EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

SFB 833, Project INF:
Heterogene Forschungsprimärdaten des SFB 833
Repräsentation und Verarbeitung

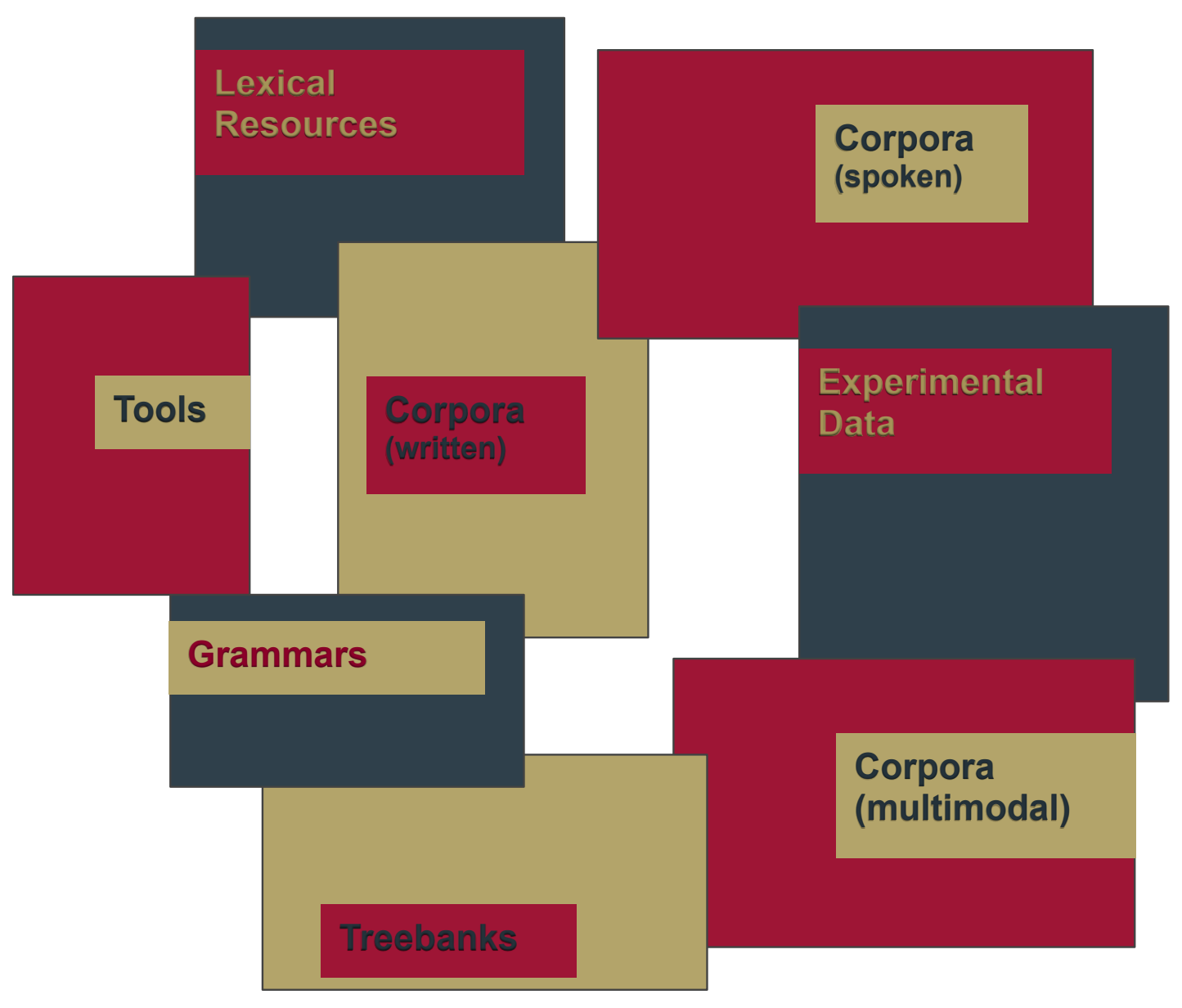
Centre for Sustainability of Linguistic Data

