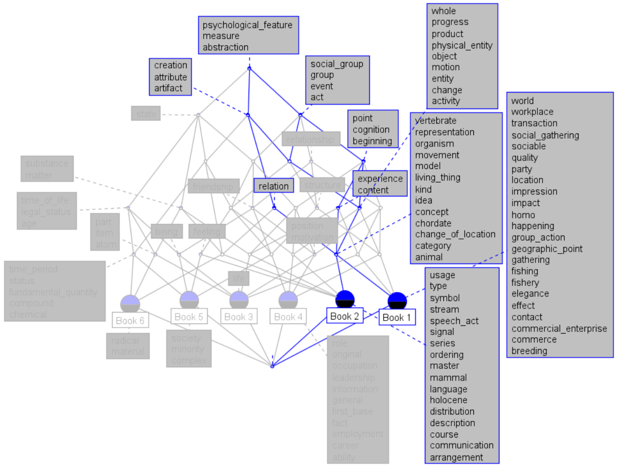
Florent Domenach, Christos Christofidis, Christian Wartena, Michael Franke-Maier

Freie Universität Berlin



Dynamic semantic subject indexing completion for classification of library items

BID16 6. BIBLIOTHEKS KONGRESS LEIPZIG 2016

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III
Medien, Information und Design

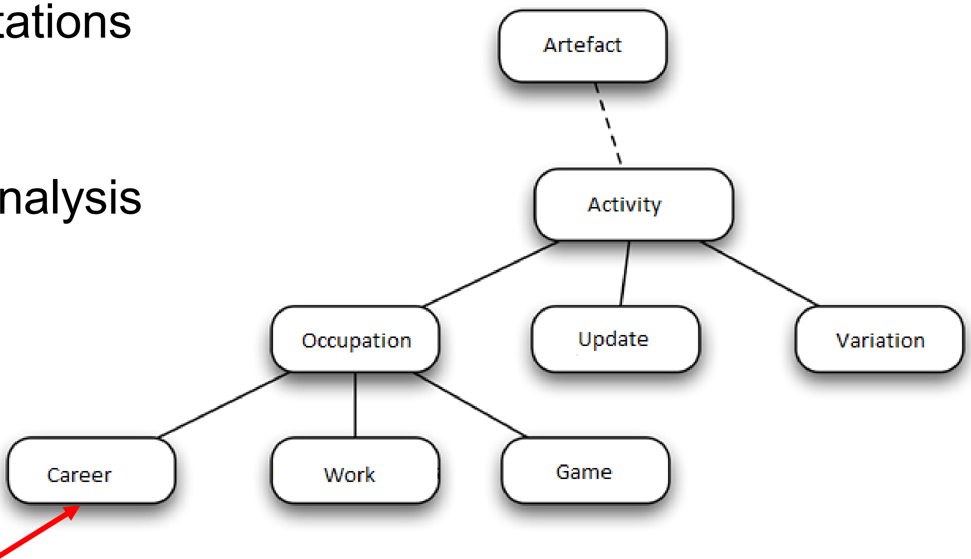
UNIVERSITY OF NICOSIA
ΠΑΝΕΠΙΣΤΗΜΙΟ ΛΕΥΚΩΣΙΑΣ

0. Agenda

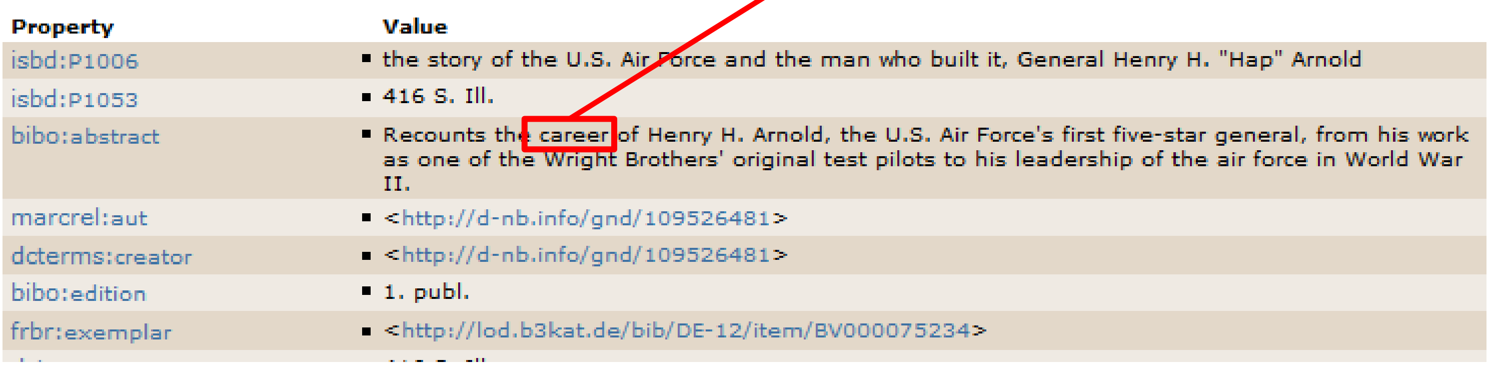
1. Objectives / Origin / Data Set
2. Formal Concept Analysis
3. WordNet Ontology
4. Clustering
5. Examples
6. Conclusion
7. Bibliography & Appendix

1.1 Objectives

We want to analyze abstracts / annotations
of bibliographic records
with Formal Concept Analysis,
a mathematical framework for data analysis



Hap at **lod.b3kat.de**
http://lod.b3kat.de/title/BV000075234



Property	Value
isbd:P1006	▪ the story of the U.S. Air Force and the man who built it, General Henry H. "Hap" Arnold
isbd:P1053	▪ 416 S. Ill.
bibo:abstract	▪ Recounts the career of Henry H. Arnold, the U.S. Air Force's first five-star general, from his work as one of the Wright Brothers' original test pilots to his leadership of the air force in World War II.
marcrel:aut	▪ <http://d-nb.info/gnd/109526481>
dcterms:creator	▪ <http://d-nb.info/gnd/109526481>
bibo:edition	▪ 1. publ.
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000075234>

1.2 Origins

- Meeting at the European Conference on Data Analysis 2015, http://ecda2015.com/
- Domenach / Christofidis: Dynamic Semantic Analysis of Tweets
  - Components:
    - Data Set: Twitter streams
    - Formal Concept Analysis (FCA)
    - WordNet: http://wordnet.princeton.edu
    - Clustering, Completion & Visualising

Are bibliographic data an alternative?

Yes!

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III
Medien, Information und Design

UNIVERSITY OF NICOSIA
ΠΑΝΕΠΙΣΤΗΜΙΟ ΛΕΥΚΩΣΙΑΣ

1.3 The Data Set

- B3Kat - Library Union Catalogue of Bavaria, Berlin and Brandenburg
- Linked open data Representation of B3Kat: http://lod.b3kat.de
- 26 Mio bibliographic entities

Selection of Data records

- Only english data records because of WordNet
- Records with title, subtitle and abtracts
- Abstracts with more than 200 characters
- At best with Library of Congress subject headings for comparison
- Set of 16.000 records as set for exemplary application

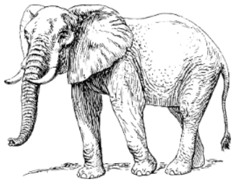
2.1 Formal Concept Analysis (FCA)

The basic procedure of Formal Concept Analysis is the following:

- Data is represented in a very basic data type, called formal context, a binary table.
- Each formal context is transformed into a mathematical structure called concept lattice
- The information contained in the formal context is preserved.
- The concept lattice is the basis for further data analysis. It may be represented graphically to support communication, or it may be investigated with algebraic methods to unravel its structure.

