Comparing human and automatic thesaurus mapping approaches in the agricultural domain

Boris Lauser, Gudrun Johannsen, Caterina Caracciolo, Johannes Keizer, Willem Robert van Hage, Philipp Mayr

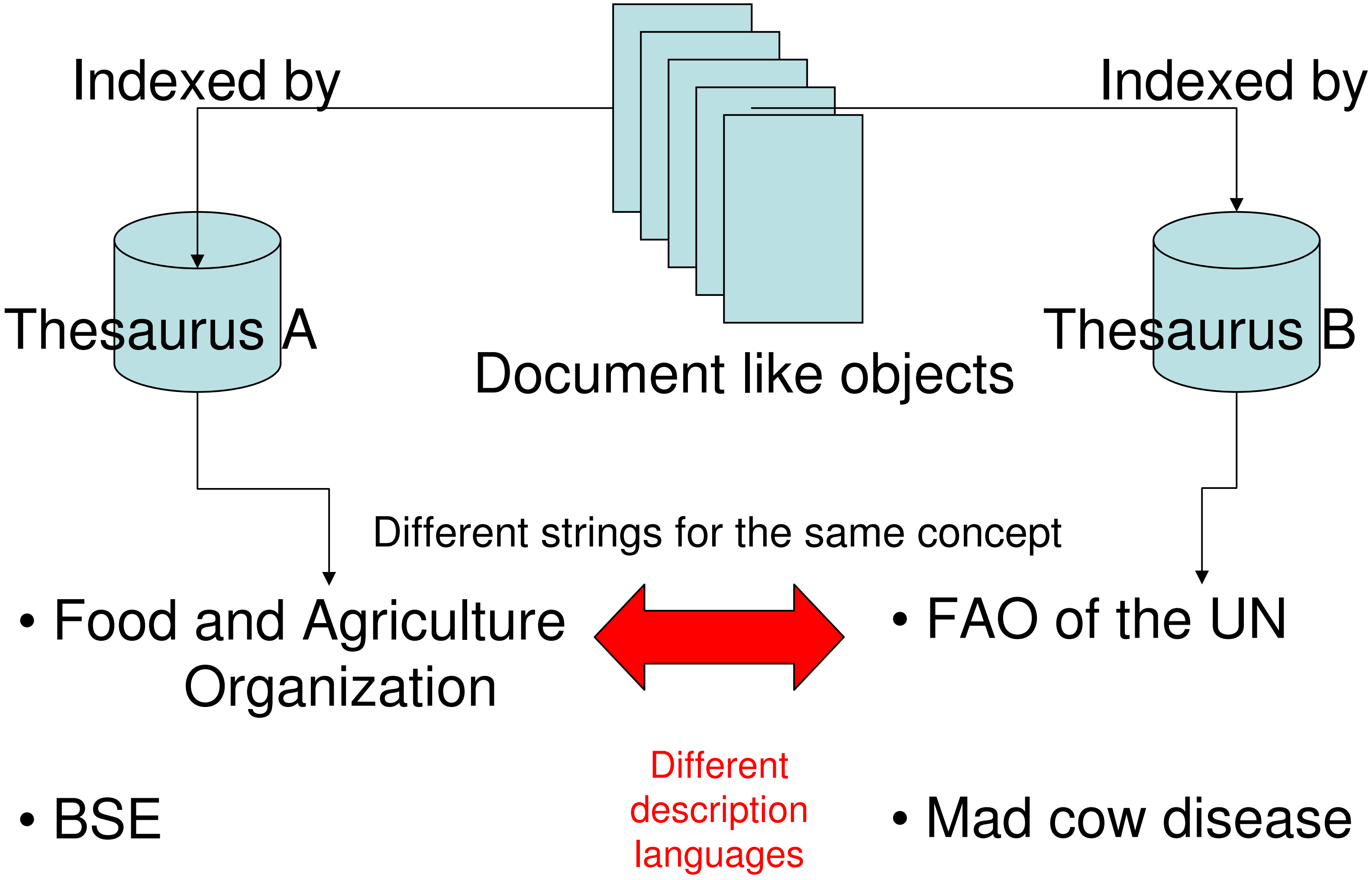
FAO, TNO Science & Industry / Vrije Universiteit Amsterdam and GESIS Social Science Information Centre

International Conference on Dublin Core and Metadata Applications 2008
Berlin, 23 September 2008

Outline

- Problem addressed by mapping
- Motivation of our work
- Experimental setup
- Results
- Conclusions

Problem Scenario: Why mapping?