Managing Linguistic Resources by Enriching Their Metadata with Linked Data

Christina Hoppermann
Thorsten Trippel
Claus Zinn

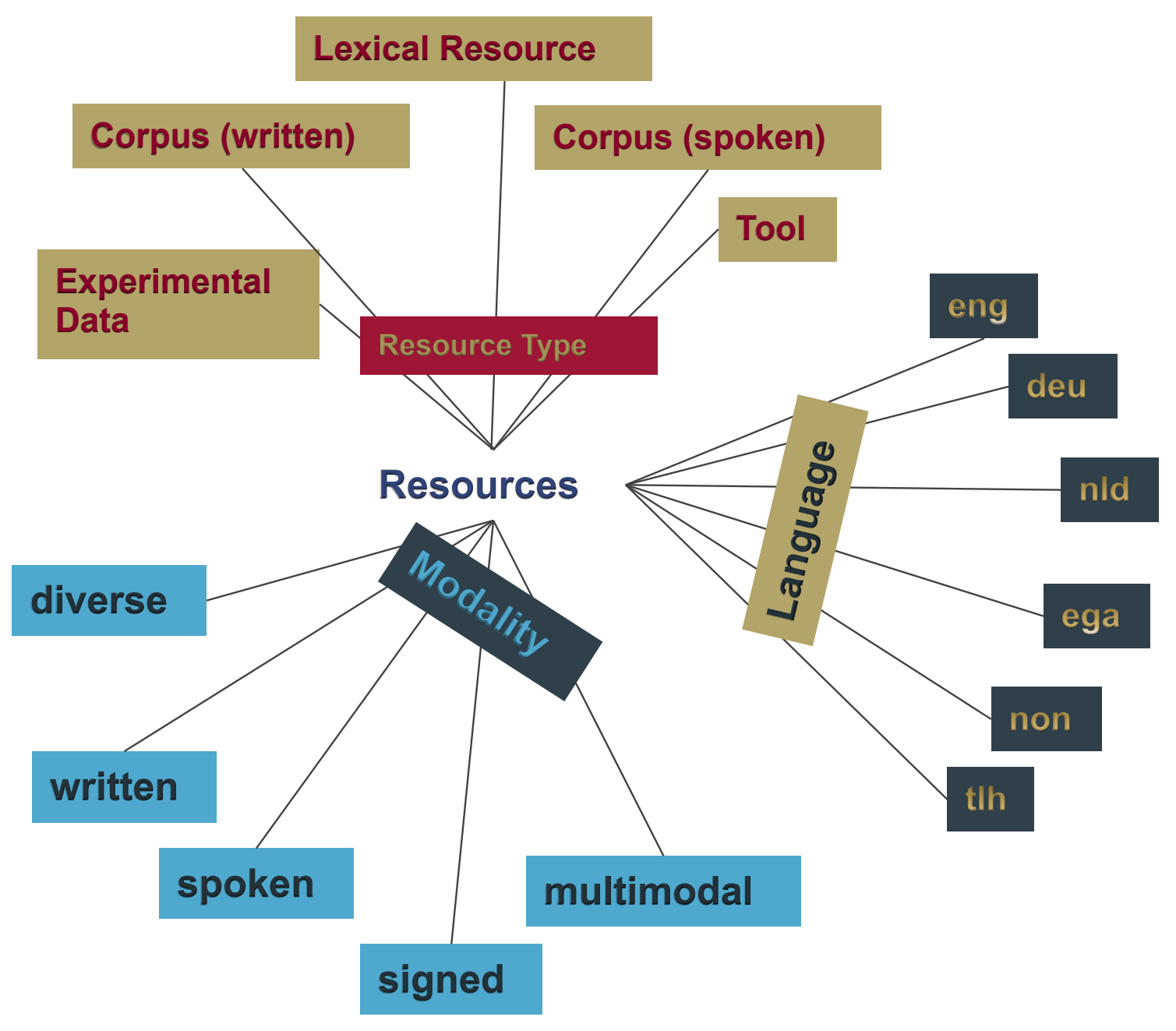
Sustainable Management of Linguistic Resources

Large Collections of Linguistic Resources



There is a large number and variety of linguistic resources such as corpora, lexicons, grammars, tools or experimental data. These are valuable data, but – so far – there is no established infrastructure in place to manage this data.
Often, research data is hardly accessible from outside the institutional boundaries, and sometimes, research is unnecessarily duplicated because of this.

Complex Descriptional Requirements



There are different types of resources. While they share a common set of descriptors, they also require type-specific metadata fields.
The Dublin Core metadata set lacks descriptional means to give a full account of research data in linguistics. A more expressive metadata framework is needed such as the Component Metadata Infrastructure (CMDI)

