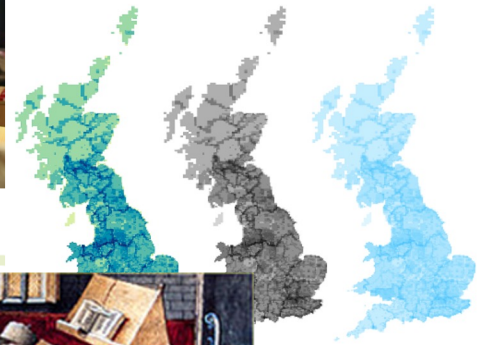
SSH VOCABULARY COMMONS

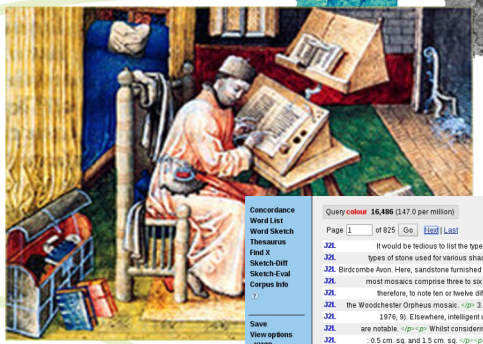
Daan Broeder - CLARIN ERIC,

TRIPLE Event Berlin March 27, 2023

Diversity in describing phenomena

- A considerable part of research concerns describing and analyzing phenomena using descriptive schemas and concepts
- Typical for the SSH is a high variety of such schema and concepts, caused by
  - wide variety of data types, sub-community specifics, schools of thought,…
  - divergent purposes and available effort
- Suitable well crafted vocabularies are essential for
  - accurate descriptions and classifications, countering interpretative vagueness (reduce ambiguity)
  - efficient information retrieval

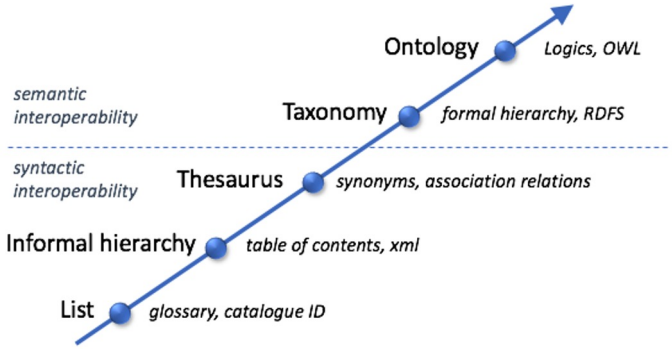


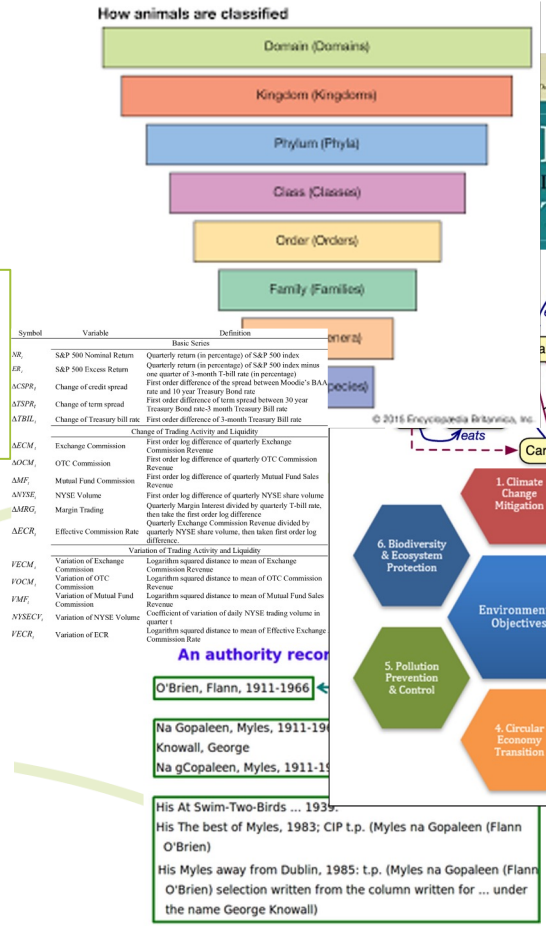


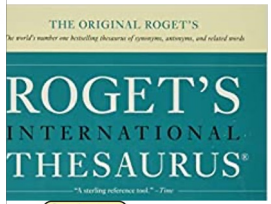
Controlled Vocabularies

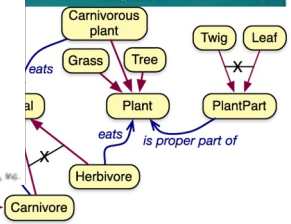
- Lists
- Glossaries
- *Thesauri*
- *Taxonomies*
- *Ontologies*

