

CV4DH

Kontrollierte Vokabulare für Digitale Geisteswissenschaften

Matej Ďurčo & Karlheinz Mörth, ICLTT, Vienna
2013-11-22, "Kulturelles Erbe in der Cloud", Graz

- Kontrollierte Vokabulare (CV)
Begriffsklärung, Nutzungsszenarien und thematische Domänen
- Aktivitäten in CLARIN, DARIAH und breiterem Umfeld
- SKOS
- OpenSKOS - Vocabulary Repository
- CLAVAS – CLARIN Arbeitsgruppe für CV
- Weitere Schritte

- **Konzept vs. Begriff – concept vs. term**
semantisch vs. lexikal – Konzept wird referenziert durch verschiedene Begriffe, aber existiert unabhängig von diesen
- **Begriffliste – term list**
flache Liste
- **Liste von Konzepten - concept list**
auch flach, aber unterscheidet zwischen der semantischen und lexikalischen Ebene
- **Taxonomie**
beinhalten (meistens hierarchische) Beziehungen zwischen Konzepten
- **Schema/Ontologie**
beide definieren/beinhalten Konzepte/Entitäten
mit Attributen und typisierten Beziehungen
Schema – XML-, DB- Welt
Ontology – Wissensmanagement, Semantic Web

- Konzept vs. Begriff – concept vs. term
semantisch vs. lexikal – Konzept wird referenziert durch verschiedene Begriffe, aber existiert unabhängig von diesen
- Begriffliste – term list
flache Liste
- Liste von Konzepten - concept list
auch flach, aber unterscheidet zwischen der semantischen und lexikalischen Ebene
- Taxonomie
beinhalten (meistens hierarchische) Beziehungen zwischen Konzepten
- Schema/Ontologie
beide definieren/beinhalten Konzepte/Entitäten mit Attributen und typisierten Beziehungen
Schema – XML-, DB- Welt
Ontology – Wissensmanagement, Semantic Web

discussed here

Nutzungsszenarien für CV

- Metadaten Erstellung und Pflege
- Daten Anreicherung, Annotation
- Suche
query expansion, autocomplete, facets etc.
- Daten Analyse und Exploration
- notwendiges Mittel um Daten ins **Semantic Web** zu bringen erlaubt Literale zu Entitäten aufzulösen
- kann Äquivalenzen zwischen Konzepten/Entitäten aus verschiedenen Vokabularien liefern.
cf. Links in Wikipedia (Seite von [J. W. Goethe](#)):
GND: 118540238 | LCCN: n79003362 |
NDL: 00441109 | VIAF: 24602065

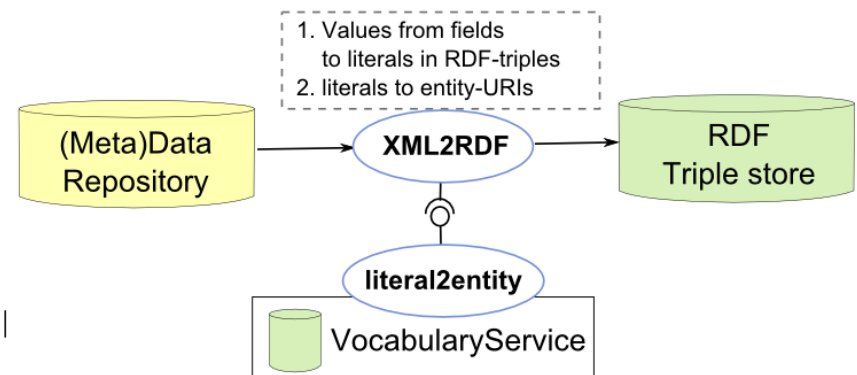
CQL

pos	=	▼	ART	✕ & OR
ana	=	▼	Art-3SG.MAS	
ana	=	▼	DF-Art.SG	
placeName	=	▼	par*	

- birth-place
- death-place
- occupation
- persId
- person
- personName
- placeName

- Paradiso |1
- Paraguay |'
- Paraguaybu
- Paris |243
- Pariser |27
- Pariseronke
- Park Monce

