Completeness-aware rule learning from knowledge graphs, *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I*, Springer, pp. 507–525. 165
- Pfisterer, F., Nitsche, M., Jameson, A. and Barbu, C. (2008). User-centered design and evaluation of interface enhancements to the Semantic MediaWiki, *Proceedings of the CHI 2008 workshop on Semantic Web User Interaction*, Florence, Italy. 151, 162
- Pickens, J., Golovchinsky, G. and Morris, M. R. (2008). Report on the 1st collaborative information retrieval workshop: Held in conjunction with the joint conference on digital libraries (jcdl) 2008. 42
- Picot, A., Reichwald, R. and Wigand, R. (2008). *The Dissolving of Hierarchies – Modularizing the Enterprise*, Springer, Berlin, Heidelberg, pp. 183–231. 19
- Pinelle, D. and Gutwin, C. (2005). A groupware design framework for loosely coupled workgroups, *ECSCW 2005: Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work, 18–22 September 2005, Paris, France*, Springer, pp. 65–82. 85
- Pipek, V., Hinrichs, J. and Wulf, V. (2003). *Sharing Expertise: Challenges for Technical Support*, in Ackerman, Pipek and Wulf (2003), pp. 111–136. 50, 85
- Podnar, I., Rajman, M., Luu, T., Klemm, F. and Aberer, K. (2007). Scalable peer-to-peer web retrieval with highly discriminative keys, *ICDE '07: Proceedings of the 23rd International Conference on Data Engineering (ICDE'07)*, IEEE Computer Society, Istanbul, Turkey, pp. 0–0. 105
- Polanyi, M. (1967). *The tacit dimension*, Doubleday, Garden City, NY. 14
- Poltrock, S., Grudin, J., Dumais, S., Fidel, R., Bruce, H. and Pejtersen, A. M. (2003). Information seeking and sharing in design teams, *GROUP '03: Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, ACM, New York, NY, USA, pp. 239–247. 47, 87
- Prasojo, R. E., Darari, F., Razniewski, S. and Nutt, W. (2016). Managing and consuming completeness information for wikidata using COOL-WD, *Proceedings of the 7th International Workshop on Consuming Linked Data co-located with 15th International Semantic Web Conference, COLD@ISWC 2015, Kobe, Japan, October 18, 2016*. 162
- Preece, J. and Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation, *AIS transactions on human-computer interaction* **1**(1): 13–32. 87, 88
- Probst, G. J. B. (1998). Practical knowledge management: A model that works, *Arthur D. Little PRISM* **2**: 17–29. 8, 9, 17
- Probst, G. J. B., Raub, S. and Romhardt, K. (2006). *Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen*, 5 edn, Gabler. 8, 9, 10, 12, 16, 19, 20, 22, 24, 31, 44, 62, 205