Indexed by

Indexed by

Thesaurus A

Document like objects

Thesaurus B

Different strings for the same concept

- Food and Agriculture Organization
- BSE

Different description languages

- FAO of the UN
- Mad cow disease

Problem: Heterogeneous collections

- Many databases:
  - document types / formats
  - vocabularies
- Controlled vocabularies:
  - internal consistency (high)
  - intersystem compatibility (low) -> (semantic heterogeneity)
- **Goal**:
  Seamless search across multiple heterogeneous collections/repositories based on semantically rich relations
- **Solution**:
  translate → cross-walks → terminology mapping

Aim of the study

Human and automatic mapping have pros & cons:

- Time, money, correctness, completeness

then

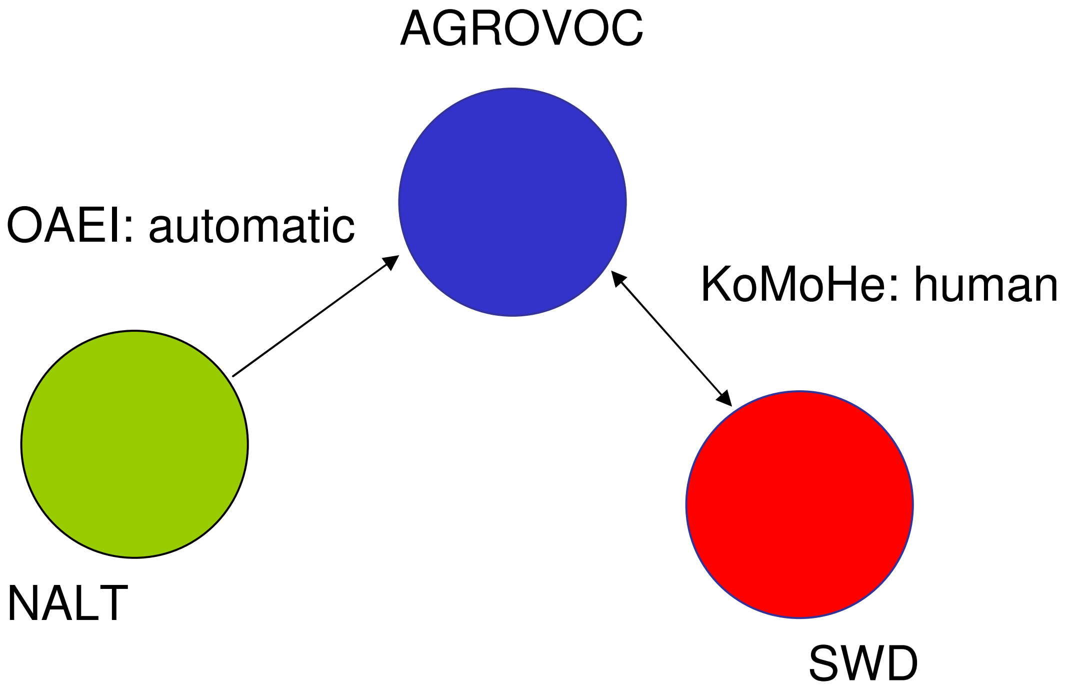
- how and when automatic is best to use automatic vs manual techniques?

