A COMMON SSH VOCABULARY INFRASTRUCTURE?

Daan Broeder, CLARIN ERIC

2022 CLARIN centre days

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

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
- Identified more opportunities during the project
  - Training & discussion: CLARIN vocabulary initiative 2020 https://www.clarin.eu/blog/clarin-sshoc-vocabulary-initiative
  - Recommendations for further common approaches e.g. Vocabulary versioning, investigating CV authoring tools
  - Vocabulary recommendations for researchers
  - Opportunity & need to represent SSH interests with other stakeholders e.g. software & service providers

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SKOSMOS

- SKOSMOS was selected as the recommended software for vocabulary publication
- Already used by some DARIAH, CESSDA centers and beyond the SSH, good relations with solid development team based at NFL, fulfills most technical requirements: API, supported formats, …
  - But note SKOSMOS is not an authoring tool,
  - its targeted audience and performance do not cover all our use cases

See:

https://www.sshopencloud.eu/news/takeaways-sshoc-webinar-series-open-source-vocabulary-hosting-and-management-platforms

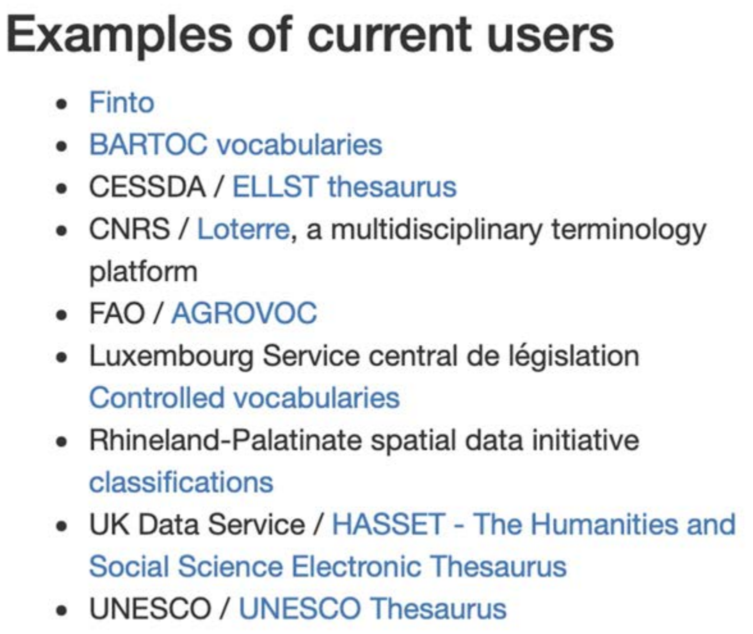
https://sshopencloud.eu/ms8-choice-vocabulary-publication-platform-sshoc

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SKOSMOS

- Out of the box SKOSMOS is
  - Jena/fuseki triple database
  - PHP / JS based frontend
  - Supports different concept schemes in one vocabulary definition file and skos collections/groups
  - Frontend can be customized
- Can be used for large vocabularies:

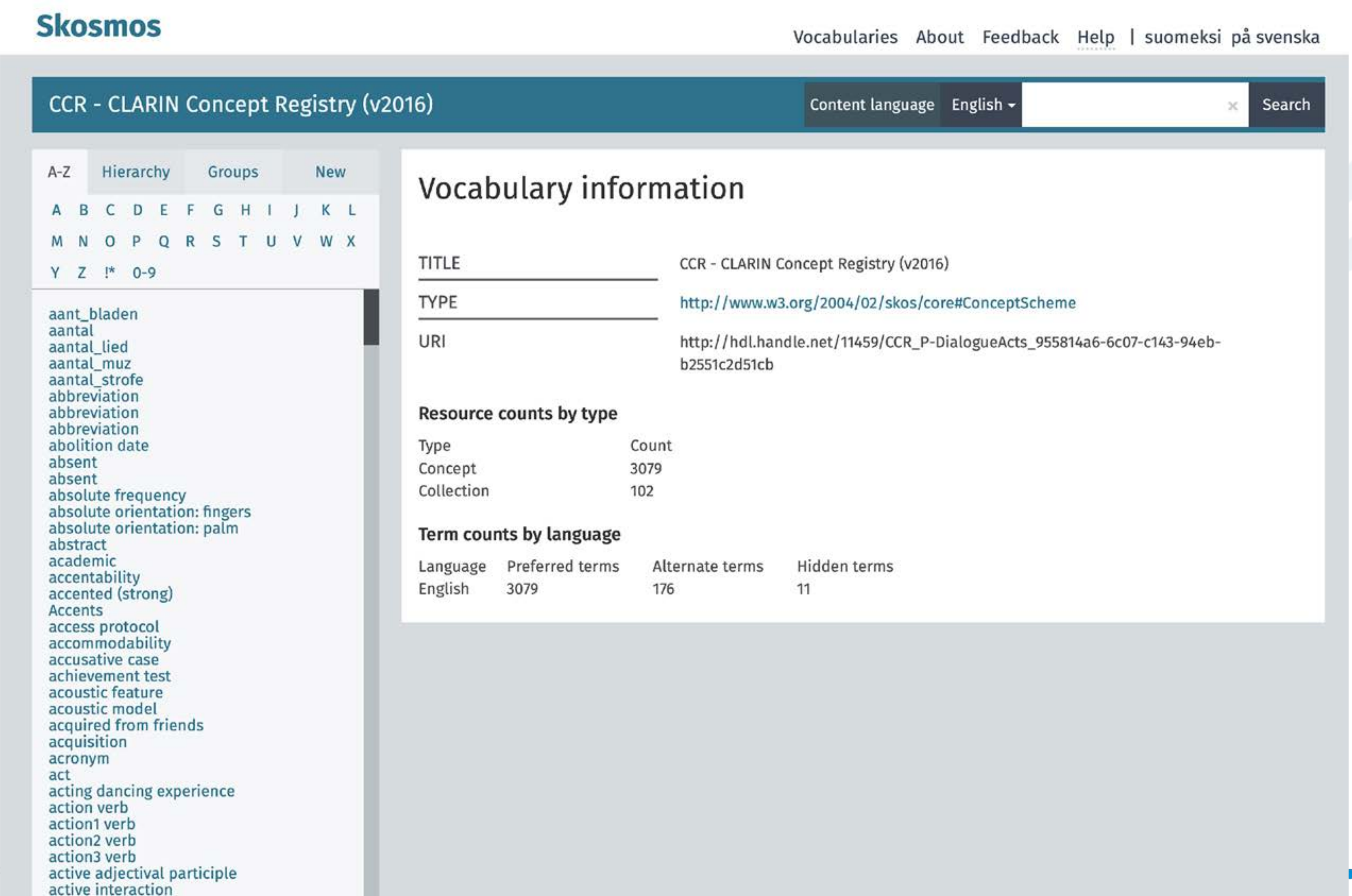
Agrovoc: 40k concepts,750k triples (https://agrovoc.fao.org/browse/agrovoc/en/), 17m load M1max cpu, but need optimized triple store for responsive UI and querying



