Metadata and the use of it

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Unterrichtseinheit Metadaten im Rahmen des Austausches von Lehrmaterial

- Zielgruppe der Lehreinheit:
  - Linguisten und fortgeschrittene Linguistik-Studierende
  - Computerlinguisten und –studierende
  - Z.T. praktizierende Wissenschaftler, die Daten archivieren wollen oder müssen
- Ziele:
  - Teilnehmer verstehen, was Metadaten zur Beschreibung von Ressourcen sind
  - Teilnehmer kennen Gründe, warum Metadaten sinnvoll und nützlich sind
  - Teilnehmer können Metadatenformulare ausfüllen
  - Teilnehmer kennen Metadateninfrastrukturen: ISOcat, Komponenten, Profile
  - Teilnehmer haben eine Vorstellung, wie Metadaten verbreitet werden
- Die Folien und Auszüge daraus dürfen für die eigene Erstellung von Lehrmaterialien verwendet werden

Aufbau dieser Unterrichtseinheit

- Der Foliensatz besteht aus zwei Versionen:
  - Foliensatz für den Dozenten: hier sind zusätzliche Informationen zum Aufbau und zur Struktur enthalten, die für Teilnehmer nicht primär hilfreich wären
  - Foliensatz für den Unterricht: die Folien dienen der sequentiellen Präsentation. Auf Effekte wird verzichtet, so dass die Präsentation auch als PDF erfolgen kann.
- Um technische Herausforderungen zu vermeiden, ist als Zielformat PDF vorgesehen. Dadurch werden Präsentationsfunktionen wie Referentenansicht, etc. nicht verwendet.
- Die Unterrichtseinheit ist in "Kapitel" untergliedert, am Ende jedes Kapitel könnten Pausen eingeschoben oder die Sitzung beendet werden. Kapitel werden als Einheit betrachtet, sie sind aber unterschiedlich lang.
- Übungsaufgaben werden angeboten, es wird empfohlen sich zumindest eine Lösung zu überlegen. Die Lösungen werden nicht immer eindeutig sein, im Dozentenfoliensatz ist immer eine Beispiellösung enthalten.
- Informationen aus Tutorien vom Projekt NaLiDa (http://www.sfs.uni-tuebingen.de/nalida/de/doku/tutorials) und CLARIN (www.clarin.eu/cmdi) werden dabei verwendet. Daher geht ein besonderer Dank für die extensive Dokumentation an Christina Hoppermann, Daan Broeder, Dieter van Uytvanck, Menzo Windhouwer.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Schlagworte zur Unterrichtseinheit "Metadaten" (unvollständig)

- Dublin Core
- Bibliographie
- Datenkategorie
- DatKat
- ISOcat
- Datenkategorien-Repositorium
- Repositorium
- OAI

- OLAC
- Component Metadata Infrastructure
- CMDI
- Komponenten
- Profile
- Instanzen
- Ressource

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Bibliographie (Auswahl)

- CMDI, http://www.clarin.eu/cmdi
- CMDI Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., and Witt, A. (2011). A pragmatic approach to XML interoperability — the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.
- I. Schuurman, M.A. Windhouwer. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? 2nd Supporting Digital Humanities conference (SDH 2011), 17-18 November 2011, Copenhagen, Denmark.
- Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 900-903). European Language Resources Association (ELRA).
- *Gavrilidou M., Labropoulou P., Piperidis S., Monachini M., Frontini F., Francopoulo G., Arranz V. and Mapelli V. (2011). A* Metadata Schema for the Description of Language Resources (LRs)*,* 5th International Joint Conference on Natural Language Processing (IJCNLP 2011).