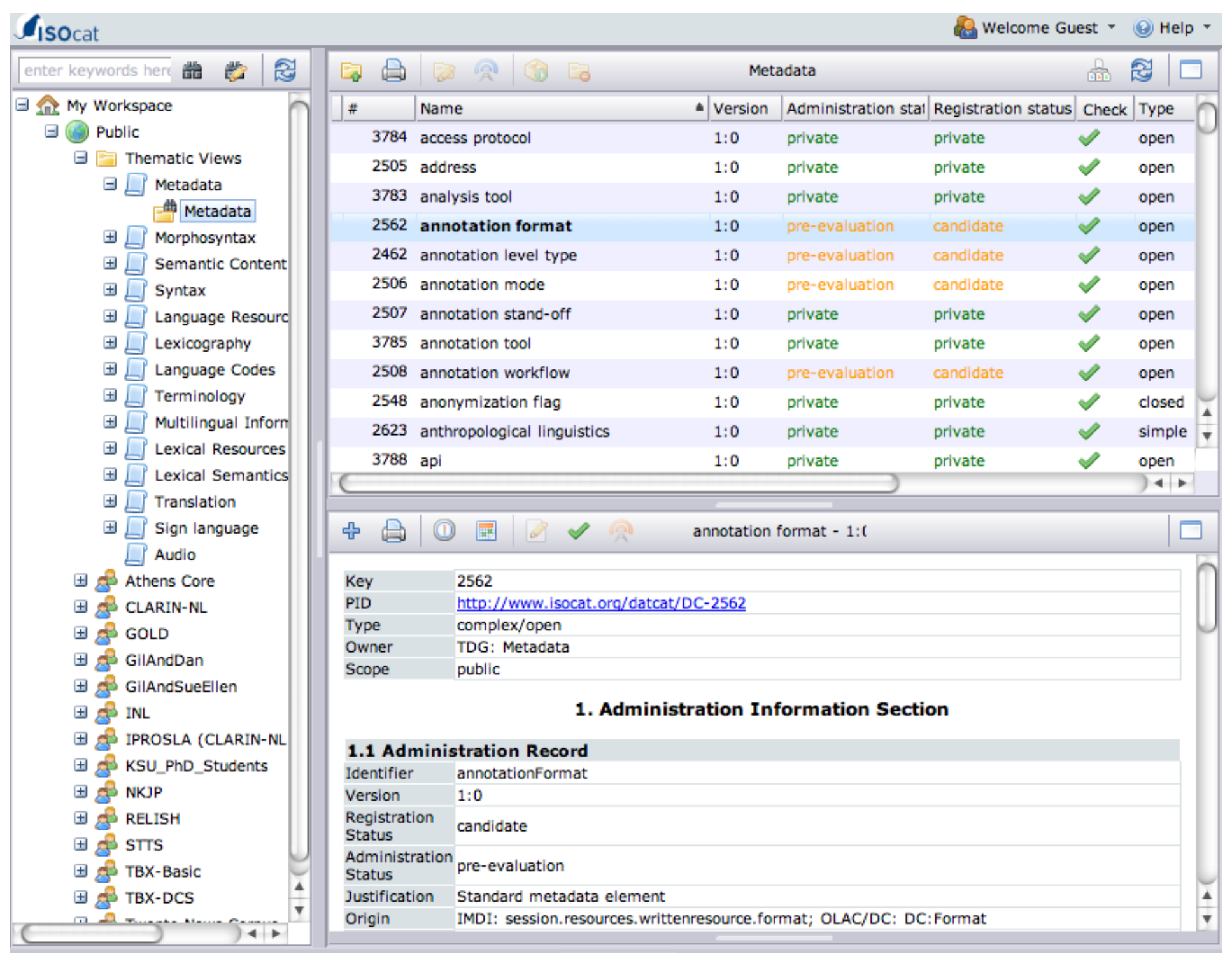
Policy of the German Research Foundation

1. Primary research data is data that has been gathered in the course of the study of sources, experiments, measurements, and census or poll activities. They constitute the foundation for scientific publication.
2. It is necessary to devise a domain-specific organizational concept that defines the sustainable management of (research) data in a given scientific discipline.
3. The management of research data must is carried out in the framework of existing standards.
4. Data mark-up must consider rights management and include the names of data creators. An open access policy is advocated.
5. Data should become public at the end of a research project (or some limited time thereafter). This holds, in particular, for publicly funded projects. When data is attractive for commercial exploitation, divergent rules apply, but in consensus with all concerned scientists.
6. Research data must be labeled, at least, following Dublin Core. The metadata description should also include aspects relevant to the creation and subsequent reuse of research data.
7. The scientific discipline producing the research data shall devise criteria and methods to assure its quality.

Original German document at http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

Component Meta Data Infrastructure (CMDI)

ISOcat Data Registry



#	Name	Version	Administration stat	Registration status	Check	Type
3784	access protocol	1:0	private	private	✓	open
2505	address	1:0	private	private	✓	open
3783	analysis tool	1:0	private	private	✓	open
2562	annotation format	1:0	pre-evaluation	candidate	✓	open
2462	annotation level type	1:0	pre-evaluation	candidate	✓	open
2506	annotation mode	1:0	pre-evaluation	candidate	✓	open
2507	annotation stand-off	1:0	private	private	✓	open
3785	annotation tool	1:0	private	private	✓	open
2508	annotation workflow	1:0	pre-evaluation	candidate	✓	open
2548	anonymization flag	1:0	private	private	✓	closed
2623	anthropological linguistics	1:0	private	private	✓	simple
3788	api	1:0	private	private	✓	open