2.2 FCA in Data Analysis

- Formal Concept Analysis (FCA) (Ganter et Wille, 1999) was developed in Darmstadt as a mathematical formalization of the notion of concept
- Formalization of the idea of concept
  - Concept = intent + extent
  - Intent: All the common attributes
    - intent of ELEPHANT = collection of all elephants' attributes (trunk, has four limbs, has ears, ...)
  - Extent: All the objects having some attributes
    - extent of ELEPHANT = collection of all elephants



- *Clustering method* that will find all those concepts, and order them
  - subconcept/superconcept relation
    - $c_1 = (extent_1, intent_1) \leq c_2 = (extent_2, intent_2) \Leftrightarrow extent_1 \subseteq extent_2 \Leftrightarrow intent_1 \supseteq intent_2$
  - ELEPHANT ≤ MAMMAL ≤ ANIMAL
- Related to association rules with 100% confidence

2.3 FCA: formally

- A context (G, M, I) is a binary table, where G is the set of objects, M the set of attributes and $I \ \subseteq G \times M$ is a binary relation
  - $(g, m) \in I$ means that "the object g is related with the attribute m through the relation I"
- Two derivation operators can be defined on sets of objects and sets of attributes, $\forall A \subseteq G, B \subseteq M$,
  - $A' = \{m \in M : \forall g \in A, (g, m) \in I\}$ (intent)
  - $B' = \{g \in G : \forall m \in B, (g, m) \in I\}$ (extent)
- A pair $(A, B), A \subseteq G, B \subseteq M$ is a formal concept if $A' = B$ and $B' = A$.
  - Maximal group of objects A sharing maximal common attributes B

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III
Medien, Information und Design

UNIVERSITY OF NICOSIA
ΠΑΝΕΠΙΣΤΗΜΙΟ ΛΕΥΚΩΣΙΑΣ

2.4 FCA: Toy Example

	Composite	Even	Odd	Prime	Square
1			X		X
2		X		X	
3			X	X	
4	X	X			X
5			X	X	
6	X	X			
7			X	X	
8	X	X			
9	X		X		X
10	X	X			

- $\{3{,}5\}' = \{Odd, Prime\}$
- $\{Odd, Prime\}' = \{3{,}5{,}7\}$
- $(\{3{,}5{,}7\}, \{Odd, Prime\})$ is a concept

Context of natural numbers

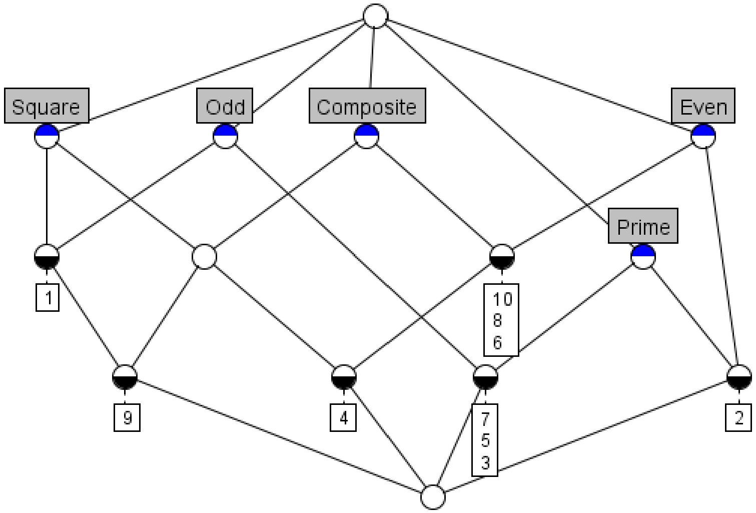
2.5 FCA: Concept Lattice

The concept lattice is the set of all concepts

- Generate and visualize hierarchies of concepts
- Ordered by $(A_1, B_1) \leq (A_2, B_2)$ iff $A_1 \subseteq A_2$ (or dually $B_2 \subseteq B_1$)

Drawing uses minimal labeling

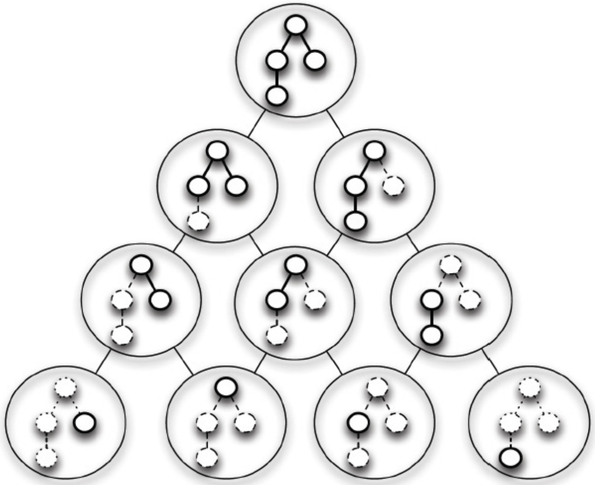
- every vertex inherits objects labels of the vertices below and attributes labels of the vertices above it
- vertex labeled $\{1\}$ represents the concept with extent $\{1{,}9\}$ and intent $\{Square, Odd\}$



Concept lattice of toy example

2.6 FCA: Pattern structures & Ontology

- Pattern structures can be thought as extension of FCA
- Pattern structures can…
  - handle non binary data
  - complete missing annotations
  - take into account domain knowledge represented within an ontology
- We need an ontology as basis for FCA with pattern structures
  - not in common librarians meaning
  - but in mathematical meaning!
  - has to be computationally doable
  - we use $\mathcal{EL}$ ontologies, see www.w3.org/TR/owl2-profiles
  - computing Least Common Subsumers



see also Appendix A to C for more details

3.1 WordNet ontology used for annotating

WordNet: online ontology containing 118,000 of terms in the English language

- Divides into nouns, verbs, adjectives, and adverbs

Synset (Synonymous Set)

- Set of terms that share a single meaning or sense (synonymy)
- Interlinked by means of conceptual-semantic and lexical relations

http://wordnet.princeton.edu

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs		

Chart:
WordNet: A Lexical Database for English, George A. Miller, November 1995/Vol. 38, No. 11 COMMUNICATIONS OF THE ACM

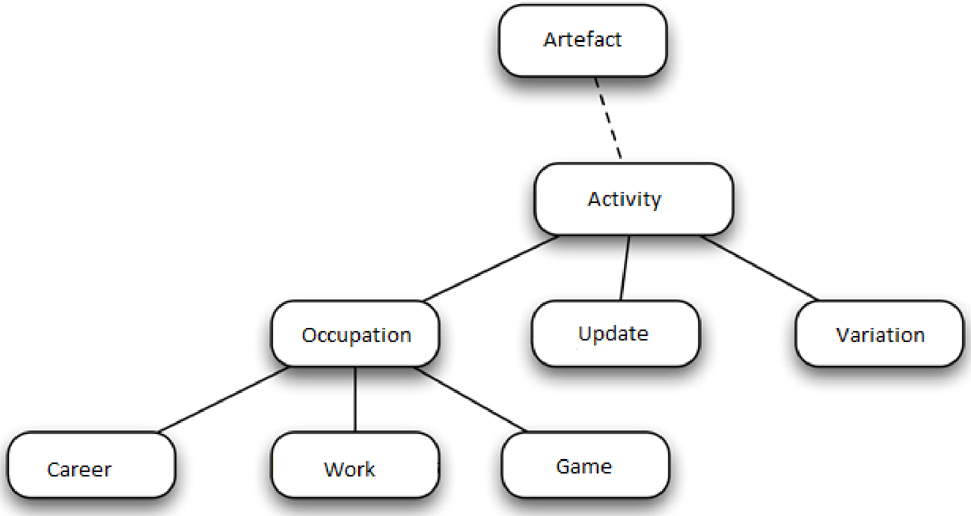
3.2 WordNet Hierarchy

Synsets

- Contained in an *is-a* hierarchy

Noun synset relationships form a complex and deep hierarchy

- subordinate children (*hyponymy*),
- superordinate parents (*hypernymy*)



