Future Generation Computer Systems 94 (2019) 879–894

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

The MASi repository service — Comprehensive, metadata-driven and multi-community research data management

Richard Grunzke [a,*], Volker Hartmann [b], Thomas Jejkal [b], Helen Kollai [c], Ajinkya Prabhune [b], Hendrik Herold [c], Aline Deicke [d], Christiane Dressler [d], Julia Dolhoff [d], Julia Stanek [e], Alexander Hoffmann [e], Ralph Müller-Pfefferkorn [a], Torsten Schrade [d], Gotthard Meinel [c], Sonja Herres-Pawlis [e], Wolfgang E. Nagel [a]

[a] *Center for Information Services and High Performance Computing, Technische Universität Dresden, Dresden, Germany*
[b] *Institute for Data Processing and Electronics, Karlsruhe Institute of Technology, Germany*
[c] *Monitoring of Settlement and Open Space Development, Institute of Ecological and Regional Development, Germany*
[d] *Digitale Akademie, Akademie der Wissenschaften und Literatur Mainz, Germany*
[e] *Institut für Anorganische Chemie, Rheinisch-Westfälische Technische Hochschule Aachen, Germany*

ARTICLE INFO

Article history:
Received 22 May 2017
Received in revised form 8 September 2017
Accepted 22 December 2017
Available online 31 January 2018

Keywords:
Metadata
Communities
Research data management

ABSTRACT

Nowadays, the daily work of many research communities is characterized by an increasing amount and complexity of data. This makes it increasingly difficult to manage, access and utilize the data to ultimately gain scientific insights based on it. At the same time, domain scientists want to focus on their science instead of IT. The solution is research data management to store data in a structured way enabling easy discovery for future reference and usage. An integral part is the use of metadata. With it, data becomes accessible by its content and context instead of its name and location only. The use of metadata shall be as automatic and seamless as possible in order to foster a high usability.

Here, we present the architecture and developments of the Metadata Management for Applied Sciences (MASi) project that is currently building a comprehensive research data management service. MASi extends the existing KIT Data Manager framework by a generic metadata programming interface and a generic graphical web interface. Furthermore, MASi is OAI compliant and supports the OAI-PMH protocol while providing support for provenance information using ProvONE, a well-established and accepted provenance model. To illustrate the practical applicability of the MASi service, we present the adoption of initial use cases within geography, chemistry and digital humanities. The MASi research data management service is currently being prepared to go into production to satisfy the complex and varying requirements in an efficient, useable and sustainable way.

1. Introduction

Today's research landscape is characterized by a steadily growing amount of data that is caused by the use of improved data recording, increasingly complex simulations and by the correlation of numerous, often heterogeneous data sources. This growth of data promises a higher amount and quality of scientific insights. When the amount and complexity of data is increasing, the requirements regarding the data structure are increasing as well. A suitable and specific data description becomes paramount. Present data handling methods are often reaching their limit. Latest methods of data management for newer and more complex data become essential. Especially important are improved data descriptions, sustainable storages, findability, pre-processing for further uses and the exploitation of existing data.

An established method to describe complex data structures is making use of metadata. It encapsulates the semantic substance of a data set in aggregated form. Metadata ("data about data") play a central role in making data available for the long-term. It is essential for the comprehension of data, its storage, preservation, curation and discovery for future use. Metadata enables an easier application of complex tasks on data by enabling the search for input data based on its content and context. Aside from a better discovery, other data management aspects, such as managing and utilizing similarities between data sets, are fostered. In diverse scientific communities very different metadata standards and data management methods exist, each incorporating community specific data characteristics [1]. This results in a limited portability to

[*] Correspondence to: Technische Universität Dresden, 01069 Dresden, Germany.
E-mail address: richard.grunzke@tu-dresden.de (R. Grunzke).

https://doi.org/10.1016/j.future.2017.12.023
0167-739X/© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

new use cases that causes established standards and methods in one scientific field to be of limited or no use in other fields.

To mitigate this situation, in the Metadata Management for Applied Sciences (MASi) [2,3] project, funded by the German Research Foundation (DFG), we developed a data management service for scientific data that is generally applicable. Along heterogeneous scientific use cases with varying data characteristics and amounts we developed the MASi service and are currently preparing it for production operation that is due in the second half of 2017 [3]. The use of metadata is not uniform across the use cases so that a suitable overarching research data management service is of fundamental advantage in order to save effort as various use cases can be served via a single service instance.

The motivating use cases are described in the following Section 2: digital maps in geography (Section 2.1), spectroscopy in chemistry (Section 2.2) and mediaeval stained glass in digital humanities (Section 2.3). Section 2.4 details the resulting requirements. Section 3 gives background on the KIT Data Manager (KIT DM) framework and describes which features it already provided and which were previously missing. Furthermore, the section delimits the KIT DM in regard to related systems. Section 4 details how the KIT DM was extended to build the MASi service: the overall goals (Section 4.1), the MetaStore and MASi API (Section 4.2), the generic web-interface (Section 4.3), the clientside graphical interface (Section 4.4), further features (Section 4.5) and the service operation and use case integration (Section 4.6). Section 5 gives an evaluation on how the MASi use cases are supported, besides detailing further use cases. Finally, a conclusion and an outlook is given (Section 6).

2. Initial use cases and their requirements

Here, we present a description of the MASi use cases (Sections 2.1 to 2.3) followed by the subsequent requirements regarding the MASi service (Section 2.4).

2.1. Historical maps

Historical topographic maps are a valuable and often the only source for reconstructing changes of land use over long periods of time. To utilize the inherently contained information within maps for large scale spatial analyses and change detection, advanced image analysis and pattern recognition algorithms have to be applied to scanned map documents. The general map processing is composed of three major components: (i) the digitalization of the paper maps through scanning, (ii) the georeferencing of the digital maps and (iii) the information extraction from the georeferenced digital map files (Fig. 1). The retrieved information can hence be used to expand the timeline of existing land use and land cover databases into the past [4,5].

Through digital map collections curated by libraries and national mapping agencies (e.g. [6,7]) amounts up to several million digitalized old maps are made available. To efficiently conduct large-scale and long-term analyses on historical maps we developed georeferencing methods [8] and information extractions procedures (e.g. for buildings [4,5,9] and settlement areas [10,11]) that can be automated to a high degree. This (semi-)automatic processing generates and necessitates a variety of metadata. Generating and organizing this metadata in a structured and standards-based way supports the objective of providing significant information interpretable by machines and human minds alike [12].

Parts of the descriptive metadata are usually derived from the original map collar and are used as bibliographic information by librarians and custodians of map archives. This primary metadata needs to be enriched with supplementary metadata to satisfy the variety of user requirements and data utilizations. The established metadata standards for geospatial data are INSPIRE [13] and ISO 19115 [14]

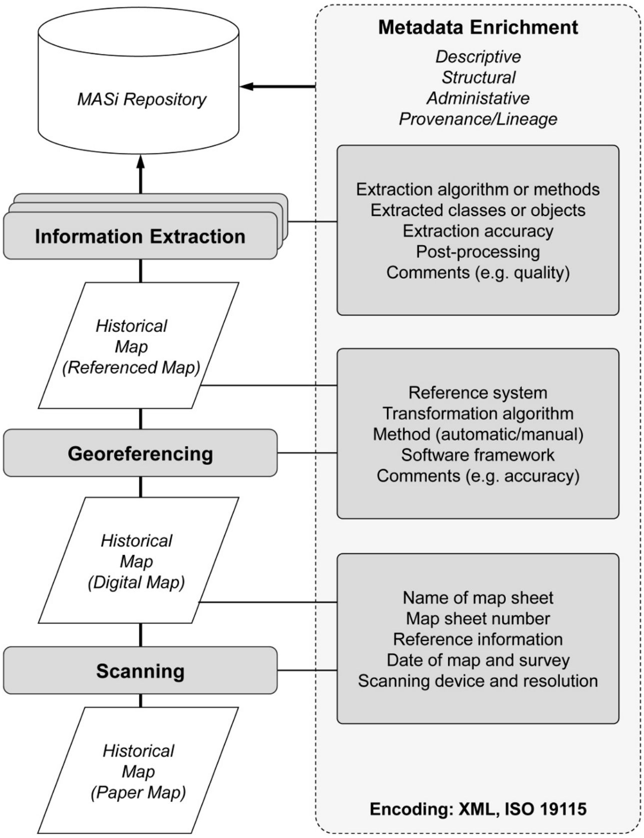


Fig. 1. General workflow of historical map processing and the linked metadata enrichment.

We applied an information extraction process and concurrent metadata generation to 5728 historical topographic maps of the map series "Topographische Messtischblätter" (Fig. 2). The data characteristics are listed in Table 1 and the subsequent requirements are detailed in Section 2.4. The maps represent the area of Germany and former territories of Eastern Prussia at a scale of 1:25000 and reference to a time between the 1860s and 1940s. Access to the digital images is provided by the Saxon State and University Library Dresden (SLUB) [15], which performed the scanning process and initiated a crowd-sourced approach to georeference the maps and crop them to their data frame [16].

As the SLUB digital map archive holds map sheet specific bibliographic data (e.g. map sheet title and number, name of the map series, date of publication), we developed a web crawler to automatically assemble descriptive metadata for each map file from the original SLUB website. In cases where this information cannot be accessed, efforts can be made to extract map collar information through optical character recognition (OCR) techniques that are currently under development. In the following steps, the annotation of additional map sheet specific metadata was performed, that is especially important for further data processing in Geographic Information Systems (GIS). Using the Geospatial Data Abstraction Library (GDAL) [17] the geographic reference system and the corner coordinates of each georeferenced file were extracted from embedded information of the GeoTIFF file. To support the map search, this boundary information was used to obtain place names that can be provided as keywords in the metadata. Therefore, the corner coordinates were used to query Open Street Map [18] with the Overpass API [19,20] to receive all place names within the

Fig. 2. Sample sheet of the historical map series "Topographische Messtischblätter".

maps data frame. All metadata was encoded in an XML (Extensible Markup Language) document. To ensure interoperability and reuse capability we decided to ensure compliance of the metadata to the most commonplace international standard for geospatial metadata ISO 19115-1:2014 [14] and its extension [21]. For each map sheet one metadata file in ISO-conform XML structure was produced using Extensible Stylesheet Language Transformation (XSLT) [22]. Each map sheet and its corresponding metadata are stored as one digital object within the MASi service.

A detailed processing history is yet missing. This limitation is addressed by publishing supplementary information on methods and parameters that were used to process the maps or their derivatives. To record this provenance metadata in an automated way, the map processing routines were customized as recommended for geospatial data in general by several approaches [23]. Provenance data is also possible to be modelled in the applied ISO 19115 metadata standard extension [24,25]. Capturing this metadata starting in early processing stages is important in regard to the transparency and reliability of a research project and thus for the reproducibility and re-usability of its results.

2.2. Spectroscopy in chemistry

In bioinorganic chemistry, metal complexes are investigated towards their role in biological systems. A complex is an inorganic compound in which an organic or inorganic ligand binds to a central metal ion and steers its properties. A multitude of information on these systems can be obtained by experimental methods such as X-ray diffraction, UV/Vis, IR, Raman, EPR and XAS spectroscopy. In most cases, these data are complemented by theoretical simulations which help to interpret the experimental data and obtain scientific insights.