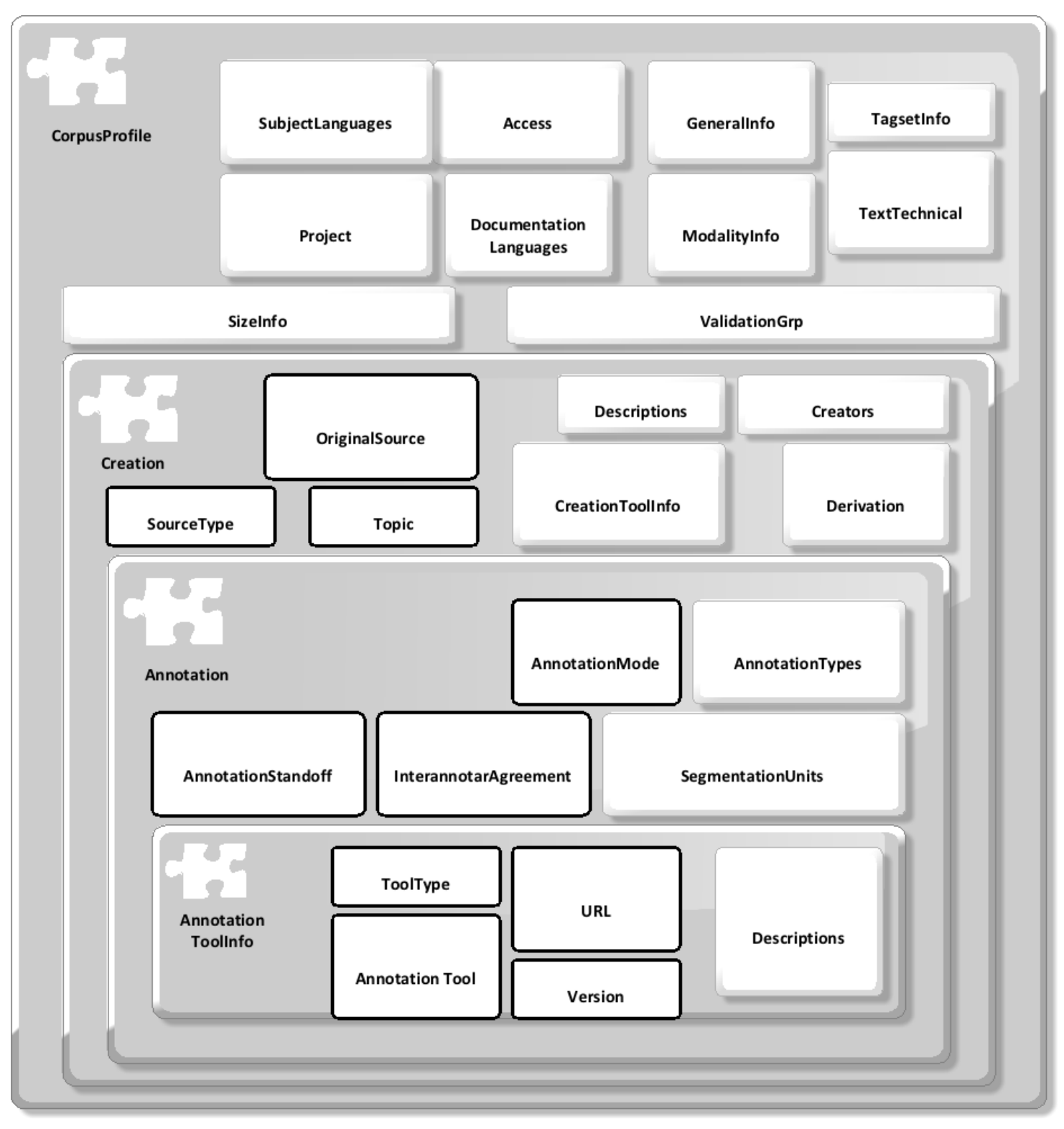
Key	2562
PID	http://www.isocat.org/datcat/DC-2562
Type	complex/open
Owner	TDG: Metadata
Scope	public

1. Administration Information Section

1.1 Administration Record	
Identifier	annotationFormat
Version	1:0
Registration Status	candidate
Administration Status	pre-evaluation
Justification	Standard metadata element
Origin	IMDI: session.resources.writtenresource.format; OLAC/DC: DC:Format

The ISOcat registry is a community-based platform for managing elementary field descriptors for research data in linguistics. The registry has a rather flat structure, with data categories being grouped into *Thematic Domains* (e.g., *Syntax*, *Semantics*, *Terminology*, *Metadata*). There are, however, data categories of type *complex* being definable in terms of data categories of type *simple*. The thematic domain *Metadata* lists about 450 data categories.
Access the ISOcat registry via http://www.isocat.org.

