

Comparing SSH vocabularies and their applications in different systems

TRIPLE Event on Vocabularies, Berlin 2023-03-28

Nina C. Rastinger, Massimiliano Carloni, Klaus Illmayer, Matej Ďurčo

Index

1. Preliminary considerations
2. Recent work at ACDH-CH
on comparing/aligning/curating vocabularies
3. Practical part (Jupyter Notebooks):
Comparing vocabularies
4. Outlook

Terminology

- Controlled vocabularies vs. Folksonomies (central bureau vs. anarchy)
- Level of control / curation
 - Autocomplete for free keywords
 - Hybrid approach: allowing new keywords as input for a curation workflow
- "Terms" / "keywords" / "keyword-strings" / "tags" / ...
- "Topics" / "subjects" / "(subject) headings" / ...

Vocabulary use

- Description of resources / metadata (indexing)
- Discovery / findability (retrieval)
- Two use cases
 - Allowed values for certain properties
 - Generic categorisation
- Specificities of more structured vocabularies
 - Some controlled terms might not be used
 - Some controlled terms might be understood only in their hierarchical context

Vocabulary interoperability

- Improving FAIRness of resources
- Harmonisation / Normalisation
As part of aggregation workflows
- Interlinking of vocabularies / Direct mapping
- Creation of a switching vocabulary
- Using multiple vocabularies for a certain property
- Autocomplete and recommender systems

User perspective: free choice

- In case of free choice for users – how do they use it?
- Some categories
 - **Time spans:** 18th century, Middle Ages, nineteenth century, ...
 - **Events:** Second World War, ...
 - **Places:** Hüyücektepe, Yeldeğirmentepe, Greece, ...
 - **Disciplines:** medieval prosopography, pottery studies, linguistics, ...
 - **Methods:** lemmatization, GPR (ground-penetrating radar), machine learning, ...
 - **Persons:** Alfred Kerr, Karl Kraus, Mozart, ...
 - **Object types:** newspapers, literary fiction, pottery fragments, ...
 - **Languages:** Latin, Latvian, Greek, ...

User perspective: metadata schemas

- The more structured the metadata schema, the more we can avoid "mixed bags"
- For example (from the [ARCHE metadata schema](#)):
 - has Applied Method
 - has Category
 - has Language
 - has Related Discipline
 - has Subject
 - has Spatial Coverage
 - has Temporal Coverage
 - ...

Vocabularies for specific aspects

Disciplines	<u>ÖFOS</u>
Methods	<u>TaDiRAH</u>
Languages	<u>ISO 639-1</u> , <u>ISO 639-3</u>
Object types	<u>ARCHE Resource Type Category</u>

User perspective: "subjects"

- The example of **"has Subject" (not controlled)**
- Keywords to supplement concepts missing in specific controlled vocabularies
 - "Geschichte des Lesens" (i.e., "History of reading") (discipline)
 - "Zeitung" (i.e., "Newspaper") (resource type)
- Keywords for internal organisation of a collection
 - Administrative units ("Province", "County", ...) in [HistoGIS](#)
 - Sometimes from project-specific controlled vocabularies (e.g., [Legal Kraus](#))
- What is a "subject"? Depends on the context
 - "CIDOC CRM" as a subject in a [project proposal](#)

Recent work at ACDH-CH

- Comparison of different "general-purpose" vocabularies used in DH context
- Starting point: Austrian DH community
- Collection from different types of sources, e.g.:
 - Repositories
 - Research Projects overview websites
 - Training resources
- Motivation: Get a deep overview of what vocabularies are being used in the field and to which extent they overlap
=> what are the most common terms/concepts used

Method: Step 1 – Data preparation

- Vocabularies available in different forms: SKOS / CSV / HTML-page
=> Collection and format transformation
- Homogenisation
 - Consideration of Language
 - here: only vocabularies with English as (main/official) language
 - Deletion of duplicates
 - Inclusion of alternative labels (if given)
 - Cleaning (e.g. lowercase, removing vocabulary-specific prefixes/numbers)