- D. Broeder, D. van Uytvanck, M. Gavrilidou, and T. Trippel. Standardizing a component metadata infrastructure. In Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012), Istanbul, Forthcoming.
- CMDI MI Search Engine. CMDI Meertens Institute Search Engine. http://www.meertens.knaw.nl/cmdi.
- M. Durco, D. Broeder, and M. Windhouwer. Semantic mapping - groundwork for query expansion and semantic search. 2012, submitted.
- ISO 12620:2009. Terminology and other language and content resources - specification of data categories and manage- ment of a data category registry for language resource. ISO, 2009.
- ISOcat. http://www.isocat.org.
- OLAC. http://www.language-archives.org/.
- VLO. http://catalog.clarin.eu/ds/vlo/.
- Dublin Core: http://dublincore.org/
- Tutorials (z. B. Component Registry): http://www.sfs.uni-tuebingen.de/nalida/de/doku/tutorials

Struktur

- Einleitung
- Motivation
- Beschreibung eines Buches: Bibliographien und Dublin Core
- Beispiele für linguistische Ressourcen: Beschreibung durch OLAC
- Datenkategorien als Beschreibungsebenen für Ressourcen
- Zentrales Verzeichnis für Datenkategorien: ISOcat und Beschreibung einer Ressource mit ISOcat Datenkategorien
- Generalisierungen von Gruppen von Datenkategorien: Komponenten
- Generalisierungen von Gruppen von Komponenten für Ressourcentypen
- Die CMDI Component Registry
- Erstellen von CMDI-Instanzen
- Zusammenfassung

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 1

Introduction of this unit

Goal of this unit

- Introducing metadata
- Creating a metadata schema
- Adjusting metadata schemas

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

At the end you can

- Describe different types of resources
- Create metadata with
  - Data categories from the CLARIN Concept Registry
  - Components from the Component Registry (including creating and using them)
- Use tools for creating metadata

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Chapter 2

Motivation for metadata

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Metadata – what is that?

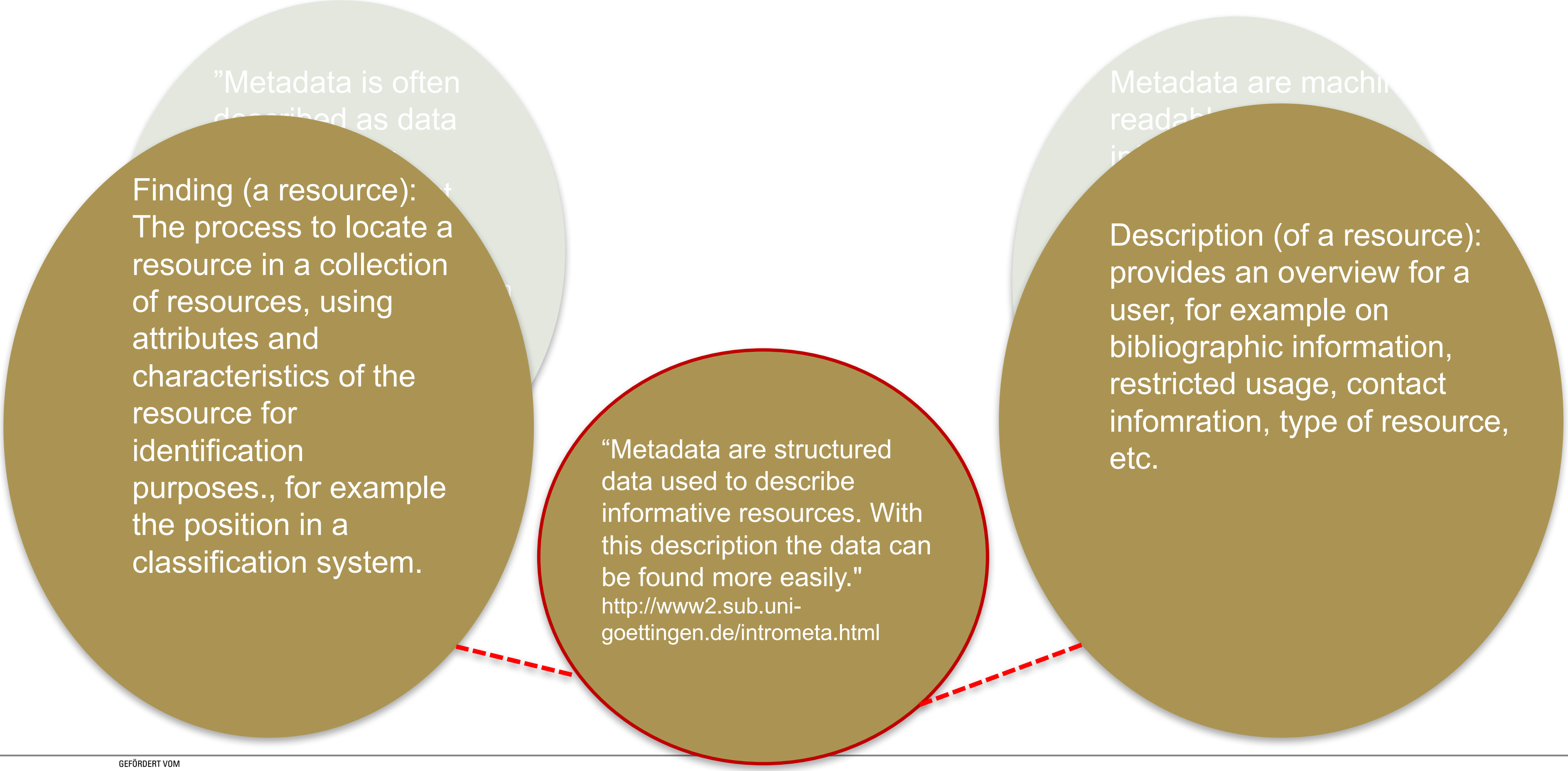
”Metadata is often described as data about data or information about information."