Fragment of WordNet Concept Hierarchy: boxes (nodes) correspond to synsets; lines (edges) indicate the hypernym/hyponym relation

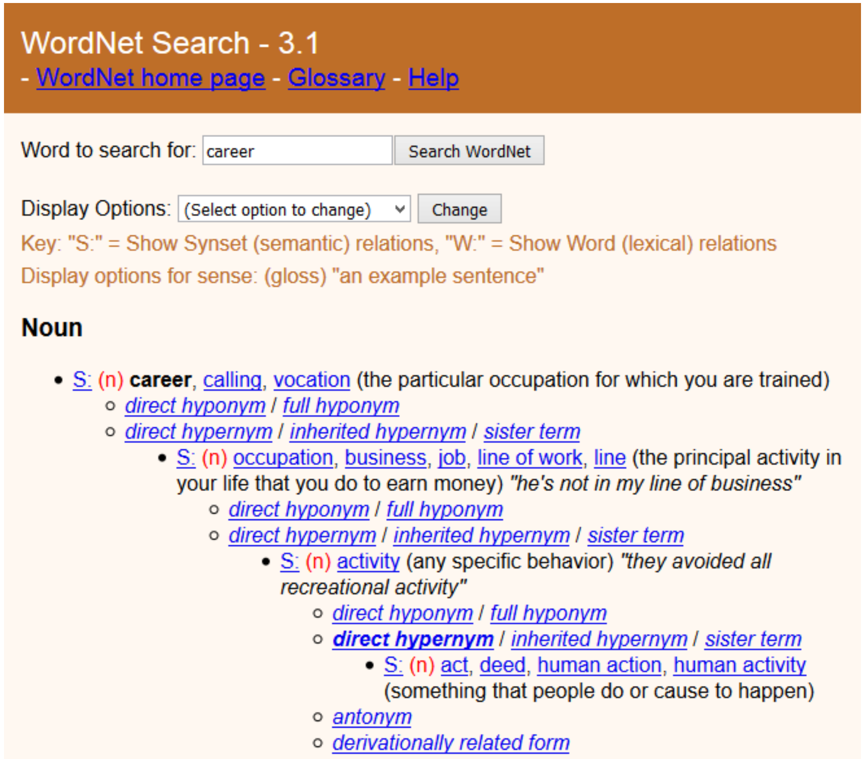
3.2 WordNet Hierarchy

Synsets

- Contained in an *is-a* hierarchy

Noun synset relationships form a complex and deep hierarchy

- subordinate children (*hyponymy*),
- superordinate parents (*hypernymy*)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: career

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **career**, calling, vocation (the particular occupation for which you are trained)
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) occupation, business, job, line of work, line (the principal activity in your life that you do to earn money) *"he's not in my line of business"*
      - *direct hyponym* / *full hyponym*
      - *direct hypernym* / *inherited hypernym* / *sister term*
        - S: (n) activity (any specific behavior) *"they avoided all recreational activity"*
          - *direct hyponym* / *full hyponym*
          - ***direct hypernym*** / *inherited hypernym* / *sister term*
            - S: (n) act, deed, human action, human activity (something that people do or cause to happen)
          - *antonym*
          - *derivationally related form*

Screenshot from WordNet Search 3.1
http://wordnetweb.princeton.edu/perl/webwn

3.3 Polysemous terms in Wordnet

Polysemous term: term is associated with many synsets

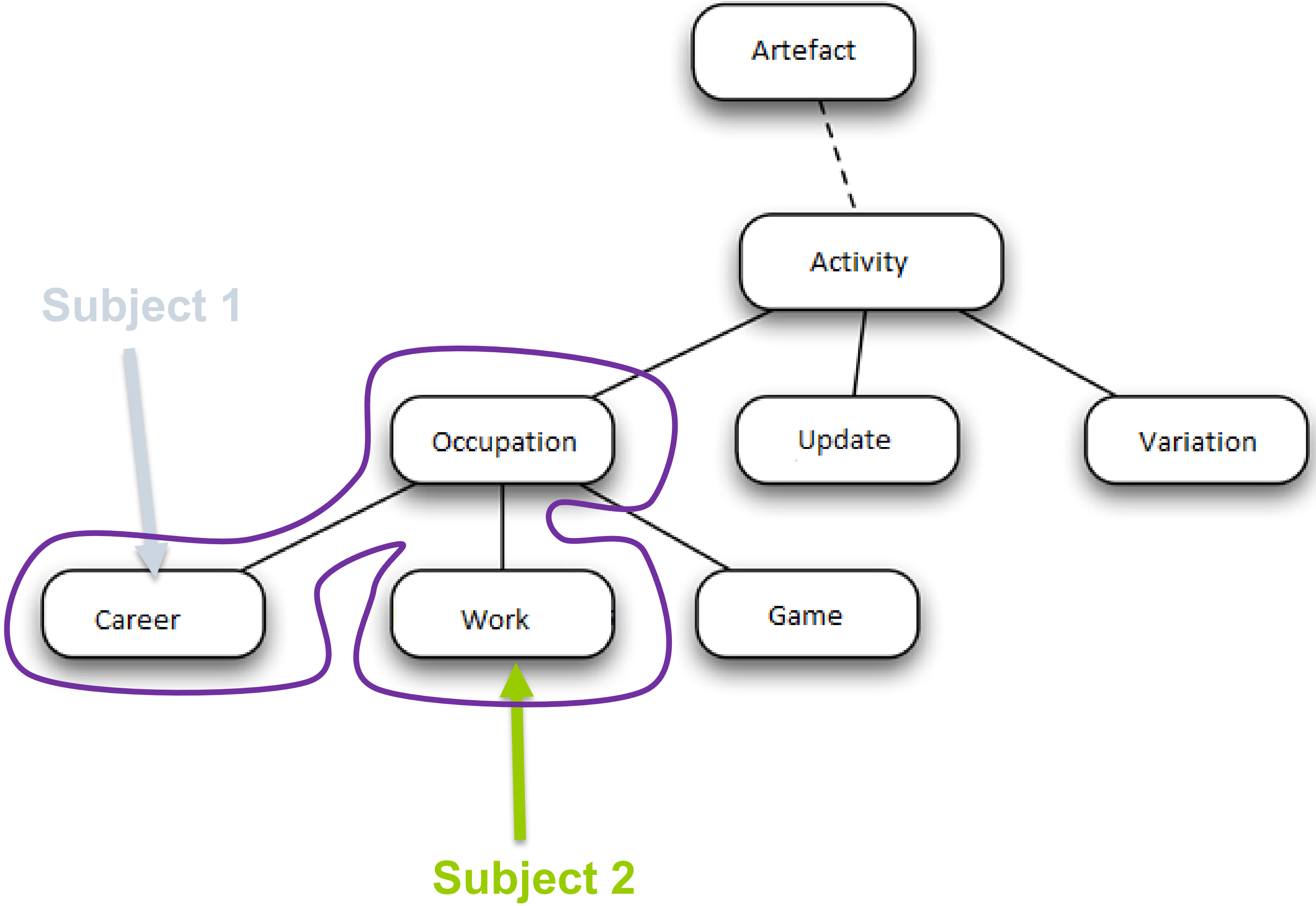
- word sense disambiguation must occur
- specific synset implied by the usage in the text must be determined
- Example: *cat* can either be a *pet* or a type of *medical scan*

