Semiautomatische Ontologiegenerierung – ein Erfahrungsbericht

M. Schwantner
FIZ Karlsruhe

Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2012)
1.-2.8.2012

Agenda

- FIZ Karlsruhe
- Projekt NanOn
- Ontologien allgemein
- Vorgehen
- Text Mining, Annotation
- NanOn-Ontologie
- Evaluation und Fazit

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

- Gegründet 1977, Mitglied der Leibniz-Gemeinschaft
- Geschäftsfelder
  - STN International
    - Kooperation mit Chemical Abstracts Services (CAS)
    - Online-Service für Forschungsinformation:
      ca. 180 Datenbanken mit über 900 Mio. Dokumenten
      (Patente, Volltexte, Metadaten, Fakten, z. B. chemische Strukturen und Reaktionen, Gensequenzen)
  - Datenbanken und Informationsdienste
    - u.a. ZBMATH, ICSD, BINE
  - KnowEsis: Verteilte digitale Infrastruktur und Services zur Unterstützung des wissenschaftlichen Wertschöpfungsprozesse

Projekt NanOn: Semiautomatische Ontologiegenerierung – ein Beitrag zum Knowledge Sharing in der Nanotechnologie

Partner AIFB und INM – zwei in der Forschung führende Institute

AIFB - Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Wissensmanagement, Ontologien, Semantic Web

Netzwerk für Nano- und Biotechnologie

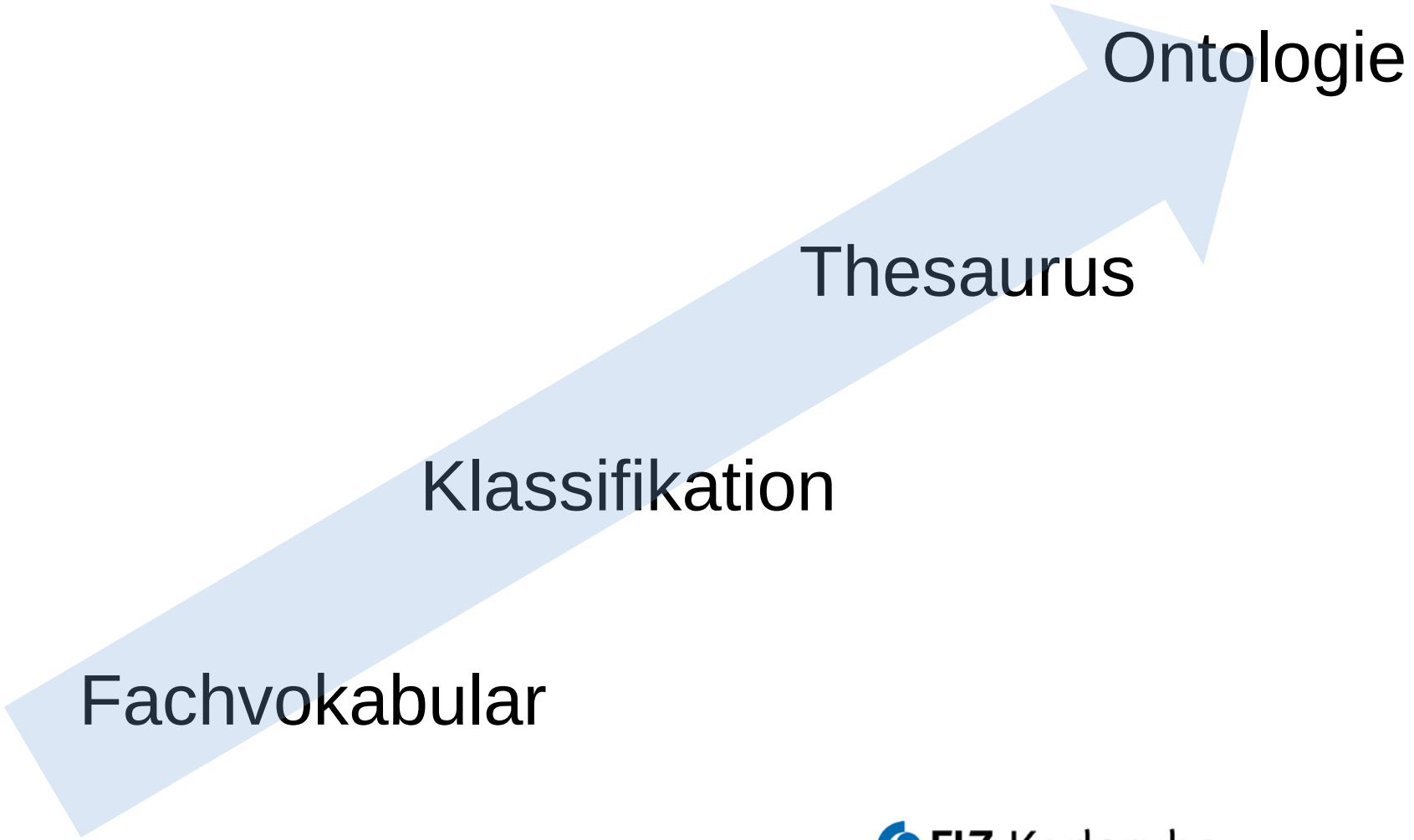
Materialforschung, Chemische Nanotechnologie Grenzflächenmaterialien