http://www.niso.org/publications/press/UnderstandingMetadata.pdf
(Metadata is often called data about data or information about information)

“Metadata are structured data used to describe informative resources. With this description the data can be found more easily."
http://www2.sub.uni-goettingen.de/intrometa.html

Metadata are machi readable bits of information for electronic resources or other things. M
(Übersetzung siehe: http://www2.sub.uni-goettingen.de/intrometa.html, Original: http://www.w3c.org/metadata)

Metadata – what is that?



”Metadata is often described as data

Metadata are machi
readab
in

Finding (a resource): The process to locate a resource in a collection of resources, using attributes and characteristics of the resource for identification purposes., for example the position in a classification system.

“Metadata are structured data used to describe informative resources. With this description the data can be found more easily."
http://www2.sub.uni-goettingen.de/intrometa.html

Description (of a resource): provides an overview for a user, for example on bibliographic information, restricted usage, contact infomration, type of resource, etc.

Data and metadata in archives and repositories

- Archive:
  - Authoritative reference copy of data
  - Long(er) time storage
- Data safety and integrity by access control
  - Identification of users by the means of central infrastructures
  - Access control lists for repository access using defined procedures
  - Defined fall back procedures for discontinued projects or organizational units
  - Availability of data
- Citability by reference copy
- Citation by referencing the description of the resource
  Description of research data: part of the archiving workflow

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Drill: That's too abstract

- Name some archives and repositories!
- What kind of resources are "saved" there?
- What could be the metadata there?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Übung: Nicht so abstrakt, bitte!

- Zählen Sie Archive und Repositorien auf, die Sie kennen!
  - Mögliche Antworten:
  - Universitätsbibliothek
  - Stadtarchiv
  - Museum
- Was für Ressourcen werden dort "gespeichert"?
  - Bücher
  - Museale Artikel
- Was könnte man dort unter Metadaten verstehen?
  - Informationen im Bibliothekskatalog
  - Museums-Bestandsliste/Inventar

