Vocabulary Alignment for archaeological Knowledge Organization Systems

14th Workshop on Networked Knowledge Organization Systems
TPDL 2015 Poznan

Lena-Luise Stahn

September 17, 2015

Summary

Introduction

- Motivation
- The German Archaeological Institute and the IR situation
- Goal
- Questions

Project

- Data
- Approach
- Conversion to SKOS
- Vocabulary Alignment with Amalgame

Conclusion

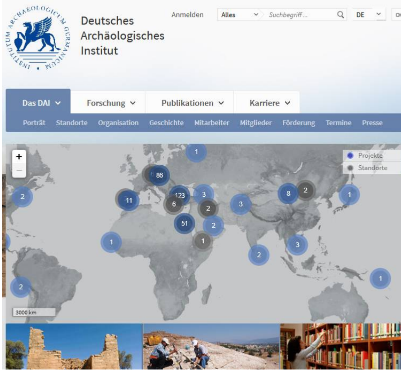
- Results
- Future work
- Conclusion

Motivation

- gap between traditional indexing instruments and scientific study at the DAI becomes bigger
- parallel to traditional thesaurus (started in 19th century) more terminologies have been developed since
- their parallel but separate existence complicates IR and has even discouraging effect
- DAI ”legacy data” prone to get out of use as it appears in several, mostly not standardised formats
- lesser capacities for intellectual indexing, questions about using automatic data mining methods instead
- interoperability and more prevalent use of archaeological KOS is needed

The German Archaeological Institute and the IR situation

- founded in the 19th century, first department in Rome
- in that time mainly focussed on ”classical” antiquity, i.e. from 2000 B.E. to 500 AD (Greeks and Romans)
- since then development to meet the diversifying interests of the archaeological scientific community
- worldwide orientation with more departments (11 + branches and further individual offices) and widely spread field work regarding all historic eras and cultures



Goal

- achieve better information retrieval results through integration of separate vocabularies
- ensure their long term usability and existence through standardised data
- to build the basic line for best practices in dealing with archaeological vocabularies

Questions

- How usable is SKOS as a schema to bring the DAI thesauri in a linked data format? How much effort is to put into the data conversion and what are the specifics of the DAI data?
- Is amalgame the right choice to do the alignment of (German-language) archaeological terminologies? Is a classification of the main errors possible?
- What kind are the matching results of? Is the alignment strategy useful? If not which parameters need to be changed?

Data

- ”Roman” thesaurus:
  - 83.053 records in MARC 21/XML
  - free available from DAI’s OAI-PHM interface
  - mainly focussed on classical antiquity
  - additional separation of thesaurus of Romano-Germanic Commission through Python script
- iDAI.gazetteer
  - 106.902 records
  - delivered as database-dump in json format
  - topographical database
- Charda
  - ”Describing Vocabulary of the Chinese Archaeology Database”
  - 604 entries
  - simple Excel file

Method

- analysis of the three vocabularies, their structure and content
- mapping to SKOS Properties via Python-Script
- feed the ”skosified” data into the alignment tool amalgame and run the label matcher
- evaluation of samples of the alignment results on correctness
- ideally get an idea about precision and recall trends of the overall results so as to adapt/change the alignment strategy