Gefördert durch die Leibniz-Gemeinschaft mittels des Paktes für Forschung und Innovation

Projekt NanOn – Ziele

- Erstellen einer Ontologie für die chemische Nanotechnologie
- Einsatzmöglichkeiten von Text Mining Methoden
  - für den Aufbau einer Ontologie
  - zur automatischen Annotation wissenschaftlicher Artikel
    - welche gibt es,
    - wo lassen sie sich einsetzen
    - Qualität der Ergebnisse
- Aufbau von Know-how

Was ist eine Ontologie ?
– Vom Fachvokabular zur Ontologie



Was ist eine Ontologie ? – Fachvokabular

carbon nanofiber, carbon nanostructures, carbon nanotube, magnetic property, material, material by structure, material property, nanoarray, nanodot array, nanohorn, nano-structures, property, physical property, single-walled carbon nanotube

Was ist eine Ontologie ? – Ontologie

Über einen Thesaurus hinausgehend: **zusätzliche semantische Relationen** erlauben logisches Schließen

Material

has property

Material by structure

- Nanostructures == nanoscale structures == nanoscale material
  - Carbon nanostructures
    - Carbon nanofiber == carbon nanofibre == CNF
      - Carbon nanotube == CNT
        - single-walled carbon nanotube == SWNT == single-walled CNT
    - . . .

Property

- Material property
- Physical property

Was ist eine Ontologie ?
– Ontologie

Über einen Thesaurus hinausgehend: Durch die Notation in einer **formalisierten Sprache (z.B. OWL)** ist die Ontologie maschinenlesbar und zwischen Systemen austauschbar.

```
<owl:Class rdf:about="&nanon;Single-walled_carbon_nanotube">
    <rdfs:subClassOf rdf:resource="&nanon;Carbon_nanotube"/>
    <rdfs:label xml:lang="en"> single-walled carbon nanotube </rdfs:la
    <rdfs:label xml:lang="de"> einwändige Kohlenstoffnanoröhre </rdfs
    <nanon:acronym rdf:datatype="&xsd;string"> SWCNT </nanon:acronym>
    <nanon:acronym xml:lang="en"> SWNT </nanon:acronym>
    <nanon:expression rdf:datatype="&xsd;string"> single-walled CNT </
    <oboInOwl:synonym xml:lang="en"> single-walled carbon nanotubes </
    <owl:equivalentClass>
        <owl:Restriction>
            <owl:someValuesFrom rdf:resource="&chebi;CHEBI_50595"/>
        </owl:Restriction> ...
```

Vorgehen:
– Anforderungsanalyse

Ontology Requirement Specification Document

gem. der Ontology Engineering Methodology (Gómez-Perez u. Suárez)

- Scope + Modularisierung: Chemische Nanotechnologie
  - Materialien
  - Eigenschaften
  - Prozesse (Synthese, Applikationen)
- Benutzer: Wissenschaftler und Produzenten, die nach Materialen, Eigenschaften, Mechanismen oder Applikationen suchen
- Competency Questions
- Level of Formality: OWL

Vorgehen:
– Competency Questions

Zusammenstellen einer beispielhaften Liste von Fragen, die mit Hilfe der zu erstellenden Ontologie beantwortet werden sollen.