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 3

Describing a book

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

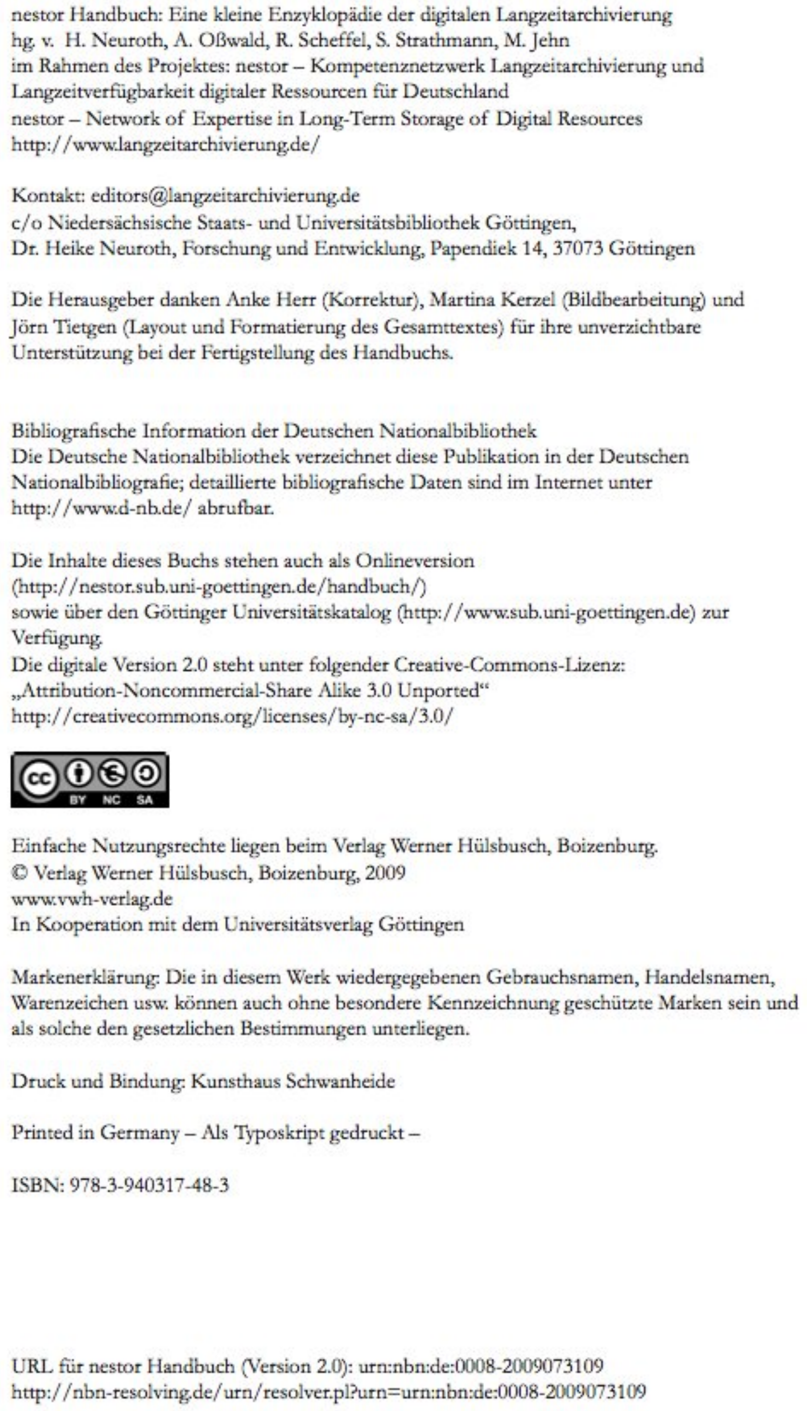
What is that?

H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg.)

nestor Handbuch

Eine kleine Enzyklopädie der digitalen Langzeitarchivierung

Version 2.0



nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion (http://nestor.sub.uni-goettingen.de/handbuch/) sowie über den Göttinger Universitätskatalog (http://www.sub.uni-goettingen.de) zur Verfügung.

www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

URL für nestor Handbuch (Version 2.0): urn:nbn:de:0008-2009073109
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2009073109

What is that?

- Describe this book, keep in mind the aspect of
  - Locating it
  - Describing the content
- Look at your description
  - Anything familiar?
  - What is it?