- ART |4782
- ADV |4073
- APPR |3101
- ADJA |3076
- ADJD |1348
- APPRART |489
- APZR |28
- APPO |12
- ADV ADJA |1



=> **Linked Data**

Domänen + existierende CVs/Services

- viele Domänen, grundlegende Unterscheidung Entitäten (Personen, Organisationen, Geographica) vs. Konzepte (Schlagwörter, Klassifizierungen, Typologien...)
- Normdateien der Bibliotheken (GND, LoC, VIAF, ...)
- allgemeine Datensammlungen (Wikipedia/dbpedia, Geonames, Getty, Yago) vs.
- spezialisierte Vokabularien (LT-World, Pleiades, ISOcat, ISO-639-*, ...)
- unterschiedliche Zugriffsmöglichkeiten: Web-Suche, Webservice, Data dumps (CSV, XML, RDF)
- [PDR](#) Persons Data Repository (@BBAW/Telota)
- Finnish Ontology Library Service [ONKI](#)
- ...

CLARIN – <http://clarin.eu>

Common Language Resources and Technologies Infrastructure

- die Nutzung von Sprachressourcen und -technologien für SSH Forschung international zu ermöglichen/vereinfachen
- stabile organisationelle und technische Strukturen
- harmonisiertes Metadaten-System (CMDI – <http://clarin.eu/cmdi>)

DARIAH – <http://dariah.eu>

Digital Research Infrastructure for Arts and Humanities

connected network of people, information, tools, and methodologies for investigating, exploring and supporting work across the broad spectrum of the digital humanities

getragen von nationalen Strukturen – so auch CLARIN/DARIAH-AT

ERIC – *European Research Infrastructure Consortium*

neue Form einer juristischen Person (legal entity) eingeführt von der Europäischen Kommission in 2009 um RIs eine stabilere Basis zu bieten

Task Force on Controlled Vocabularies

gestartet beim DARIAH VCC meeting in Wien 2012-11, vertieft in Copenhagen 2013-09

- Ziel: establish a comprehensive infrastructure for harmonized provision and collaborative maintenance of controlled vocabularies and reference data for the DARIAH (and Digital Humanities) community
- über 10 Institutionen beteiligt
verwandt mit den Aktivitäten zu **Linked Data** und zu Ontologie von Forschungsmethoden

zwei grundlegende Arbeitsbereiche:

1. Inventarisieren und Harmonisieren

CVs technisch auf gemeinsame Grundlage stellen

2. Vocabulary (**/ontology**) alignment

- Links zwischen Konzepten/Entitäten verschiedener CVs/Ontologien erstellen

- spezialisierte Arbeitsgruppe (lead: IEG Mainz) für historische Ortsnamen

Workshop in Mainz 2013-10

Hauptidee (am Beispiel der Typen von Orten): Verlinken basierend auf Eigenschaften/Funktionen von Konzepten/Entitäten

- Hauptaugenmerk auf Inhalt statt auf technische Entwicklung im Einklang mit der allgemeinen DARIAH Strategie
- aber gleichzeitig nicht das Rad (die Daten) neu erfinden
=> wiederverwenden existierender CVs
die allerdings oft zu allgemein sind und nie vollständig (VIAF, GND)
also brauchen wir eine Möglichkeit Konzepte hinzuzufügen, oder allgemeiner:
+ Verwalten/Pflegen von Vokabularien
- daher ein eigenes Vocabulary Repository
das kollaboratives Erstellen und Pflegen von CVs ermöglicht
- Versuchen das erzeugte Material wieder in die Normdateien zurückzuspeisen (Kommunikation mit Nationalbibliotheken)
prinzipiell möglich (GND Regelwerk), aber ein langwieriger Prozess
DARIAH könnte vermitteln, Druck ausüben, aber es darf nicht die Nutzung behindern
= unabhängige parallele Aktivität