In the concrete case, we investigate metal complexes and their redox behaviour with oxidants (electron-taking reagents) and reductants (electron-delivering reagents) by UV/Vis spectroscopic measurements. This serves for the determination of the electron transfer speed. Electron transfer systems mediate in all organisms the oxidation of glucose, breathing and also photosynthesis. Mostly, the detailed electron transfer is mediated by iron or copper complexes [26]. We focus on bis(chelate) copper complexes which denotes that two ligands coordinate to one metal (Fig. 3, upper right corner). The ligands comprise guanidine and quinoline groups which are useful for the complex formation and stabilization. In earlier studies, we found that these systems serve as models for biological electron transfer systems [27,28].

Here, for instance, a copper(II) complex is treated with the reductant decamethylferrocene yielding the resulting copper(I) complex and decamethylferrocinium under exchange of an electron. The copper(II) spectroscopic features decay and those of copper(I) form (Fig. 3, upper left corner).

The speed of this development is monitored every 1.5 ms for some seconds producing a large amount of raw data. These raw data are accumulated for every collection of measurements as ksd files (machine specific) which are automatically converted into csv and xls files. This data trio will be stored as raw data. In the next step, the researcher reduces these data by choosing a suitable wavelength and hence generating absorption time traces (Fig. 3, left middle). Hereby, a smaller xls file is generated which is used for fitting in Origin [29] generating an opj file. For fast overview, mostly a png file of the resulting fit graph is provided. In Section 2.4 the subsequent data management requirements are described. Table 1 details the characteristics of the data sets utilized in this context.

From these time traces, the kinetic decay constants are determined. This analysis is performed for several ratios between oxidant and reductant to resolve the second-order kinetics of the electron transfer. This second-order kinetic constant is determined at different temperatures yielding the enthalpy, entropy and free energy of the electron transfer by an Eyring plot (Fig. 3, lower left corner).

In parallel, the researcher models the electron transfer by density functional methods [30–32]. Hereby, the molecular structure is modelled by suited codes such as Gaussian [33] or NWChem [34]. At first, the geometry of the structure is optimized using suited functionals and basis sets. These data are ingested via the GUI introduced in Section 4.4. The metadata are a mixture of manually given information and metadata which shall be automatically extracted from the input (com files) and output files (out files) [35]. Data to be parsed are: functional, basis set, multiplicity, charge and solvent model. Manually, metadata such as the operator, the metal, the ligand and further details shall be given during ingest. These data are later combined to obtain the reorganization energy. In principle, the data can be also combined in workflows and metaworkflows [36,37].

2.3. Mediaeval stained glass

"Corpus Vitrearum Deutschland" (CVD) [38] is part of an international long-term research project, the "Corpus Vitrearum Medii Aevi" (CVMA). The Academy of Sciences and Literature Mainz and the Berlin-Brandenburg Academy of Sciences and Humanities is funding its German project part. The CVMA aims at recording, cataloguing and analysing mediaeval stained glass that is preserved in church windows, museums, galleries and other places all over Europe, the USA and Canada (see Fig. 4 for an example).

Due to its fragile nature, mediaeval stained glass is greatly affected by environmental factors. Thus, in order to preserve its illustrations, all windowpanes are photographed and then documented in schematic drawings. This documentation forms the basis for further research e.g. in order to analyse the history of each window's glazing and any changes or restoration activities that might have been carried out throughout the centuries. Finally, the iconography and the religious context of each window within its ecclesiastical space are interpreted. The results of these studies

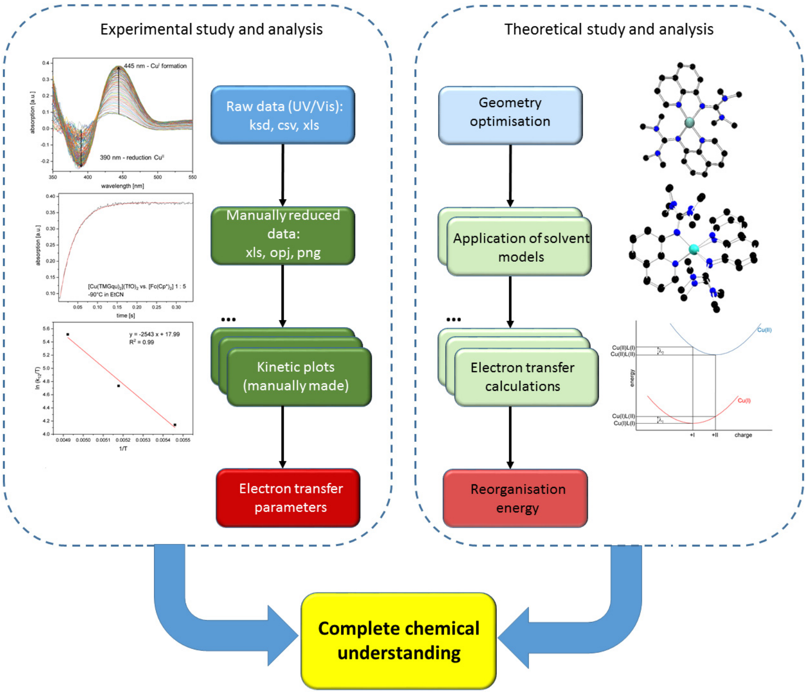

Experimental study and analysis
445 nm - Cu^I formation
390 nm - reduction Cu^{II}
absorption [a.u.]
wavelength [nm]
350
400
450
500
550
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
absorption [a.u.]
time [s]
[Cu(TMGqu)$_2$](TfO)$_2$ vs. [Fc(Cp*)$_2$] 1 : 5
-90°C in EtCN
y = -2543 x + 17.99
R^2 = 0.99
ln (k$_{12}$/T)
1/T
Raw data (UV/Vis): ksd, csv, xls
Manually reduced data: xls, opj, png
...
Kinetic plots (manually made)
Electron transfer parameters
Theoretical study and analysis
Geometry optimisation
Application of solvent models
...
Electron transfer calculations
Reorganisation energy
Cu(II)
Cu(I)
Cu(II)L(I)
Cu(II)L(II)
Cu(I)L(II)
Cu(I)L(I)
energy
λ$_2$
λ$_1$
+I
+II
charge
Complete chemical understanding

Fig. 3. Workflow of experimental and theoretical data in electron transfer studies in the chemistry use case.

are published in a printed edition that so far includes 24 of 40 volumes [39].

Since 2010 the CVD is curating a digital image archive [40] of the photographs taken. So far 5086 digital images are available online on the projects web database. For each image, an extensive set of metadata is provided, as described in the project specific XMP metadata specification [41] that are modelled according to the guidelines of the internationally acknowledged XMP metadata standard. This includes information about

- the general context such as the title of the digital image and the depiction on it,
- the creation of the image such as the date and time the resource was created (XMP-Spezifikation [41]),
- the location of the church/building and window,
- the specific window of the single pane,
- persons and institutions that are connected to the image,
- the depicted scene or figure, which is done with the help of the ICONCLASS vocabulary [42],
- the rights-management of the photography and
- its status of publication.

All XMP information is directly embedded in the TIFF files. However, because of the highly individual information that is embedded in each file, this process has to be done manually, which is error-prone. A circumstance that requires several validation routines on different levels. In the following a routine that validates the submitted iconographic information is described as an example.

As mentioned above, the ICONCLASS vocabulary is used to describe and classify the iconography of each image. "Iconclass is a classification system designed for art and iconography. It is the most widely accepted scientific tool for the description and retrieval of subjects represented in images" [42].

In order to correct data that contains faulty information a PHP script was developed that reads and manipulates the metadata of the CVMA image archive using ExifTool [43]. From a data curation perspective, a problematic area resulting from the manual metadata editing process are iconclass notations that should have been entered as separate values but inadvertently have been inserted as a single string. Therefore, a tree-step routine for disambiguation was implemented. First, every iconclass in the string is identified. Second, a uniform separator character is inserted. In the last step, the iconclass string is split by the separator and sent to the data curation script to overwrite the incorrect iconclass information with the correct values. Section 2.4 describes how this use case scenario influenced the design of the MASi service whereas Table 1 lists the data characteristics.

2.4. Subsequent overall and specific requirements

This section details the requirements based on the use cases. A large portion of the requirements are common across the use cases while some are specific. The requirements motivate the MASi developments to extend the KIT DM in order to build the MASi repository service.

A basic requirement in the MASi use cases and beyond is a system that safely stores the respective scientific data and enables a search functionality to easily find and access data with a high performance. Closely coupled with that is the requirement to keep the reusability and sustainability of the data at a high level even if the scientist that produced the data is unavailable. This is commonly enabled by annotating the data with metadata in order to store related information about the data. At the same time, metadata enables the also required search functionality.

To manage data in a structured way it needs to be possible for the owner of the data to assign specific rights to specific people that access the data. For example student workers might only be able to add data but the team lead should also be able to modify data. Sharing data needs to be possible as well, either internally with other users or globally with everybody by publishing it as Open Access. Especially the latter necessitates that the data is archived for the long-term and that the metadata can be easily accessed externally.

For a high user acceptance, convenient user interfaces are required that expose the capabilities to the users. This includes ingesting data, annotating it with metadata, viewing and editing

Table 1
Data characteristics.

Community data set	Size in GB	# of files	# of files per digital object	Average size of file in MB	Average size of a digital object in MB	# of digital objects
Historic maps	234	11456	2	20.42	40.85	5728
Georeferenced historic maps	599	11404	2	52.53	105.05	5702
Spectroscopy	19	3477	3	5.47	16.39	1159
Church window images	136	5675	1	23.97	23.97	5675

Fig. 4. An example of mediaeval stained glass showing Aaron and an unnamed prophet. Image: Andrea Gössel, CVMA Deutschland/Freiburg, CC BY-NC 4.0, Link: http://id.corpusvitrearum.de/images/2165.html.

metadata where necessary and searching through the data. The generation of metadata needs to be as automated as possible to keep the burden on the user to a minimum. This includes automatic metadata validation to ensure a high metadata quality for the long-term usefulness of the data.

Further requirements include the support for handling provenance information to increase the reproducibility of the data, a programming interface to access data and metadata from other applications and the ability to automatically run pre-processing tasks.

Although many requirements are the same across different communities, some are different. These include varying data characteristics of the communities (see Table 1) ranging from small overall and average sizes and small number of images (spectroscopy) to large sizes and number (georeferenced historic maps). Another significant difference is the heterogeneity of data formats.

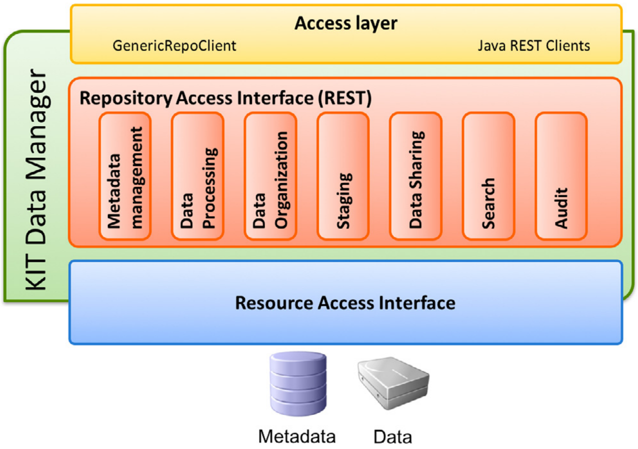


Fig. 5. The architecture of the KIT DM repository framework that is the basis the for MASi repository service [44].