W3C std. RDF, SKOS, OWL:
widely accepted and supported formal way to represent vocabularies









Vocabularies in the SSHOC project

- Coordination wrt. vocabularies: originally a limited effort
  - Investigation of a common recommended platform for publishing and sharing vocabularies
  - Testing machine translation for vocabularies
  - Flexible integration of vocabularies in tools: e.g. SSHOC Dataverse
  - Identifying & creating proper vocabularies for SSH Marketplace and others
- Identified more opportunities during the project
  - Inventory and registration of relevant SSH vocabularies
  - Recommendations for further common approaches e.g. CV authoring tools
  - Opportunity & need to represent SSH interests with other stakeholders e.g. software & service providers
- Continued discussion about collaboration in the SSH Vocabulary Commons

The importance of multilinguality for the SSH and its vocabularies

- SSH researchers often produce culturally and socially relevant work in their local languages
- Deposit and search facilities (for data and papers) are still mostly available only using English
- Translate metadata, keywords, terminologies and their definition can **enable multilingual search** and **facilitate access**
- Multilingual metadata terms and vocabularies are meant to provide multilingual access to content across different languages and to improve discovery of resources and tools by non-native English speakers.

Vocabulary visibility and discovery

- Vocabularies not always FAIR yet; they need proper registration and publication so researchers & infrastructure providers can discover and reuse -> see the **FAIRsFAIR project report for FAIR semantic artefacts and now its follow-up the FAIR IMPACT project**
- Need a SSH Vocabulary catalogue or a general one that supports sufficient discipline specificity e.g. Bartoc.org (3300 entries whereof 1200 SSH)
- Vocabulary search function, that searches in vocabulary metadata **but also** the vocabulary terms themselves.
- *Note that providing optimal recommendations for researchers can be complicated e.g. also aspects of context and user profile play a role*

Vocabularies & Interoperability

For data reuse and data integration we have to look at interoperability of vocabularies

- **Technical / Format interoperability.** SKOS and OWL are broadly accepted
  - but many projects use spreadsheets and tables and are locked in silos using highly specific software to manage and use these
  - Specific recommendations for vocabulary versioning are needed
  - NOTE there is a small but persistent call for **MORE THEN SKOSS, a topic to discuss**
- **Semantic interoperability**. Coming from different traditions different organizations and projects have developed different vocabularies to describe similar data. Normalization or conversion needed; the vocabularies involved can be huge and expertise expensive (Ariadne Vocabulary Matching Tool and others).
- **Cultural & Human interoperability** aspects. Multilingual vocabularies, localization aspects. -> MT technology + network of human experts

A SSH Vocabulary Commons

Common interest by SSH RIs: CESSDA, CLARIN, DARIAH and E-RIHs

Providing recommendations and infrastructure enabling

- common collaborative use and management of vocabularies
- have vocabularies as first-class citizens / FAIR data objects in their own right
- Interoperability beyond the SKOS format
  - Versioning recommendations
  - Minimal vocabulary metadata

Wider scope for collaboration in the Humanities:
- national SSH projects (CLARIAH, NFDI), TRIPLE project, …
- new relevant EU projects submitted

Vocabulary Commons Charter available here

Priorities for the Vocabulary Commons

- Operating a Vocabulary repository for SSHOC results and ‘orphaned’ SSH vocabularies
  - Accomplished
- Vocabulary recommendations
  - SSH vocabulary overview of all relevant SSH vocabularies
  - Vocabulary federated content search “Deep” search also in the terms
  - Some tests made using SKOSMOS SPARQL endpoint for querying elsewhere hosted vocabularies, was discouraged by NFL
- Recommendations for versioning of vocabularies
  - Good document from Darren Bell , but still (26-3-2023) to be finalized
- Minimal vocabulary metadata
  - Inventory of existing practices made, but no selection yet
- Foster exchange between users and developers (eg. SKOSMOS developers at NFL)
  - Advise available from NFL
  - OntoPortal is manifesting & pushing, but is perhaps overkill for our purpose