Aktivitäten in CLARIN

- ISOcat – Data Category Registry
 - Registry zum Definieren von (linguistischen) Konzepten ("flach"=(fast) keine Beziehungen)
 - Implementierung des ISO standard ISO12620:2009
 - Grundstein für [CMDI](#) –semantische Verlinkung für MD Schemata

www.isocat.org
- Relation Registry
 - begleitend zu *ISOcat* Beziehungen zwischen Datenkategorien auszudrücken
 - erste Version verfügbar: lux13.mpi.nl/relcat/
- Taskforce für Metadatenpflege (metadata curation)
 - innerhalb des SCCTC (Standing Committee for CLARIN Technical Centres)
- Vorkehrungen für Integration von Vocabulary Services in CMD 1.2
- CLAVAS
 - Vocabulary Alignment Service for CLARIN
 - Initiative ursprünglich aus CLARIN-NL
 - Ziel: das Vocabulary Repository OpenSKOS für CLARIN Bedürfnisse zu adaptieren
- OpenSKOS and controlled vocabularies meeting
in Utrecht, 2013-05-17
www.clarin.eu/node/3780

Weitere Aktivitäten anderswo

zahlreiche Daten/Vokabularien (und einige Services) existieren bereits, hauptsächlich in der LIS Welt

- [VIAF](#) - Virtual International Authority File
 - betrieben von Nationalbibliotheken + OCLC
 - Ziel : Harmonisieren und Clustern der (nationalen) Normdateien
 - bietet Webservices, Suchinterface und Daten-Dumps
 - Normdateien für : Personen, Körperschaften, Geographika usw.
- [The European Library](#) (48 Nationalbibliotheken)
 - Datenanreicherung basierend auf Vokabularien
 - MACS – Multilingual Access to Subjects (semi-automatic alignment)
 - Abgleichen von DDC und UDC via CERIF
 - Abgleichen/Verlinken mit anderen Ontologien (Geonames, VIAF)
 - Suchservice: <http://www.theeuropeanlibrary.org/tel4/apisearch>
- [Library of Congress](#) - LCSH, MADS, ...
- [Getty Thesauri](#)
- [Geonames](#) – Suchinterface, Service, Dumps
- [LT-World @DFKI](#) – volle Ontologie, eher Kandidat für LOD-linking
- uvm ...

Anforderungen / Ansatz

- Konzepte/Entitäten identifiziert mit einem PID ([coolURI](#) reicht)
- Unterstützung für Multilingualität
- Pluralität der Konzeptualisierungen
mehrere (evtl. konfliktierende) Vokabularien für denselben Themenbereich erlaubt
- Verwaltung der CVs, Datenpflege als kollaborativer laufender Prozess
- Semantic Web kompatibel
- erstellte CVs gemeinsam nutzen (mit anderen teilen)
- existierende Daten und Services nutzen/wiederverwenden
- aber mit einem harmonisierten Zugriff (/technischer Lösung)
= „**one stop shop**“ für CVs
von Vorteil für Anbieter, Benutzer und Entwickler
- Hauptfunktionen: Explore/Search, Edit/Manage, Integrate as lookup

Vorteile (und Risiken) eines harmonisierten Systems

- für Betreiber:
 - vereinfacht die Bereitstellung/Veröffentlichung von CVs
 - vereinfacht die Wiederverwendung von (eigenen) CVs in Applikationen von Dritten
 - vereinfacht den Abgleich und die Verlinkung von Konzepten zwischen CVs
- für Benutzer
 - vereinfacht das Entdecken, Evaluieren und Nutzen von CVs
(was die Notwendig reduziert, eigene zu erstellen)
 - neue Möglichkeiten der Interaktion (Search & Browse)
 - online Vokabularien sind immer aktuell
- für Entwickler
 - keine Anpassung für individuelle CVs notwendig
 - erleichterte Wiederverwendung existierender Tools und Module
- Risiken
 - Babylon Szenario – zu unterschiedliche konzeptuelle Domänen treffen auf einander
 - Überlastung des System – zu viele Anfragen, zu großes Datenvolumen
 - Überforderung der Benutzer – zu viel Information (zu viele CVs verfügbar)