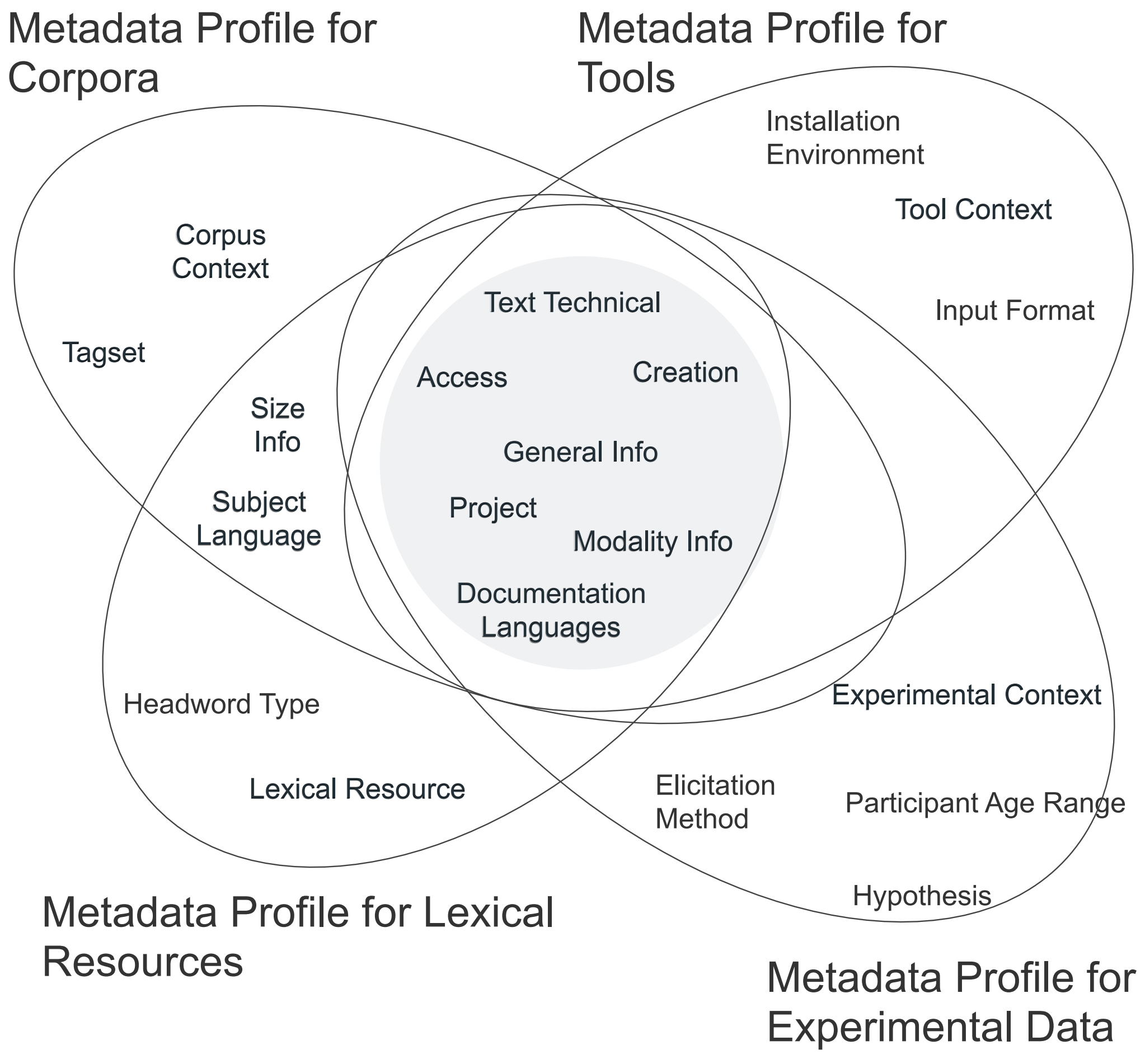
Component Registry



The Component Registry is a community-based platform for managing metadata structures, grouping together elementary field descriptors from the ISOcat registry or other (predefined) components. Components serve as complex building blocks for *profiles*. A profile is a schema for the description of a single type of linguistic resource.

Access the Component Registry via http://catalog.clarin.eu/ds/ComponentRegistry/#