The SSH operate a multitude of vocabulary platform instances at EU, national and institute organization levels

This may be needed from

- an organizational perspective ie. separate responsibility and control
- possibility to configure and change the platform at will
- …

But would sharing also operational effort ie running the platform(s) be usefull next to sharing vocabularies themselves

Vocabulary recommendations

Many vocabularies exist and are registered and published on one or more general or thematic hosting platforms and registries such as schema.org, Bartoc vocabulary registry, Open metadata registry (RDA), BioPortal, …

To support vocabulary recommendations we can use an existing registry eg. Bartoc, or create our own instance. Requirements are:

- specific metadata supporting vocabulary finding
- support for searching in the vocabulary terms
- adequate recall vs. precision when searching in a large general registry

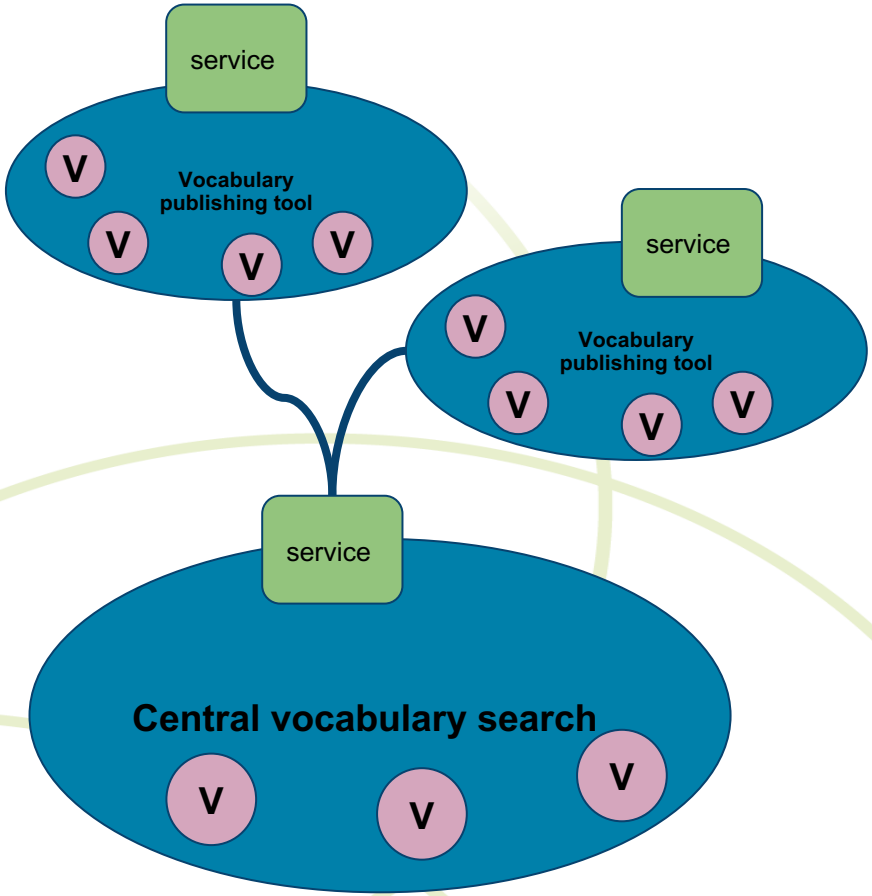
Vocabulary Search

Effective vocabulary recommendations for researchers require:

- Browsing/searching via metadata AND
- Searching in the vocabulary content

Considering two strategies:

- Central metadata and content harvesting and indexing (-> update/syncing problem)
- Federated search, possibly on the basis of
  - SKOSMOS instances (SPARQL endpoint performance problem)
  - Bartoc registry software offers alternative
  - Build our own?