Vocabularies used

- ACDH-CH website: <https://www.oeaw.ac.at/acdh>
- ACDH-CH HowTo: <https://howto.acdh.oeaw.ac.at>
- ARCHE (Property hasSubject): <https://arche.acdh.oeaw.ac.at>
- DARIAH Campus: <https://campus.dariah.eu/tags>
- dha website: <https://digital-humanities.at/de/dha/projects>
- dha taxonomy: https://vocabs.acdh.oeaw.ac.at/dha_taxonomy
- GAMS: <https://gams.uni-graz.at/context:gams.projekte>
- KONDE Zotero: <https://www.zotero.org/groups/1332658/konde/library>
- SSH Open Marketplace: <https://marketplace.sshopencloud.eu/>

Overview of investigated vocabularies

Resource	Naming	Tagged objects	Controlled?
ACDH-CH website	"tags"	projects, tools	no
ACDH-CH HowTo	-	tutorials	no
ARCHE (Property: hasSubject)	"subjects"	resources	no
DARIAH Campus	"topics"	tutorials	no
dha website	-	projects	no
dha taxonomy	"subjects"	open	yes
GAMS	"keywords"	resources	yes
SSH Open Marketplace	"keywords"	resources	no
Zotero-Library of KONDE	"tags"	publications	no

Method: Step 2 – Comparison

- Pair-wise matching of keywords
 - Strict (exact) vs. fuzzy matching

Strict vs. fuzzy matching

- challenges of matching: different spelling, inflections, typos, ...
- levensthein distance ([polyleven](#)) as one approach (e.g. < 3):

visualizations – visualisation

User centered design – user-centered design



tei – mei

(*wrong match*)

geodata – geographical data

(*not matched*)

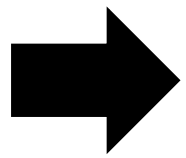


Method: Step 2 – Comparison

- Pair-wise matching of keywords
 - Strict (exact) vs. fuzzy matching
- Result:
 - Overlaps between vocabularies, hints for similarity/relatedness
 - Popularity of keywords
 - In how many vocabularies is a keyword used?
 - (If usage counts are given: How often is a keyword used?)

Method Step 3 – Expansion of comparison

- Expanding the set of vocabularies (further keywords)
- Broadening the scope (general + specific vocabularies)
- Comparing subject vocabularies with specific vocabularies to identify terms in general vocabularies, which actually describe a certain aspect



Getting information on the contents of the potentially 'mixed bags' of keywords
(e.g. how frequently do they contain disciplines, languages, places?)

First results: unique vs. recurring keywords

Comparison of 9 vocabularies with 1705 distinct keywords

	Strict Matching	Fuzzy Matching (< 2)	Fuzzy Matching (< 3)
Unique	1486	1392	1224
Overlap between 2 vocabularies	139	187	222
Overlap between 3 vocabularies	50	74	105
Overlap between 4 vocabularies	18	29	58
Overlap between 5 vocabularies	8	15	43
Overlap between 6 vocabularies	3	6	24
Overlap between 7 vocabularies	0	0	13
Overlap between 8 vocabularies	1	2	14
Overlap between 9 vocabularies	0	0	2