BIBLIOGRAPHY

- Probst, G., Raub, S. and Romhardt, K. (1999). *Managing Knowledge: Building Blocks for Success*, Wiley. 8
- Prud'Hommeaux, E. and Seaborne, A. (2008). SPARQL query language for RDF, World Wide Web Consortium, Recommendation REC-rdf-sparql-query-20080115. 135, 141, 142, 144
- Purcell, K. (2011). Search and email still top the list of most popular online activities, *Technical report*, Pew Internet Project.
URL: <http://pewinternet.org/Reports/2011/Search-and-email.aspx> 174
- Quigley, N. R., Tesluk, P. E., Locke, E. A. and Bartol, K. M. (2007). A multilevel investigation of the motivational mechanisms underlying knowledge sharing and performance, *Organization Science* **18**(1): 71–88. 44
- Rader, E. (2010). The effect of audience design on labeling, organizing, and finding shared files, *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, ACM, New York, NY, USA, pp. 777–786. 46, 48, 57
- Rader, E. J. (2009). *Social Influences on User Behavior in Group Information Repositories.*, PhD thesis, University of Michigan. 106, 173
- Rafes, K., Cohen-Boulakia, S. and Abiteboul, S. (2017). Une autocomplément générique de SPARQL dans un contexte multi-services, *BDA 2017 - 33ème conférence sur la "Gestion de Données - Principes, Technologies et Applications"*, Nancy, France. 160
- Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches, *IEEE Data Eng. Bull.* **23**(4): 3–13. 164
- Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R. and Riedl, J. (2006). Motivating participation by displaying the value of contribution, *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, New York, NY, USA, pp. 955–958. 72, 84
- Razavi, M. N. and Iverson, L. (2007). Designing for privacy in personal learning spaces, *New Review of Hypermedia and Multimedia* **13**(2): 163–185. 49, 107
- Razniewski, S. and Nutt, W. (2011). Completeness of queries over incomplete databases, *PVLDB* **4**(11): 749–760. 164
- Razniewski, S., Savkovic, O. and Nutt, W. (2016). Turning the partial-closed world assumption upside down, *Proceedings of the 10th Alberto Mendelzon International Workshop on Foundations of Data Management, Panama City, Panama, May 8-10, 2016*. 165
- Razniewski, S., Suchanek, F. M. and Nutt, W. (2016). But What Do We Actually Know?, *AKBC workshop*. 133
- Redman, T. C. (1998). The impact of poor data quality on the typical enterprise, *Commun. ACM* **41**(2): 79–82. 164
- Richardson, M. (2008). Learning about the world through long-term query logs, *ACM Trans. Web* **2**(4): 21:1–21:27. 83

- Rioux, K. S. (2004). *Information acquiring-and-sharing in Internet-based environments: an exploratory study of individual user behaviors*, PhD thesis, University of Texas at Austin. 87
- Rioux, K. S. (2005). Information acquiring-and-sharing, in Fisher et al. (2005), chapter 29, pp. 179–184. 47, 87
- Robillard, M., Walker, R. and Zimmermann, T. (2010). Recommendation systems for software engineering, *Software, IEEE* **27**(4): 80–86. 81
- Romberg, B. (2010). Intopedia: Personalisierte beitragsempfehlungen für die wikipedia, *Technical report*, Hochschule Karlsruhe and FZI Forschungszentrum Informatik. 130, 132, 172
- Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search, *Proceedings of the 13th International Conference on World Wide Web, WWW '04*, ACM, New York, NY, USA, pp. 13–19. 75
- Roth, D. (2009). The answer factory: Demand media and the fast, disposable, and profitable as hell media model, *Wired*. 83
- Ruiz, E. J., Hristidis, V. and Ipeirotis, P. G. (2014). Facilitating document annotation using content and querying value, *IEEE Trans. on Knowl. and Data Eng.* **26**(2): 336–349. 161
- Sarasua, C., Simperl, E., Noy, N., Bernstein, A. and Leimeister, J. M. (2015). Crowdsourcing and the semantic web: A research manifesto, *Human Computation* **2**. 168
- Sarasua, C., Simperl, E. and Noy, N. F. (2012). Crowdmap: Crowdsourcing ontology alignment with microtasks, *The Semantic Web – ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11–15, 2012, Proceedings, Part I*, Springer, Berlin, Heidelberg, pp. 525–541. 168
- Sauermann, L., Bernardi, A. and Dengel, A. (2005). Overview and outlook on the semantic desktop, *Proceedings of the 2005 International Conference on Semantic Desktop Workshop: Next Generation Information Management & Collaboration Infrastructure - Volume 175*, sdw'05, CEUR-WS.org, Aachen, Germany, pp. 74–91. 163
- Schilling, M. A. and Steensma, H. K. (2001). The use of modular organizational forms: An industry-level analysis, *The Academy of Management Journal* **44**(6): 1149–1168. 19
- Schirmer, A. L. (2003). Privacy and knowledge management: Challenges in the design of the lotus discovery server, *IBM Systems Journal* **42**(3): 519–531. 94
- Sekine, S. and Suzuki, H. (2007). Acquiring ontological knowledge from query logs, *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, ACM, New York, NY, USA, pp. 1223–1224. 83
- Selke, J., Lofi, C. and Balke, W.-T. (2012). Pushing the boundaries of crowd-enabled databases with query-driven schema expansion, *Proc. VLDB Endow.* **5**(6): 538–549. 168, 169