Controlled vocabularies in the study

- AGROVOC by FAO
  - Multilingual, structured thesaurus
  - 28,718 descriptors (Engl. version)
- NALT by National Agricultural Lib.
  - Thesaurus
  - 42,326 descriptors (Engl. version)
- SWD by German National Lib.
  - Subject authority file, flat structure
  - 5,350 German terms in agricultural subsection

Initiatives

- OAEI : AGROVOC-NALT mapping (automatic)
- KoMoHe : AGROVOC-SWD mapping (human)



Corresponding mappings within the initiatives

OAEI 2007 food task

- the OAEI (Ontology Alignment Evaluation Initiative)
  - a comparative evaluation initiative for automatic ontology-mapping systems
  - six tasks in 2007: benchmark, anatomy, directory, library, environment, and food
- the OAEI 2007 food task (AGROVOC-NALT)
  - Six mapping systems
    - Falcon-AO - South East University
    - RiMOM - Tsinghua University
    - X-SOM - Polytechnic of Milan
    - DSSim - Open University
    - SCARLET - Open University

    see http://www.few.vu.nl/~wrvhage/oaei2007/food.html

KoMoHe Project (2004-2007)

KoMoHE (Competence Center Modeling and Treatment of Semantic Heterogeneity)

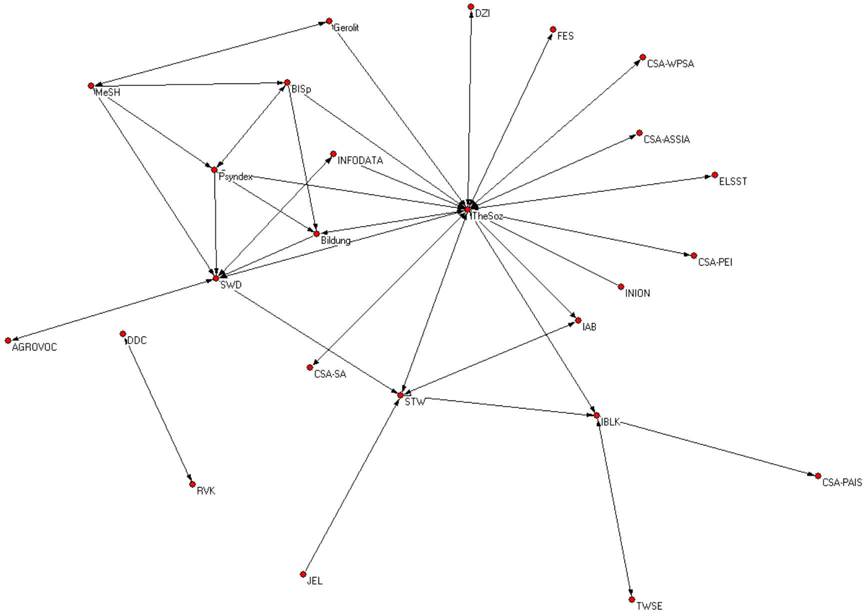
Goals:

- Models for searching heterogeneous collections
- Development, organization & management of cross-walks between controlled vocabularies
- IR evaluation of the mappings (effectiveness of intellectual mapping)

KoMoHe : Cross-concordances

= manually created, directed relations between controlled terms of two knowledge organization systems (KOS)

- 25 Vocabularies in 64 cross-concordances
  - Thesauri (16)
  - Descriptor lists (4)
  - Classifications (3)
  - Subject heading lists (2)