Collaborative Vocabulary Management

Imagine that vocabulary management is “solved” and reuse by means of copy too

Still:

Reuse and management of a vocabulary across multiple organisational scopes is not

- Ownership in a distributed (loosely coupled) setup => authority
- Procedures for evolution and agreement (new concepts, semantic shift)
- Synchronisation with source/target applications
  Ideally, a user wants to adjust vocabulary (add new concepts) from the point of use so how can vocabulary updates be initiated and authorized from the user side

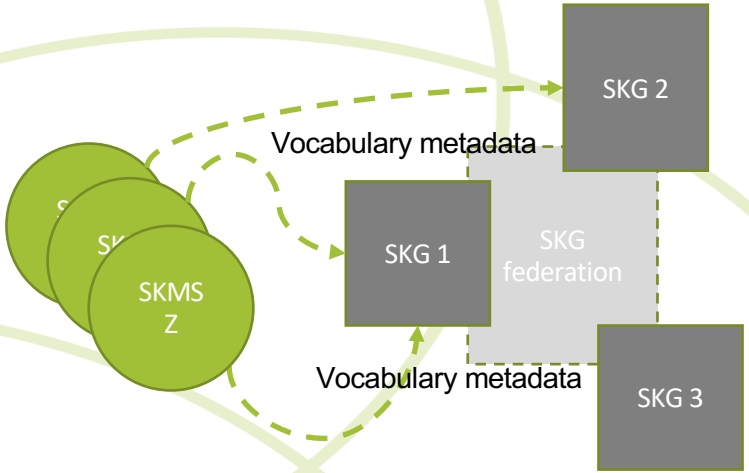
For this the organisational aspect of authorization for ‘create’ and ‘change’ operations should be addressed

Goals wrt. federating and integrating in EOSC

- Federating existing separate SKOSMOS vocabulary platform instances
  - Showing a consolidated catalogue
  - Giving recommendations on the basis of metadata + content
- Tested federated search on basis of using the SPARQL endpoints of the federated SKOSMOS instances
  - However we noted performance problems
  - SKOSMOS team advised against this
  - No solution for searching federated metadata
- Alternative is using consolidated index of vocabulary metadata and content
- Note that federating on the basis of a single implementation technology (eg. SKOSMOS) is considered not the preferred way but can be less complex

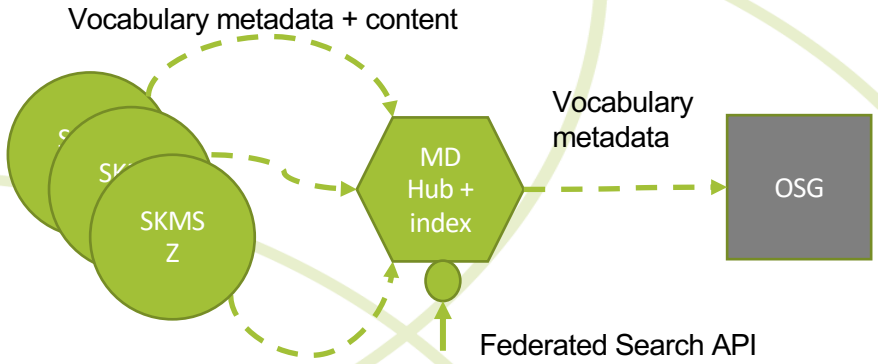
Potential integration SKG federation and vocabulary platforms

- EOSC projects activities
  - EOSC portal and research products catalogue -> EOSCfuture
  - EOSC federation and integration activities
- EC topic SKG interoperability (EOSC infra 01-03 call) and integration
  - OpenAIRE Research Graph (ORG), PID Graph, thematic graphs, NI graphs, …
- Different architectures possible, how to obtain additionally useful functionality also for SSH users
- Problem is that vocabulary platforms and content, however rich the metadata, are not currently seen as a full SKG peer, but only noted as support making the various SKGs interoperable



Look for opportunity to create vocabulary federated search scenario

- The SKOS vocabularies esp. with all the relations to other concepts also provide a graph
- However, note that vocabulary platforms and content, however rich the metadata, are not seen as a full SKG peer, but only as support for making the various SKGs interoperable
- Is there a way to present a Vocabulary Commons Hub as a SKG???



Added value for EOSC SKGs

In general, its not difficult to see a need for having vocabulary metadata in the SKGs

- Vocabularies are first class citizens and their use and need to be documented in machine actionable manner

But what about the vocabulary terms themselves

- Establishing similarity between research data
- Establishing similarity relations between vocabulary content can prove interesting to establish also similarities between methodologies (tagging texts with different but similar vocabularies)

Some resources mentioned

- CLARIN & SSHOC Vocabulary Initiative
- SSH Vocabulary Commons, vocabulary commons charter
- Bartoc.org , Bartoc federated search
- Open metadata registry
- Bioportal
- Recommendations for FAIR semantic artefacts

Thank you for your attention