These are the source of the metadata to be extracted for indexing and search. Example formats are XML (historic maps and spectroscopy) and TIFF (church window images) with further formats to be relevant in the future. See Section 4.5 for the currently support formats. Further differences are the circumstances from which the data are coming and if the data are already annotated with metadata. For example, in the spectroscopy use case the newly generated data comes directly from the device without any metadata. In the historic maps use case the data already exists but the metadata needs to be gathered from different sources. In the church window case the images are manually curated and annotated with metadata directly embedded in the images before it is handled by MASi.

3. Background

The MASi research data management service is being built using the KIT DM repository framework (see Section 3.1). Utilizing and extending the KIT DM enables MASi to offer elaborate metadata functionality with a large degree of automation and flexibility. Such metadata management capabilities are a high level abstraction based on basic storage devices and data management systems in the data life cycle hierarchy [45]. Other systems including a delimitation regarding the KIT DM are described in Section 3.2.

3.1. KIT data manager - a repository framework

The KIT Data Manager (KIT DM) [46,47] is a generic and highly customizable Open Source software framework for building research data repository systems. Horizontally, it is organized into a number of well-defined high-level services providing functionalities for data and metadata management and sharing as well as administrative services for user and group management (Fig. 5). Due to its focus on research data, KIT DM also provides features in addition to typical repository systems, namely a flexible data transfer service being able to support every data transfer protocol and a data workflow service to locally or remotely trigger the automatic execution of data processing tasks. These are configured in the repository system and include the data transfer to the processing environment, data ingest of the processing results back into the repository as well as provenance tracking. High-level services can be accessed either via the Java API, e.g. to integrate with other applications or to extend the basic framework by additional functionalities, or via RESTful (Representational state transfer) service interfaces, e.g. to access KIT DM based repository systems remotely using any programming language.

To illustrate an example interaction in principle, Fig. 6 depicts the sequence of steps that are performed during a data ingest into a KIT DM instance. "After selecting the data (1) a Digital Object is registered (2) and a new ingest is scheduled. As soon as the transfer is prepared, the data can be transferred (3). Finally, the ingest is marked for finalization (4). During finalization the cached data is copied to the archive (5), the Data Organization is obtained and content metadata might be extracted automatically. Finally, extracted content metadata is made accessible e.g. by a search index (6)" [44]. These steps are supported by the KIT DM client reference implementation, the Generic Repository Client [48] (GRC).

Vertically, KIT DM is organized into different layers where the middle layer is formed by the high-level services described before. For repository systems based on KIT DM this layer provides reliable and well-defined extension points on the one hand and a high degree of abstraction from underlying technologies on the other hand. The lowest layer of the architecture interfaces these technologies by defining a basic set of functionalities, e.g. to store and restore a digital object. This offers a high degree of sustainability as changing technologies only affect the lower layer whereas upper layers are unaffected.

In regard to the motivating requirements detailed in Section 2.4, the KIT DM did already partly fulfil them. It is able to store and access data with the GRC, extract metadata automatically, index the metadata via a search index and thus enable the command line search for the data via its metadata. It supports rights management to enable different roles and groups for different users and communities. Furthermore, KIT DM already supported the automatic triggering of pre-processing tasks such as format conversions or applying analysis algorithms.

The following features were previously missing in KIT DM; (1) a graphical user interfaces for ingesting and searching for data, (2) publishing data as Open Access, (3) automatic metadata validation, (4) programming interfaces that expose all metadata and related capabilities and (5) handling of provenance information. In the MASi project, the KIT DM was extended to close these gaps in order to enable building of the MASi service. See Section 4 for a detailed description of these extensions.

3.2. Other systems and delimitation

In the following, other systems to manage data with metadata are shorty described and then delimited against KIT DM.

DSpace [49] is a mature and ready-to-use Open Source software solution for institutional repositories. It is adaptable to fit the needs of individual institutions and fosters Open Access to all kinds of content. It supports submission workflows and various ingest and export methods. Various file types, persistent IDs and PostgreSQL and Oracle databases are supported. Search capabilities via metadata (descriptive, administrative, structural) are provided that foster the long-term preservation and accessibility of the data it manages.

Fedora (Flexible Extensible Digital Object Repository Architecture) [50] provides a framework with individual basic components to build repositories. It is Open Source and aims at being robust and modular. The main use case is to provide specialized services that may be integrated with existing environments and technologies. A central goal is to foster digital content preservation for complex and large datasets. Metadata for data organization is supported as well as descriptions of relationships between and linking of datasets.

EUDAT [51] is a European project aiming to create a generically applicable infrastructure to manage, access and preserve research data. Various EUDAT services exist. For example, B2SHARE is for storing and sharing research data via a web portal. It is also the central mean to upload data. This is done via the web portal or REST API and metadata has to be given with community specific profiles being definable. B2FIND enables to access datasets via their metadata and to annotate them with comments. A B2NOTE service is planned which aims at enabling an automatic annotation of metadata [52].

iRODS [53] as a distributed data management systems is not focused on metadata management although it offers some basic metadata functionality. Metadata can be attached to data as attribute-value-unit triples on a per file basis which can be used for searching. Integrated capabilities for metadata extraction, annotation, provenance are missing.

The ICAT system [54] aims at enabling data management for photon science facilities [55]. This includes supporting beamline proposals, access rights, experiments, studies and instruments that produce the actual data. This data is collected as datasets which can then be published. The attaching of metadata such as experiment parameters, instrument parameters and sample descriptions is supported, closely following the physics requirements but at the same time making it hard to adapt for other use cases. Technically, ICAT relies on a Java EE application server with Glassfish being the standard and Oracle and MySQL databases supported. It offers a web service interface to support, for example, the ICAT download manager TopCat. Authentication and authorization mechanisms are supported via LDAP, local database or anonymous access. A plugin interface is available.

In contrast to the above introduced systems, the KIT DM is more flexible and more advanced. It can specifically and in-depth be adapted to arbitrary target communities with a close integration into the respective community workflows. It also enables far reaching automation in extracting metadata and beyond for high usage efficiency and with ready-made generic capabilities to be adapted to specific communities.

Another delimitation dimension of the introduced systems regarding the KIT DM is in respect to highly use case specific repository software that often focus on one use case. Here, the previously introduced ICAT system serves as one example. Such systems are often more advanced in regard to their specific use case, since they are very specifically tailored to it. At the same time, this means that such solutions are hardly, if at all, adaptable to other use cases. The KIT DM, on the other hand, is generic and flexible with a focus on realizing synergy effects. First, its capabilities are developed to be beneficial across community boundaries. With this effort-saving approach the KIT DM is able to become continuously more advanced for the benefit of many communities at the same time. In contrast, developments in community specific solutions have to be done for each solution individually. In KIT DM, once a feature exists it is immediately useable for all communities. Second, building, adapting and maintaining repository systems requires significant effort. This holds true for all mentioned systems including the KIT DM. However, in contrast to the other systems the KIT DM is able to serve many communities at the same time with advanced capabilities within a single installation. Although, the other systems except ICAT can serve multiple communities, they either have a focus on generic institutional repositories without community specific adaptations (DSpace, Fedora) or lack advanced capabilities such as automated metadata extraction (EUDAT, iRODS). The ability of KIT DM to serve multiple communities within one instance while

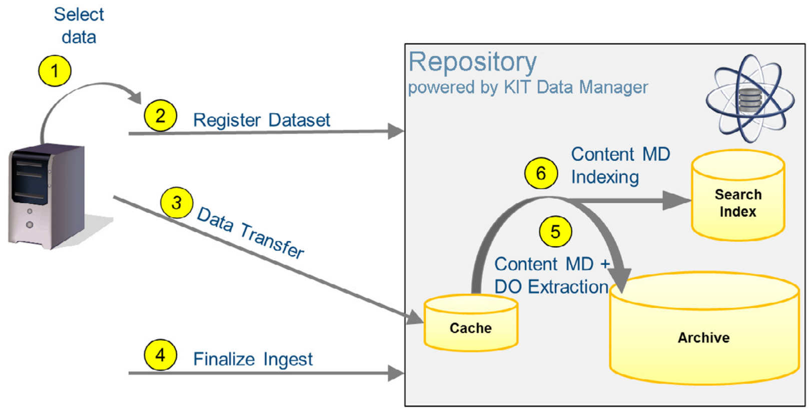


Fig. 6. The principal sequence of steps that are performed during an ingest of a digital object utilizing the KIT DM framework [44].

enabling specific adaptations and advanced features significantly reduces the average effort to provide such a repository service for a community. With traditional solutions this would require an additional repository solution instance and respective administrator training for each new use case. Based on the KIT DM, the MASi service is an advanced example of the "many use cases in one repository" approach. For additional use cases the MASi service can be specifically adapted with limited effort while the overall maintenance effort only slightly increases.

4. MASi research data management service

4.1. Overarching goals

The MASi project is building a generic and sustainable research data management service. It will be sustainably operated for communities to fully handle their data management requirements by utilizing metadata. This generic service offer is aiming at arbitrary community use cases also beyond the ones in MASi. The MASi service is based on the KIT DM repository framework that was significantly extended within the MASi project to fill capability gaps (see Section 3.1). On the one hand, this includes the creation of a generic API and backend (storage system based on a flexible data model), the MetaStore, to support modelling of heterogeneous metadata models (see Section 4.2). On the other hand, we implemented generic graphical interfaces, both on the client and web side, to fundamentally lower the effort to adapt MASi to current and new use cases (see Sections 4.3 and 4.4). Further goals are to

- enable the convenient data sharing of both restricted and Open Access data,
- automatically extract and validate metadata,
- automatically process ingested data with examples being format conversions or thumbnail creation,
- enable automatic assignment of persistent identifiers,
- establish metadata as the central information source in the data life cycle of a community use case,
- publish all MASi developments as Open Source to increase its re-use potential and sustainability and
- provide a best practice implementation guide to foster (1) KIT DM adoptions based on MASi experiences and (2) the integration of new communities into the MASi service.

Finally, we closely collaborate within the Research Data Alliance (RDA) with other data researchers to develop recommendations and we aim at implementing relevant ones.

4.2. The MetaStore metadata framework and the MASi API

In this section, we describe the multi-layered architecture of the MetaStore framework and the MASi REST API that exposes it. The MetaStore framework, both significantly extend the KIT DM framework (see Section 3.1) and are pivotal to the functionalities of the MASi service. The MetaStore is the technical implementation of the core metadata handling of the MASi service. In the following, we present the MetaStore components (Sections 4.2.1 to 4.2.7). A first application was within the Nanoscopy Open Reference Data Repository (NORDR) in the biology domain (see Section 5.2).

The metadata in MASi is being handled as METS (Metadata Encoding and Transmission Standard) [57] documents that are structured as XML. METS/XML was chosen as it is mature, well supported and flexible making it ideal for MASi as a production service aims at arbitrary communities. The metadata handled by MASi itself is split in several packages (Fig. 7). Some of the packages are very similar with the sections defined in METS. Others are allocated in a way most suitable for MASi. Each package is responsible for a special purpose (administrative, structural, content, bit preservation, provenance and annotation metadata). As not all packages are needed by every community, the structure of the METS document may slightly differ.