KoMoHe : Relations

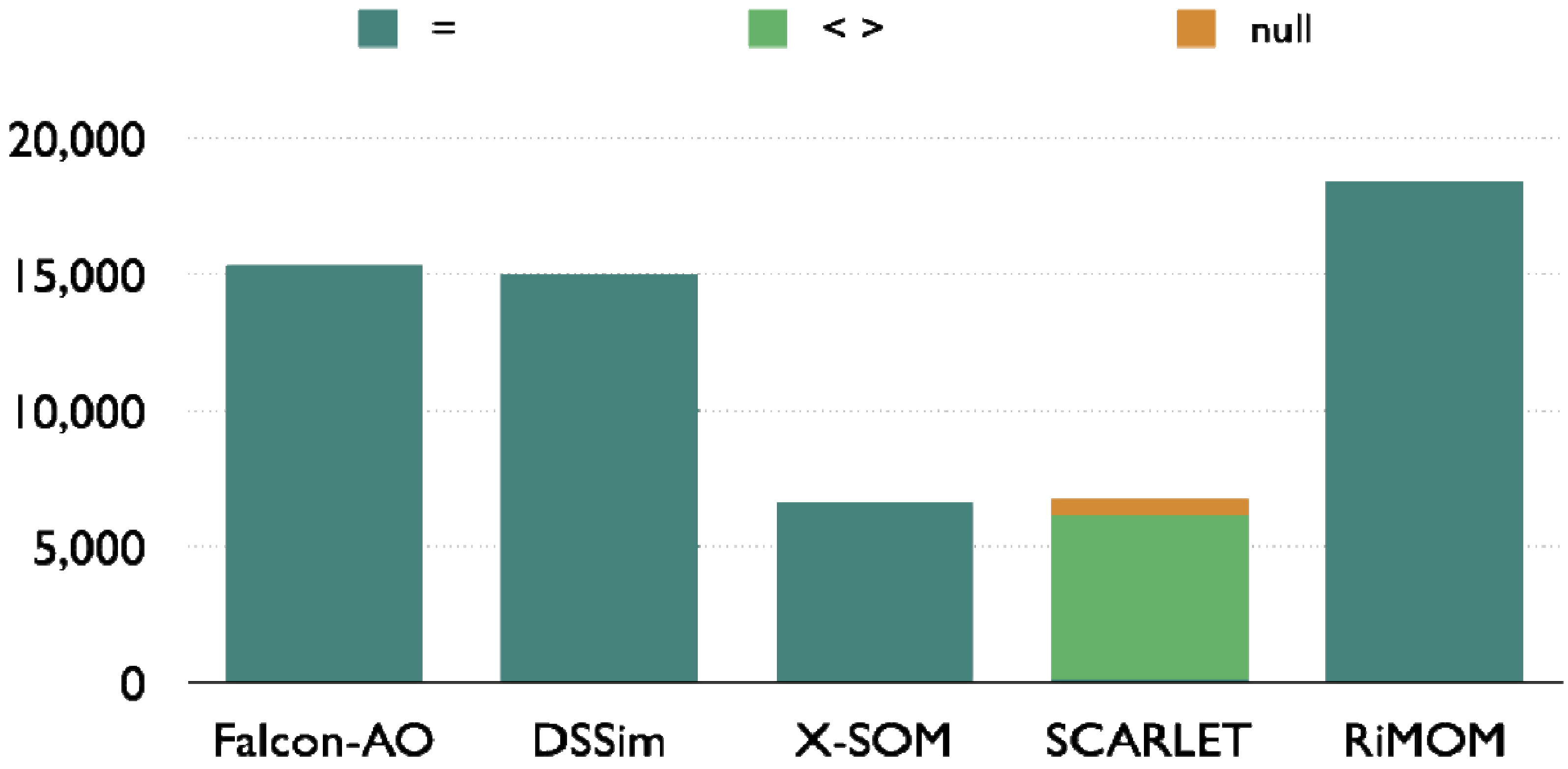
KOS 1	Relation	KOS 2
Library	= equivalence	Bibliothèque
Library	> Narrower term	Special library
Thesaurus	< Broader term	KOS
Hacker	^ Related term	Computers + Security
Virus	0 No mapping	