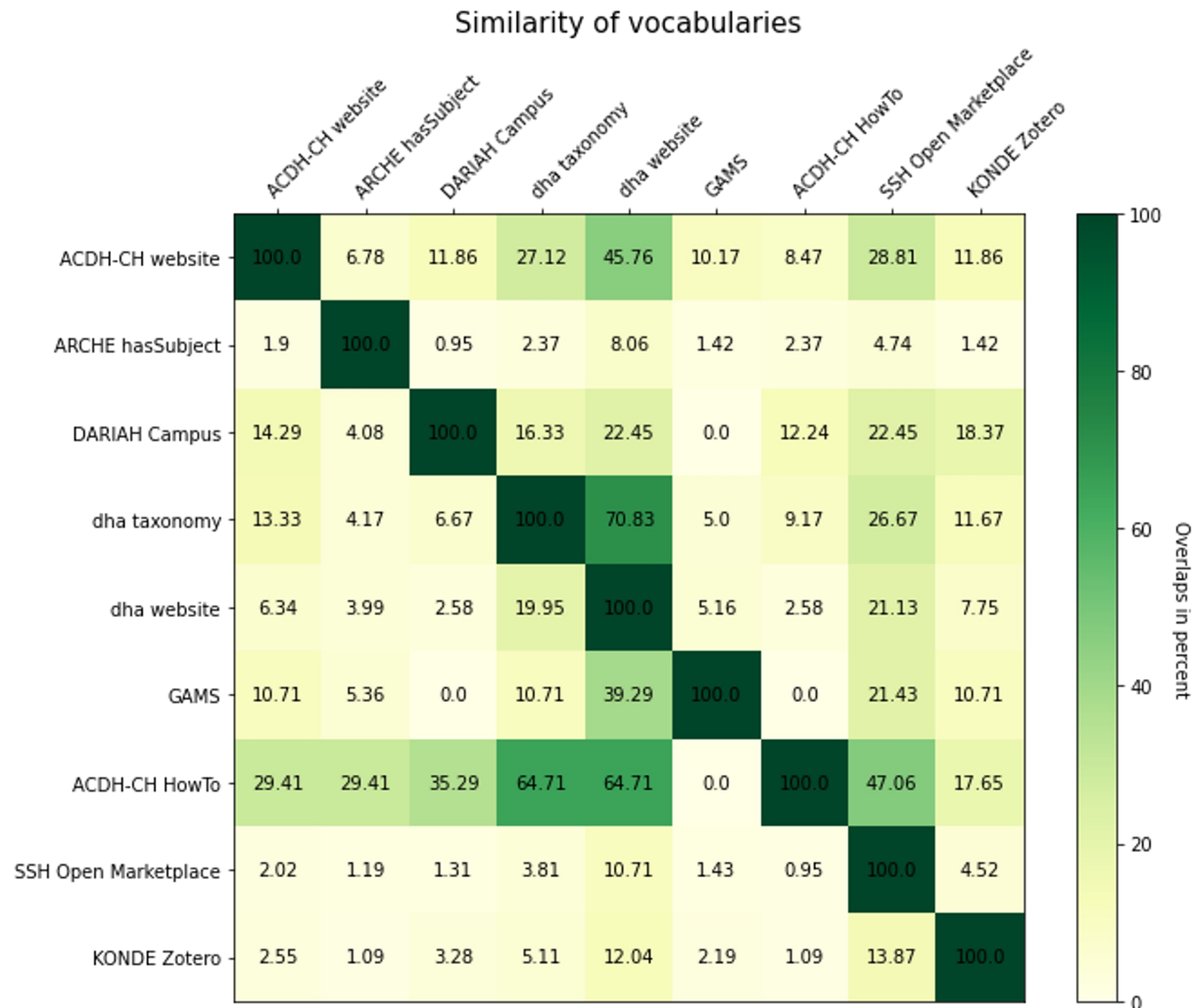
First results: recurring keywords

Keywords recurring in multiple vocabularies -> hints at high relevance for DH

Keyword	Number of vocabularies	Vocabularies
tei	8	'ACDH-CH website', 'ARCHE', 'DARIAH Campus', 'dha taxonomy', 'dha website', 'ACDH-CH HowTo', 'SSH Open Marketplace', 'Zotero (KONDE)'
xml	6	'ARCHE', 'DARIAH Campus', 'dha taxonomy', 'ACDH-CH HowTo', 'SSH Open Marketplace', 'Zotero (KONDE)'
semantic web	6	'ACDH-CH website', 'DARIAH Campus', 'dha taxonomy', 'dha website', 'ACDH-CH HowTo', 'Zotero (KONDE)'
metadata	6	'ACDH-CH website', 'DARIAH Campus', 'dha taxonomy', 'dha website', 'SSH Open Marketplace', 'Zotero (KONDE)'