Word sense can be evaluated through similarity measures of the different possible synsets of a term in its context

- Measures of similarity quantify how much two concepts are alike
- Can be based on the *is-a* hierarchy or on information content
- We used *Wu-Palmer* similarity

$$sim(w_1, w_2) = \frac{2 \times depth(lcs(w_1, w_2))}{depth(w_1) + \ depth(w_2)}$$

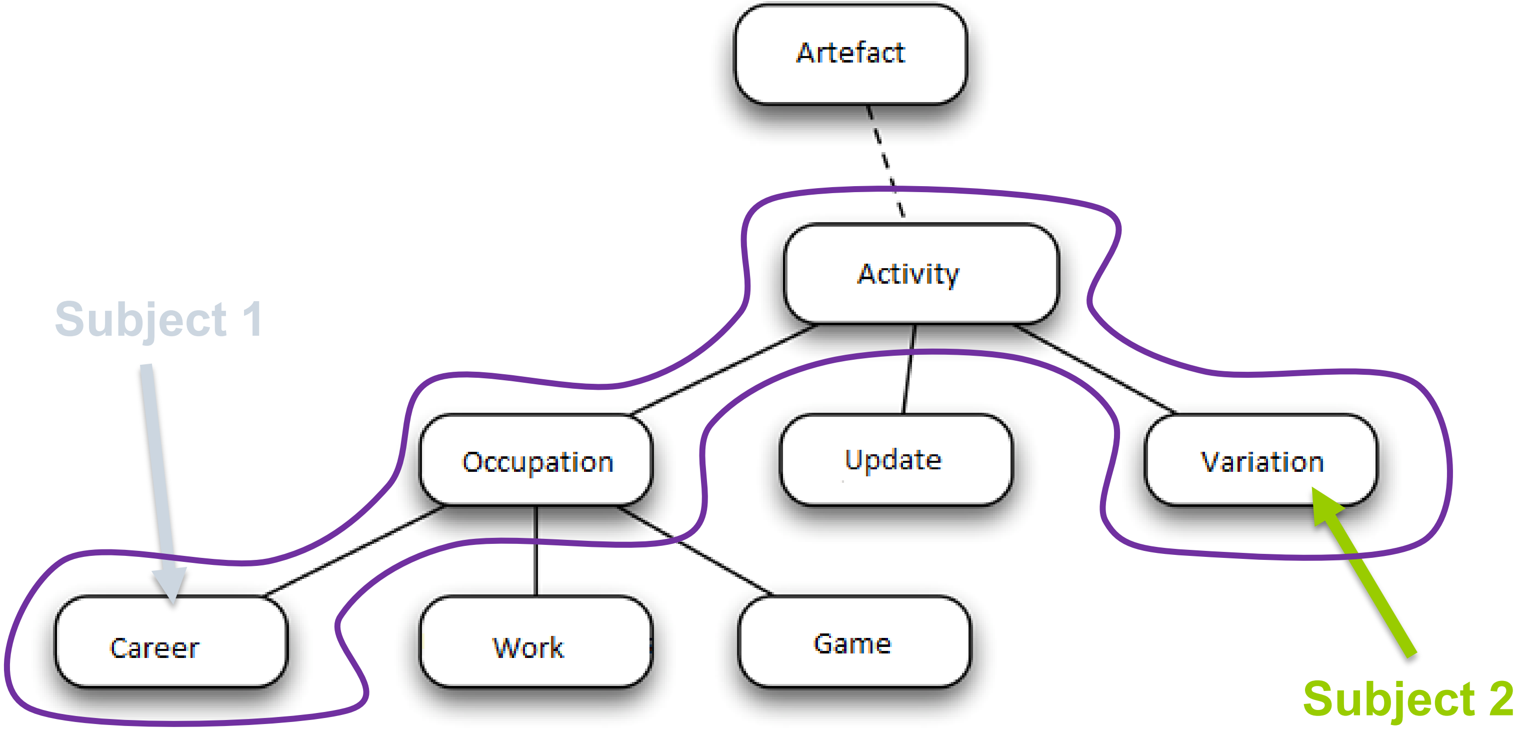
3.4 Completion of Descriptions



HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III
Medien, Information und Design

UNIVERSITY OF NICOSIA
ΠΑΝΕΠΙΣΤΗΜΙΟ ΛΕΥΚΩΣΙΑΣ

3.4 Completion of Descriptions



4. Clustering to create semantic categories

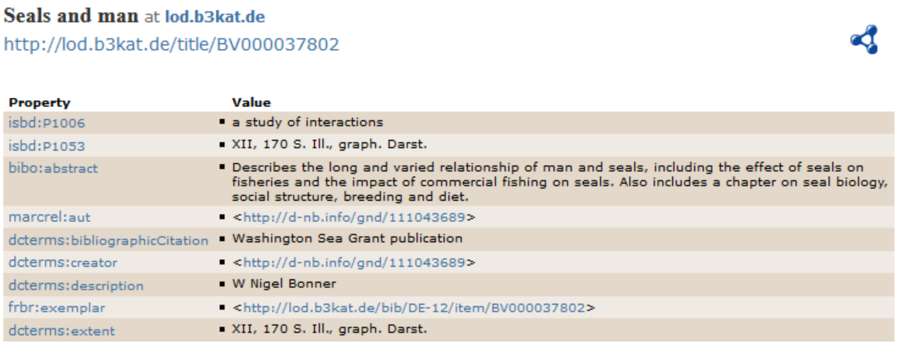
Filtering of keywords

- Put to lower case
- Change plural to singular
- Eliminates duplicates
- Check if it exists in WordNet

Clustering on keywords

- Subjects grouped together should not be too dissimilar
  - Else the completion of descriptions is too large
- Using Wikipedia dataset
  - 6 billions tokens, 400K vocabulary, 300 dimensions vector representation
  - Available at http://nlp.stanford.edu/projects/glove/
- Hierarchical clustering
  - Using Scipy Python library
  - Average linkage
  - Adjusted threshold
- Keep the subjects (relatively) semantically close