BIBLIOGRAPHY

- Sellam, T. and Kersten, M. (2016). Cluster-driven navigation of the query space, *IEEE Transactions on Knowledge and Data Engineering* **28**(5): 1118–1131. 160
- Sellam, T. and Kersten, M. L. (2013). Meet charles, big data query advisor, *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*. 160
- Sellam, T. and Kersten, M. L. (2017). 80 new packages to mine database query logs, *CoRR* **abs/1703.08732**. 160
- Shami, N. S., Muller, M. and Millen, D. (2011). Browse and discover: Social file sharing in the enterprise, *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11, ACM, New York, NY, USA*, pp. 295–304. 106, 117
- Shenton, A. K. and Johnson, A. (2007). Library suggestions and typologies of information needs, *Library and Information Research* **31**(98): 3–15. 83
- Si, X., Chang, E. Y., Gyöngyi, Z. and Sun, M. (2010). Confucius and its intelligent disciples: Integrating social with search, *Proc. VLDB Endow.* **3**(1-2): 1505–1516. 129
- Simperl, E., Norton, B. and Vrandečić, D. (2012). Crowdsourcing tasks in open query answering, *AAAI Spring Symposium: Wisdom of the Crowd*. 168
- Singer, J., Sim, S. E. and Lethbridge, T. C. (2008). *Software Engineering Data Collection for Field Studies*, Springer London, London, pp. 9–34. 24
- Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M. and Leskovec, J. (2017). Why we read wikipedia, *Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland*, pp. 1591–1600. 87
- Singh, J., Hoffart, J. and Anand, A. (2016). Discovering entities with just a little help from you, *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, ACM, New York, NY, USA*, pp. 1331–1340. 161
- Siorpaes, K. and Hepp, M. (2008a). Games with a purpose for the semantic web, *IEEE Intelligent Systems* **23**(3): 50–60. 161
- Siorpaes, K. and Hepp, M. (2008b). Ontogame: weaving the semantic web by online games, *ESWC'08: Proceedings of the 5th European semantic web conference on The semantic web*, Springer, Berlin, Heidelberg, pp. 751–766. 137, 151
- Siorpaes, K. and Hepp, M. (2008c). Ontogame: Weaving the semantic web by online games, *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings*, Springer, Berlin, Heidelberg, pp. 751–766. 161
- Small, C. T. and Sage, A. P. (2006). Knowledge management and knowledge sharing: A review, *Information, Knowledge, Systems Management* **5**(3): 153–169. 43, 44