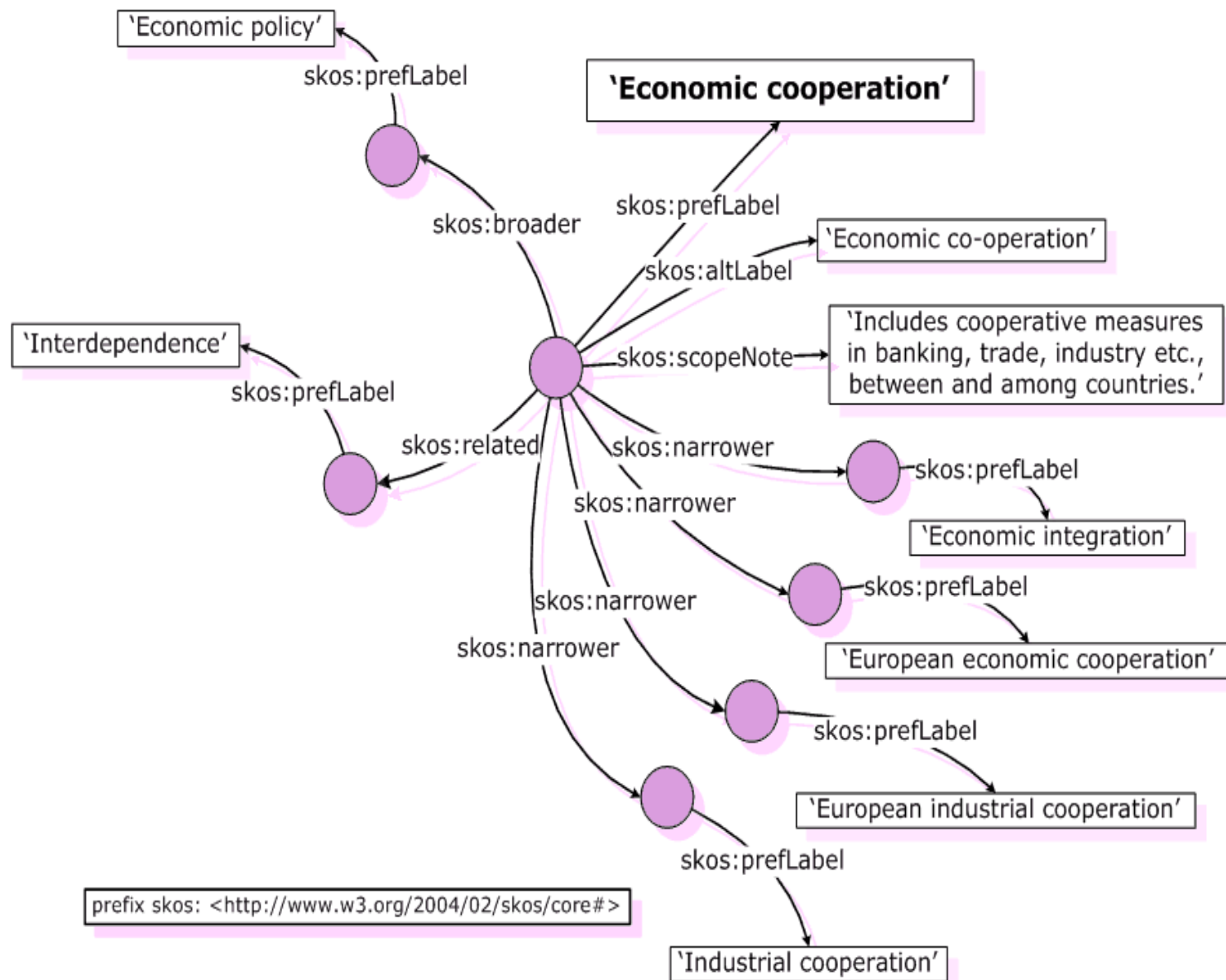
Simple Knowledge Organization System

<http://www.w3.org/TR/skos-reference/>

- SKOS knowledge structures consist of **Concepts** grouped in **ConceptSchemes**
- Concepts are identified by a **URI**
- Concepts have **labels** in 1 or more languages

`skos:prefLabel@lang, skos:altLabel@lang` => multilinguality

- Concepts can be documented with **'notes'**
- Concepts have mutual **semantic relations**
`broader, narrower, related` => taxonomy construction
- Concept in different ConceptSchemes can have **matching relations**
- Concepts can be part of multiple ConceptSchemes