Zweck: Richtschnur während des Erstellens und Benchmark

Auszug:

- Materials
  - Which metals show surface plasmon resonance?
  - Which surfactants are used in surface modification?
- Prozesse:
  - Which methods are used for conducting process X of TiO2?
  - Which process is used for achieving property P of material M?
  - Which process with material M leads to mixture Mx with property P?

Vorgehen:
– Aufbau der Ontologie

- Zusammenstellen der Begrifflichkeit
  - Intellektuell
    - Übernahme aus bestehenden Ontologien
    - Sammeln von Begriffen
    - manuelle Annotation von Texten
  - Text Mining
- intellektuelles Festlegen der Hierarchien
- intellektuelle Definition weiterer Relationen
- Erweiterung des Vokabulars und der Relationen um Synonyme
  - Intellektuell mittels Annotation
  - Unterstützung durch Text Mining
- automatische Annotation

Vorgehen:
– vorhandenes Material nutzen

Einbeziehen bereits existierender Ontologien:

- CMO (Chemical Methods Ontology der Royal Society of Chemistry)
- ChEBI (Chemical Entities of Biological Interest)
- ChemAxiomMetrology (ontology of common measurement techniques in the chemistry and materials science domain)

Text Mining auf

- Patenttexten (Detailed Description)
- Abstracts
- Volltexten wissenschaftlicher Artikel

Text Mining:
– Identifikation signifikanter Terme /1

- für die Zusammenstellung des Vokabulars
- basierend auf OpenNLP-Werkzeugen und tf/idf
  - Tokenizer
  - PoS-Tagger
  - Noun Phrase Detection
  - Morphologische Analyse

    ➔ u.a. Zusammenführen von Singular- und Pluralformen
  - Elimination allgemeinsprachlicher Terme mittels kontrastivem Korpus
  - Anwendung von tf / idf (modifiziert)

Text Mining:
– Identifikation signifikanter Terme /2

Aus einem Textkorpus mit ca. 107 Mio lfd Worten wurden 6653 Terme extrahiert, davon waren 27% sehr relevant, 39% relevant, 34% nicht relevant

Die ersten 10 …		und	… die Positionen 2001-2010	
film	0.616		*X-ray scattering*	0.020
diameter	0.554		*AAO membrane*	0.019
nanotube	0.512		*amorphous silicon*	0.019
coating	0.458		*antioxidant*	0.019
room temperature	0.457		*asymmetry*	0.019
clay	0.427		*catalyst layer*	0.019
morphology	0.410		*cell density*	0.019
substrate	0.404		*ceramic phase*	0.019
matrix	0.374		*clay mineral*	0.019
thermal stability	0.366		*clay tactoids*	0.019

blau: Begriffe wurden in die Ontologie aufgenommen

Text Mining
– für Synonyme und Hierarchien

Eigene Entwicklungen

- Parser für Akronyme
  - *… and X-ray diffraction (XRD) at …*
- Schreibvarianten
  - Singular / Plural
  - Gemeinsame, formale Grundform
  - Englisch / Amerikanisch
- Hearst Patterns (für Unter- / Oberbegriffe)
  - *NP such as* NP, NP *and other NP*, *NP including* NP
    ⇒ Kandidaten für *Oberbegriff* - Unterbegriff
  - Bsp.: *… their molecular properties such as* molecular mass, molecular weight *and* bond angle *could also …*

Text Mining – Relationen

- Definition der Relationen: intellektuell, unterstützt durch manuelle Annotation
- Text Mining zur Identifikation synonymer Formulierungen
- Tools
  - AntConc – interaktive Exploration eines Korpus, KWIC, Häufigkeiten
  - Sketch Engine – statistische Auswertung eines Korpus, Kollokationen, grammatikalische Kategorien
- Conditional Random Fields
  - graph-basiertes maschinelles Lernverfahren, u.a. zur Aufdeckung unspezifizierter Relationen
  - Master-Arbeit, Performance-Probleme

Text Mining
– Relationen Beispiele

- Material has Chemical Role Chemical Role
  - as
  - serves as
  - acting as
  - act as
- Prozeß has Process Property Eigenschaft
  - with
  - under
  - was produced at
  - synthesised at
  - conducted in
  - sowie weitere 9 synonyme Formulierungen