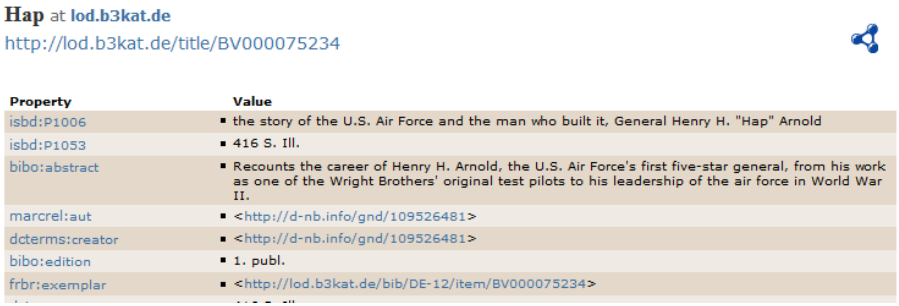
5.1 Example: 6 randomly picked books



Seals and man at **lod.b3kat.de**

http://lod.b3kat.de/title/BV000037802

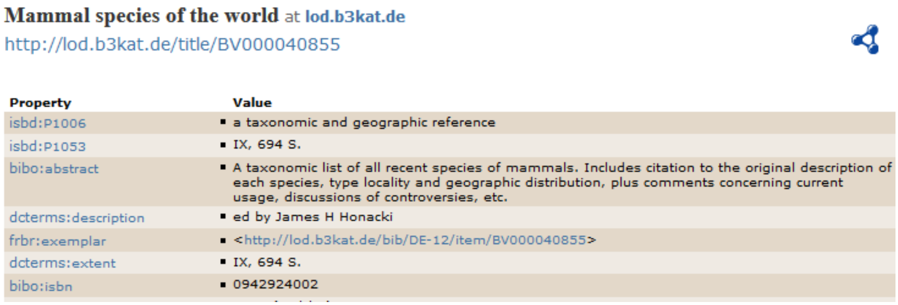
Property	Value
isbd:P1006	▪ a study of interactions
isbd:P1053	▪ XII, 170 S. Ill., graph. Darst.
bibo:abstract	▪ Describes the long and varied relationship of man and seals, including the effect of seals on fisheries and the impact of commercial fishing on seals. Also includes a chapter on seal biology, social structure, breeding and diet.
marcrel:aut	▪ <http://d-nb.info/gnd/111043689>
dcterms:bibliographicCitation	▪ Washington Sea Grant publication
dcterms:creator	▪ <http://d-nb.info/gnd/111043689>
dcterms:description	▪ W Nigel Bonner
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000037802>
dcterms:extent	▪ XII, 170 S. Ill., graph. Darst.



Hap at **lod.b3kat.de**

http://lod.b3kat.de/title/BV000075234

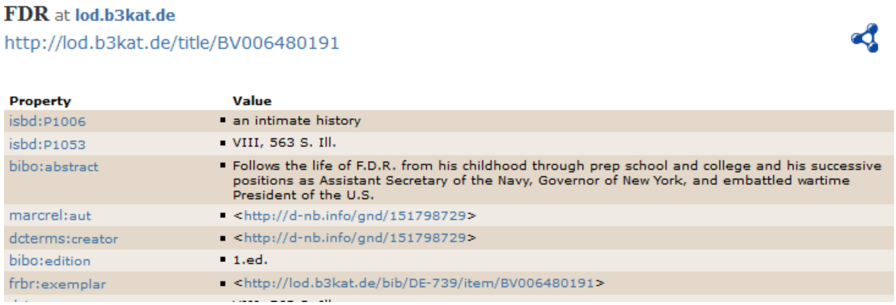
Property	Value
isbd:P1006	▪ the story of the U.S. Air Force and the man who built it, General Henry H. "Hap" Arnold
isbd:P1053	▪ 416 S. Ill.
bibo:abstract	▪ Recounts the career of Henry H. Arnold, the U.S. Air Force's first five-star general, from his work as one of the Wright Brothers' original test pilots to his leadership of the air force in World War II.
marcrel:aut	▪ <http://d-nb.info/gnd/109526481>
dcterms:creator	▪ <http://d-nb.info/gnd/109526481>
bibo:edition	▪ 1. publ.
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000075234>



Mammal species of the world at **lod.b3kat.de**

http://lod.b3kat.de/title/BV000040855

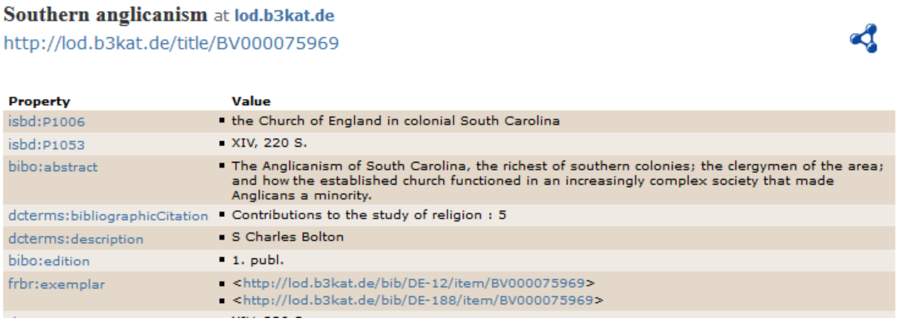
Property	Value
isbd:P1006	▪ a taxonomic and geographic reference
isbd:P1053	▪ IX, 694 S.
bibo:abstract	▪ A taxonomic list of all recent species of mammals. Includes citation to the original description of each species, type locality and geographic distribution, plus comments concerning current usage, discussions of controversies, etc.
dcterms:description	▪ ed by James H Honacki
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000040855>
dcterms:extent	▪ IX, 694 S.
bibo:isbn	▪ 0942924002



FDR at **lod.b3kat.de**

http://lod.b3kat.de/title/BV006480191

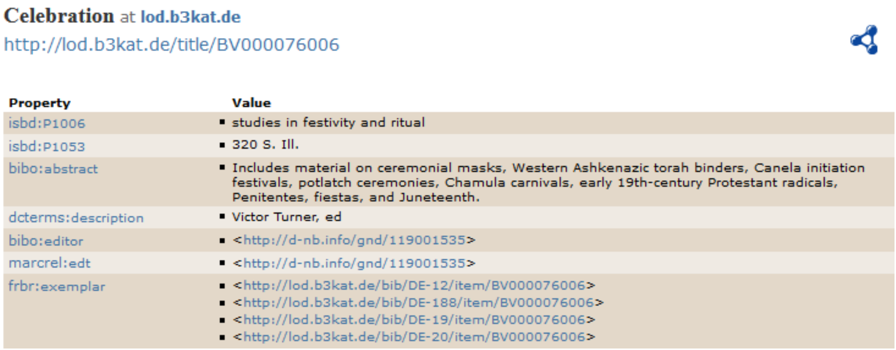
Property	Value
isbd:P1006	▪ an intimate history
isbd:P1053	▪ VIII, 563 S. Ill.
bibo:abstract	▪ Follows the life of F.D.R. from his childhood through prep school and college and his successive positions as Assistant Secretary of the Navy, Governor of New York, and embattled wartime President of the U.S.
marcrel:aut	▪ <http://d-nb.info/gnd/151798729>
dcterms:creator	▪ <http://d-nb.info/gnd/151798729>
bibo:edition	▪ 1.ed.
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-739/item/BV006480191>



Southern anglicanism at **lod.b3kat.de**

http://lod.b3kat.de/title/BV000075969

Property	Value
isbd:P1006	▪ the Church of England in colonial South Carolina
isbd:P1053	▪ XIV, 220 S.
bibo:abstract	▪ The Anglicanism of South Carolina, the richest of southern colonies; the clergymen of the area; and how the established church functioned in an increasingly complex society that made Anglicans a minority.
dcterms:bibliographicCitation	▪ Contributions to the study of religion : 5
dcterms:description	▪ S Charles Bolton
bibo:edition	▪ 1. publ.
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000075969> ▪ <http://lod.b3kat.de/bib/DE-188/item/BV000075969>