Vocabulary repository & service
openskos.org

- Daten im SKOS Format
- verteilte Architektur
- RESTful API
- Linked Data
- Dateneinspeisung (ingest) mittels Hochladen oder über OAI-PMH beziehen ("harvesting")
- Eigener integrierter administrativer Bereich
- Unterstützt das Verlinken von Konzepten
- Propagieren von offenen Datenbanklizenzen
- Neuerdings auch Datenpflege über den integrierten Editor



- entwickelt im Rahmen des NL Projekts [CATCHplus](#)
- kommerzielle SW-Firma (Picturae), aber open source
- zur Zeit 3 Instanzen laufend: Meertens Institute, NISL, ICLTT (Testphase) (Picturae betreibt weitere 7 Instanzen für ihre Kunden)

OpenSKOS Management Search, browse and edit Logged in as matej durco

Search, browse and edit Manage institution Manage collections Manage users Manage jobs Manage concept schemes

Academy*

Concept scheme

- open source
- OpenSKOS-CLARIN Organizations
- Organizations - need curation
- participant presence
- participant sex

Search result (6)

Add to selection Export Relate

(Op)	Academy of Sciences		
(Op)	Academy of Sciences at Goettingen, Germany		
(Op)	Academy of Sciences at Latvia		
(Op)	Academy of sciences of the Estonian SSR		
(Op)	Academy of Sciences of the USSR		
(Op)	Institute of Language and Literature, Academy of sciences of the Estonian SSR		

Institute of Language and Literature, Academy of sciences of the Estonian SSR

Status: approved
To be checked: No Switch to edit mode Export

EN OpenSKOS-CLARIN Organizations

Preferred label
Institute of Language and Literature, Academy of sciences of the Estonian SSR

Alternative label
Academy of sciences of the Estonian SSR, Institute of Language and Literature

Has broader (1)
(Op) Academy of sciences of the Estonian SSR

URI: <http://openskos.meertens.knaw.nl/Organisations/288ab4e6-5a14-4835-9db0-18cdcebf6aa6>