Profiles



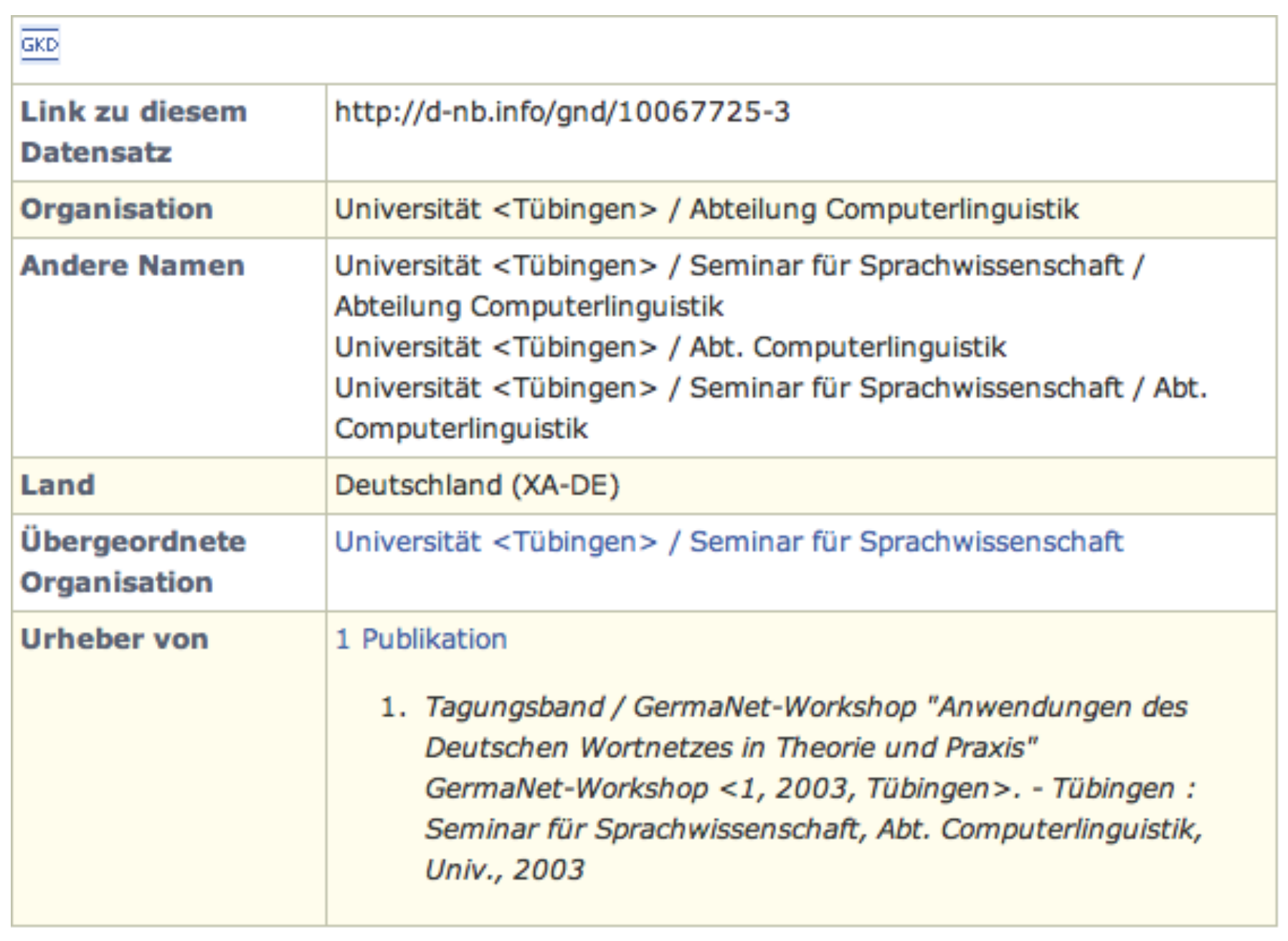
In NaLiDa/SFB833, we have defined 6 profiles, and contributed many entries to the ISOcat registry and the Component Registry.

Use of Linked Data

Authority Files of the German National Library

The Gemeinsame Normdatei (GND) dataset pools *three German authority files:*
- Personennamendatei (**Personal Name Authority File**, PND)
- *Gemeinsame Körperschaftsdatei* (**Corporate Body Authority File**, *GKD)*
- *Schlagwortdatei* (**Subject Headings Authority File**, *SWD)*

together into a single universal authority file (GND) with > 40 Mio. triples, > 40k links to dbpedia, 38k links to lcsh, and approx. 1.8 mio links to viaf.