- Smetters, D. K. and Good, N. (2009). How users use access control, *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, ACM, New York, NY, USA, pp. 15:1–15:12. 94
- Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M. and Boydell, O. (2005). Exploiting query repetition and regularity in an adaptive community-based web search engine, *User Modeling and User-Adapted Interaction* **14**(5): 383–423. 67
- Smyth, B., Boydell, O., Balfe, E., Bradley, K., Briggs, P., Coyle, M. and Freyne, J. (2005). A live-user evaluation of collaborative web search, *IJ-CAI'05: Proceedings of the 19th international joint conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1419–1424. 65, 67
- Staab, S., Studer, R., Schnurr, H.-P. and Sure, Y. (2001). Knowledge processes and ontologies, *IEEE Intelligent Systems* **16**(1): 26–34. 135
- Stamou, S. and Efthimiadis, E. (2009). Queries without clicks: Successful or failed searches?, *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, IR Publications, Amsterdam, pp. 13–14. 65
- Stamou, S. and Efthimiadis, E. N. (2010). Interpreting user inactivity on search results, *ECIR*, pp. 100–113. 65
- Statistisches Bundesamt (2016). Unternehmen und arbeitsstätten: Nutzung von informations- und kommunikationstechnologien in unternehmen, *Technical report*, Statistisches Bundesamt, Wiesbaden. 28
- Stehr, N. (1994). *Knowledge Societies*, Sage Publications (CA). 24
- Stocker, A. and Tochtermann, K. (2011). *Wissenstransfer mit Wikis und Weblogs: Fallstudien zum erfolgreichen Einsatz von Web 2.0 in Unternehmen*, Gabler. 28
- Stroh, F., Winter, R. and Wortmann, F. (2011). Method support of information requirements analysis for analytical information systems, *Business & Information Systems Engineering* **3**(1): 33–43. 82
- Strohmaier, M. and Kröll, M. (2009). Studying databases of intentions: Do search query logs capture knowledge about common human goals?, *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, ACM, New York, NY, USA, pp. 89–96. 83
- Sundarmurthy, B., Koutris, P., Lang, W., Naughton, J. and Tannen, V. (2017). m-tables: Representing Missing Data, in M. Benedikt and G. Orsi (eds), *20th International Conference on Database Theory (ICDT 2017)*, Vol. 68 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 21:1–21:20. 164
- Swartz, A. (2002). Musicbrainz: a semantic web service, *IEEE Intelligent Systems* **17**(1): 76–77. 85
- Tang, J. C., Drews, C., Smith, M., Wu, F., Sue, A. and Lau, T. (2007). Exploring patterns of social commonality among file directories at work, *Pro-*

BIBLIOGRAPHY

- ceedings of the SIGCHI conference on Human factors in computing systems, CHI '07*, ACM, New York, NY, USA, pp. 951–960. 94
- Tang, L., Austin, S., Zhao, Y. and Culley, S. (2006). Immortal information and through life knowledge management (kim): how can valuable information be available in the future?, *The 3rd Asia-Pacific International Conference on Knowledge Management (KMAP 2006)* . 14, 87, 205
- Tapscott, D. and Williams, A. D. (2006). *Wikinomics: How Mass Collaboration Changes Everything*, Portfolio Hardcover. 28
- Taylor, F. W. (1911). *The Principles of Scientific Management*, Harper, New York. 24
- Taylor, R. S. (1962). The process of asking questions, *American Documentation* **13**(4): 391–396. 34, 35, 83
- Taylor, R. S. (1968). Question-Negotiation and information seeking in libraries, *College and Research Libraries* (29): 178–194. 34, 39
- Teasley, S., Covi, L., Krishnan, M. S. and Olson, J. S. (2000). How does radical collocation help a team succeed?, *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work*, pp. 339–346. 26
- Teevan, J., Adar, E., Jones, R. and Potts, M. (2006). History repeats itself: repeat queries in yahoo's logs, *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 703–704. 66, 67
- Teevan, J., Adar, E., Jones, R. and Potts, M. A. S. (2007). Information retrieval: Repeat queries in yahoo's logs, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, ACM, New York, NY, USA, pp. 151–158. 66
- Teevan, J., Jones, W. and Bederson, B. B. (2006). Introduction, *Commun. ACM* **49**(1): 40–43. 23
- ten Cate, B., Civili, C., Sherkhonov, E. and Tan, W.-C. (2015). High-level why-not explanations using ontologies, *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS '15*, ACM, New York, NY, USA, pp. 31–43. 166
- Thaler, S., Simperl, E. and Siorpaes, K. (2011). Spothelink: Playful alignment of ontologies, *Proceedings of the 2011 ACM Symposium on Applied Computing, SAC '11*, ACM, New York, NY, USA, pp. 1711–1712. 161
- Thaler, S., Siorpaes, K., Mear, D., Simperl, E. and Goodman, C. (2011). Seafish: A game for collaborative and visual image annotation and interlinking, *The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29 – June 2, 2011, Proceedings, Part II*, Springer, Berlin, Heidelberg, pp. 466–470. 161
- Thomas, C. F. and Griffin, L. S. (1998). Who will create the metadata for the internet?, *First Monday* **3**(12). 133, 136, 137