MetaStore is a generic framework that provides handling of metadata standards via individual schemas. It provides a dedicated metadata schema registry for uniquely registering metadata standard schemas and versions of them tracking their evolution. To enable full-text search over the metadata, MetaStore supports the automated creation of indexes during the metadata registration step that includes it into the integrated ArangoDB database. Based on the registered metadata schema, for each ingest of metadata the MetaStore performs well-formedness and syntax validation quality checks. For large-scale metadata harvesting the MetaStore is OAI complaint and supports OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

The MetaStore Service Layer is a collection of web services that comprises various components, which collectively build the functionalities of the MetaStore (Fig. 8). This layer is further divided into task specific components that individually or collectively build the functionality of the MetaStore framework. For example, to completely automate the handling of a newly registered metadata standard, the Metadata Indexer component together with the Metadata Registry and Metadata Management components provides the registration, indexing and automated creation of services for handling the newly registered metadata standard. For allowing flexible and scalable data storage, MetaStore utilizes the NoSQL database ArangoDB [58] as a core component. Via the MASi REST API, the MetaStore features are exposed to research community use cases.

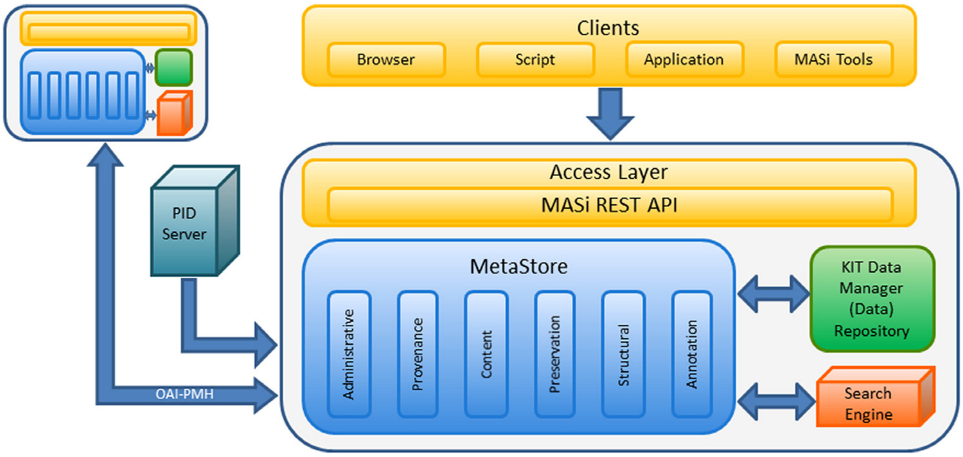


Fig. 7. The architecture of the generic MASi research data management service.

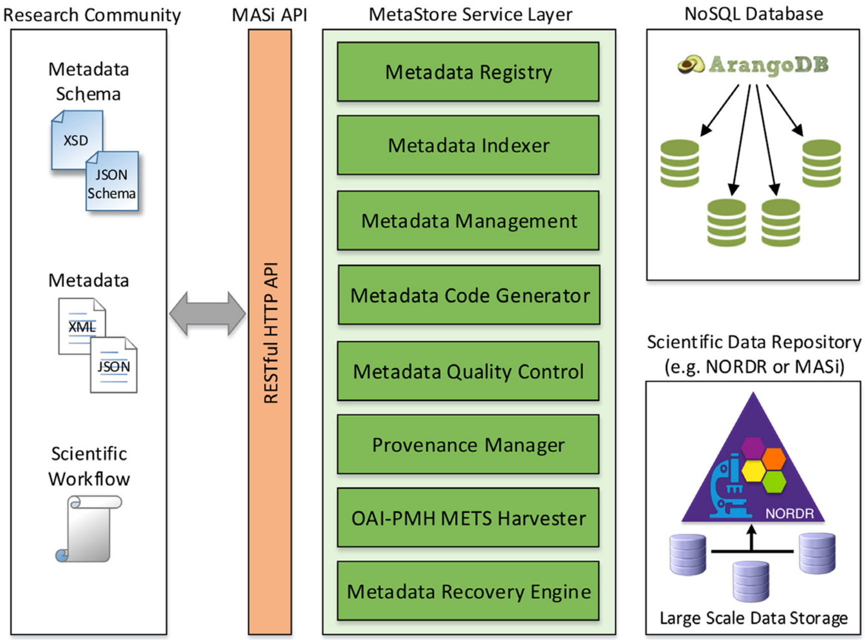


Fig. 8. The multilayered architecture of the MetaStore framework, developed in and utilized by MASi [56].

The architecture design decision to use ArangoDB as database system is motivated as follows. First, ArangoDB is a multi-model database that supports three data models namely key–value, document, and graph within a single database. Hence, metadata schema can be systematically modelled and queried using the appropriate data model. In the realization of the MetaStore framework, the descriptive metadata is modelled in the document model for allowing full-text search and executing user-defined queries, the provenance metadata involving complex relationships is structured in graph data model for allowing efficient execution of graph pattern matching queries, and the key–value model is employed for registering the metadata schema. Additionally, with the automated indexing of the entire schema for enabling full-text search, an auxiliary search engine like ElasticSearch is not required. Second, to minimize the architectural complexity of the MetaStore framework, integrating the ArangoDB offered a consolidated database solution that provides polyglot persistence. An alternative solution would be to use application specific database systems, for example, MongoDB for storing the descriptive metadata, Neo4j a graph database for storing provenance metadata, and RIAK or REDIS as a metadata registry. However, adopting these databases would not only increase the architecture complexity of the framework but will also increase the total cost of ownership in maintaining three different database systems. Third, with the utilization of ArangoDB, only a single query language needs to be adopted, i.e., the Arango Query Language (AQL) for creating queries, instead of multiple database-specific query languages. Thus, the effort required to utilize and integrate MetaStore for research communities is significantly reduced.

In the following we briefly describe the features of the MetaStore framework while details are presented in our previous work [56].

4.2.1. Metadata registry

The first step for a research community is the registration of one or multiple metadata standards that are required to describe their data. As we use ArangoDB that supports multiple data models, we leverage its support of key–value data models for registering metadata standards. For each metadata standard, the Metadata Registry creates a unique key–value pair in ArangoDB, wherein the key is the combination of the metadata standard namespace and version and the value is the complete metadata schema. Metadata schemas with multiple versions are also supported by the Metadata Registry. Moreover, the Metadata Registry enables the registration of complex metadata standards such as METS.

4.2.2. Metadata indexer

When a metadata schema is registered, a unique collection for storing the subsequent metadata is created in the document data model of the ArangoDB instance. For example, in the use case of mediaeval stained glass (see Section 2.3), where the metadata is defined based on the XMP metadata standard, an XMP collection is created for storing the metadata that is extracted from the TIFF images. On the one hand, dedicated collections provide the research communities with isolated metadata storages, but on the other hand this metadata still needs to be queried and retrieved. For this, there are two options. First, a set of queries could be implemented using the ArangoDB Query Language (AQL) for retrieving the metadata. Second, the metadata could be indexed for allowing full-text search over the entire metadata. Comparing the two options, the second options is more feasible as creating custom queries is a time consuming and labour-intensive task and any changes in the metadata schema will necessitate an update of the queries. Hence, to avoid this potential downside, the Metadata Indexer component extracts all the index-terms from the registered metadata schema and creates the indexes in the respective collections. When a metadata schema is modified the Metadata Indexer component also updates the indexes, thus, at any given instance the full-text search can be performed over the entire up-to-date metadata. Moreover, this means that the overhead of integrating and maintaining a dedicated search engine is avoided.

4.2.3. Metadata code generator

The minimum set of functions required for working with metadata are the CRUD (create, retrieve, update, delete) operations. However, it is a redundant task for creating these functions every time a new metadata schema is registered. To completely avoid the manual creation of services, the Metadata Code Generator component automatically generates these services and updates the MetaStore functionality. When a metadata schema is successfully registered, this component creates the services for handling this newly registered metadata and appends them to the already available services.

4.2.4. Metadata quality control

For enabling automated quality control, this component performs the syntax validation and well-formedness check of the metadata against its corresponding registered schema. After a successful quality control, the metadata is inserted into its assigned collection. In the case of complex metadata standard like METS, each section of the metadata is extracted, validated against the registered schema and individually inserted in the assigned collection.

4.2.5. Provenance manager

Currently, the MetaStore supports two provenance models, namely ProvONE [59] and PREMIS [60]. ProvONE allows modelling of both the prospective provenance (workflow definition) and the retrospective provenance (run-time execution events of workflows). PREMIS is the *de facto* standard supported by METS. Principally, for each execution of a workflow, the provenance is automatically captured in the ProvONE model using the Prov2ONE algorithm and stored as a graph data model in the ArangoDB [61]. However, for provenance interoperability, the ProvONE retrospective provenance is translated into the PREMIS provenance standard for enabling compatibility with the *digiProvMD* section of the METS standard.

4.2.6. OAI-PMH METS provider and harvester

For sharing metadata across data repositories, OAI-PMH is the standard protocol for enabling metadata harvesting. The MetaStore supports harvesting of metadata through the MASi METS profile. This component implements the six verbs recommended by the OAI-PMH specifications as AQL queries and exposes them through the MASi API. The OAI-PMH METS interface constructs the entire METS on the fly, i.e., the descriptive or the content metadata in the *dmdSec* section, the administrative metadata in the *amdSec* section, the technical metadata in the *techMD* section, the rights pertaining to a digital object in the *rightsMD* section and the provenance metadata in the *digiProvMD* section. They are queried from the different collections in the ArangoDB and assembled according to the MASi METS profile. This METS document is serialized as XML and enclosed within the OAI-PMH protocol.

4.2.7. Metadata recovery engine

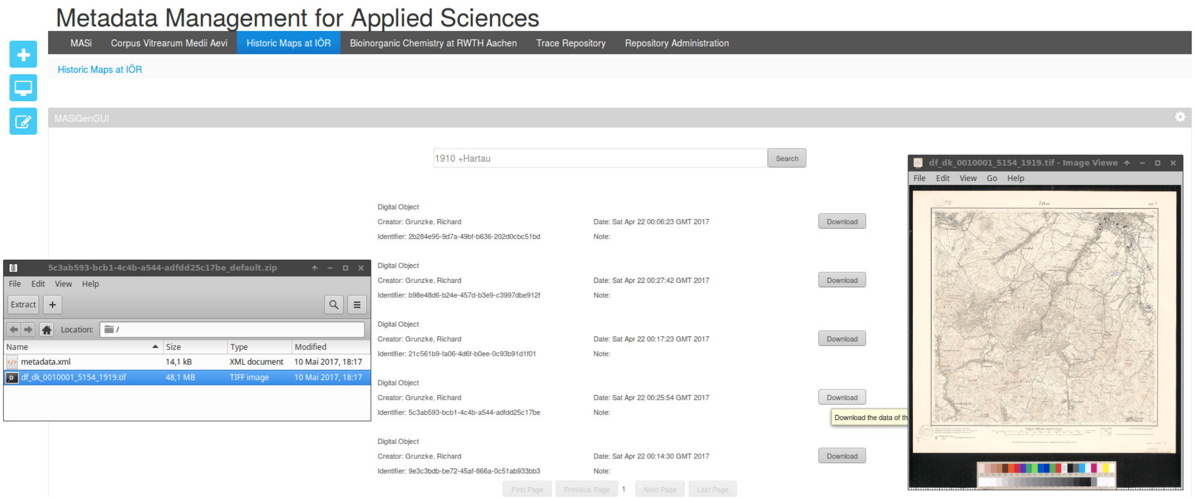
The Metadata Recovery Engine component is responsible for performing a complete recovery of the metadata storage in case the ArangoDB cluster crashes and the entire metadata is lost. Principally, this component collects all the METS files from the MASi service and reconstructs the metadata storage. Each METS file is decomposed according to the different sections it contains and the metadata is modelled in its respective data-model. For example, the descriptive metadata from the *dmdSec* section would be extracted and with prior schema registration and validation newly inserted into the ArangoDB document store. As another example, the provenance metadata comprising the workflow definition as XML and the PREMIS retrospective provenance from the *digiprovMD* section would be extracted, transformed into the ProvONE model and again stored in the ArangoDB graph store.