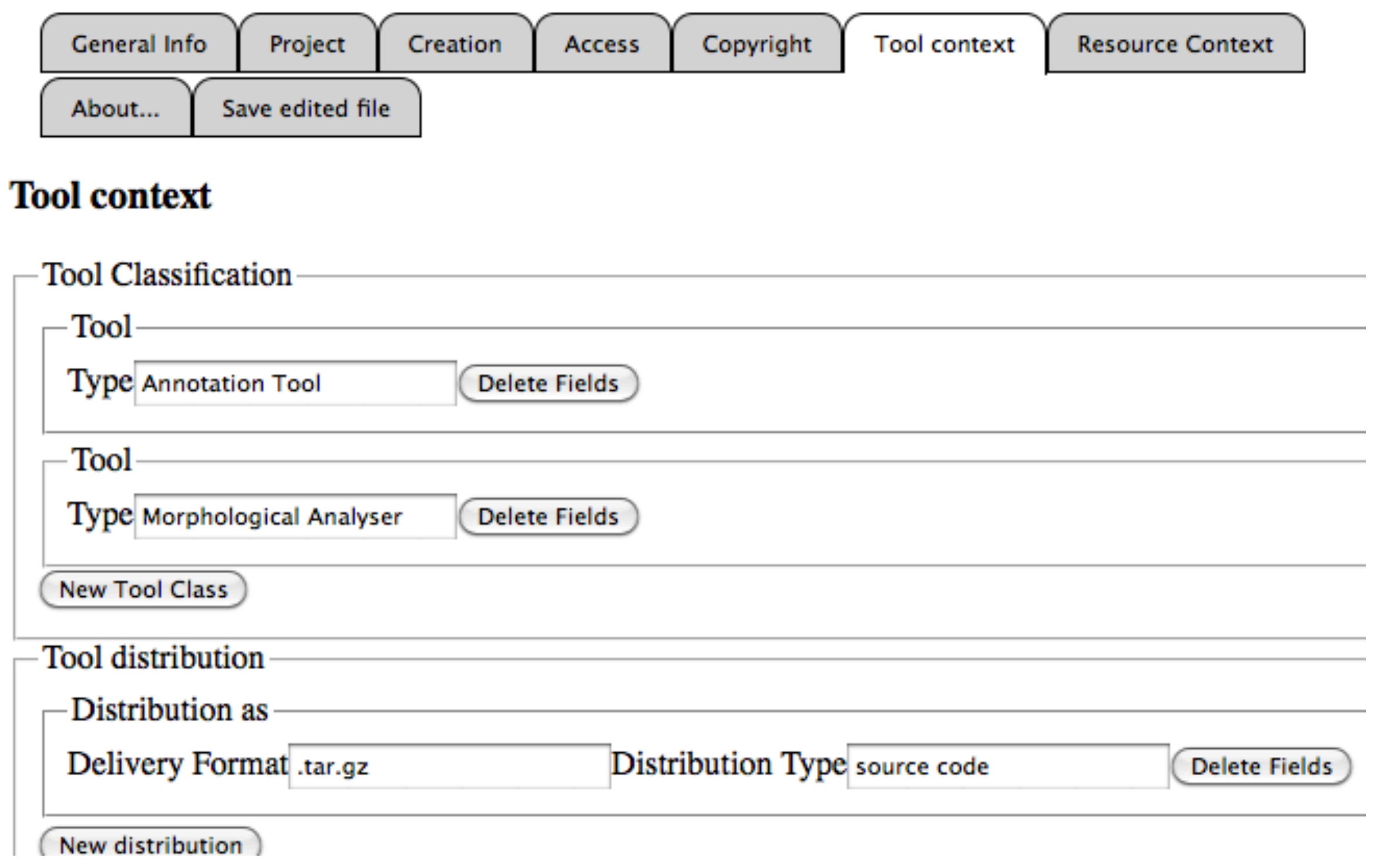


GKD

Link zu diesem Datensatz	http://d-nb.info/gnd/10067725-3
Organisation	Universität <Tübingen> / Abteilung Computerlinguistik
Andere Namen	Universität <Tübingen> / Seminar für Sprachwissenschaft / Abteilung Computerlinguistik Universität <Tübingen> / Abt. Computerlinguistik Universität <Tübingen> / Seminar für Sprachwissenschaft / Abt. Computerlinguistik
Land	Deutschland (XA-DE)
Übergeordnete Organisation	Universität <Tübingen> / Seminar für Sprachwissenschaft
Urheber von	1 Publikation 1. *Tagungsband / GermaNet-Workshop "Anwendungen des Deutschen Wortnetzes in Theorie und Praxis" GermaNet-Workshop <1, 2003, Tübingen>. - Tübingen : Seminar für Sprachwissenschaft, Abt. Computerlinguistik, Univ., 2003*

- **GKD**: 915.000 records of institutions (2300 entries for Tübingen institutions)
- **PND**: 3.600000 records for persons (1.800000 individualised entries)
- **SWD**: 600.000 descriptors and 700.000 synonyms, 115.000 hierarchical and 26.000 associative relations; terms grouped into 500 classes (36 clusters); 40 different subfields for linguistics; mapping of SWD entries to Dewey classification