Examples of current users

- Finto
- BARTOC vocabularies
- CESSDA / ELLST thesaurus
- CNRS / Loterre, a multidisciplinary terminology platform
- FAO / AGROVOC
- Luxembourg Service central de législation Controlled vocabularies
- Rhineland-Palatinate spatial data initiative classifications
- UK Data Service / HASSET - The Humanities and Social Science Electronic Thesaurus
- UNESCO / UNESCO Thesaurus

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782



Skosmos

Vocabularies About Feedback Help | suomeksi på svenska

CCR - CLARIN Concept Registry (v2016)

Content language English Search

A-Z Hierarchy Groups New

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z !* 0-9

aant_bladen
aantal
aantal_lied
aantal_muz
aantal_strofe
abbreviation
abbreviation
abbreviation
abolition date
absent
absent
absolute frequency
absolute orientation: fingers
absolute orientation: palm
abstract
academic
accentability
accented (strong)
Accents
access protocol
accommodability
accusative case
achievement test
acoustic feature
acoustic model
acquired from friends
acquisition
acronym
act
acting dancing experience
action verb
action1 verb
action2 verb
action3 verb
active adjectival participle
active interaction

Vocabulary information

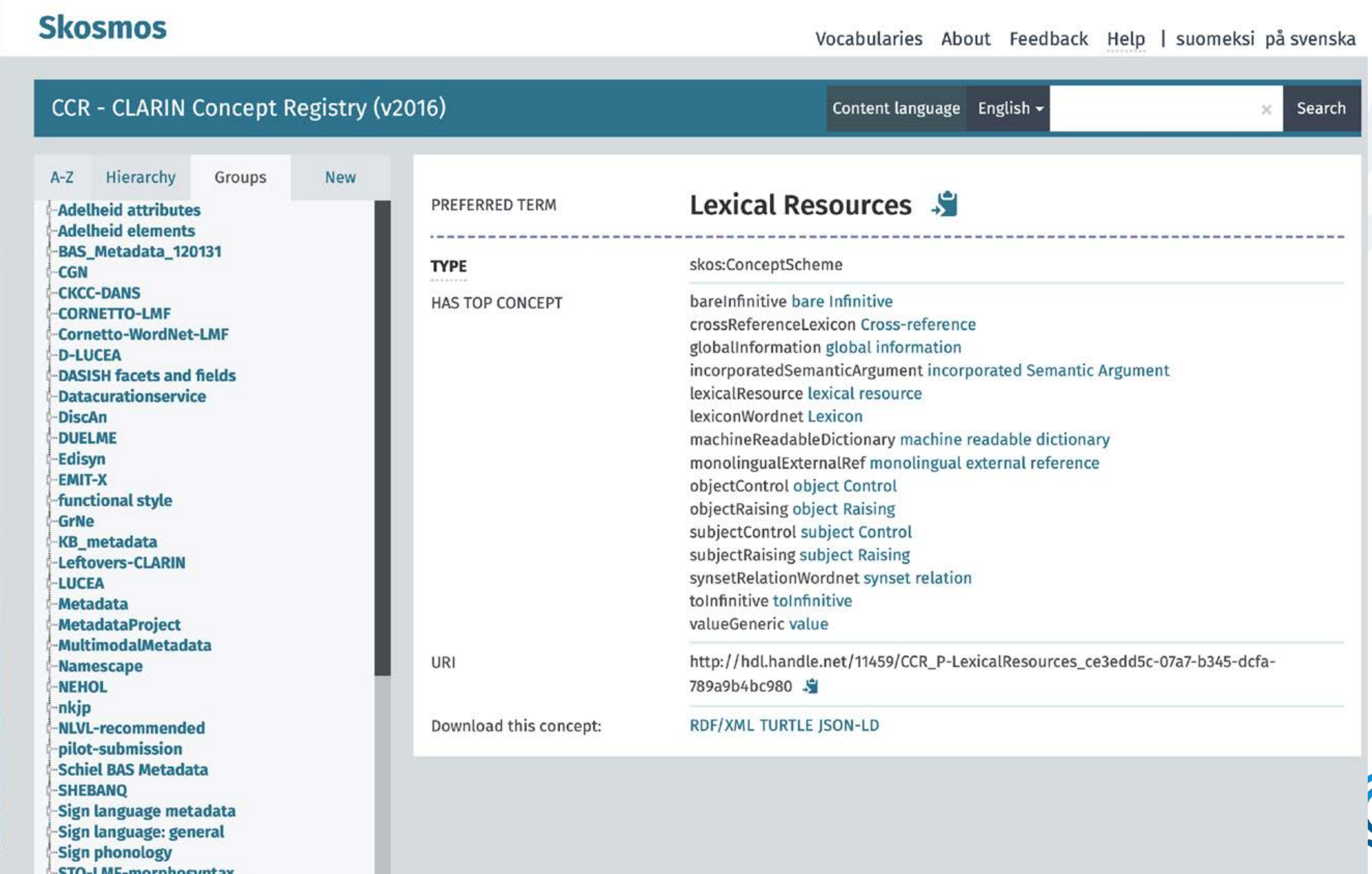
TITLE	CCR - CLARIN Concept Registry (v2016)
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
URI	http://hdl.handle.net/11459/CCR_P-DialogueActs_955814a6-6c07-c143-94eb-b2551c2d51cb