Celebration at **lod.b3kat.de**

http://lod.b3kat.de/title/BV000076006

Property	Value
isbd:P1006	▪ studies in festivity and ritual
isbd:P1053	▪ 320 S. Ill.
bibo:abstract	▪ Includes material on ceremonial masks, Western Ashkenazic torah binders, Canela initiation festivals, potlatch ceremonies, Chamula carnivals, early 19th-century Protestant radicals, Penitentes, fiestas, and Juneteenth.
dcterms:description	▪ Victor Turner, ed
bibo:editor	▪ <http://d-nb.info/gnd/119001535>
marcrel:edt	▪ <http://d-nb.info/gnd/119001535>
frbr:exemplar	▪ <http://lod.b3kat.de/bib/DE-12/item/BV000076006> ▪ <http://lod.b3kat.de/bib/DE-188/item/BV000076006> ▪ <http://lod.b3kat.de/bib/DE-19/item/BV000076006> ▪ <http://lod.b3kat.de/bib/DE-20/item/BV000076006>

5.2 Example: Clusters of 6 randomly picked books

15 Clusters of size > 1

- ['varied', 'complex', 'usage', 'type', 'description', 'material', 'structure', 'distribution', 'geographic', 'original']
- ['college', 'church', 'school']
- ['chapter', 'list']
- ['area', 'southern', 'commercial', 'western']
- ['force', 'pilot', 'air', 'test']
- ['life', 'impact', 'long', 'current', 'recent', 'relationship', 'effect', 'man', 'society', 'career', 'work', 'social', 'first']
- ['comment', 'discussion', 'controversy']
- ['general', 'minority', 'radical', 'leadership', 'position']
- ['ceremonial', 'ceremony']
- ['festival', 'carnival', 'fiesta']
- ['fishing', 'fishery', 'breeding', 'mammal']
- ['initiation', 'torah']
- ['taxonomic', 'biology']
- ['childhood', 'diet']
- ['mask', 'seal']

5.3 Extracted Partial Context

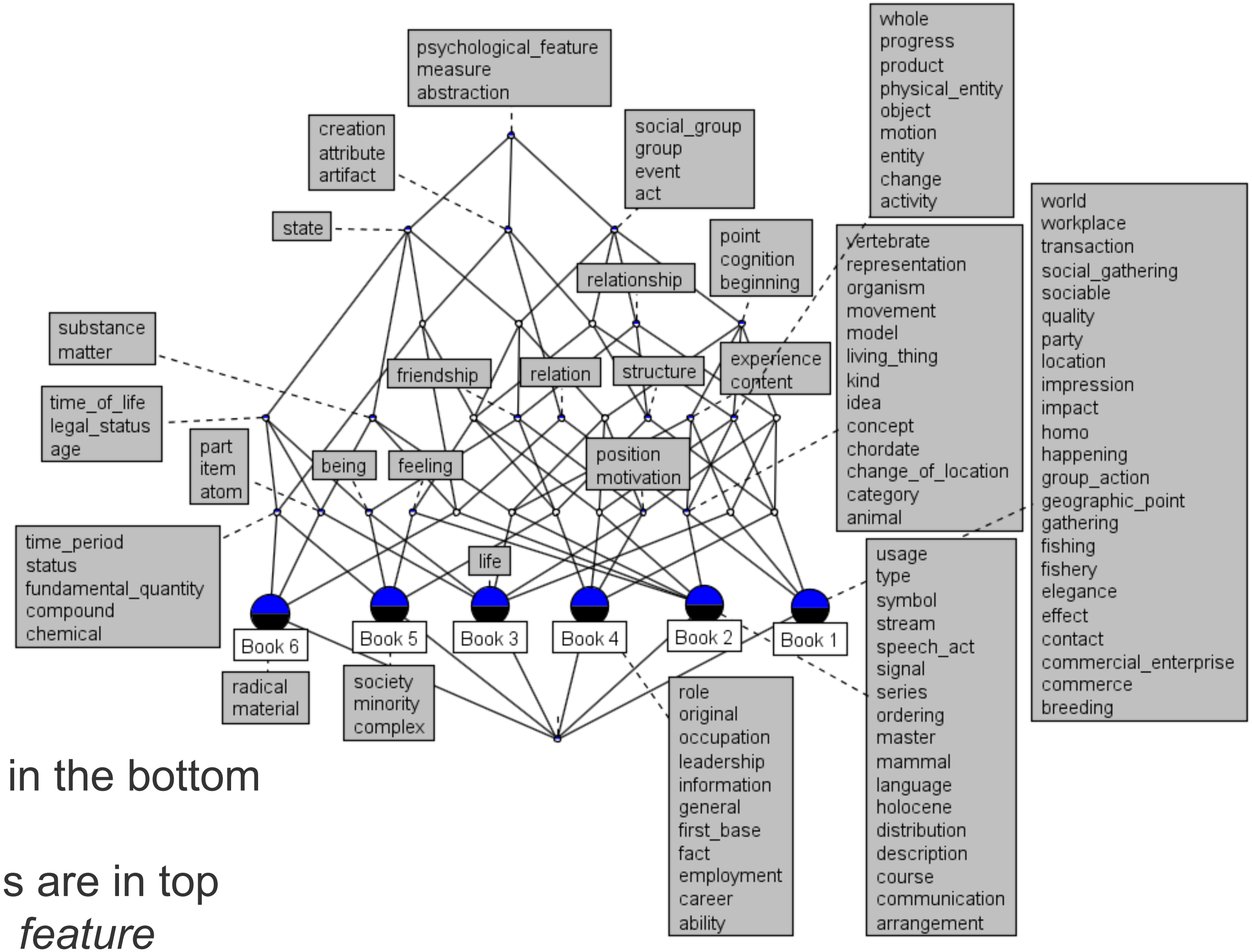
Book	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Book 1 (Seals and Man)	'structure',	'impact', 'relationship', 'effect', 'man', 'social'		'fishing', 'fishery', 'breeding',
Book 2 (Mammal species of the world)	'usage', 'type', 'description', 'distribution', 'original'	'current', 'recent',		'mammal'
Book 3 (FDR)		'life',	'position'	
Book 4 (Hap)	'original'	'career', 'work', 'first'	'general', 'leadership',	
Book 5 (Southern Anglicanism)	'complex',	'society',	'minority',	
Book 6 (Celebration)	'material',		'radical'	

5.3 Extracted Partial Context

Book	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Book 1 (Seals and Man)	'structure',	'impact', 'relationship', 'effect', 'man', 'social'		'fishing', 'fishery', 'breeding',
Book 2 (Mammal species of the world)	'usage', 'type', 'description', 'distribution', 'original'	'current', 'recent',		'mammal'
Book 3 (FDR)		'life',	'position'	
Book 4 (Hap)	'original'	'career', 'work', 'first'	'general', 'leadership',	
Book 5 (Southern Anglicanism)	'complex',	'society		
Book 6 (Celebration)	'material',		'radical'	