Metadata Editor



Tool context

Tool Classification

Tool
Type: Annotation Tool

Tool
Type: Morphological Analyser

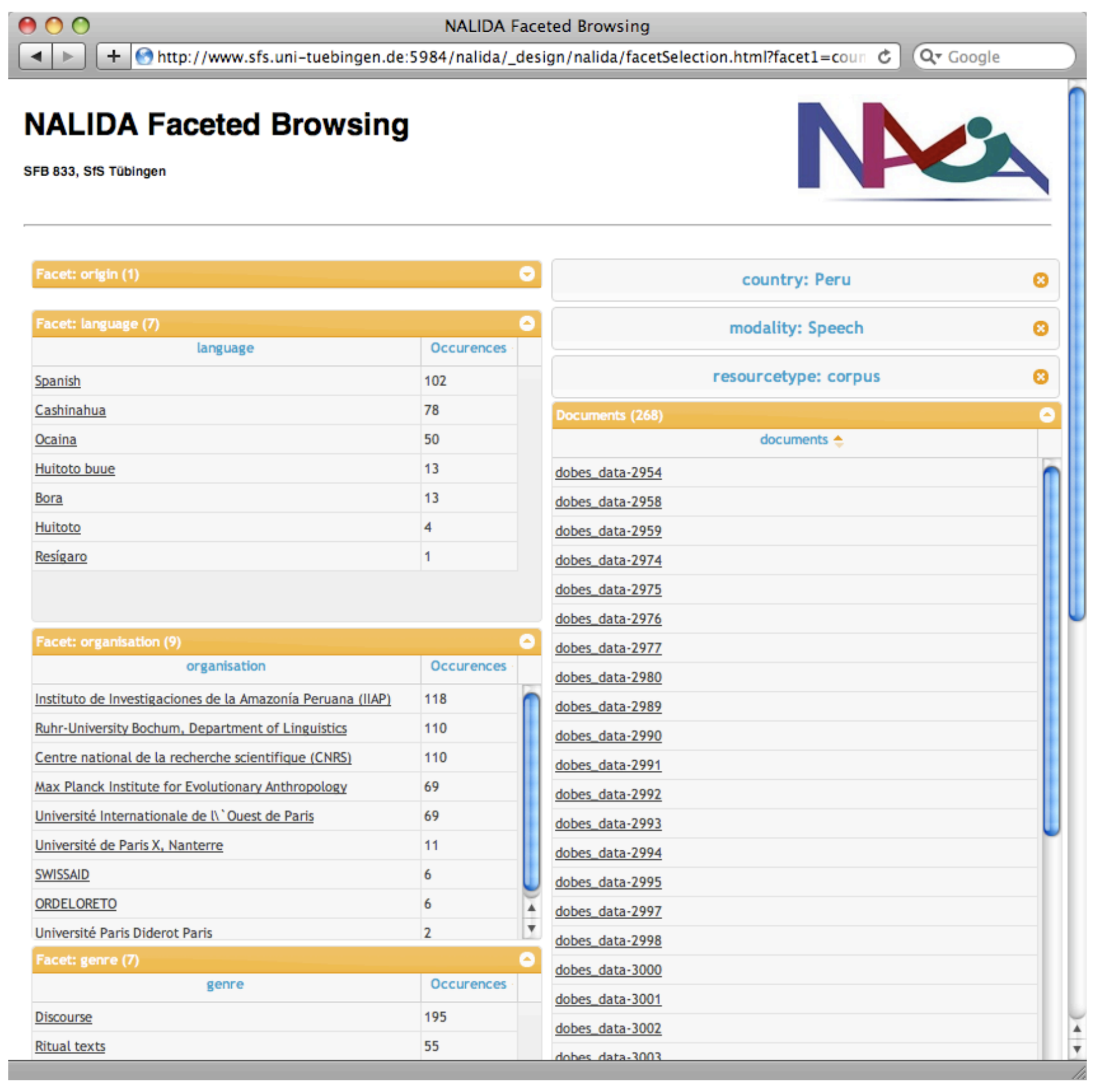
Tool distribution

Distribution as
Delivery Format: .tar.gz
Distribution Type: source code

The Metadata editor is based on XForms. All XForms specifications are automatically generated from a resource's CMDI profile and a configuration file (using XSLT).
The Metadata editor offers GUI elements where values can be entered using auto-completion on PND and GKD data. Any value selected will get associated to the persistent URL of the German National Library.

See http://www.sfs.uni-tuebingen.de/nalida/en/ for the catalogue (faceted search) and more information about the NaLiDa project.

Faceted Search


NALIDA Faceted Browsing
http://www.sfs.uni-tuebingen.de:5984/nalida/_design/nalida/facetSelection.html?facet1=coun
Google

NALIDA Faceted Browsing

SFB 833, SfS Tübingen

Facet: origin (1)

Facet: language (7)

language	Occurences
Spanish	102
Cashinahua	78
Ocaina	50
Huitoto buue	13
Bora	13
Huitoto	4
Resígaro	1

Facet: organisation (9)

organisation	Occurences
Instituto de Investigaciones de la Amazonía Peruana (IIAP)	118
Ruhr-University Bochum, Department of Linguistics	110
Centre national de la recherche scientifique (CNRS)	110
Max Planck Institute for Evolutionary Anthropology	69
Université Internationale de l\`Ouest de Paris	69
Université de Paris X, Nanterre	11
SWISSAID	6
ORDELORETO	6
Université Paris Diderot Paris	2

Facet: genre (7)

genre	Occurences
Discourse	195
Ritual texts	55

country: Peru
modality: Speech
resourcetype: corpus

Documents (268)

documents
dobes_data-2954
dobes_data-2958
dobes_data-2959
dobes_data-2974
dobes_data-2975
dobes_data-2976
dobes_data-2977
dobes_data-2980
dobes_data-2989
dobes_data-2990
dobes_data-2991
dobes_data-2992
dobes_data-2993
dobes_data-2994
dobes_data-2995
dobes_data-2997
dobes_data-2998
dobes_data-3000
dobes_data-3001
dobes_data-3002
dobes_data-3003

For the NaLiDa faceted browser, we have curated existing metadata collections in a semi-automatically manner using the GND triples.

Acknowledgements

The NaLiDa project is funded by the German Research Foundation (DFG): *Wissenschaftliche Literaturversorgungs- und Informationssysteme* (LIS). Parts of this work is carried out within the Collaborative Research Centre 833 (SFB 833). Poster and demo presented at ISWC 2011 on Oct 25, 2011 in Bonn, Germany (C. Zinn).