More details in the presentation on Thursday, 15:30

Mappings in the experiment

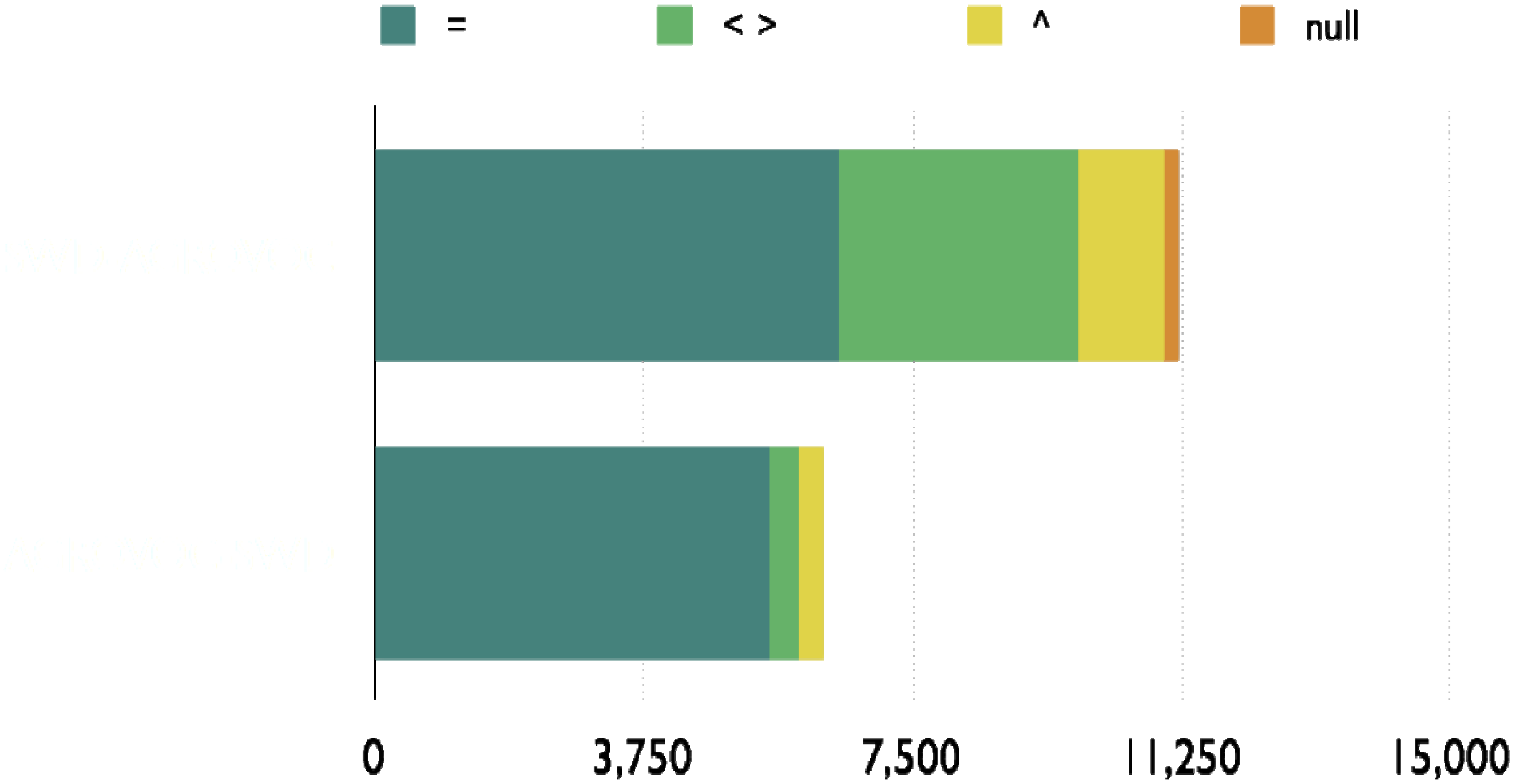
AGROVOC-NALT mapping (OAEI,automatic)

- Number of mapping results and systems involved



AGROVOC-SWD mapping (KoMoHe, human)

- Two crosswalks (SWD-AG and AG-SWD) and number of mappings build



Our hypothesis

1. **Machines are humans' equals in domains with clear naming schemes** (e.g. taxonomy and geography). For other domains, machines are inferior.