Resource counts by type

Type	Count
Concept	3079
Collection	102

Term counts by language

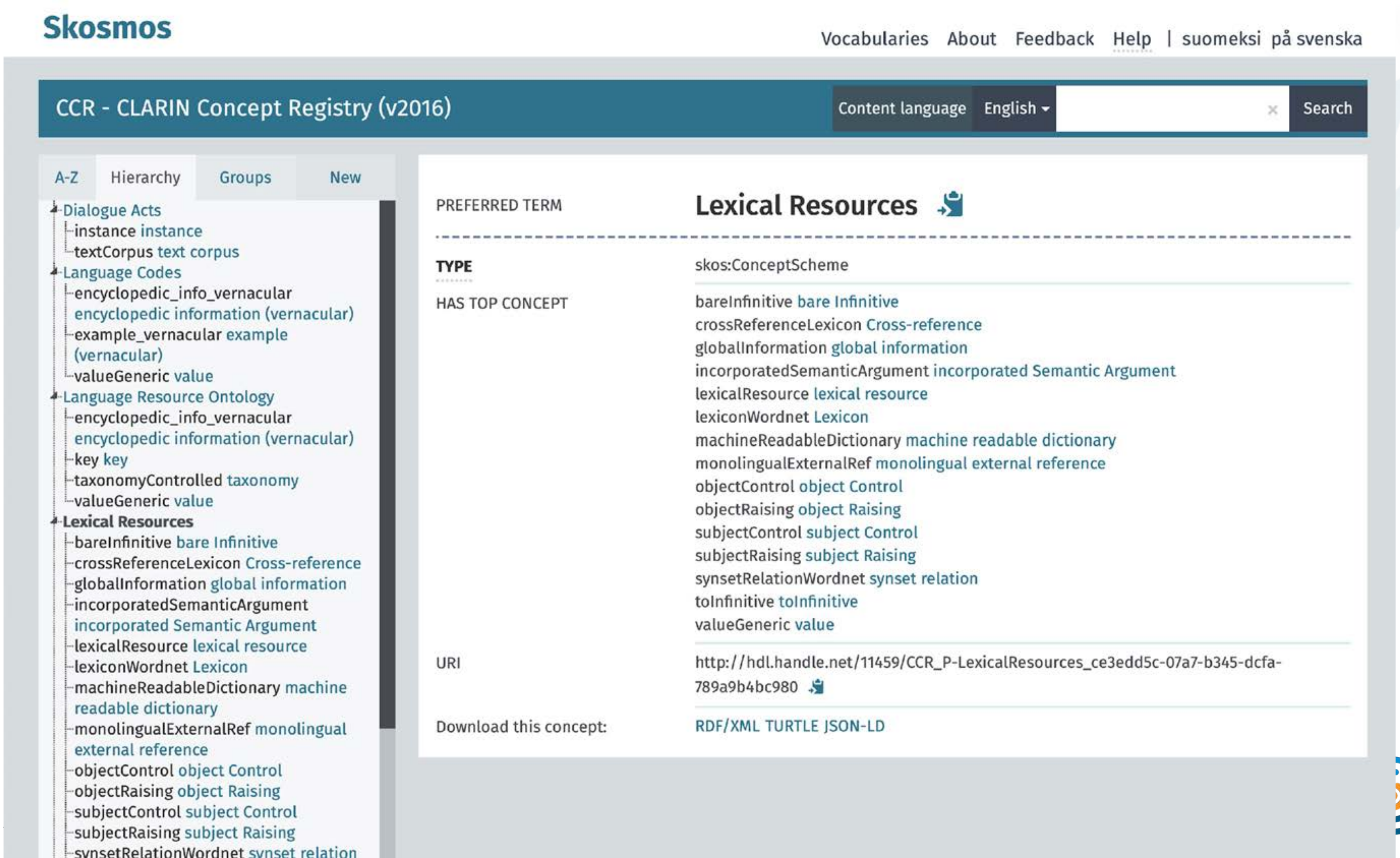
Language	Preferred terms	Alternate terms	Hidden terms
English	3079	176	11



CCR - CLARIN Concept Registry (v2016)

- Adelheid attributes
- Adelheid elements
- BAS_Metadata_120131
- CGN
- CKCC-DANS
- CORNETTO-LMF
- Cornetto-WordNet-LMF
- D-LUCEA
- DASISH facets and fields
- Datacurationservice
- DiscAn
- DUELME
- Edisyn
- EMIT-X
- functional style
- GrNe
- KB_metadata
- Leftovers-CLARIN
- LUCEA
- Metadata
- MetadataProject
- MultimodalMetadata
- Namescape
- NEHOL
- nkjp
- NLVL-recommended
- pilot-submission
- Schiel BAS Metadata
- SHEBANQ
- Sign language metadata
- Sign language: general
- Sign phonology
- STO-LMF-morphosyntax

PREFERRED TERM	Lexical Resources
TYPE	skos:ConceptScheme
HAS TOP CONCEPT	bareInfinitive bare Infinitive crossReferenceLexicon Cross-reference globalInformation global information incorporatedSemanticArgument incorporated Semantic Argument lexicalResource lexical resource lexiconWordnet Lexicon machineReadableDictionary machine readable dictionary monolingualExternalRef monolingual external reference objectControl object Control objectRaising object Raising subjectControl subject Control subjectRaising subject Raising synsetRelationWordnet synset relation toInfinitive toInfinitive valueGeneric value
URI	http://hdl.handle.net/11459/CCR_P-LexicalResources_ce3edd5c-07a7-b345-dcfa-789a9b4bc980
Download this concept:	RDF/XML TURTLE JSON-LD