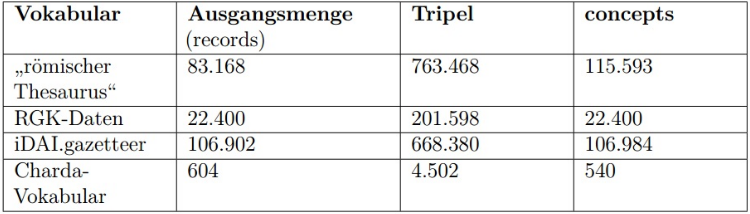
Mapping to the SKOS Properties

SKOS Property	"Roman" Thesaurus (MARC 21 fields)	Gazetteer/ json-record key	Charda table (column)
skos:Concept skos:inScheme	001	'_id'	German term (B)
skos:prefLabel	551.a	'prefName' and all 'names'	B (German) C (English term) D (Chinese term)
skos:altLabel	-	-	Alalternative German terms (K)
skos:hiddenLabel	553.a	'ids' im Kontext „zenon-thesaurus“	-
skos:broader	554.b OR	'parent' OR	Broader German Term (A) OR
skos:topConceptOf respectively skos:hasTopConcept	In case of no entry in 554.b	Falls kein Eintrag in 'parent'	In case of no Broader Term (A)
skos:related	-	'relatedPlaces'	-
skos:definition	-	'types'	-
skos:scopeNote	-	'comments'	-
skos:Concept skos:inScheme skos:prefLabel skos:broader	552.r or 552.m or 552.e	'tags'	-
owl:sameAs	-	'ids'	-

Output

```
<rdf:Description rdf:about="https://gazetteer.dainst.org/place/2296437">
  <skos:definition>archaeological-site</skos:definition>
  <owl:sameAs rdf:resource="http://arachne.uni-koeln.de/entity/1208422"/>
  <skos:prefLabel>Amarna</skos:prefLabel>
  <skos:prefLabel xml:lang="pol">Tell el-Amarna</skos:prefLabel>
  <skos:hiddenLabel>zTopogAsienVordeSyrieTell Amar</skos:hiddenLabel>
  <owl:sameAs rdf:resource="http://sws.geonames.org/347585"/>
  <owl:sameAs rdf:resource="http://zenon.dainst.org/000074457"/>
  <skos:inScheme rdf:resource="https://gazetteer.dainst.org/place/thesaurus"/>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:prefLabel xml:lang="por">Amarna</skos:prefLabel>
  <skos:prefLabel xml:lang="eng">Amarna</skos:prefLabel>
  <skos:prefLabel xml:lang="ita">Amarna</skos:prefLabel>
  <skos:prefLabel xml:lang="ara">أخت  أتون</skos:prefLabel>
  <skos:definition>populated-place</skos:definition>
  <skos:related rdf:resource="https://gazetteer.dainst.org/place/2296228"/>
  <skos:prefLabel xml:lang="fra">Tell el-Amarna</skos:prefLabel>
  <skos:broader rdf:resource="https://gazetteer.dainst.org/place/2086499"/>
  <skos:related rdf:resource="https://gazetteer.dainst.org/place/2281769"/>
  <skos:prefLabel xml:lang="rus">Телль-эль-Амарна</skos:prefLabel>
  <skos:scopeNote xml:lang="eng">Near Tall al-Amarna</skos:scopeNote>
  <skos:related rdf:resource="https://gazetteer.dainst.org/place/2296229"/>
  <skos:prefLabel xml:lang="spa">Tell el-Amarna</skos:prefLabel>
  <owl:sameAs rdf:resource="http://arachne.uni-koeln.de/place/6332"/>
  <skos:prefLabel xml:lang="deu">Tall ʿamarna</skos:prefLabel>
</rdf:Description>
```

Output quantity

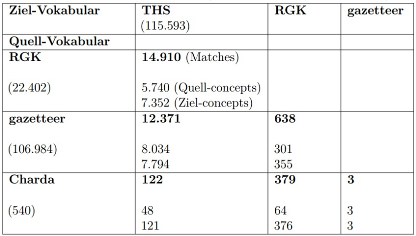


Vokabular	**Ausgangsmenge** (records)	**Tripel**	**concepts**
„römischer Thesaurus“	83.168	763.468	115.593
RGK-Daten	22.400	201.598	22.400
iDAI.gazetteer	106.902	668.380	106.984
Charda-Vokabular	604	4.502	540