- Thompson, J. (1995). *The Media and Modernity: A Social Theory of the Media*, Stanford University Press. 54
- Thompson, J. D. (1967). *Organizations in Action*, McGraw-Hill, New York. 27
- Tiwana, A. (2003). Affinity to infinity in peer-to-peer knowledge platforms, *Commun. ACM* **46**(5): 76–80. 20
- Troumpoukis, A., Konstantopoulos, S. and Charalambidis, A. (2017). An extension of sparql for expressing qualitative preferences, *The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I*, Springer, pp. 711–727. 166
- Tsui, E. (2002). Technologies for personal and peer-to-peer (p2p) knowledge management, *Technical report*, CSC Leading Edge Forum (LEF). 20
- Tungare, M. and Perez-Quinones, M. (2008). Thinking outside the (beige) box: Personal information management beyond the desktop, *Proceedings of the 3rd Invitational Workshop on Personal Information Management*, p. 8. 18, 23, 85
- Tyler, S. K. and Teevan, J. (2010). Large scale query log analysis of re-finding, *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, ACM, New York, NY, USA, pp. 191–200. 66
- U.S. Bureau of Labor Statistics (2005). Computer and internet use at work in 2003, *Technical report*, U.S. Bureau of Labor Statistics, Washington, DC. 28
- Uyar, A. and Aliyu, F. M. (2015). Evaluating search features of google knowledge graph and bing satori: Entity types, list searches and query interfaces, *Online Information Review* **39**(2): 197–213. 160
- Valle, E. D. and Ceri, S. (2011). *Querying the Semantic Web: SPARQL*, in Domingue et al. (2011), 1st edn, pp. 299–363. 135
- van Elst, L. and Abecker, A. (2004). Agent-based knowledge management, *KI* **18**(2): 11–16. 20, 217
- van Elst, L., Dignum, V. and Abecker, A. (2003). Towards agent-mediated knowledge management, *AMKM*, pp. 1–30. 20
- Varga, J., Dobrokhotova, E., Romero, O., Pedersen, T. B. and Thomsen, C. (2017). Sm4mq: A semantic model for multidimensional queries, *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 – June 1, 2017, Proceedings, Part I*, Springer, pp. 449–464. 160
- Vasilyeva, E. (2017). *Why-Query Support in Graph Databases*, PhD thesis, Dresden University of Technology, Germany. 166
- Vasilyeva, E., Heinze, T., Thiele, M. and Lehner, W. (2016). Debeaq - debugging empty-answer queries on large data graphs, *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 1402–1405. 166