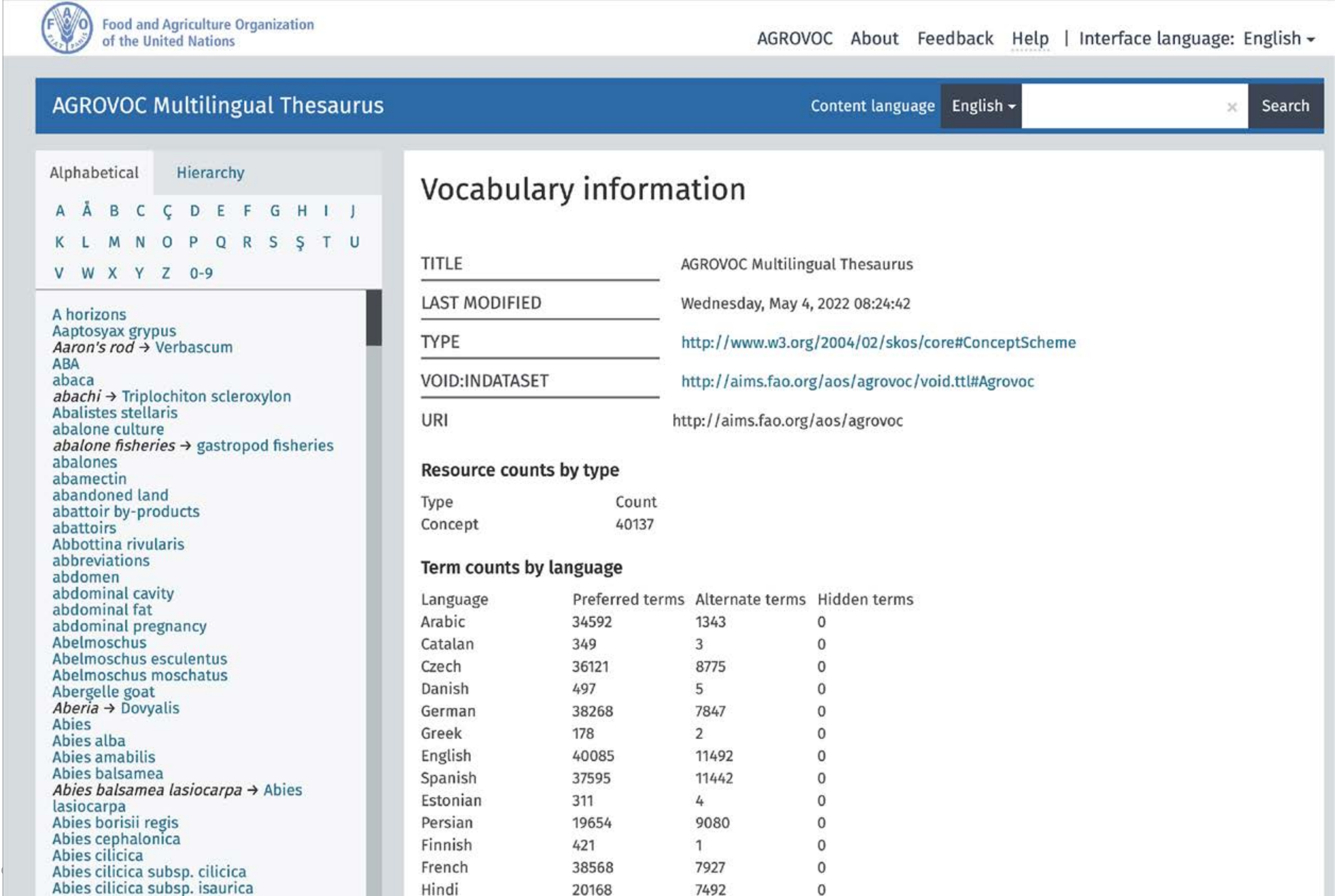
Skosmos

Vocabularies About Feedback Help | suomeksi på svenska

CCR - CLARIN Concept Registry (v2016)

- Dialogue Acts
  - instance instance
  - textCorpus text corpus
- Language Codes
  - encyclopedic_info_vernacular encyclopedic information (vernacular)
  - example_vernacular example (vernacular)
  - valueGeneric value
- Language Resource Ontology
  - encyclopedic_info_vernacular encyclopedic information (vernacular)
  - key key
  - taxonomyControlled taxonomy
  - valueGeneric value
- **Lexical Resources**
  - bareInfinitive bare Infinitive
  - crossReferenceLexicon Cross-reference
  - globalInformation global information
  - incorporatedSemanticArgument incorporated Semantic Argument
  - lexicalResource lexical resource
  - lexiconWordnet Lexicon
  - machineReadableDictionary machine readable dictionary
  - monolingualExternalRef monolingual external reference
  - objectControl object Control
  - objectRaising object Raising
  - subjectControl subject Control
  - subjectRaising subject Raising
  - synsetRelationWordnet synset relation

PREFERRED TERM	**Lexical Resources**
TYPE	skos:ConceptScheme
HAS TOP CONCEPT	bareInfinitive bare Infinitive crossReferenceLexicon Cross-reference globalInformation global information incorporatedSemanticArgument incorporated Semantic Argument lexicalResource lexical resource lexiconWordnet Lexicon machineReadableDictionary machine readable dictionary monolingualExternalRef monolingual external reference objectControl object Control objectRaising object Raising subjectControl subject Control subjectRaising subject Raising synsetRelationWordnet synset relation toInfinitive toInfinitive valueGeneric value
URI	http://hdl.handle.net/11459/CCR_P-LexicalResources_ce3edd5c-07a7-b345-dcfa-789a9b4bc980
Download this concept:	RDF/XML TURTLE JSON-LD



A horizons
Aaptosyax grypus
Aaron's rod → Verbascum
ABA
abaca
abachi → Triplochiton scleroxylon
Abalistes stellaris
abalone culture
abalone fisheries → gastropod fisheries
abalones
abamectin
abandoned land
abattoir by-products
abattoirs
Abbottina rivularis
abbreviations
abdomen
abdominal cavity
abdominal fat
abdominal pregnancy
Abelmoschus
Abelmoschus esculentus
Abelmoschus moschatus
Abergelle goat
Aberia → Dovyalis
Abies
Abies alba
Abies amabilis
Abies balsamea
Abies balsamea lasiocarpa → Abies lasiocarpa
Abies borisii regis
Abies cephalonica
Abies cilicica
Abies cilicica subsp. cilicica
Abies cilicica subsp. isaurica

Vocabulary information

TITLE	AGROVOC Multilingual Thesaurus
LAST MODIFIED	Wednesday, May 4, 2022 08:24:42
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
VOID:INDATASET	http://aims.fao.org/aos/agrovoc/void.ttl#Agrovoc
URI	http://aims.fao.org/aos/agrovoc