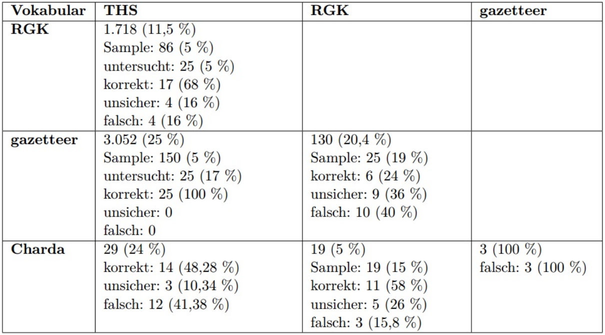
Amalgame

- developed at the Free University of Amsterdam as part of the ClioPatria rdf-environment and triple store
- written in Prolog
- can deal with SKOS data, whereas most alignment tools only work on OWL data: main point for choice
- unfortunately scarce documentation, infos via direct communication with developers:
- ”[...] But the exact match is really simple: - it really only matches if the two labels are identical - it does case-insensitive by default, you can switch this in the settings - it will match ”foobar”@en to ”foobar”@de unless you say do not match cross language.”
- thus matching is done on string level only; ok in study intended as starting point
- strategy variations: match across languages

Quantity and Quality of found matches

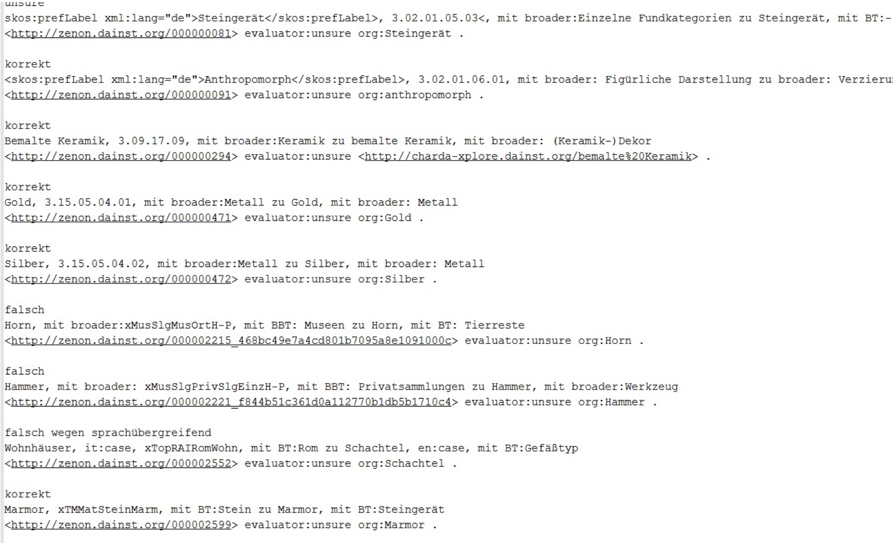


Ziel-Vokabular	**THS** (115.593)	**RGK**	**gazetteer**
Quell-Vokabular			
RGK (22.402)	**14.910** (Matches) 5.740 (Quell-concepts) 7.352 (Ziel-concepts)		
gazetteer (106.984)	**12.371** 8.034 7.794	**638** 301 355	
Charda (540)	**122** 48 121	**379** 64 376	**3** 3 3



Vokabular	**THS**	**RGK**	**gazetteer**
RGK	1.718 (11,5 %) Sample: 86 (5 %) untersucht: 25 (5 %) korrekt: 17 (68 %) unsicher: 4 (16 %) falsch: 4 (16 %)		
gazetteer	3.052 (25 %) Sample: 150 (5 %) untersucht: 25 (17 %) korrekt: 25 (100 %) unsicher: 0 falsch: 0	130 (20,4 %) Sample: 25 (19 %) korrekt: 6 (24 %) unsicher: 9 (36 %) falsch: 10 (40 %)	
Charda	29 (24 %) korrekt: 14 (48,28 %) unsicher: 3 (10,34 %) falsch: 12 (41,38 %)	19 (5 %) Sample: 19 (15 %) korrekt: 11 (58 %) unsicher: 5 (26 %) falsch: 3 (15,8 %)	3 (100 %) falsch: 3 (100 %)