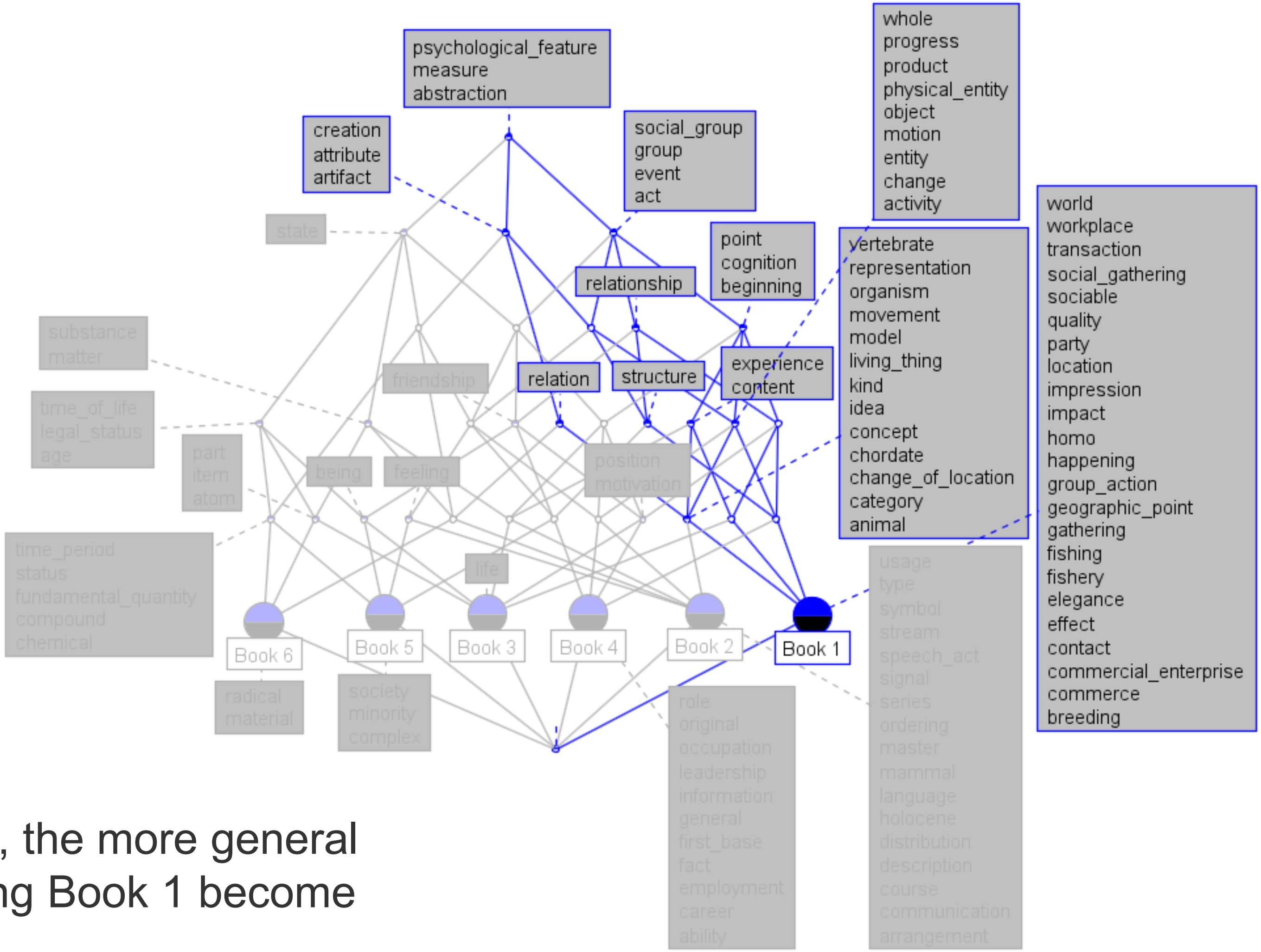
Sense disambiguation:
synset *first* refers to *first_base* (baseball)
in the context of *career* and *work*

5.4 Lattice of the Partial Context



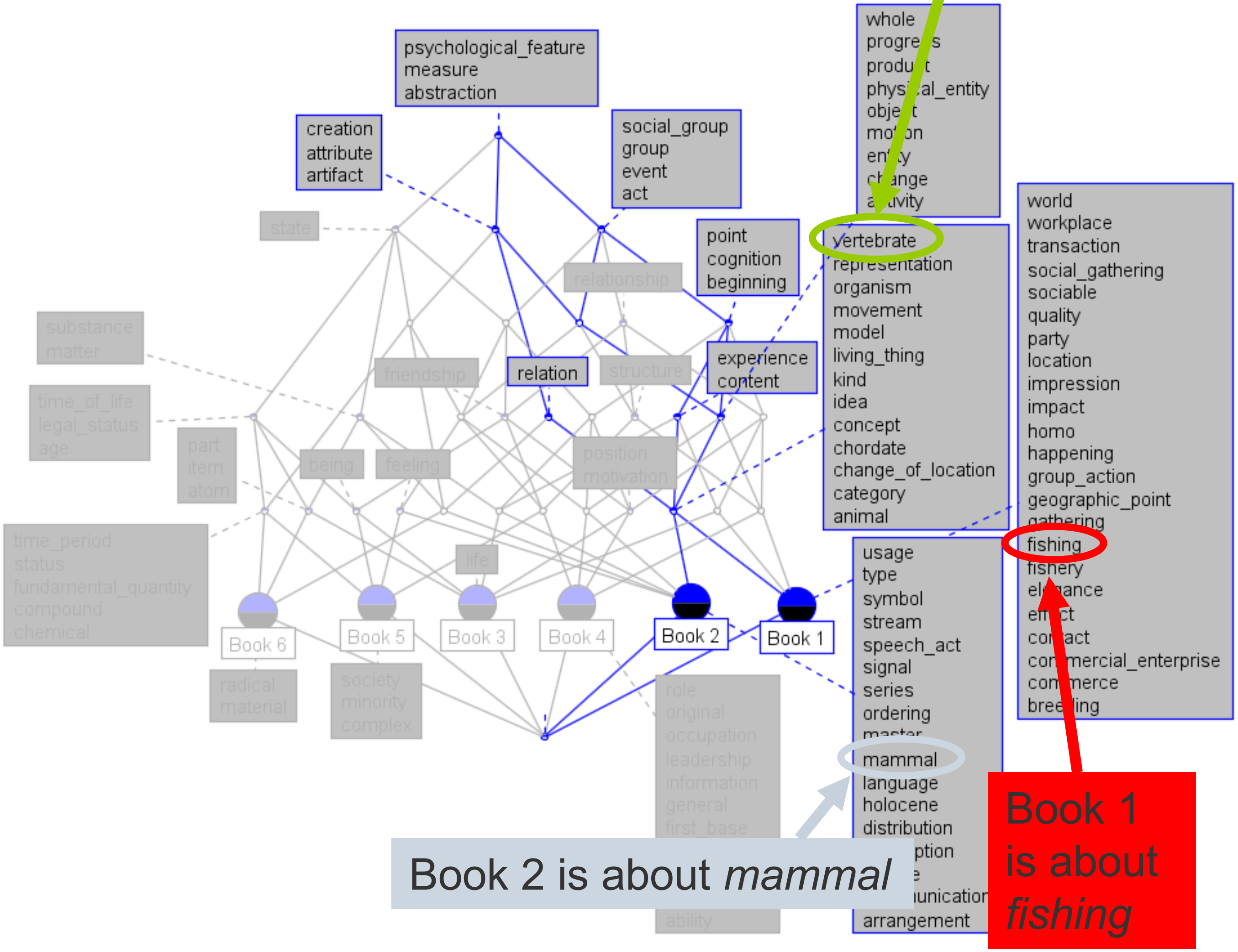
- Specific terms are in the bottom
  - *fishing*
- More general terms are in top
  - *psychological_feature*

5.5 Navigating Book 1



- Highlighted in blue
- The “upper” we go, the more general the terms describing Book 1 become