4.3. Generic graphical web interface

We developed a generic web interface for MASi to free user communities from completely re-developing new user interfaces for every new user case. The MASi user interface in its unaltered state can either be directly utilized without any development effort at all. Or, it can be used as a basis to optionally extend it to suit the individual use case requirements. Both saves time and significantly lowers the barrier of entry by significantly lowering the time and effort required to integrate the MASi service with further use cases.

As first part, the generic web interface was built on the basis of the Liferay portal framework [62] (Fig. 9) that provides ready-made capabilities such as plugins, menus, groups, roles, separable areas and user authentication management with the integration of systems such as LDAP. This enables the integration of identity federations such as DFN-AAI [63] in Germany and eduGAIN [64] in Europe for the easy and safe re-use of existing and trusted institute logins. Liferay is Open Source, mature and widely used. It is, for example, utilized as a fundamental building block for various Science Gateways [65,66]. For identity management to seamlessly work within the MASi service, we are currently developing an extension to integrate the Liferay user and group management with the one of the KIT DM repository framework. The extension will ensure the consistency between the user management systems of Liferay and MASi by automatically synchronizing new users. This integration will enable the MASi service to transparently support authentication systems that are already supported by Liferay.

The second main part is the development of a generic graphical interface portlet with common functionality. This includes basic search and download capabilities besides the following further features that shall be available in the MASi production service.


Metadata Management for Applied Sciences
MASi
Corpus Vitrearum Medii Aevi
Historic Maps at IÖR
Bioinorganic Chemistry at RWTH Aachen
Trace Repository
Repository Administration
Historic Maps at IÖR
MASiGenGUI
Search
Download
Download the data of th
First Page
Previous Page
1
Next Page
Last Page
5c3ab593-bcb1-4c4b-a544-adfdd25c17be_default.zip
File Edit View Help
Extract
Location:
df_dk_0010001_5154_1919.tif - Image Viewe
File Edit View Go Help

1910 +Hartau

Digital Object
Creator: Grunzke, Richard
Identifier: 2b284e95-9d7a-49bf-b636-202d0cbc51bd
Date: Sat Apr 22 00:06:23 GMT 2017
Note:

Digital Object
Creator: Grunzke, Richard
Identifier: b98e48d6-b24e-457d-b3e9-c3997dbe912f
Date: Sat Apr 22 00:27:42 GMT 2017
Note:

Digital Object
Creator: Grunzke, Richard
Identifier: 21c561b9-fa06-4d6f-b0ee-0c93b91d1f01
Date: Sat Apr 22 00:17:23 GMT 2017
Note:

Digital Object
Creator: Grunzke, Richard
Identifier: 5c3ab593-bcb1-4c4b-a544-adfdd25c17be
Date: Sat Apr 22 00:25:54 GMT 2017
Note:

Digital Object
Creator: Grunzke, Richard
Identifier: 9e3c3bdb-be72-45af-866a-0c51ab933bb3
Date: Sat Apr 22 00:14:30 GMT 2017
Note:

Name	Size	Type	Modified
metadata.xml	14,1 kB	XML document	10 Mai 2017, 18:17
df_dk_0010001_5154_1919.tif	48,1 MB	TIFF image	10 Mai 2017, 18:17

Fig. 9. The web interface and current pre-production GUI of the MASi service showing a search within the historic map use case data with the map downloaded and displayed.

- The integration with the MetaStore component enables advanced search features, viewing and editing of metadata and directly incorporating metadata updates into an search index. See Section 4.2 for further MetaStore details and features.
- A requirement from all communities is a web-based metadata viewer and editor. Building on top of the MetaStore backend, it shall be based on the Metawidget framework [67] that enables the automatic generation of user interfaces based on objects in various forms.
- The functionality to display thumbnails of images in the search results is required by all communities as well. After clicking on that thumbnail a full-sized version of the image shall open. This feature significantly enhances the usefulness of the search as it better enables users to judge the relevancy of search results.
- The ability to redirect users to the original source of the data instead of downloading the digital object. This is a requirement from the historic maps use case, where it is allowed to ingest, manage and enable searching through the original data but only to directly download it as administrative user.
- The capability to download a file with a list of links to data resulting from a search. With this list users can easily download digital objects in large numbers at once to systems such as their workstation or a supercomputer.

The generic web interface is currently being prepared for release as Open Source. This includes implementing the just listed features and increasing the maintainability of the code by adding detailed comments, cleaning up and structuring the code as clearly as possible. To fundamentally increase the impact of these developments, we will create an extensive documentation to enable developers to easily adapt the portlet for their specific use case requirements. The documentation will include everything from code checkout, development project configuration, adaptation examples to compilation. A main goal of the extensive documentation is to shorten the training period as much as possible and significantly foster re-usability. All binaries, source code and documentation will be Open Source and will become part of the KIT DM framework. In Fig. 9 the current pre-production GUI within the MASi service is depicted.

After release of the production version of the MASi service in August 2017, the generic GUI will be further extended. Possible candidate features include the ability to search according to geographical coordinates that are specified either manually or by visually defining a geographical bounding box on a map. The results are then the objects that semantically reside within that geographical area. Currently, digital objects are uploaded via the commandline or the graphical interface of the GRC. We envision to enable ingesting via the web client as well. The feasibility shall be explored as uploading a directory or multiple files via a web interface is not straight-forward. Furthermore, ways to visualize provenance metadata shall be explored.

4.4. Graphical user interface for manual metadata input on the client

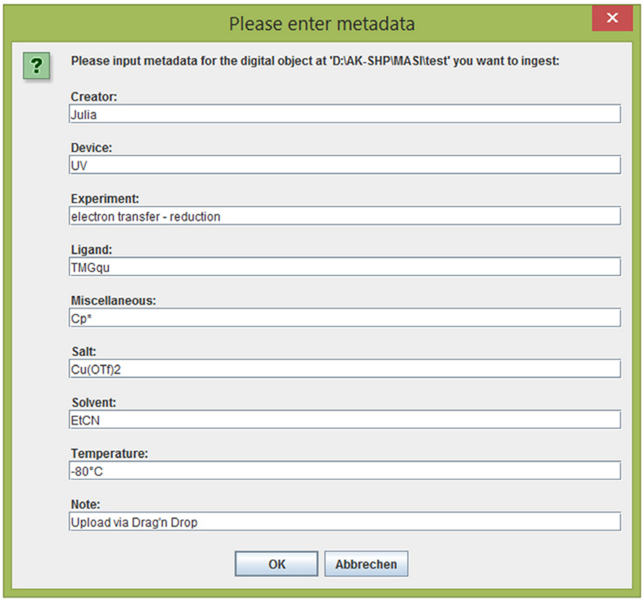
In MASi the modus operandi is to automate the extraction of metadata, besides other things. The reasons are multitude. First, this is for the convenience of the users to save the labour of manually entering metadata. Second, automation is an absolute necessity when metadata of thousands or millions of files need to be included. Third, in the case when information needs to be attached as metadata but the information is only available by processing the data via appropriate algorithms.

However, in some cases metadata cannot be automatically extracted as the information is not present in the data itself. This holds true for example when a measurement device is operated by a person and the name of the person is not entered into the device. Thus, the metadata written by the device does not contain the name of the operator. This is the case in the chemistry use case (see Section 2.2). Thus, such metadata has to be manually given by the scientist before the data is ingested into the MASi repository.

This can be configured as mandatory, meaning that if the metadata is not entered on the client side by the scientist the ingest fails. To efficiently support this we extended the GRC by a graphical user interface to enable users to manually enter metadata, which only they know and that is unavailable by other means. By utilizing the GRC plugin interface, it was straight forward to create the user interface. See Section 2.2 for a use case in chemistry and Fig. 10 for a screenshot. After the metadata is entered it is stored in an XML file besides the data and ingested together with the data. On the server, the metadata is then automatically extracted from the XML file and further handled in order to be available for searching purposes. The code of the GUI will be part of the upcoming GRC release. This way, adapting the existing plugin to create further plugins for other use cases is a simple matter.

4.5. Further features

Further significant developments within MASi include the following capabilities that are seamlessly integrated with the MASi service;



Please input metadata for the digital object at 'D:\AK-SHP\MASI\test' you want to ingest:

Creator: Julia

Device: UV

Experiment: electron transfer - reduction

Ligand: TMGqu

Miscellaneous: Cp*

Salt: Cu(OTf)2

Solvent: EtCN

Temperature: -80°C

Note: Upload via Drag'n Drop

Fig. 10. The GUI within the chemistry use case for the exceptional case of manually entering metadata.

- Metadata extractors for the XML [68], XMP/TIFF [69] and HDF5 [70] data formats and Apache Tika toolkit [71] with its support for extracting metadata from numerous data formats [72].
- The ability to directly publish data as Open Access including a landing page by simple assigning it to the respective group.
- An XSD tool that enables a community to graphically create well-formed XML schemas to organize their metadata based on a list of key–value pairs. It is included in the MASi service as well as being downloadable as Open Source soon.
- Data and its metadata can be uniquely identified via persistent identifiers (PIDs). The PIDs can be custom ones as long the data is only available internally. Functionality was implemented so that as soon as the metadata is available to a wider circle, a PID service such as the Handle system can be queried to get a worldwide unique PID.

4.6. Service operation and use case integration

The MASi service will be professionally and sustainably run in the long-term with the assurance that users can rely on it to safely and efficiently manage their data. Currently, it is in pre-production state and due to go public in the second half of 2017 as a service at the TU Dresden [3]. In order to use the service every user has to accept the terms of use that define the access, obligations, acceptable content, privacy, blocking and compensation.

The service currently runs on a virtual machine with 24 core (E5-2620 v3, 2.40 GHz), 64 GB RAM and 20 TB of initial storage. An additional VM for development and testing is maintained in parallel. For a first HPC integration the GRC is installed on the ZIH supercomputer Taurus [73].

A quota feature enables that every user can only use a defined amount of storage space. This ensures the stability of the service when the user managed data amount grows. The storage space can then safely be increased and subsequently new users can be accepted or the quota of existing user can be increased. The quota functionality also ensures the stability in case a user runs an erroneous ingest script. Then instead of filling up the whole storage space of the service and rendering it non-functional, the script can only fill up the allotted storage space of the user. A default quota for every user will be active.

The internal access rights management is group based. If two users are in the same group they can access each others files. The reason is that research groups internally often share data per default. If that is not the case, either a separate group can be used or the group access rights can be restricted. Security is provided via the use of the HTTPS protocol for encrypting network communication and considering the IT baseline protection guidelines of the German Federal Office for Information Security (FSI). Besides an administration documentation, a backup and monitoring concept for storage, virtual machines, configurations, database and index will ensure the reliability of the service.