2. **Machines cannot find mappings that require background (domain) knowledge.**

Experimental set-up

1. A random sample of 644 mappings from the (union of the) AGROVOC-NALT mappings
2. Mappings classified by their topic:

   taxonomical, biological & chemical, geographical, and miscellaneous
3. false mappings were filtered out
4. to each mapping we added the corresponding SWD-AGROVOC mapping(s)
5. the difficulty of each mapping was judged manually

Mappings by topic

1. Taxonomical

‘Rubus plicatus’ ; ‘Rubus fruticosus’

2. biological & chemical

‘hexachlorobenzene‘ ; ‘Hch’

3. geographical

‘Eastern Africa’ ; ‘East Africa’

4. miscellaneous

‘shelterbelts’ ; ‘Windbreaks’

Classes of mapping according to difficulty

1. **simple**: the preferred terms are literally the same
   → Ananas comosus ; Ananas comosus
2. **alt label**: there is a literal match with an alternative term
   → Lipids ; Fats
3. **easy lexical**: the terms are so close that any layman can see that they match

   → Rocks ; Rock
4. **hard lexical**: the labels are very close, but expert knowledge is needed to see that they match

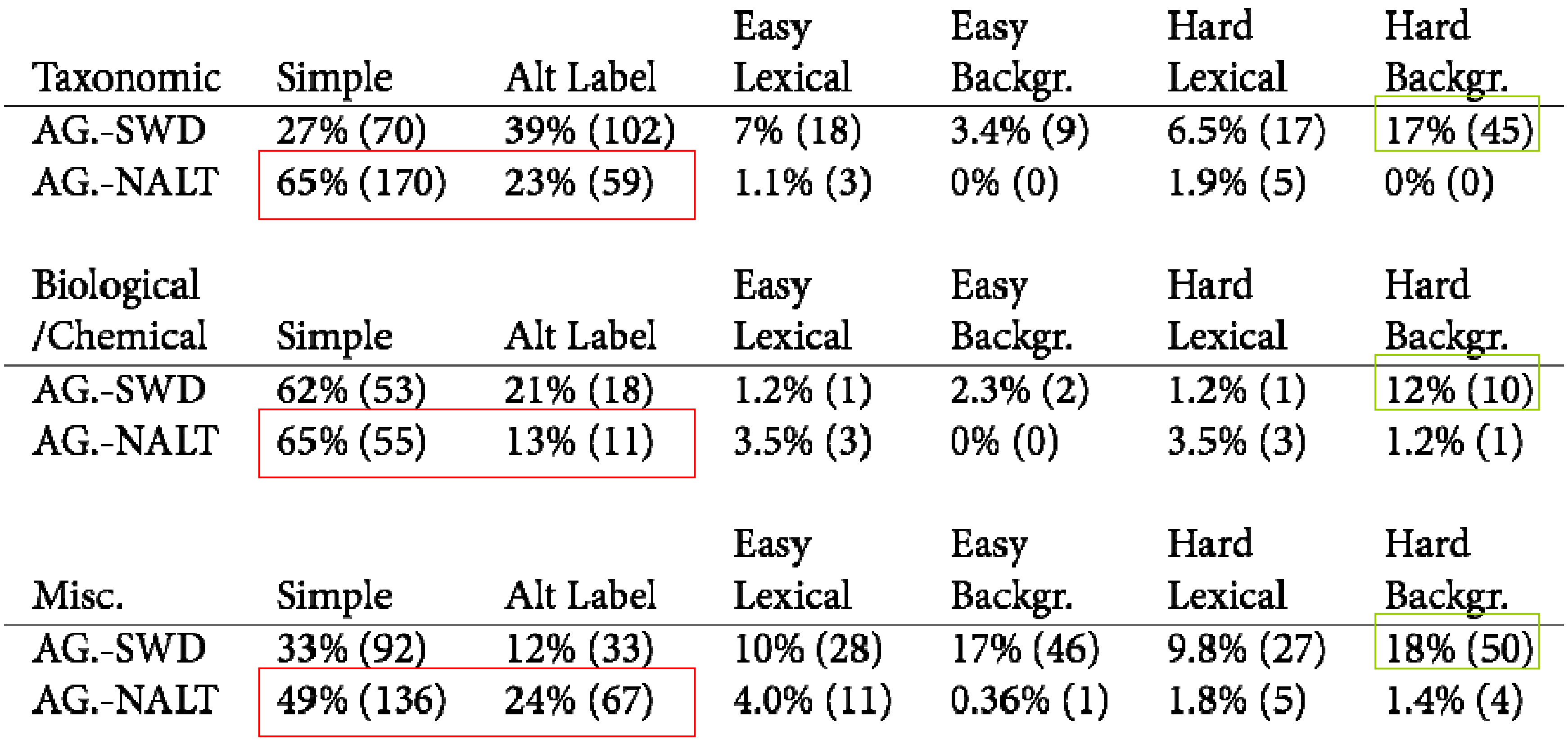
   → Smut diseases ; Smuts
5. **easy background knowledge**: there are no clues as in point 1-4, but general common knowledge suffices to see that the terms match