Resource counts by type

Type	Count
Concept	40137

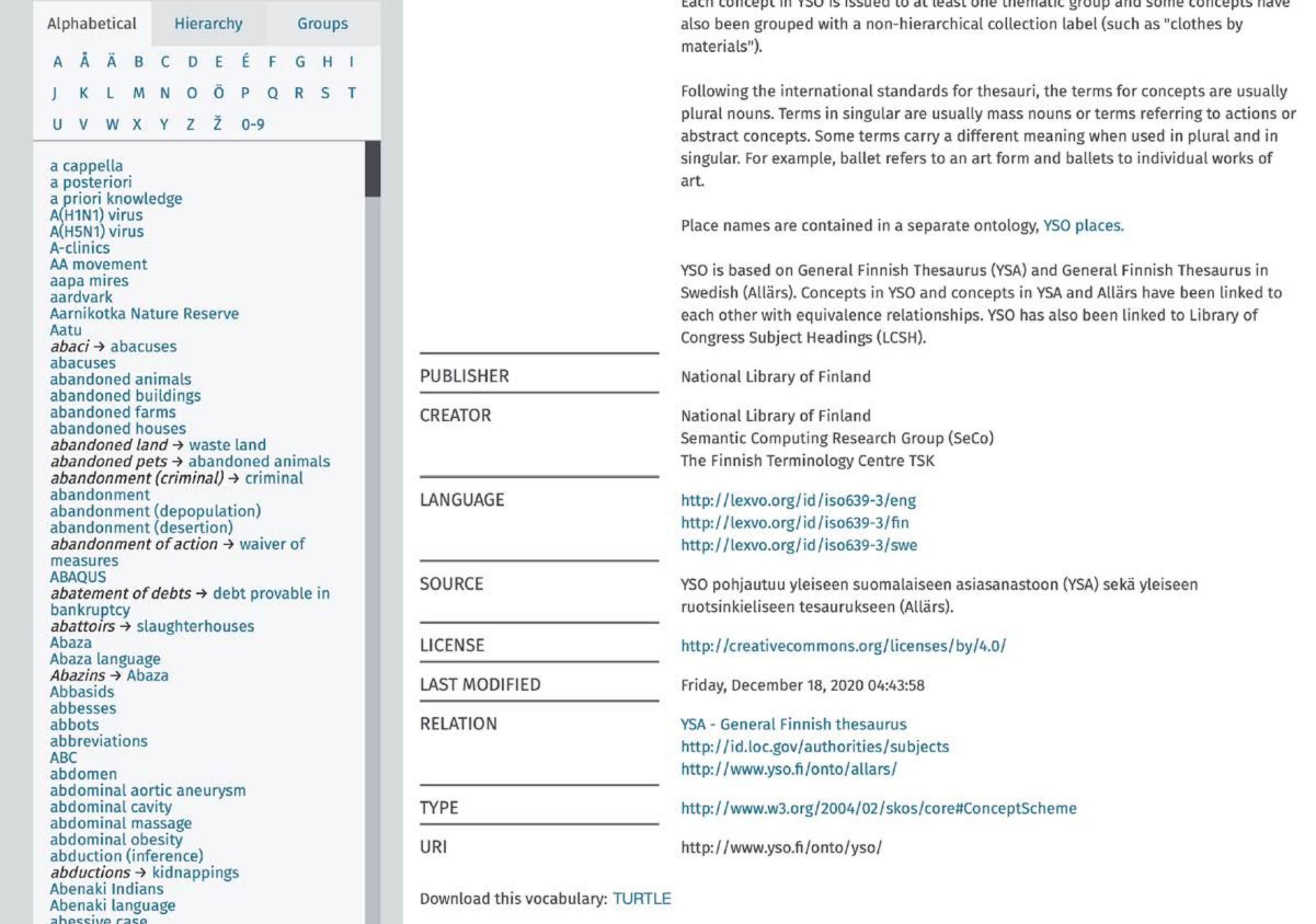
Term counts by language

Language	Preferred terms	Alternate terms	Hidden terms
Arabic	34592	1343	0
Catalan	349	3	0
Czech	36121	8775	0
Danish	497	5	0
German	38268	7847	0
Greek	178	2	0
English	40085	11492	0
Spanish	37595	11442	0
Estonian	311	4	0
Persian	19654	9080	0
Finnish	421	1	0
French	38568	7927	0
Hindi	20168	7492	0

But wrt. vocabulary discovery & researcher recommendations

- SKOSMOS does not provide useful registry functionality
- Only a scrollable list of vocabularies displaying vocabulary metadata
- The UI offers no metadata search, no facets, no keyword search, although for the content (terms) extensive search & filter options are offered

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782



a cappella
a posteriori
a priori knowledge
A(H1N1) virus
A(H5N1) virus
A-clinics
AA movement
aapa mires
aardvark
Aarnikotka Nature Reserve
Aatu
abaci → abacuses
abacuses
abandoned animals
abandoned buildings
abandoned farms
abandoned houses
abandoned land → waste land
abandoned pets → abandoned animals
abandonment (criminal) → criminal abandonment
abandonment (depopulation)
abandonment (desertion)
abandonment of action → waiver of measures
ABAQUS
abatement of debts → debt provable in bankruptcy
abattoirs → slaughterhouses
Abaza
Abaza language
Abazins → Abaza
Abbasids
abbesses
abbots
abbreviations
ABC
abdomen
abdominal aortic aneurysm
abdominal cavity
abdominal massage
abdominal obesity
abduction (inference)
abductions → kidnappings
Abenaki Indians
Abenaki language
abessive case

Each concept in YSO is issued to at least one thematic group and some concepts have also been grouped with a non-hierarchical collection label (such as "clothes by materials").