Notation: 106574

Is top concept: No

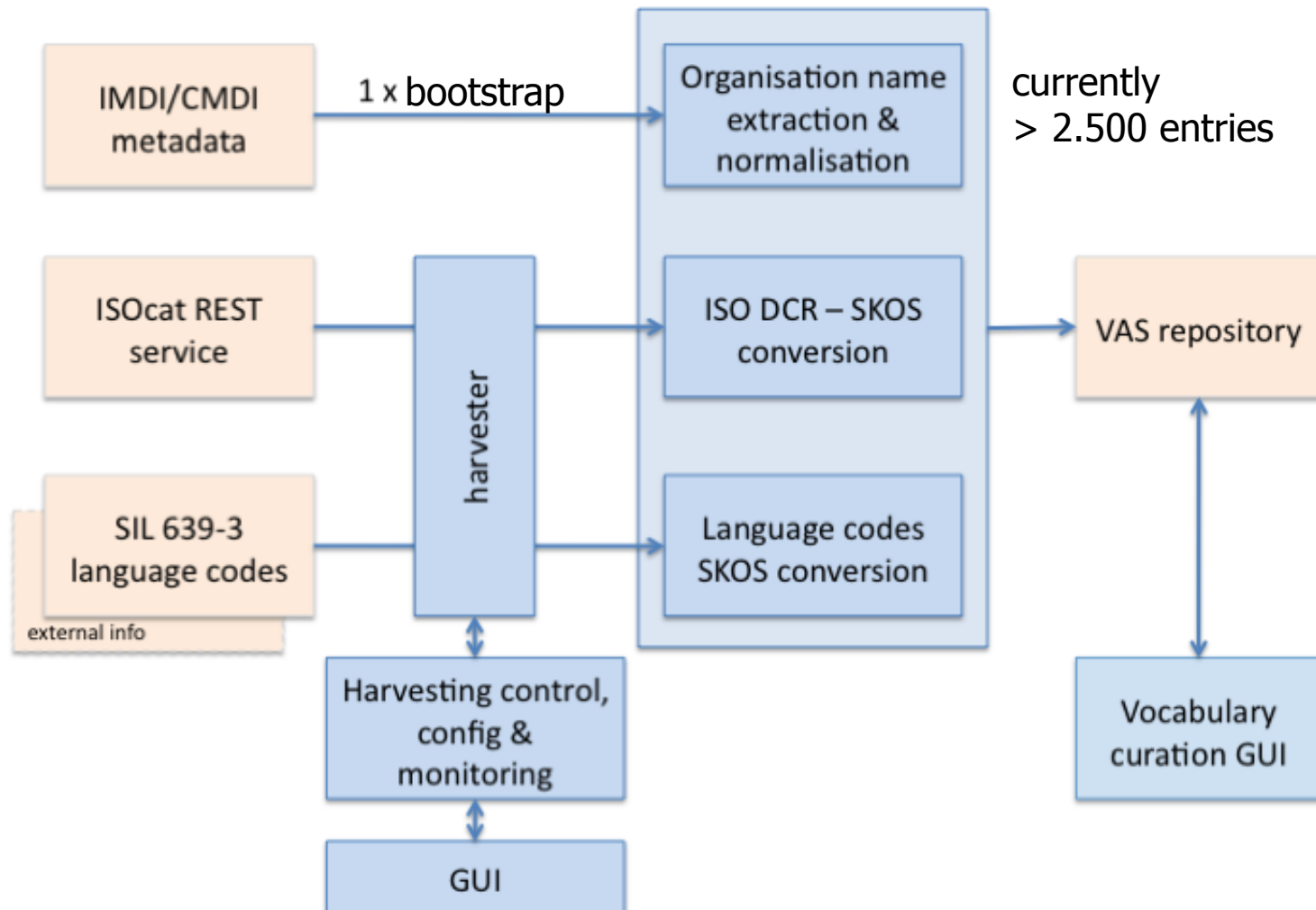
Created	Modified	Approved
26-06-2013 15:00:05	26-06-2013 15:00:32	N/A
N/A	N/A	N/A

History (22) Selection (0)