Percentual overlaps

- Heatmaps for visualising (potential) similarity of vocabularies
- Here: values from strict matching
- In general: rather low coverage rates, with a few exceptions



Taking a look into the mixed bag(s) of keywords

- Exemplary bag: SSH Open Marketplace keywords
- Strict matching with specific vocabularies (e.g. dnet countries, dnet languages, ISO-639-3):
 - 18 activities (although specific property exists)
 - 20 cases of countries
 - at least 79 cases of languages

Vocabulary	SSH Open Marketplace	dnet countries	dnet languages	dnet publication resource	ISO-639-3	TaDiRAH
SSH Open Marketplace	840	20	79	4	77	18
dnet countries	20	256	1	1	4	0
dnet languages	79	1	415	0	284	0
dnet publication resource	4	1	0	48	0	0
ISO-639-3	77	4	284	0	7863	1
TaDiRAH	18	0	0	0	1	178

Practical Part: Let's compare our vocabularies

- Use the above method to align/compare vocabularies "around the table"
- Jupyter Notebook available on Zenodo (can be run on Google Colab)
- Further vocabularies prepared

Further subject vocabularies added

Resource	Naming	Tagged objects	Controlled?
Getty AAT	-	open	yes
GoTriple Vocabulary	-	open	yes
OpenAIRE subjects of research community "DARIAH" - scheme: ACM	"subjects"	publications	yes
OpenAIRE subjects of research community "DARIAH" - scheme: arxiv	"subjects"	publications	yes
OpenAIRE subjects of research community "DARIAH" - scheme: icsh	"subjects"	publications	yes
OpenAIRE subjects of research community "DARIAH" - scheme: keyword	"subjects"	publications	no
OpenAIRE subjects of research community "DARIAH" - scheme: MAG	"subjects"	publications	yes

Further specific vocabularies added

Disciplines	FOS
Methods	TaDiRAH , ARCHE hasAppliedMethod
Languages	ISO 639-1 , ISO 639-3 , dnet languages
Resource types	ARCHE Resource Type Category , dataCite resource , dnet result type , dnet publication resource
Countries	dnet countries
+ various archaeology-specific vocabularies from Heritage Data	

Practical Part: Let's compare our vocabularies

- Use the above method to align/compare vocabularies "around the table"
- [Jupyter Notebook](#) available on Zenodo (can be run on Google Colab)
- Further vocabularies prepared
- 39 prepared vocabularies (controlled/keyword-strings)
- 71.208 distinct keywords

Outlook

- Alignment of vocabularies between partners?
- Understanding users' use of free text fields for 'topics', 'subjects' etc.?
- Candidates for pruning?
- Matching keywords to specific vocabularies (e.g. discipline, method)?

Links & References

- [ACDH-CH Vocabs](#) / [DARIAH Vocabs](#)
- [Triple Vocabulary: an SSH multilingual vocabulary based in LCSH](#)
- [OpenAIRE Vocabularies](#)
- [Getty Art & Architecture Thesaurus](#)
- Harpring, Patricia. *Introduction to Controlled Vocabularies*. 2nd ed. Los Angeles: Getty Research Institute, 2013.
- Lippell, Helen, ed. *Taxonomies: Practical Approaches to Developing and Managing Vocabularies for Digital Information*. London: Facet Publishing, 2022.
- Pomerantz, Jeffrey. *Metadata*. Cambridge, MA: MIT Press, 2015, pp. 48-54 («Metadata Gone Wild!»)

Thank you for your attention!

ninaclaudia.rastinger@oeaw.ac.at

massimiliano.carloni@oeaw.ac.at

klaus.illmayer@oeaw.ac.at

matej.durco@oeaw.ac.at