Everybody with an institutional login at the TU Dresden will be able to login and use the service via an identity management integration (either via LDAP or Shibboleth). At a later stage access for all German researchers is planned to be enabled via the DFN-AAI identity provider federation. This automatic access is for convenient testing with standard metadata extraction such as Apache Tika [71] and getting acquainted with the service. Basic community adaptations such as creating new metadata extraction plugins or modifying the graphical user interface according to community requirements will be either performed by the MASi project itself or via collaborations with other projects. When extensive and time-consuming adaptations are required, the model is to see about acquiring dedicated funding to perform the respective design and implementation.

5. Evaluation and further use cases

Here, it is evaluated how MASi supports the research done in the MASi use cases. This was enabled by filling the KIT DM capability gaps listed in Section 3.1. Furthermore, other use cases are detailed that currently use or plan to use MASi and KIT DM to support them.

5.1. Initial use cases

In the historical maps use case (see Section 2.1), the MASi service significantly supports the aim of reproducible work via its provenance capabilities. At the same time it lowers the obstacles of interconnection between scientific communities. It provides a central access point while it reduces the limitations caused by incompatibilities of metadata. Hence, historical maps and the derived land use information can be safely stored in a well-structured way and also easily accessed by other disciplines (e.g. history, planning and ecology) that share a spatial interest. The metadata-driven and query-able storage of data in MASi furthermore assists well-scalable map processing for large volumes of data and thus contributes to the extensive availability of the resulting historical land use information.

In the chemistry use case (see Section 2.2) data produced by a spectroscopy measurement device is stored and managed while a GUI to ingest the produced collection of data (ksd, csv, xls1, xls2, opj, png) is offered. Via the graphical interface (Fig. 10) the researchers manually define the following metadata upon ingesting: name of operator, type of device and of measurement, ligand, metal salt, solvent and temperature. Further two fields are possible to be filled for more details; here for instance the reductant used. With these metadata, the ingested data can be easily found via the MASi search interface. Also important for the chemistry use case, MASi enables to define access rights to enable distinctions between groups of people and the publication of data sets as Open Access data via the OAI-PMH interface. The next steps here are the integrating of further devices and support for simulation data. The aim is also to make it easy to correlate experimental and theoretical data sets by making them accessible in the same way via the MASi service.

The MASi service is highly beneficial to the use case of mediaeval stained glass (see Section 2.3) in various aspects. It provides (1) much easier sharing capabilities of the images with other repositories such as Foto Marburg or EUROPEANA via the OAI-PMH interface, (2) a long term repository infrastructure, (3) an advanced data ingest service with automatic metadata validation that will significantly increase the metadata quality, (4) a REST API to access images from within other applications and (5) a web-based metadata editor that will also be of great benefit to other projects in the humanities domain and beyond.

MASi is inherently cross-disciplinary, meaning that synergy effects with other disciplines can easier be realized. A example related is between the historic maps and mediaeval stained glass use cases (see also Sections 2.1 and 2.3) that both have a geographical context. The ability to access historical maps via the same search as mediaeval stained glass has the potential to open up new avenues to knowledge. From the perspective of the mediaeval stained glass use case, information about the geographical context and the date of creation of each pane is very important. In order to be as precise as possible, the GPS coordinates and time information of each church or building are embedded into the images. Taken together, this will allows to create spatio-temporal analyses either on the full range or on smaller sets of images. Questions like what type of iconographic content was predominant at what time in a given area can then be visualized and analysed using common tools from the digital humanities (like the DARIAH-Geobrowser or Stanford's Palladio). In addition to the GPS information, each image also contains two geographic identifiers: one for the current location and one for the original location. Both identifiers are URIs from Geonames [74]. The use of Geonames URIs will allow to connect the images with normdata repositories and other place identifiers via a Linked Open Data (LOD). Thus, the data becomes available for exchange and further research. Via the *seeAlso* property from RDF Schema, each CVMA Geonames URI is connected to other identifiers from national and international authority files such as the Virtual International Authority File [75], the Gemeinsame Normdatei (Integrated Authority File) [76] and many more.

5.2. Further use cases

The NFFA [77] H2020 project currently establishes a distributed infrastructure for 24 large-scale nanoscience facilities across Europa. It supports data access, publication of data and the integration of various data systems (ICAT, KIT DM, NoMaD, iRODS, ...). It uses and extends MASi and the KIT DM in order to build an overarching search and access layer.

The NORDR [48] nanoscopy repository is being build up in a cooperation with the LSDMA (Large-Scale Data Management and Analysis) project, the Kirchhoff-Institute for Physics (KIP) in Heidelberg and the Institute of Molecular Biology (IMB) in Mainz. NORDR is data repository specifically built for handling extremely large datasets produced by nanoscopy investigations in biology (each investigation generates 150–200 terabytes of data) [48]. NORDR builds on the KIT DM and MASi developments to provide a variety of advanced services. For example, specific metadata extraction modules and algorithms for processing data on high-performance computing clusters are deployed.

The DFG Collaborative Research Centre (CRC) 980: "Episteme in Motion - Transfer of Knowledge from the Ancient World to the Early Modern Period" [78] investigates the processes of knowledge transfer in European and non-European cultures. In the information infrastructure project, MASi and KIT DM will be used to build up a research data management and analysis infrastructure. eCodicology [79] is a project about the "development, testing and optimization" of "algorithms for the automatic tagging of mediaeval manuscripts". Therein, dynamically generated metadata will be associated with digital objects and their relationships investigated. eCodicology was previously funded by the German Federal Ministry of Education and Research and is currently a guest project in the CRC 980.

The goal of the CRC 940 "Volition and Cognitive Control: Mechanisms, Modulators and Dysfunctions" [80] "is to elucidate cognitive and neural mechanisms underlying adaptive volitional control as well as impaired control in selected mental disorders". In its information infrastructure project MASi and KIT DM are evaluated to be used to manage research data and integrate various data management systems to enable overarching search and access capabilities. The aim is to provide an overall system that is easy to use, highly integrated, efficient and metadata-driven to manage projects, subjects and research data (MRI, EEG, genetic and behavioural data) as well being integrated with high performance computing resources for data analysis.

Performance analysts of highly parallel applications are recording event traces of the runtime of such an application to analyse them and optimized their performance. Numerous and complex of such traces are created. To easily find and access these according to specific characteristics, the Trace Repository [81] within the MASi service was created. "With this, teams working with many very large trace data sets become able to organize them in a structured and collaborative manner and identify interesting ones while avoiding the need to scan them in their entirety again and again" [81].

The GeRDI [82,83] DFG project aims at building a German-wide interconnected research data management infrastructure including search capabilities for all German research data repositories with additional added-value services. MASi and the KIT DM are being evaluated as a recommended reference implementation for the repository functionality in GeRDI to handle arbitrary research data.

6. Conclusion and outlook

The MASi research data management service provides a solution for the highly relevant challenge of managing large amounts of complex data. It builds on our substantial previous work, the KIT Data Manager research data management framework that was further extended and broadened in MASi based on the use case requirements. The framework was substantially improved by the MetaStore component with its API to enable the integration of arbitrary metadata models and web- and client-GUIs to significantly ease the use of MASi besides other features. The MASi service is currently in pre-production state and will go into production in the second half of 2017 [3]. MASi is able to seamlessly integrate with highly diverse use cases in geography, chemistry and the digital humanities, besides others. It is able to safely and efficiently manage their data, automatically annotate it with metadata and make the data accessible by various means.

As future work, we are planing the integration of MASi with the UNICORE HPC middleware [84] and to further develop and extend the existing use cases and integrate further ones. The service shall be consolidated, maintained and extended. The KIT DM repository framework is planned to be further extended to provide even more advanced features such as integrated ontologies, schema hierarchies, controlled vocabularies and support for GridFTP [85] and UFTP [86] high-performance file transfer methods. To increase the interoperability of the MASi service and add to the existing OAI-PMH support we aim at evaluating the ResourceSync protocol [87] for implementation. We also plan to formalize and extend support for defined policies (such as publishing data as Open Access after an embargo period) and their (semi-)automatic enforcement. The

MASi service itself will directly benefit from these developments. We will importantly continue to work within RDA and contribute our own expertise to create RDA recommendations on how to best handle various aspects of research data management. At the same time, we aim at implementing the resulting joint recommendations within MASi. Possible examples are the support for collections of digital objects, the migration of digital objects between repositories and a metadata schema catalogue. This will enable us to benefit from work done within RDA and at the same time it benefits RDA and connected communities as we will feed back our adoption experiences. This future work will contribute to the extension of the MASi research data management service to be efficient, future-proof and with a high user acceptance.

Acknowledgements

This work was supported by the German Research Foundation via the MASi project (NA 711/9-1, GE 2481/1-1, HE 5480/11-1, ME 1592/4-1, STO 397/4-1) and the Helmholtz Association of German Research Centres via the LSDMA project.

References

[1] B. Landesman, Seeing standards: A visualization of the metadata universe, Tech Serv. Q. 28 (4) (2011) 459–460. http://dx.doi.org/10.1080/07317131.2011.598072.

[2] R. Grunzke, V. Hartmann, T. Jejkal, A. Prabhune, S. Herres-Pawlis, A. Hoffmann, A. Deicke, T. Schrade, H. Herold, G. Meinel, R. Stotzka, W.E. Nagel, Towards a metadata-driven multi-community research data management service, in: Proc. of IWSG 2016 (8th International Workshop on Science Gateways), 8–10 June 2016, Rome, Italy, 2017. http://ceur-ws.org/Vol-1871/paper12.pdf.

[3] The MASi research data management service, 2017. https://masi.zih.tu-dresden.de.

[4] G. Meinel, R. Hecht, H. Herold, Analyzing building stock using topographic maps and GIS, Build. Res. Inf. 37 (5–6) (2009) 468–482.

[5] H. Herold, G. Meinel, R. Hecht, E. Csaplovics, A GEOBIA approach to map interpretation-multitemporal building footprint retrieval for high resolution monitoring of spatial urban dynamics, in: International Conference on Geographic Object-Based Image Analysis, 2012, pp. 252–256.

[6] Old maps online - The search engine for historical maps, 2017. http://www.oldmapsonline.org.

[7] Deutsche fotothek kartenforum, 2017. http://www.deutschefotothek.de/kartenforum.

[8] P. Röhm, H. Herold, G. Meinel, Automatische georeferenzierung gescannter deutscher topographischer karten im maßstab 1: 25.000, Kartographische Nachrichten 62 (4) (2012) 195–199.

[9] R. Hecht, H. Herold, G. Meinel, M. Buchroithner, Automatic derivation of urban structure types from topographic maps by means of image analysis and machine learning, in: 26th International Cartographic Conference, 2013.

[10] S. Muhs, H. Herold, G. Meinel, D. Burghardt, O. Kretschmer, Automatic delineation of built-up area at urban block level from topographic maps, Comput. Environ. Urban Syst. 58 (2016) 71–84.

[11] D. Schemala, D. Schlesinger, P. Winkler, H. Herold, G. Meinel, Semantic segmentation of settlement patterns in gray-scale map images using RF and CRF within an HPC environment, in: GEOBIA 2016: Solutions and Synergies, 2016.

[12] Commission high level expert group on the european open science cloud - Realising the european open science cloud, 2016. http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf.

[13] Directive 2007/2/EC of the european parliament and of the council of 14 March 2007 establishing an infrastructure for spatial information in the european community (INSPIRE), Official J. Eur. Union, L 108(1) (2007) 50.