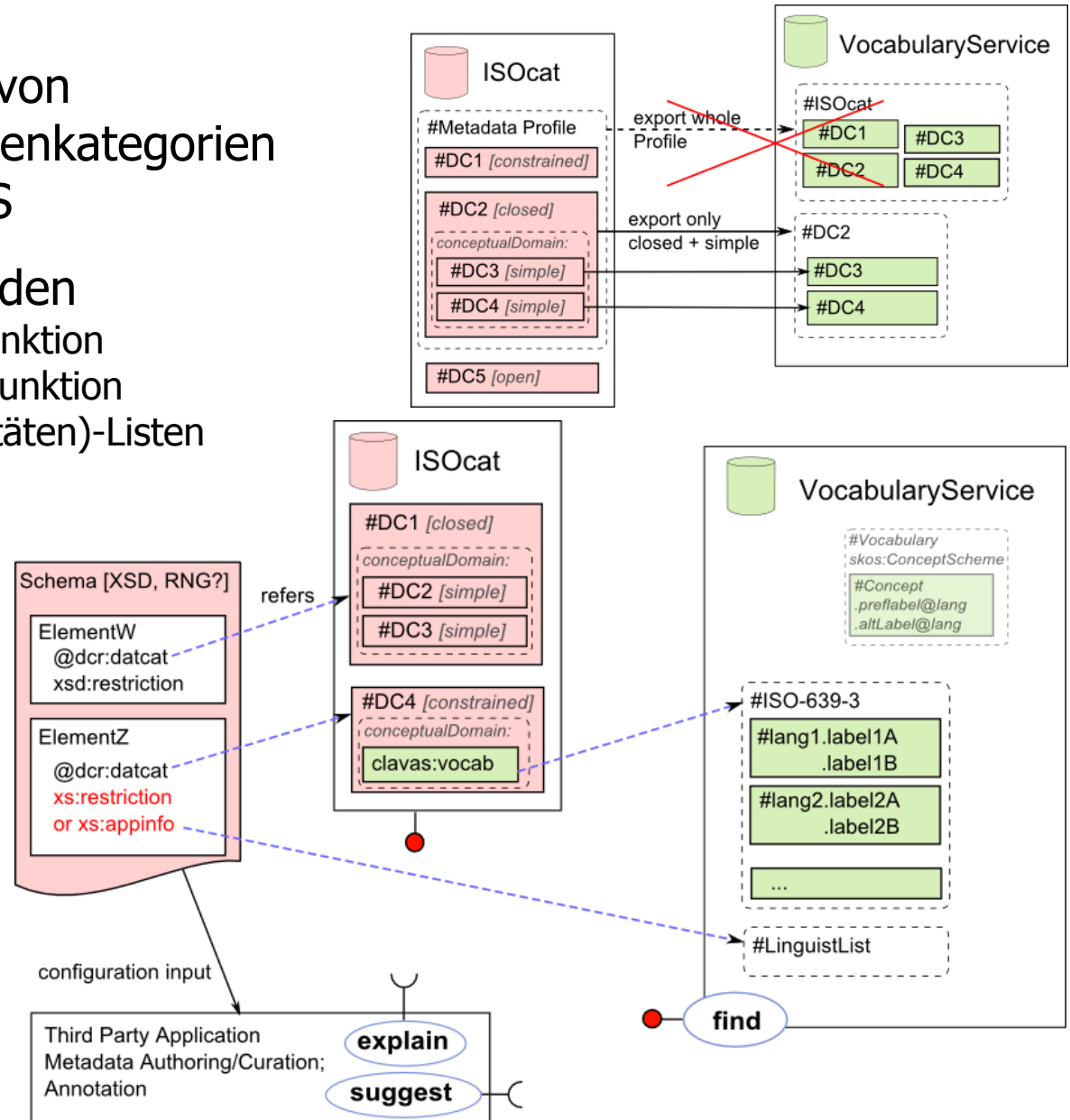
Clear history Export

(Op)	Institute of Language and Literature, Academy of sciences of the Estonian SSR	
(Op)	Academy of Sciences	
(Op)	Academy of Sciences at Goettingen, Germany	
(Op)	Academia	
(Op)	A	
(Op)	Akademie der Wissenschaften	
(Op)	Akademie der Wissenschaften	
(Op)	Akademie	
(IS)	German	
(ac)(an)(t)	unspecified	
(ac)(an)(e)	unknown	
(ne)	CAT	
(pa)	group	
(Op)	BBAW	
(Op)	Bulgarische Akademie der Wissenschaften	
(Op)	Berlin-Brandenburg Academy of Sciences and Humanities	

Adaptierung von OpenSKOS für CLARIN Bedürfnisse
= eigene Instanz + eigene Vokabularien



- automatischer Export von von ausgewählten Datenkategorien aus ISOcat ins CLAVAS
- Anwendungen verwenden
 - *ISOcat* für explain() Funktion
 - *CLAVAS* für suggest() Funktion = Werte (/Entitäten)-Listen (autocomplete)



Weitere Schritte

- CLARIN/CLAVAS:
 - weitere Bearbeitung/Pflege des Vokabulars für Organisationen
 - integrieren des Vocabulary Services mit anderen Modulen der Infrastruktur (zB MD editor)
 - weitere CVs aufnehmen
 - Ausarbeiten wie existierende CVs und Services integriert werden können => **proxy?**
- DARIAH
 - Inventarisierung: Kandidat-CVs mit Gruppen, die sie brauchen/nutzen
 - Testen der Nutzbarkeit von OpenSKOS
 - Ausarbeiten konkreter Szenarien und Workflows für den Einsatz von CVs
 - Arbeitsgruppe für CVs für Typologien von historischen Orten (IEG Mainz)
- Ausarbeiten der Beziehung zu den **Semantic Web** Aktivitäten
 - Transformieren der Daten in Linked Data (RDF)
 - Verlinken von CVs/Ontologien (dbpedia als [LOD-pivot](#))

Vocabulary Proxy