Die NanOn Ontologie: – Klassen /1

Reihenfolge der Prozessierung bedingt jeweiligen Zuwachs

	CMO u.a.	Intellektuell	Manuelle Annotation	Text Mining	Gesamt
Eigenschaften	**114**	**502**	**95**	**1801**	**2512**
Prozesse	**1676**	**122**	**98**	**954**	**2850**
Materialien	**29**	**511**	**94**	**538**	**1172**
Gegenstände	**84**	**78**	**62**	**174**	**398**
Hilfsklassen				**456**	**456**
Summe	**1807**	**1213**	**349**	**3896**	**7388**

Die NanOn Ontologie: – Klassen /2

7.388 Klassen

11.149 zusätzliche sprachliche Ausdrücke

- **9.405** Synonyme
  - **3.783** Synonyme durch Text Mining
  - **5.622** Synonyme und Akronyme aus CMO
- **196** automatisch bestimmte Akronyme
- **342** einfache chemische Formeln
- **1.206** Expressions (sprachliche Ausdrücke wie Verben etc.)

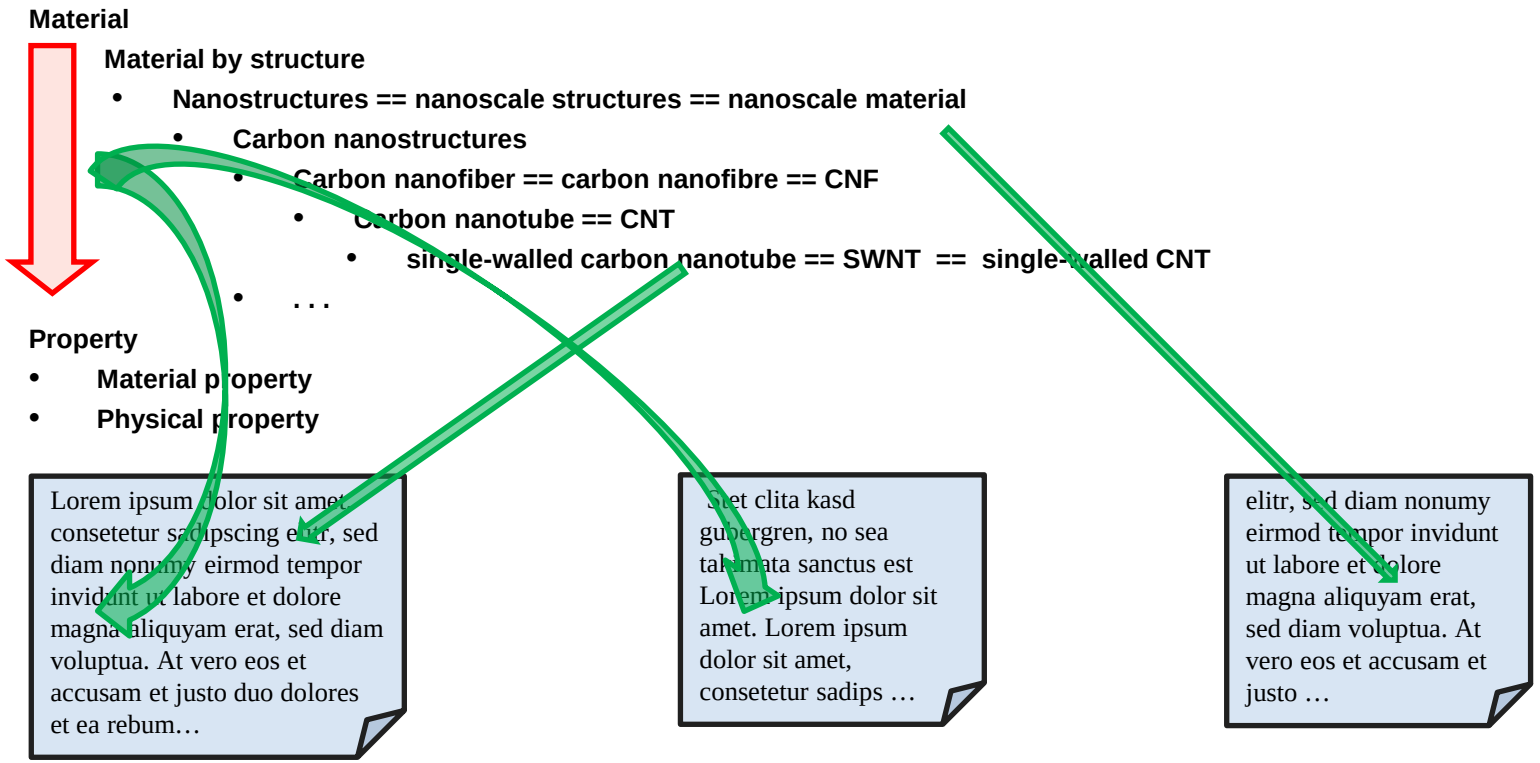
Die NanOn Ontologie: – Relationen

50 Relationen (einschl. Umkehrrelationen), darunter

Material	has Application	Anwendung
Material	has Chemical Role	Chemische Rolle
Material	has Microstructure	Struktur
Material	has Analytical Process	analyt. Prozeß
Material	has Fabrication Process	Fabrikationsprozeß
Material	is Input Material Of	Fabrikationsprozeß
Material	material Has Process	Fabrikationsprozeß
Material	has Material Property	Eigenschaft
Prozeß	has Process Property	Eigenschaft
Material	has Starting Material	Material

Automatische Annotation /1

- Verknüpfung der Ontologie mit Instanzen (hier Dokumente)



Automatische Annotation /2

- Entwicklung eines Plug-In für GATE
  - Klassen
    - ➔ String-Identität
  - Relationen
    - ➔ Grammatikalische Muster
    - ➔ String-basierte Muster
    - ➔ Abstandsmaß

Evaluation:
– Annotation für Relationen

Automatische Annotation von 29 Texten

Relation	Synonyme Ausdrücke	Anzahl	Richtige
has Starting Material	7	5711	0%
is Application Of	1	254	46%
has Process Property	15	677	51%
has Material Property	12	5985	52%
has Application	14	103	65%
has Fabrication Process	44	715	71%
has Analytical Process	12	32	72%
is Chemical Role	1	376	79%
has Chemical Role	4	75	89%

Evaluation:
– Competence Questions