BIBLIOGRAPHY

- Voida, S., Edwards, W. K., Newman, M. W., Grinter, R. E. and Ducheneaut, N. (2006). Share and share alike: exploring the user interface affordances of file sharing, *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, ACM, New York, NY, USA, pp. 221–230. 47, 94, 107
- von Hippel, E. (2001). User toolkits for innovation, *Journal of Product Innovation Management* **18**(4): 247–257. 174
- Vrandečić, D. (2009). Towards automatic content quality checks in semantic wikis, *Social Semantic Web: Where Web 2.0 Meets Web 3.0, Papers from the 2009 AAAI Spring Symposium, Technical Report SS-09-08, Stanford, California, USA, March 23-25, 2009*, pp. 69–70. 162, 169
- Vrandečić, D. and Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase, *Commun. ACM* **57**(10): 78–85. 162
- Wahl, A. M., Endler, G., Schwab, P. K., Herbst, S. and Lenz, R. (2017). We can query more than we can tell: Facilitating collaboration through query-driven knowledge-sharing, *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17 Companion*, ACM, New York, NY, USA, pp. 335–338. 160
- Walk, S. and Strohmaier, M. (2014). Characterizing and predicting activity in semantic mediawiki communities, *Proceedings of the Third International Conference on Semantic Web Collaborative Spaces - Volume 1275, SWCS'14, CEUR-WS.org, Aachen, Germany*, pp. 53–67. 162
- Walkerdine, J. and Rodden, T. (2001). Sharing searches: Developing open support for collaborative searching., in M. Hirose (ed.), *Human-computer Interaction: INTERACT '01 : IFIP TC.13 International Conference on Human-Computer Interaction, 9th-13th July 2001, Tokyo, Japan*, IOS Press, pp. 140–147. 128
- Walsh, J. P. and Maloney, N. G. (2002). *Computer Network Use, Collaboration Structures, and Productivity*, in Hinds and Kiesler (2002b), chapter 18, pp. 433–458. 25
- Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers, *J. Manage. Inf. Syst.* **12**(4): 5–33. 164
- Wang, X., Tan, B., Shakery, A. and Zhai, C. (2009). Beyond hyperlinks: Organizing information footprints in search logs to support effective browsing, *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, ACM, New York, NY, USA, pp. 1237–1246. 128
- Wasko, M. M. and Faraj, S. (2005). Why should i share? examining social capital and knowledge contribution in electronic networks of practice, *MIS Quarterly* **29**(1): 35–57. 44, 49, 137
- Wegner, D. M. (1986). Transactive memory: A contemporary analysis of the group mind., in B. Mullen and G. R. Goethals (eds), *Theories of group behavior*, Springer, pp. 185–208. 16

- Weikum, G., Hoffart, J. and Suchanek, F. M. (2016). Ten years of knowledge harvesting: Lessons and challenges, *IEEE Data Eng. Bull.* **39**(3): 41–50. 133
- Wen, J.-R., Nie, J.-Y. and Zhang, H.-J. (2001). Clustering user queries of a search engine, *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, ACM, New York, NY, USA, pp. 162–168. 82
- Whalen, T., Smetters, D. and Churchill, E. F. (2006). User experiences with sharing and access control, *CHI '06 Extended Abstracts on Human Factors in Computing Systems, CHI EA '06*, ACM, New York, NY, USA, pp. 1517–1522. 107
- Whalen, T., Toms, E. and Blustein, J. (2008a). File sharing and group information management, *PIM 2008: Proceedings of the CHI 2008 Workshop on Personal Information Management.* 94
- Whalen, T., Toms, E. G. and Blustein, J. (2008b). Information displays for managing shared files, *Proceedings of the 2Nd ACM Symposium on Computer Human Interaction for Management of Information Technology, CHiMiT '08*, ACM, New York, NY, USA, pp. 5:1–5:10. 47, 107
- White, M. (2015). *Enterprise Search: Enhancing Business Performance*, O'Reilly Media. 92
- Whittaker, S., Bellotti, V. and Gwizdka, J. (2006). Email in personal information management, *Commun. ACM* **49**(1): 68–73. 174
- Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives, *ACM Trans. Comput.-Hum. Interact.* **8**(2): 150–170. 24
- Wiegand, M. (1998). *Prozesse organisationalen Lernens*, Gabler, Wiesbaden. 15
- Wilson, T. D. (1981). On user studies and information needs., *Journal of Documentation* **37**(1): 3–15. 34, 35, 36, 37, 39, 205
- Wilson, T. D. (2010). Information sharing: an exploration of the literature and some propositions, *Information Research* **15**(4). 44
- Wouters, P. and Gerbec, D. (2003). Interactive internet? studying mediated interaction with publicly available search engines, *Journal of Computer-Mediated Communication* **8**(4): 0–0. 54, 205
- Wright, K. (2005). Personal knowledge management: supporting individual knowledge worker performance, *Knowledge Management Research & Practice* **3**(3): 156–165. 23
- Wyman, B. (2005). Prospective vs retrospective search, OASIS XML-DEV mailing list. 38, 40
- Yang, B. and Jeh, G. (2006). Retroactive answering of search queries, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 457–466. 66, 98, 99, 106, 111