matching results sample rdf/xml file



```
unsure
skos:prefLabel xml:lang="de">Steingerät</skos:prefLabel>, 3.02.01.05.03<, mit broader:Einzelne Fundkategorien zu Steingerät, mit BT:-
<http://zenon.dainst.org/000000081> evaluator:unsure org:Steingerät .

korrekt
<skos:prefLabel xml:lang="de">Anthropomorph</skos:prefLabel>, 3.02.01.06.01, mit broader: Figürliche Darstellung zu broader: Verzieru
<http://zenon.dainst.org/000000091> evaluator:unsure org:anthropomorph .

korrekt
Bemalte Keramik, 3.09.17.09, mit broader:Keramik zu bemalte Keramik, mit broader: (Keramik-)Dekor
<http://zenon.dainst.org/000000294> evaluator:unsure <http://charda-xplore.dainst.org/bemalte%20Keramik> .

korrekt
Gold, 3.15.05.04.01, mit broader:Metall zu Gold, mit broader: Metall
<http://zenon.dainst.org/000000471> evaluator:unsure org:Gold .

korrekt
Silber, 3.15.05.04.02, mit broader:Metall zu Silber, mit broader: Metall
<http://zenon.dainst.org/000000472> evaluator:unsure org:Silber .

falsch
Horn, mit broader:xMusSlgMusOrtH-P, mit BBT: Museen zu Horn, mit BT: Tierreste
<http://zenon.dainst.org/000002215_468bc49e7a4cd801b7095a8e1091000c> evaluator:unsure org:Horn .

falsch
Hammer, mit broader: xMusSlgPrivSlgEinzH-P, mit BBT: Privatsammlungen zu Hammer, mit broader:Werkzeug
<http://zenon.dainst.org/000002221_f844b51c361d0a112770b1db5b1710c4> evaluator:unsure org:Hammer .

falsch wegen sprachübergreifend
Wohnhäuser, it:case, xTopRAIRomWohn, mit BT:Rom zu Schachtel, en:case, mit BT:Gefäßtyp
<http://zenon.dainst.org/000002552> evaluator:unsure org:Schachtel .

korrekt
Marmor, xTMMatSteinMarm, mit BT:Stein zu Marmor, mit BT:Steingerät
<http://zenon.dainst.org/000002599> evaluator:unsure org:Marmor .
```

Results

- conversion to SKOS worked fine: provided Properties met the DAI-data's requirements
- data itself brought on bigger problems: considerable amount of manual adjustments and cleaning was necessary
- big differences in coverage and dimension of the DAI-data caused great deal of wrong matches,
- Amalgame unable to recognize specifics of the German language (e.g. Umlauts), therefore future use of this tool needs to be reconsidered
- results showed that sensible selection of source vocabularies is necessary (e.g. Charda and gazetteer)
- however Alignment results show almost 50 % correctness, which can be considered as good, factoring only simple label exact matching algorithm as well as very dissimilar source vocabularies

Future Work

- adapt alignment strategy (better selection and adaptation of source vocabularies, additional matching algorithms etc.)
- use further alignment tools to get comparable, and as of that, more reliable results, especially in those cases where corrections of the strategy are necessary
- ’skosification’ and alignment of more DAI vocabularies
  - maintenance tool and workflow for ’skosified’ vocabularies needed
- connect the data to the LOD cloud

Conclusion

lessons learned

- SKOS useful and flexible enough for the DAI-data
- data too diverse in coverage and dimension, separation and selection needed
- additional alignment algorithms and tools need to be tested for more comparable data

Conclusion

what can you get from this very individual case?

- can only serve as starting point for Ontology Matching strategy on archaeological vocabularies
- use case for standardising heterogeneous ’legacy data’ to improve their long term usability
- base line for workflow for data interoperability and long term usability to improve information retrieval situation in the classical studies at large

Thank you!

Questions?