[14] ISO, ISO 19115-2: 2009 Geographic Information—Metadata—Part 1: Fundamentals, 2014.

[15] Virtuelles Kartenforum 2.0, 2017. http://kartenforum.slub-dresden.de.

[16] R. Bill, K. Walter, Crowdsourcing zur Georeferenzierung alter topographischer Karten–Ansatz, Erfahrungen und Qualitätsanalyse, Zeitschrift für Geodäsie, Geoinformation und Landmanagement http://dx.doi.org/10.12902/zfv-0060-2015.

[17] F. Warmerdam, The geospatial data abstraction library, in: Open Source Approaches in Spatial Data HandLing, Springer, 2008, pp. 87–104.

[18] M. Haklay, P. Weber, OpenStreetMap: user-generated street maps, IEEE Pervas. Comput. 7 (4) (2008) 12–18.

[19] Overpass API public instance, 2017. http://overpass-api.de/.

[20] R.M. Olbricht, Data retrieval for small spatial regions in openstreetmap, in: OpenStreetMap in GIScience, Springer, 2015, pp. 101–122.

[21] ISO, ISO 19115-2: 2009 Geographic information—metadata—Part 2: Extensions for imagery and gridded data, 2009.

[22] J. Mize, C.F. Robertson, A solution to metadata: Using XML transformations to automate metadata, in: OCEANS 2009 , MTS/IEEE Biloxi-Marine Technology for Our Future: Global and Local Challenges, IEEE, 2009, pp. 1–7.

[23] L. Di, P. Yue, H.K. Ramapriyan, R.L. King, Geoscience data provenance: An overview, IEEE Trans. Geosci. Remote Sens. 51 (11) (2013) 5065–5072.

[24] L. Di, Y. Shao, L. Kang, Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 Lineage Model, IEEE Trans. Geosci. Remote Sens. 51 (11) (2013) 5082–5089.

[25] C. Henzen, S. Mäs, L. Bernard, Provenance information in geodata infrastructures, in: Geographic Information Science at the Heart of Europe, Springer, 2013, pp. 133–151.

[26] J. Liu, S. Chakraborty, P. Hosseinzadeh, Y. Yu, S. Tian, I. Petrik, A. Bhagi, Y. Lu, Metalloproteins containing cytochrome, iron–sulfur, or copper redox centers, Chem. Rev. 114 (8) (2014) 4366–4469.

[27] A. Hoffmann, S. Binder, A. Jesser, R. Haase, U. Flörke, M. Gnida, M. Salomon Stagni, W. Meyer-Klaucke, B. Lebsanft, L.E. Grünig, S. Schneider, M. Hashemi, A. Goos, A. Wetzel, M. Rübhausen, S. Herres-Pawlis, Catching an entatic state—A pair of copper complexes, Angew. Chem., Int. Ed. 53 (1) (2014) 299–304. http://dx.doi.org/10.1002/anie.201306061.

[28] A. Hoffmann, J. Stanek, B. Dicke, L. Peters, B. Grimm-Lebsanft, A. Wetzel, A. Jesser, M. Bauer, M. Gnida, W. Meyer-Klaucke, M. Rübhausen, S. Herres-Pawlis, Implications of guanidine substitution on copper complexes as entatic-state models, Eur. J. Inorg. Chem. 2016 (29) (2016) 4731–4743. http://dx.doi.org/10.1002/ejic.201600655.

[29] P.M. Edwards, Origin 7.0: Scientific graphing and data analysis software, J. Chem. Inf. Comput. Sci. 42 (5) (2002) 1270–1271.

[30] A. Hoffmann, R. Grunzke, S. Herres-Pawlis, Insights into the influence of dispersion correction in the theoretical treatment of guanidine-quinoline copper(i) complexes, J. Comput. Chem. 35 (27) (2014) 1943–1950. http://dx.doi.org/10.1002/jcc.23706.

[31] A. Jesser, M. Rohrmüller, W.G. Schmidt, S. Herres-Pawlis, Geometrical and optical benchmarking of copper guanidine–quinoline complexes: Insights from TD-DFT and many-body perturbation theory, J. Comput. Chem. 35 (1) (2014) 1–17. http://dx.doi.org/10.1002/jcc.23449.

[32] A. Hoffmann, M. Rohrmüller, A. Jesser, I.d.S. Vieira, W.G. Schmidt, S. Herres-Pawlis, Corrigendum: Geometrical and optical benchmarking of copper(II) guanidine–quinoline complexes: Insights from TD-DFT and many-body perturbation theory (Part II), J. Comput. Chem. 36 (4) (2015). http://dx.doi.org/10.1002/jcc.23793. 272–272.

[33] F. Ogliaro, M. Bearpark, J. Heyd, E. Brothers, K. Kudin, V. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, et al., Gaussian 09, revision a. 02. Gaussian, Wallingford, CT, Inc.

[34] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, et al., Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations, Comput. Phys. Comm. 181 (9) (2010) 1477–1489.

[35] R. Grunzke, S. Breuers, S. Gesing, S. Herres-Pawlis, M. Kruse, D. Blunk, L. de la Garza, L. Packschies, P. Schäfer, C. Schärfe, T. Schlemmer, T. Steinke, B. Schuller, R. Müller-Pfefferkorn, R. Jäkel, W.E. Nagel, M. Atkinson, J. Krüger, Standards-based metadata management for molecular simulations, Concurr. Comput.: Pract. Exper. 26 (10) (2014) 1744–1759. http://dx.doi.org/10.1002/cpe.3116.

[36] S. Herres-Pawlis, A. Hoffmann, A. Balaskó, P. Kacsuk, G. Birkenheuer, A. Brinkmann, L. de la Garza, J. Krüger, S. Gesing, R. Grunzke, G. Terstyansky, N. Weingarten, Quantum chemical meta-workflows in mosgrid, Concurr. Comput.: Pract. Exper. 27 (2) (2015) 344–357. http://dx.doi.org/10.1002/cpe.3292.

[37] J. Arshad, A. Hoffmann, S. Gesing, R. Grunzke, J. Krüger, T. Kiss, S. Herres-Pawlis, G. Terstyanszky, Multi-level meta-workflows: new concept for regularly occurring tasks in quantum chemistry, J. Cheminformat. 8 (1) (2016). http://dx.doi.org/10.1186/s13321-016-0169-8.

[38] Corpus vitrearum deutschland, http://www.corpusvitrearum.de/.

[39] Corpus vitrearum deutschland publications. http://www.corpusvitrearum.de/projekt/publikationen.html.

[40] Corpus vitrearum international digitales bildarchiv, 2017. http://www.corpusvitrearum.de/cvma-digital/bildarchiv.html.

[41] Corpus vitrearum deutschland XMP specification. http://www.corpusvitrearum.de/cvma-digital/spezifikationen/cvma-xmp/1.1.html.

[42] The iconclass classification system. http://www.iconclass.nl/.

[43] P. Harvey, ExifTool: Read, write and edit meta information, Software package available at https://www.sno.phy.queensu.ca/~phil/exiftool/.

[44] KIT data manager manual, 2017. http://datamanager.kit.edu/dama/manual/index.html.

[45] R. Grunzke, J. Krüger, S. Gesing, S. Herres-Pawlis, A. Hoffmann, A. Aguilera, W.E. Nagel, Managing complexity in distributed data life cycles enhancing scientific discovery, in: IEEE 11th International Conference on e-Science, 2015, pp. 371–380. http://dx.doi.org/10.1109/eScience.2015.72.

[46] T. Jejkal, A. Vondrous, A. Kopmann, R. Stotzka, V. Hartmann, KIT data manager: the repository architecture enabling cross-disciplinary research, in: Large-Scale Data Management and Analysis - Big Data in Science - 1st Edition, 2014, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000043270.

[47] KIT Data Manager (Website), 2017. http://datamanager.kit.edu.

[48] A. Prabhune, R. Stotzka, T. Jejkal, V. Hartmann, M. Bach, E. Schmitt, M. Hausmann, J. Hesser, An optimized generic client service api for managing large datasets within a data repository, in: 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015, pp. 44–51. http://dx.doi.org/10.1109/BigDataService.2015.25.

[49] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, J.H. Walker, DSpace: An open source dynamic digital repository. http://hdl.handle.net/1721.1/29465.

[50] C. Lagoze, S. Payette, E. Shin, C. Wilper, Fedora: an architecture for complex objects and their relationships, Int. J Digit. Libr. 6 (2) (2006) 124–138.

[51] D. Lecarpentier, P. Wittenburg, W. Elbers, A. Michelini, R. Kanso, P. Coveney, R. Baxter, Eudat: a new cross-disciplinary data infrastructure for science, Int J. Digit. Curation 8 (1) (2013) 279–287.

[52] EUDAT, EUDAT semantics working group, 2017. http://eudat.eu/semantics.

[53] A. Rajasekar, R. Moore, C.-y. Hou, C.A. Lee, R. Marciano, A. de Torcy, M. Wan, W. Schroeder, S.-Y. Chen, L. Gilbert, et al., iRODS primer: Integrated rule-oriented data system, Synth. Lect. Inf. Concepts Retr. Serv. 2 (1) (2010) 1–143.

[54] D. Flannery, B. Matthews, T. Griffin, J. Bicarregui, M. Gleaves, L. Lerusse, R. Downing, A. Ashton, S. Sufi, G. Drinkwater, et al., Icat: integrating data infrastructure for facilities based science, in: E-Science, 2009 E-Science'09 Fifth IEEE International Conference on, IEEE, 2009, pp. 201–207.

[55] R. Grunzke, J. Hesser, J. Starek, N. Kepper, S. Gesing, M. Hardt, V. Hartmann, S. Kindermann, J. Potthoff, M. Hausmann, R. Müller-Pfefferkorn, R. Jäkel, Device-driven metadata management solutions for scientific big data use cases, in: 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2014, 2014. http://dx.doi.org/10.1109/PDP.2014.119.

[56] A. Prabhune, H. Ansari, A. Keshav, R. Stotzka, M. Gertz, J. Hesser, Metastore: A metadata framework for scientific data repositories, in: Big Data (Big Data), 2016 IEEE International Conference on, IEEE, 2016, pp. 3026–3035.

[57] M.V. Cundiff, An introduction to the metadata encoding and transmission standard (METS), Library Hi Tech 22 (1) (2004) 52–64.

[58] ArangoDB, Production ready highly available Multi-Model NoSQL database, 2017. https://www.arangodb.com/.

[59] V. Cuevas-Vicenttín, B. Ludäscher, P. Missier, K. Belhajjame, F. Chirigati, Y. Wei, S. Dey, P. Kianmajd, D. Koop, S. Bowers, I. Altintas, ProvONE: A PROV extension data model for scientific workflow provenance, in: DataONE Provenance Working Group, 2014.

[60] PREMIS, The PREMIS provenance model, 2017. http://www.loc.gov/standards/premis/.

[61] A. Prabhune, A. Zweig, R. Stotzka, M. Gertz, J. Hesser, Prov2ONE: An algorithm for automatically constructing ProvONE provenance graphs, in: International Provenance and Annotation Workshop, Springer, 2016, pp. 204–208.

[62] Liferay, Enterprise open source portal and collaboration software. http://www.liferay.com/.

[63] DFN-AAI - authentication and authorization infrastructure, 2017. https://www.aai.dfn.de/.

[64] Geant, eduGAIN - Interconnecting federations to link services and users worldwide. http://www.geant.net/service/eduGAIN/Pages/home.aspx.