BIBLIOGRAPHY

- Yang, J., Hauff, C., Bozzon, A. and Houben, G.-J. (2014). Asking the right question in collaborative q&a systems, *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, ACM, New York, NY, USA, pp. 179–189. 130
- Yao, S., Liu, J., Wang, M., Wei, B. and Chen, X. (2015). ANNA: answering why-not questions for SPARQL, *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015)*, Bethlehem, PA, USA, October 11, 2015. 166
- Yates, D., Wagner, C. and Majchrzak, A. (2010). Factors affecting shapers of organizational wikis, *J. Am. Soc. Inf. Sci. Technol.* **61**(3): 543–554. 131
- Zack, M. H. (1999). Managing codified knowledge, **40**: 45–58. 1
- Zhai, K., Kozareva, Z., Hu, Y., Li, Q. and Guo, W. (2016). Query to knowledge: Unsupervised entity extraction from shopping queries using adaptor grammars, *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, ACM, New York, NY, USA, pp. 255–264. 161
- Zhang, A. X., Verou, L. and Karger, D. (2017). Wikum: Bridging discussion forums and wikis using recursive summarization, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, ACM, New York, NY, USA, pp. 2082–2096. 88
- Zhang, D. and Lu, J. (2009). What queries are likely to recur in web search?, *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, pp. 827–828. 67
- Zhang, J., Jie, L., Rahman, A., Xie, S., Chang, Y. and Yu, P. S. (2015). Learning entity types from query logs via graph-based modeling, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, ACM, New York, NY, USA, pp. 603–612. 161
- Zhang, P. (2008). Motivational affordances: Reasons for ICT design and use, *Communications of the ACM* **51**(11): 145–147. 84

Index

Numbers written in *italic* refer to the page where the corresponding entry is described; numbers underlined refer to the definition; numbers in *roman* refer to the pages where the entry is used.

- Anticipation, 9, 13, 17, 21, 22
- Barriers, 9, 17, 20
- Codification, 21
- Communities of practice, 15, 86
- Context, 81
- Coordination, 26
- Data, Information, and Knowledge, 12
- Decision, 17
- Decoupling, 9, 19, 22, 72
- Desktop search, 92
- Distributed knowledge management, 20
- Distributed Work, 25
- Document and content sharing, 56
- Email, 69
- Emergent needs, 61
- Enterprise 2.0, 27
- Enterprise search, 41, 92
- Explicit knowledge, 14
- Face-to-face communication, 55
- Future use, 9, 17
- Future work, 62, 69, 74, 76, 79, 102, 118, 122, 125, 147, 150, 152
- Implicit knowledge, 14
- Incompleteness, 17
- Information gap, 9
- Information literacy, 61
- Information management, 7, 35, 82
- Information need, 9
- personal, 34, 79
- unsatisfied, 63
- Information space, 69
- Inverse Search, 96
- IR system, 40, 80
- Keeping, 23, 47
- Knowledge distribution, 16
- Knowledge Management, 7
- Knowledge maturing, 17, 51, 54
- Knowledge sharing, 27
- Knowledge transfer, 13
- Knowledge Work, 24
- Long tail, 28
- Media richness, 54
- Mediation, 52, 72, 74
- Mediation service, 52, 73
- Mediation space, 52, 73
- Natural sharing context, 9
- Organizational information gap, 99
- Organizational information need, 99
- Organizational knowledge, 15
- Organizational knowledge base, 16
- Organizational knowledge communities, 15
- Organizational learning, 7, 16
- Organizational memory, 16
- Personal communication, *see* Face-to-face communication
- Personal information management, 22
- Personal knowledge management, 22
- Personalization, 21
- Private information space, 73
- Prospective, 38
- Prospective search, 40
- Pull/push, 9, 19, 22, 52, 74
- Receiver, 13
- Retrospective search, 38, 40
- Selection, 9
- Semantic desktop, 147
- Social search, 38, 42
- Standing interest, 98
- Tacit knowledge, 14
- Transactional memory systems, 16
- Value, 9