   → Sewage treatment ; Wastewater treatment
6. **hard background knowledge**: there are no clues as in point 1-4, and domain expertise is needed to see that the terms match

   → Probability analysis ; Statistical methods

Results (difficulties)

All 20 geographical mappings were “Simple”.



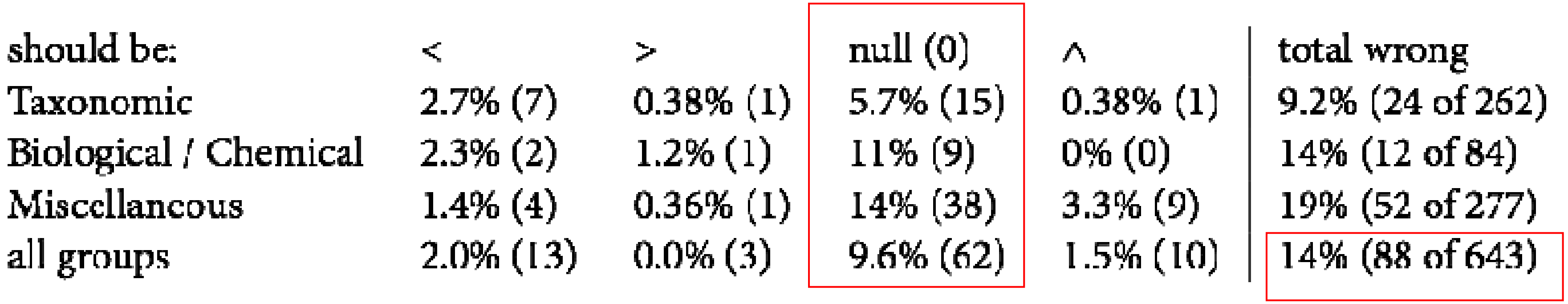
Taxonomic	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	27% (70)	39% (102)	7% (18)	3.4% (9)	6.5% (17)	17% (45)
AG.-NALT	65% (170)	23% (59)	1.1% (3)	0% (0)	1.9% (5)	0% (0)

Biological /Chemical	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	62% (53)	21% (18)	1.2% (1)	2.3% (2)	1.2% (1)	12% (10)
AG.-NALT	65% (55)	13% (11)	3.5% (3)	0% (0)	3.5% (3)	1.2% (1)