- Which metals show surface plasmon resonance ?
  - Mit Booleschem Retrieval schwierig recherchierbar –
    nur 22 Elemente des Periodensystems sind Nicht-Metalle:
    (METALS OR ALUMINIUM OR LEAD OR IRON OR GOLD OR COPPER OR …)
    AND SURFACE PLASMON RESONANCE
  - Erleichterung durch Thesaurus:
    METALS+NT/CT AND SURFACE PLASMON RESONANCE
- Which are the applications for apolar materials ?
  - mit Booleschem Retrieval nicht mehr sinnvoll recherchierbar

⇒ **Boolesches Retrieval alleine führt zu Einbußen an Precision und Recall !**

Ontologien bieten hier Mehrwert, vorausgesetzt:

- Qualitativ hochwertige Annotation
- Große Anzahl annotierter Texte

Evaluation:
– Mehrwert durch Ontologie

Which are the applications for materials produced with *microwave plasma-assisted chemical vapor deposition (MPVCD)* ?

Mit Booleschem Retrieval nicht sinnvoll recherchierbar, jedoch mittels Ontologie:

MPVCD is_Fabrication_Process_of X AND
X has_Application Y

(SPARQL-ähnliche-Anfrage)

Die Information kann auf mehrere Dokumente verteilt sein:

We study conditions for ***microwave plasma-assisted chemical vapor deposition*** of high-quality single-crystal **diamond films** in a CVD reactor …

Diamond films are used to coat magnetic tapes and diamond fibers are used in composites for reinforcement …

… the deposited monocrystalline **diamond films** are used for optical windows and heat sinks …

Fazit

- Hoher intellektueller Aufwand
- Einsatz automatischer Verfahren sorgen für größere Vollständigkeit
- Automatische Annotation von Begriffen hängt stark von der Vollständigkeit der Ontologie bzgl. der Synonyme ab
- Automatische Annotation von Relationen in der Qualität sehr unterschiedlich und verbesserungswürdig
- Benutzungsoberfläche fehlt
- Weiterer Einsatz der Ontologie im Forschungsverbund Nanotechnologie der Leibniz-Gemeinschaft geplant

Das NanOn-Team

- FIZ Karlsruhe
  - Nils Elsner, Helmut Müller, Silke Rehme, Michael Schwantner, Andrea Zielinski
- AIFB
  - Nadejda Nikitina, Achim Rettinger
- INM
  - Elke Galli, Peter König, Mario Quilitz
- NanoBioNet
  - Matthias Mallmann

Danke !