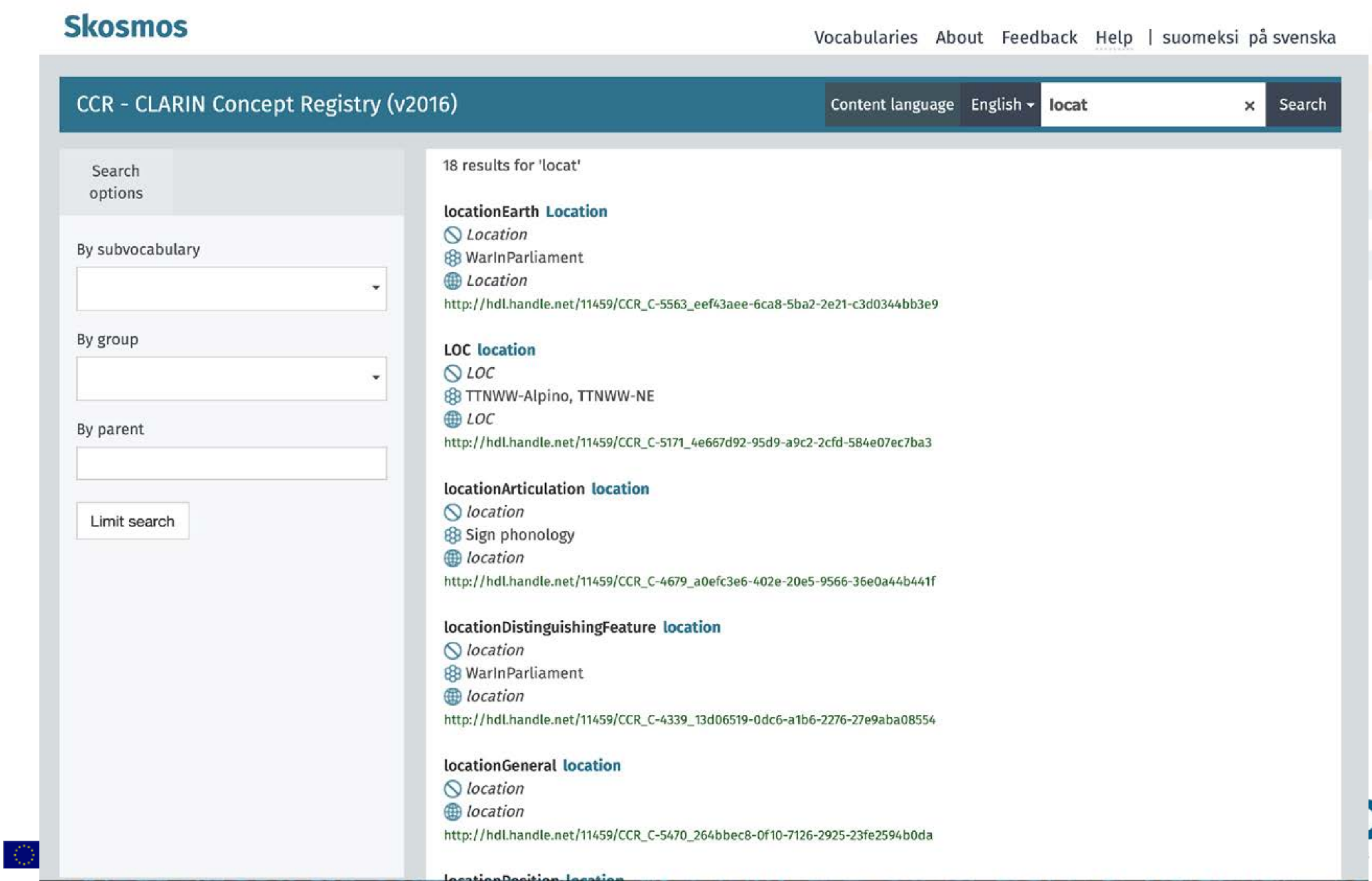
Following the international standards for thesauri, the terms for concepts are usually plural nouns. Terms in singular are usually mass nouns or terms referring to actions or abstract concepts. Some terms carry a different meaning when used in plural and in singular. For example, ballet refers to an art form and ballets to individual works of art.

Place names are contained in a separate ontology, YSO places.

YSO is based on General Finnish Thesaurus (YSA) and General Finnish Thesaurus in Swedish (Allärs). Concepts in YSO and concepts in YSA and Allärs have been linked to each other with equivalence relationships. YSO has also been linked to Library of Congress Subject Headings (LCSH).

PUBLISHER	National Library of Finland
CREATOR	National Library of Finland Semantic Computing Research Group (SeCo) The Finnish Terminology Centre TSK
LANGUAGE	http://lexvo.org/id/iso639-3/eng http://lexvo.org/id/iso639-3/fin http://lexvo.org/id/iso639-3/swe
SOURCE	YSO pohjautuu yleiseen suomalaiseen asiasanastoon (YSA) sekä yleiseen ruotsinkieliseen tesaurukseen (Allärs).
LICENSE	http://creativecommons.org/licenses/by/4.0/
LAST MODIFIED	Friday, December 18, 2020 04:43:58
RELATION	YSA - General Finnish thesaurus http://id.loc.gov/authorities/subjects http://www.yso.fi/onto/allars/
TYPE	http://www.w3.org/2004/02/skos/core#ConceptScheme
URI	http://www.yso.fi/onto/yso/

Download this vocabulary: TURTLE



CCR - CLARIN Concept Registry (v2016)

locat

18 results for 'locat'

locationEarth Location
⊘ *Location*
⊞ WarInParliament
🌐 *Location*
http://hdl.handle.net/11459/CCR_C-5563_eef43aee-6ca8-5ba2-2e21-c3d0344bb3e9

LOC location
⊘ *LOC*
⊞ TTNWW-Alpino, TTNWW-NE
🌐 *LOC*
http://hdl.handle.net/11459/CCR_C-5171_4e667d92-95d9-a9c2-2cfd-584e07ec7ba3

locationArticulation location
⊘ *location*
⊞ Sign phonology
🌐 *location*
http://hdl.handle.net/11459/CCR_C-4679_a0efc3e6-402e-20e5-9566-36e0a44b441f

locationDistinguishingFeature location
⊘ *location*
⊞ WarInParliament
🌐 *location*
http://hdl.handle.net/11459/CCR_C-4339_13d06519-0dc6-a1b6-2276-27e9aba08554

locationGeneral location
⊘ *location*
🌐 *location*
http://hdl.handle.net/11459/CCR_C-5470_264bbec8-0f10-7126-2925-23fe2594b0da

locationPosition location

Vocabulary visibility and discovery

- Vocabularies not always FAIR yet; they should be properly registered and published, researchers & infrastructure providers should be able to find and reuse -> **FAIR semantic artefacts**
- SSH Vocabulary registry or a general one that supports sufficient discipline specificity e.g. **Bartoc (3300 entries whereof 1200 SSH)**
- Vocabulary search facility, 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*