Misc.	Simple	Alt Label	Easy Lexical	Easy Backgr.	Hard Lexical	Hard Backgr.
AG.-SWD	33% (92)	12% (33)	10% (28)	17% (46)	9.8% (27)	18% (50)
AG.-NALT	49% (136)	24% (67)	4.0% (11)	0.36% (1)	1.8% (5)	1.4% (4)

errors in the AGROVOC-NALT mappings

- ‘Viola’ in AGROVOC is not a music instrument (should be a 0).
- ‘Sex differentiation disorders’ ; ‘Seed certification’ (should be 0).
- ‘Kater’ (tomcat) is a ‘männliches Individuum’ (male individual).
- ‘Heckstapler’ (rear stapler) is some kind of ‘Handhabungsgeraet’ (handling equipment).



should be:	<	>	null (0)	∧	total wrong
Taxonomic	2.7% (7)	0.38% (1)	5.7% (15)	0.38% (1)	9.2% (24 of 262)
Biological / Chemical	2.3% (2)	1.2% (1)	11% (9)	0% (0)	14% (12 of 84)
Miscellaneous	1.4% (4)	0.36% (1)	14% (38)	3.3% (9)	19% (52 of 277)
all groups	2.0% (13)	0.0% (3)	9.6% (62)	1.5% (10)	14% (88 of 643)

Our hypothesis

1. **Machines are humans' equals in domains with clear naming schemes** (e.g. taxonomy and geography). For other domains, machines are inferior.

2. **Machines cannot find mappings that require background (domain) knowledge.**

Conclusion I: **Hypothesis 1 does not hold as strictly as we phrased it**

- Biological/chemical like geographical terminology is fairly easy to map (over 60% rated as Simple).
- If you include alternative labels, this statement also holds for taxonomic terminology.
- The ‘Miscellaneous’ group is the most difficult.
- BUT, with the exception of geographical terminology, machines are not as good as humans, even in domains with clear naming schemes (error rate 14% in our sample).

Conclusion II: **Hypothesis 2 holds**

- Most systems rely on (lexical) clues from within the thesauri and do not have background knowledge. This is necessary to find most < > relations.
- Therefore, machines have great difficulty to find the same kind of hierarchical mappings (< >) as humans.
- Of course, machines have difficulty to disqualify or exclude a mapping (0 relation).

Conclusion III: summing up...

- Machines might not be humans’ equals, but they can take care of a large portion of the tedious work.
- Further problems appear if you match different disciplines automatically. Especially ‘softer’ sciences are hard to map automatically (e.g. social sciences).

Consequences

- Bi-lingual or interdisciplinary mappings are even more difficult to process automatically
- One need well-structured KOS to get automatic mapping being effective
- Correctness of automatic mapping has to be checked
- More quality measurement aspects: completeness, consistency

OAEI : Systems descriptions I

- Falcon-AO - South East University
  - lexical matcher (V-Doc, similar to edit distance)
  - iterative structural matcher
  - ontology partitioner
  - try harder to find mappings where few obvious mappings are found
- RiMOM - Tsinghua University
  - lexical matcher (edit distance)
  - structural similarity propagation
  - strategy selector (rely more on lexical or structural matches)
  - remove unlikely matches by heuristics
- X-SOM - Polytechnic of Milan
  - lexical matcher (Jaro similarity, Levenshtein, and WordNet Leacock-Chodorow distance)
  - partitioning using SWOOP ontology editing framework
  - no other matchers due to scalability issues

OAEI : Systems descriptions II

- DSSim - Open University
  - lexical matcher (Monger-Elkan, similar to edit distance, plus Jaccard of term token sets)
  - manual partitioning
  - belief combination with Dempster’s rule of combination
- SCARLET - Open University
  - literal matching to third party ontologies in the Watson semantic web search engine
  - Description Logic reasoning over third party ontologies to find relations

Publications

Mayr, Philipp; Petras, Vivien (2008): Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: 74th IFLA World Library and Information Congress. Québec, Canada- http://www.ifla.org/IV/ifla74/papers/129-Mayr_Petras-en.pdf

Mayr, Philipp; Petras, Vivien (2008 to appear): Building a terminology network for search: the KoMoHe project. In: International Conference on Dublin Core and Metadata Applications.

Thank you for your attention!

E-mail: philipp.mayr@gesis.org