5.6 Common Descriptions of Books 1 & 2

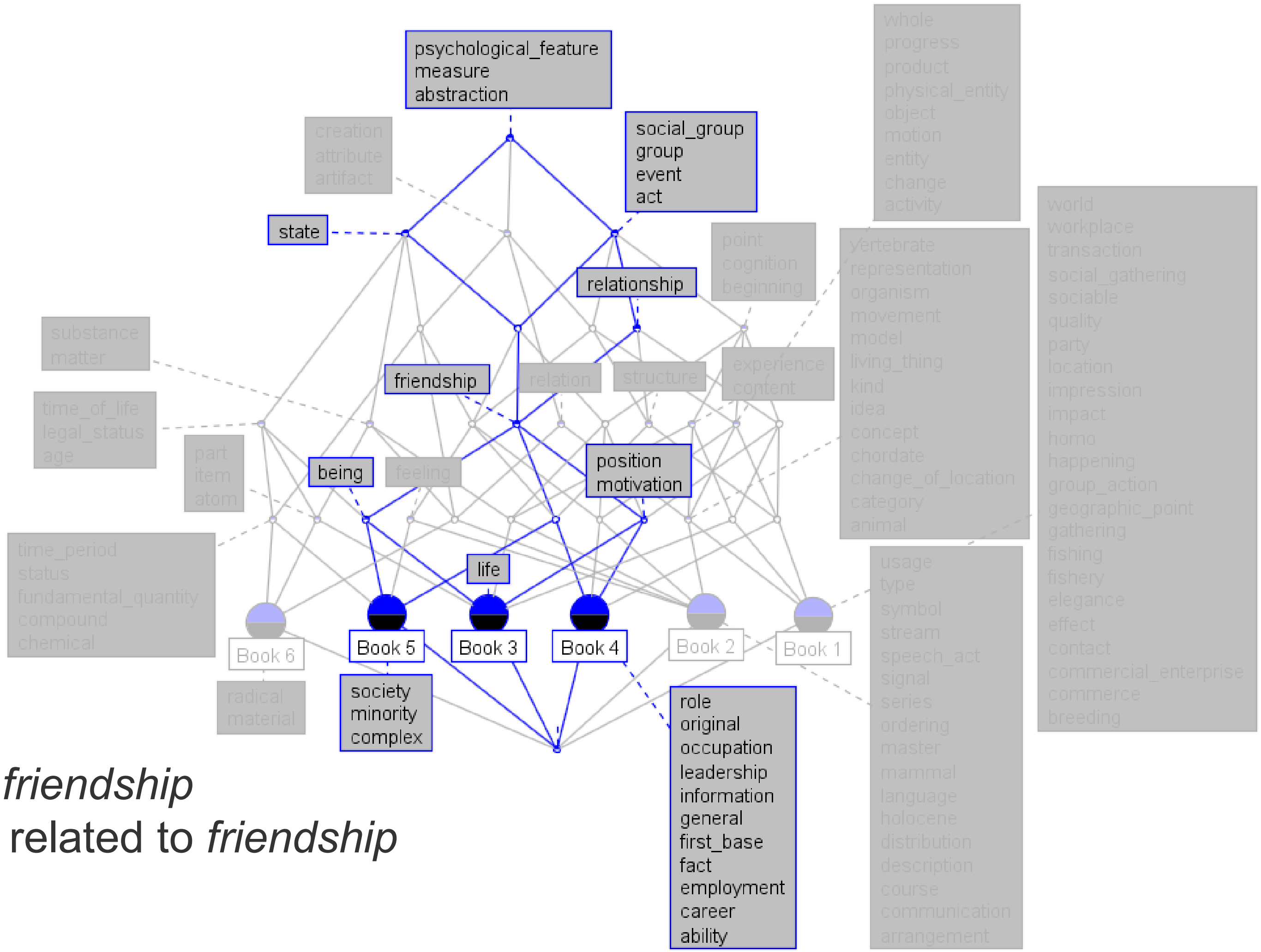


Book 1 and Book 2 both are about *vertebrate*

Book 2 is about *mammal*

Book 1 is about *fishing*

5.7 Exploring a subject



- Looking at subject *friendship*
- Books 3, 4 & 5 are related to *friendship*

6. Conclusion

- Current completion of descriptions is only for nouns
  - Only nouns are ordered in a *is-a* hierarchy on WordNet
  - Planning to expand it to verbs, adjectives and adverbs through RDF graphs
- Needs of visualization tools, not only for examples but for the whole data set
- Plan to use the lattice for recommendation of books
  - Semantically
  - Based on the description completion
- Plan to use the lattice for indexing of books
  - Instead of WordNet using authority files of subject headings like LCSH
- Plan to use wordnets in other languages

Thank you for your attention!

Dr. Florent Domenach
Associate Professor
University of Nicosia – Computer Science
domenach.f@unic.ac.cy

Christos Christofidis
University of Nicosia – Computer Science
christofidis.c@student.unic.ac.cy

Prof. Dr. Christian Wartena
Hochschule Hannover (HsH)
Fakultät III – Medien, Information u. Design
christian.wartena@hs-hannover.de

Michael Franke-Maier
Freie Universität Berlin
Universitätsbibliothek
franke@ub.fu-berlin.de

7. Bibliography

- A. Coulet, F. Domenach, M. Kaytoue and A. Napoli (2013), Using pattern structures for analyzing ontology-based annotations, Formal Concept Analysis, P. Cellier, F. Distel and B. Ganter (eds), LNCS/LNAI vol. 7880, pp 76-91.
- Ganter, B., Kuznetsov, S. O.: Pattern Structures and Their Projections. ICCS, Springer, 2001, 2120, 129-142.
- Ganter, B., Wille, R.: Formal Concept Analysis, Mathematical Foundations, Springer Verlag (1999).

7. Appendix A: Pattern structure (formally)

- A pattern structure is denoted by $(G, (D, \sqcap), \delta)$ with
  - G a set of objects,
  - $(D, \sqcap)$ a set of descriptions,
  - $\delta : G \rightarrow D$ maps each object with its descriptions.
- The derivation operators are defined such as $\forall A \subseteq G,\ d \in (D, \sqcap)$:
  - $A^{\blacksquare} = \sqcap_{\{g \in A\}} \delta(g)$ (smallest description common to all elements of A)
  - $d^{\blacksquare} = \{g \in G \colon d \sqsubseteq \delta(g)\}$ (all objects having a description more general than d)
- The order of pattern concepts, ordered by inclusion of extents is denoted by
  - $(A_1, d_1) \leq (A_2, d_2)$ iff $A_1 \subseteq A_2$ $(\Leftrightarrow d_2 \sqsubseteq d_1)$

7. Appendix B:

To use a pattern structure based on annotations, we need…

- An adapted pattern structure where
  - objects are documents and
  - object descriptions are ontological concepts that annotate documents
- An order (≼) between object descriptions
  - The subsumption relation defines an order between concepts of an ontology,
  - so we can use this order between concepts of an ontology to order object descriptions
- A similarity operation (⊓) between descriptions
  - we can use the order defined by an ontology as a base to the similarity operation.

… an ontology!

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS
–
Fakultät III Medien, Information und Design

UNIVERSITY OF NICOSIA
ΠΑΝΕΠΙΣΤΗΜΙΟ ΛΕΥΚΩΣΙΑΣ

7. Appendix C: Ontology

- DL ontologies
  - represented in Description Logic (DL)
  - set of concepts, relations and individuals
  - set of axioms: concept inclusion (e.g., $C \preccurlyeq D$), concept definition (e.g., $C \equiv D, \exists r.c$), relation inclusion and relation definition
  - concepts can either be atomic or defined
- We consider $\mathcal{EL}$ ontologies
  - concept definitions include only conjunctions ($\wedge$) and existential restrictions ($\exists r.c$)
  - the Least Common Subsumer of two concepts, denoted by $lcs(\{c_1, c_2\})$, always exists
  - the computation of the lcs is tractable, i.e., works with very large ontologies