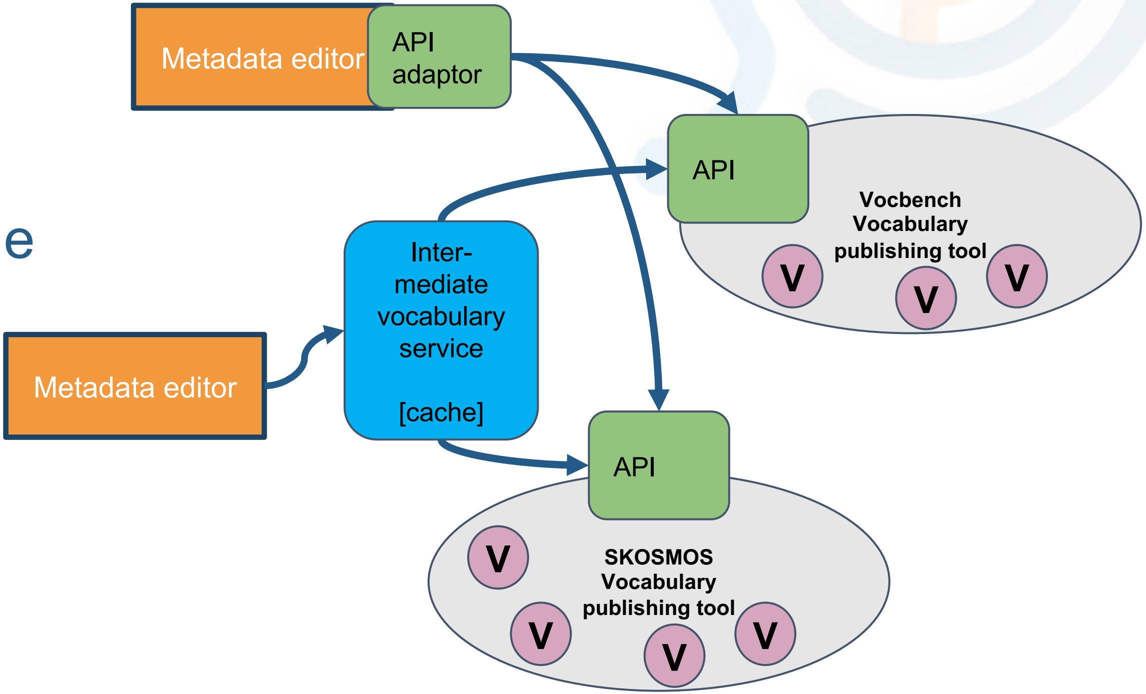
Bartoc registry: https://bartoc.org/vocabularies

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

FAIR Vocabularies - tool access to vocabularies via APIs

Two strategies

- Every tool knows multiple APIs
  - Scalable approach?
- Intermediate broker allows also caching vocabularies?
  - Added maintenance
  - Caching seems only useful in case of unreliable providers



Metadata editor
API adaptor
API
Vocbench Vocabulary publishing tool
V
V
V
Inter-mediate vocabulary service [cache]
Metadata editor
API
SKOSMOS Vocabulary publishing tool
V
V
V
V

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Towards a SSH Vocabulary Commons

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

Collaborative use and management of vocabularies
vocabularies as first-class citizens / FAIR data objects in their own right.

Mission: vocabulary commons charter

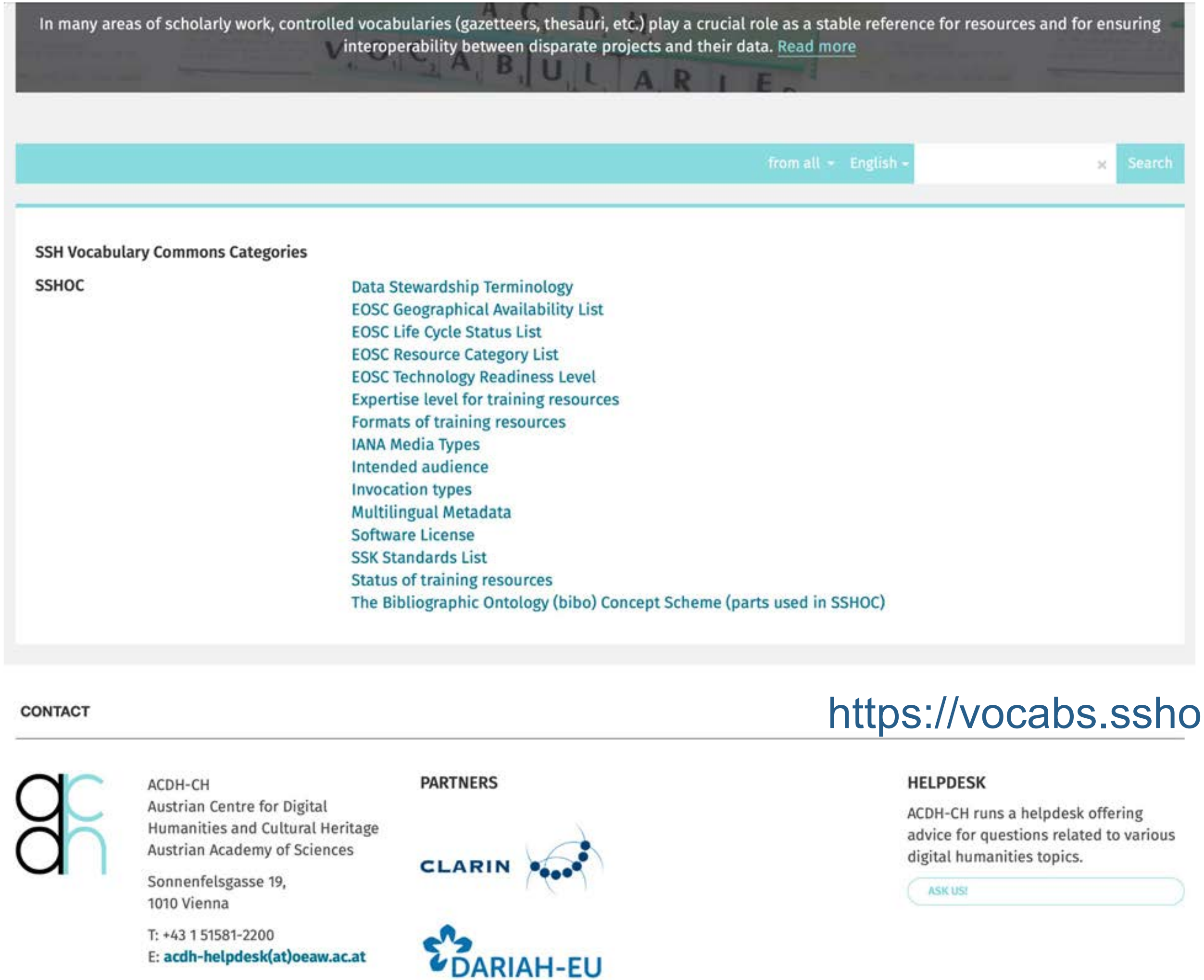
Include other projects initiatives EU eg. TRIPLE and national

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Priorities for the Vocabulary Commons

- Operating a Vocabulary repository
- SSH vocabulary overview
  - of all relevant SSH vocabularies
- Vocabulary federated content search
  - “Deep” search **also** in the terms (essential for recommendations)
- Harmonized versioning of vocabularies
  - yet no agreed way of how to manage update of vocabularies and individual vocabulary terms
- API standardisation:
  - to enable tools to interact smoothly with different vocabulary platforms
- Foster exchange between users and developers (eg SKOSMOS)