[65] P. Kacsuk, Science Gateways for Distributed Computing Infrastructures, Springer, 2014.

[66] J. Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W.E. Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K.D. Warzecha, A. Zink, S. Herres-Pawlis, The mosgrid science gateway - a complete solution for molecular simulations, J. Chem. Theory Comput. 10 (6) (2014) 2232–2245. http://dx.doi.org/10.1021/ct500159h.

[67] Metawidget, 2017. http://metawidget.org/.

[68] T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, F. Yergeau, Extensible markup language (XML), World Wide Web J. 2 (4) (1997) 27–66.

[69] Adobe, Extensible metadata platform (XMP), 2017. http://www.adobe.com/products/xmp.html.

[70] M. Folk, A. Cheng, K. Yates, HDF5: A file format and I/O library for high performance computing applications, in: Proceedings of Supercomputing, 1999.

[71] C. Mattmann, J. Zitting, Tika in Action, Manning Publications Co., 2011.

[72] A. Tika, Supported Document Formats, 2017. https://tika.apache.org/1.14/formats.html.

[73] Taurus supercomputer at ZIH, 2017. https://tu-dresden.de/zih/hochleistungsrechnen/hpc.

[74] The GeoNames geographical database, http://www.geonames.org/.

[75] Virtual international authority file, 2017. https://viaf.org/.

[76] Gemeinsame normdatei, 2017. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html.

[77] NFFA - Nanoscience Foundries & Fine Analysis, 2017. http://www.nffa.eu/.

[78] CRC 980 - Episteme in motion. Transfer of knowledge from the ancient world to the early modern period, 2017. http://www.sfb-episteme.de/.

[79] eCodicology - Algorithms for the automatic tagging of medieval manuscripts, 2017. http://www.ecodicology.org.

[80] CRC 940 - Volition and cognitive control: Mechanisms, modulators and dysfunctions, 2017. http://sfb940.de.

[81] R. Grunzke, M. Neumann, T. Ilsche, V. Hartmann, T. Jejkal, R. Stotzka, A. Knüpfer, W.E. Nagel, Design evaluation of a performance analysis trace repository, in: Tools for Program Development and Analysis in Computational Science, 2017, in press, https://doi.org/10.1016/j.procs.2017.05.190.

[82] R. Grunzke, T. Adolph, C. Biardzki, A. Bode, T. Borst, H.-J. Bungartz, A. Busch, A. Frank, C. Grimm, W. Hasselbring, A. Kazakova, A. Latif, F. Limani, M. Neumann, N.T. d. Sousa, J. Tendel, I. Thomsen, K. Tochtermann, R. Müller-Pfefferkorn, W.E. Nagel, Challenges in creating a sustainable generic research data infrastructure, in: 4th Collaborative Workshop on Evolution and Maintenance of Long-Living Software Systems, 2017, in press, http://pi.informatik.uni-siegen.de/gi/stt/37_2/03_Technische_Beitraege/EMLS2017/paper-4-grunzke.pdf.

[83] GeRDI - Generic Research Data Infrastructure, 2017. http://www.gerdi-project.de/.

[84] K. Benedyczak, B. Schuller, M. Petrova, J. Rybicki, R. Grunzke, UNICORE 7 - Middleware services for distributed and federated computing, in: International Conference on High Performance Computing Simulation, HPCS, 2016. http://dx.doi.org/10.1109/HPCSim.2016.7568392.

[85] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, I. Foster, The Globus striped GridFTP framework and server, in: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing, IEEE Computer Society, 2005, p. 54.

[86] B. Schuller, T. Pohlmann, UFTP: High-performance data transfer for UNICORE, in: UNICORE Summit 2011 Proceedings, 2011, pp. 135–142. http://hdl.handle.net/2128/4518.

[87] B. Haslhofer, S. Warner, C. Lagoze, M. Klein, R. Sanderson, M.L. Nelson, H. Van de Sompel, Resourcesync: leveraging sitemaps for resource synchronization, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 11–14.

Richard Grunzke is group leader for Virtual Research Environments at the Center for Information Services and High Performance Computing in Dresden, Germany. After studying Computer Science at TU Dresden, he received a Doctor of Engineering in the research data management domain. His further interests include HPC, workflows, big data and data life cycle management. Richard published 49 papers and is active as referee and in program committees. He is leading the MASi project that is building a universal research data management service. Furthermore, he serves in the UNICORE technical advisory board and advises the VAVID, GeRDI and CRC 940 projects.

Volker Hartmann received his graduate engineer degree in electrical engineering from the University of Karlsruhe in 1993. From 1993 to 1995 he made a postgraduate study of computer science at the University of Karlsruhe. Since 1995 he works as a scientific engineer at Institute for Data Processing and Electronics (IPE) at Karlsruhe Institute of Technology (KIT). In LSDMA his main topic is the management of the metadata for different communities. In MASi he implements generic metadata management methods suitable for many communities based on RDA recommendations.

Thomas Jejkal is software engineer at the Institute for Data Processing and Electronics (IPE) at Karlsruhe Institute of Technology (KIT) in Karlsruhe, Germany. He received his graduate engineer degree at Baden-Würtemberg Cooperative State University in 2004. He started working at KIT in the field of distributed computing and building up service oriented architectures. Currently, his main focus is on research data management. Thomas is responsible for managing the KIT Data Manager development and keeping it in line with international research data management efforts. Thomas has published more than 40 papers and is chairing a RDA working group.

Helen Kollai is a research associate at the Leibniz Institute of Ecological Urban and Regional Development in Dresden, Germany. She received the M.Sc. in International Area Studies in 2014 and focused her studies on geospatial analysis and remote sensing. Her current research interests cover (meta)data management and extraction methods in the field of land use change monitoring.

Ajinkya Prabhune is a researcher in the Supercomputing and Big Data group at the Karlsruhe Institute of Technology, Germany. His research includes diverse topics, in the areas of big data and metadata management, scientific workflow management systems, provenance, provenance interoperability and process algebra. Furthermore, his research extends in the field of big data in medical infrastructures. His expertise lies in software architecture designs for big data in research and scalable information management systems. He is currently contributing to the MASi project and the Helmholtz LSDMA project in the Key Technologies sub-project.

Hendrik Herold obtained a Ph.D. from the Technical University of Dresden, Germany. He currently is a postdoctoral researcher in the land use monitoring group at the Leibniz Institute of Ecological Urban and Regional Development in Dresden. His current research interests are focused on methods and applications of geoinformatics and image analysis in the context of historical map processing for long-term land use change monitoring.

Aline Deicke is deputy head of the Digital Academy of the Academy of Sciences and Literature Mainz, Germany. She has studied in Mainz, Pécs and Budapest and graduated from the Johannes Gutenberg-University Mainz with a master's degree in Pre- and Protohistory, Classical Archaeology and Anthropology. She has worked and interned on several excavations and museums. Her academic interests lie in elite burials of the Late Bronze Age, social archaeology, network analysis, web technologies and data modelling in the domain of cultural heritage.

Christiane Dreßler is research assistant within the digital humanities at the Academy of Sciences and Literature in Mainz, Germany. She graduated with a magister degree in German studies, sociology and psychoanalysis and with the first state examination for teaching English and German in secondary schools at the Goethe University of Frankfurt. She was employed as a trainee for public relations & marketing and for digital humanities at the Academy of Science and Literature. Along with other research projects she is assigned to the Corpus Vitrearum Deutschland, one of the use cases of the MASi project.

Julia Dolhoff is student assistant at the Digital Academy of the Academy of Sciences and Literature Mainz, Germany. She completed her bachelor's degree in sociology and computer science at University of Tübingen and conducts her postgraduate studies in digital methods in the humanities and cultural studies at the University of Mainz in cooperation with the University of Applied Sciences Mainz. Her main assignment at the Digital Academy is programming in the context of digital humanities projects.

Julia Stanek studied Chemistry at LMU Munich and finished her master thesis on copper complexes for electron transfer reactions in march 2015 in the Herres-Pawlis group. During her subsequent Ph.D. studies, she moved to RWTH Aachen University. Here, she synthesizes copper guanidine complexes under inert gas conditions as entatic state models and electron transfer systems. Moreover, she studies their reactivity experimentally by stopped-flow techniques and theoretical by density functional theory. The target is to steer the electron transfer rate by the constraints of the ligand.

Alexander Hoffmann is holding a permanent position in the Herres-Pawlis group at RWTH Aachen University for density functional calculations and X-ray crystallography. He studied Chemistry at the University of Paderborn and finished his Ph.D. thesis at TU Dortmund in 2011 on the coordination chemistry of late transition metals for catalysis and bioinorganic chemistry. Then, he moved to LMU Munich and turned to the theoretical description of N donor transition metal systems with focus on copper-dioxygen chemistry. In 2013, he was awarded with the Römer-Postdoc Prize. He is coordinator of the DFG-research unit FOR1405 ("Charge transfer dynamics in bioinorganic copper complexes").

Ralph Müller-Pfefferkorn is the head of the department "Distributed and Data Intensive Computing" at the Center for Information Services and High Performance Computing of the Technische Universität Dresden. He a has been active in a wide variety of research topics in High Performance Computing, Grid Computing, and Cloud Computing — always with a focus on data. His current interests are research data management, metadata and Big Data.

Torsten Schrade is professor for digital humanities at the University of Applied Sciences Mainz and heads the Digital Academy, the digital humanities department of the Academy of Sciences and Literature Mainz. His academic background is in history, his practical skills and experiences lie in software engineering, information architecture and programming. His main research interests focus on methods for data curation in the humanities, agile development, web technologies, web standards and approaches to linked open data in the digital humanities.

Gotthard Meinel received the M.Sc. in Information Technology in 1981, respectively the Ph.D. degree in image processing from Dresden University of Technology in 1987. Later he has been a postdoctoral researcher in biomathematics und technical mathematics. Since 1992 he is project leader in informatics, GIS and remote sensing at Leibniz Institute of Ecological and Regional Development in Dresden and since 2009 head of the research area "Monitoring of settlement and open space development". He is specialist in the field of monitoring land use development on base of indicators, automated analysis of large spatial datasets, topographic maps and visualization technologies.

Sonja Herres-Pawlis leads since 2015 the Chair of Bioinorganic Chemistry at RWTH Aachen University. She studied chemistry at Paderborn, Germany, and Montpellier, France. After her Ph.D. thesis (Paderborn 2005) and Postdoc (Stanford University) in bioinorganic chemistry, habilitation on sustainable polymerization catalysts at TU Dortmund, she worked as associate professor for coordination chemistry at LMU Munich. She received the Innovation Prize of the state of Northrhine-Westphalia and the Arnold-Sommerfeld Prize of the Bavarian Academy of Arts and Sciences. She is speaker of the interdisciplinary DFG-research unit FOR1405 and has published about 130 original papers and 4 patents.

Wolfgang E. Nagel holds the chair for computer architecture at TU Dresden and is director of the ZIH. His research covers programming concepts and software tools to support the development of scalable and data intensive applications, analysis of computer architectures, and development of efficient parallel algorithms and methods. Prof. Nagel is chairman of the Gauß-Allianz e.V. and member of the international Big Data and Extreme-scale Computing (BDEC) project. He is leading the Big Data competence center ‘‘ScaDS - Competence Center for Scalable Data Services and Solutions Dresden/Leipzig’’, funded by the German Federal Ministry of Education and Research.