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782



In many areas of scholarly work, controlled vocabularies (gazetteers, thesauri, etc.) play a crucial role as a stable reference for resources and for ensuring interoperability between disparate projects and their data. Read more

SSH Vocabulary Commons Categories

SSHOC

- Data Stewardship Terminology
- EOSC Geographical Availability List
- EOSC Life Cycle Status List
- EOSC Resource Category List
- EOSC Technology Readiness Level
- Expertise level for training resources
- Formats of training resources
- IANA Media Types
- Intended audience
- Invocation types
- Multilingual Metadata
- Software License
- SSK Standards List
- Status of training resources
- The Bibliographic Ontology (bibo) Concept Scheme (parts used in SSHOC)

CONTACT

ACDH-CH
Austrian Centre for Digital
Humanities and Cultural Heritage
Austrian Academy of Sciences

Sonnenfelsgasse 19,
1010 Vienna

T: +43 1 51581-2200
E: acdh-helpdesk(at)oeaw.ac.at

PARTNERS

HELPDESK

ACDH-CH runs a helpdesk offering advice for questions related to various digital humanities topics.

SKOSMOS instance to publish the SSHOC vocabulary results and other SSH vocabularies

https://vocabs.sshopencloud.eu/vocabularies/

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

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**
- 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*

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Vocabulary Search

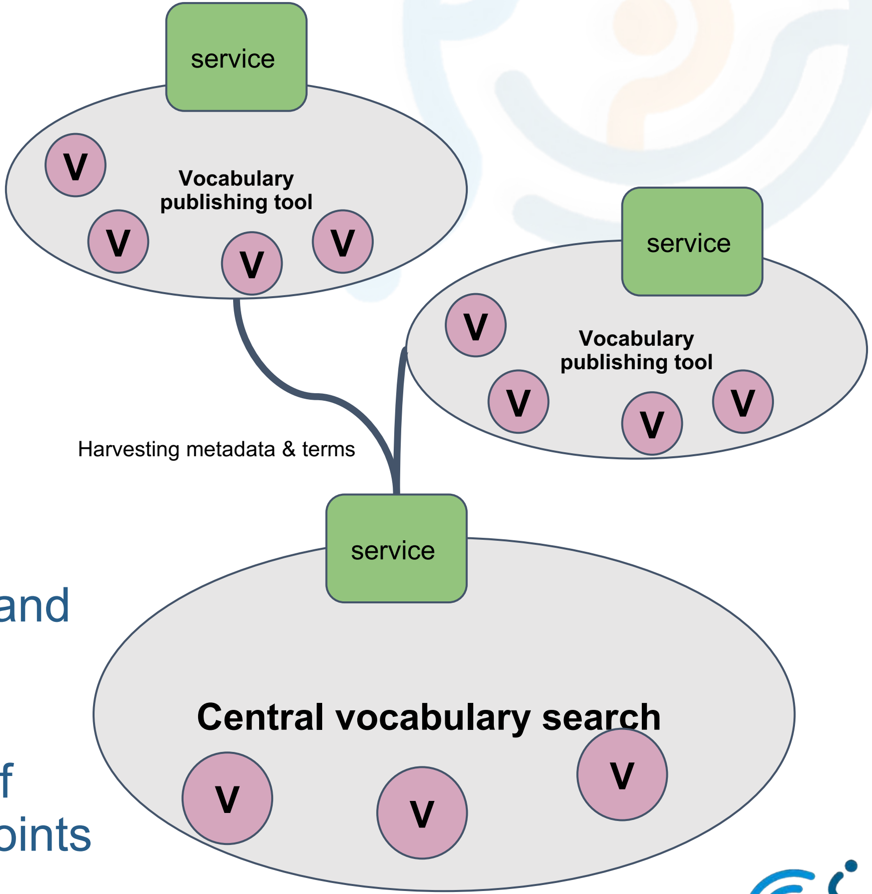
Effective vocabulary recommendations for researchers require:

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

Considering two strategies:

- Central metadata and content harvesting and indexing
  (update/syncing problem)
- Federated search, possibly on the basis of SKOSMOS instances and SPARQL endpoints

  (performance problem)


service
Vocabulary publishing tool
V
V
V
V
service
Vocabulary publishing tool
V
V
V
V
Harvesting metadata & terms
service
Central vocabulary search
V
V
V

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Collaborative Vocabulary Management

Vocabulary management is “solved” and reuse by means of copy too

BUT:

Reuse and management of a vocabulary across multiple organisational scopes not

- Ownership in a distributed (loosely coupled) setup => authority
- Procedures for evolution and agreement
  (new concepts, semantic shift)
- Synchronisation with source/target applications
  Ideally, user wants to adjust vocabulary (add new concepts) at point of use

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Towards a SSH Vocabulary Commons

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

Collaborative use and management of vocabularies
vocabularies as first-class citizens / FAIR data objects in their own right.

Mission: vocabulary commons charter

SSHOC activities wrt vocabularies:

- tasks 3.1 Multilingual terminologies, 3.5 Interoperability, WP7 SSH Open Marketplace usage of vocabularies etc.
- CLARIN Vocabulary Initiative inventorizing relevant SSH vocabularies and organizing info sessions on vocabulary management software and SSH requirements
- ICTeSSH workshop on vocabulary use in the SSH
- Wider scope for collaboration in the Humanities eg. TRIPLE project

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Collaborative Vocabulary Management

Vocabulary management is “solved” and reuse by means of copy too

BUT NOT

Reuse and management of a vocabulary across multiple organisational scopes

- Ownership in a distributed (loosely coupled) setup => authority
- Procedures for evolution and agreement
  (new concepts, semantic shift)
- Synchronisation with source/target applications
  Ideally, user wants to adjust vocabulary (add new concepts) at point of use

Note also that changes in vocabularies should be made explicit and documented beyond versioning eg.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Note: In the CMDI TF meeting tomorrow Menzo will present what happens with regard to vocabularies in CLARIAH-NL

Thank you for your attention

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Vocabularies & Interoperability

- **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
- **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, good tools exist (Ariadne VMT)
- **Organizational interop.** when sharing also responsibility for vocabulary maintenance, what model for (non-)agreement can we have
- **Cultural & Human interop** aspects. Multilingual vocabularies, localization aspects.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Vocabularies & Interoperability

For data reuse and data integration we have to look at interopreability 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
- **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).
- **Cultural & Human interoperability** aspects. Multilingual vocabularies, localization aspects. -> MT technology + network of human experts

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782