H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg

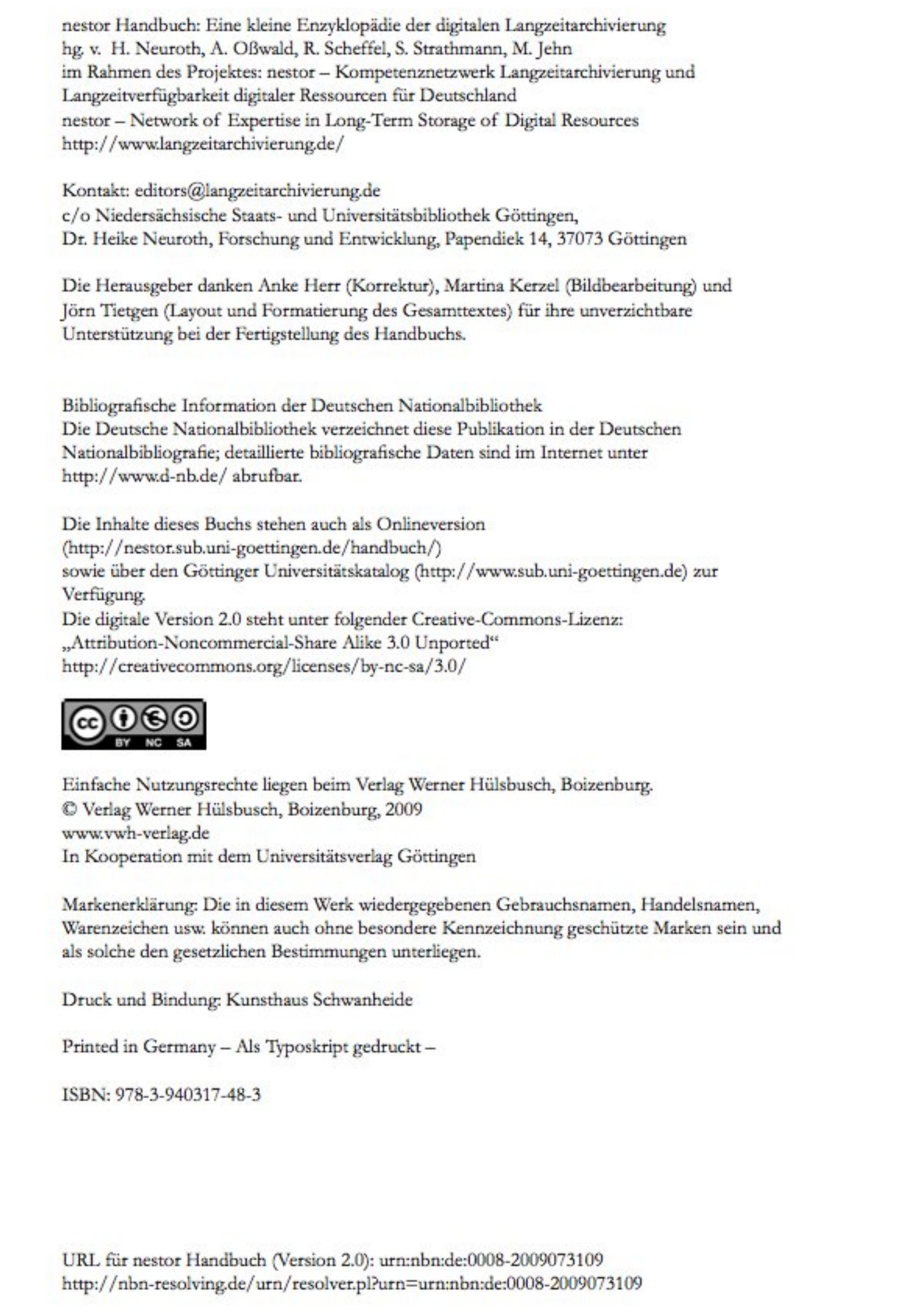
nestor Handbuch

Eine kleine Enzyklopädi
der digitalen Langzeitarchivierun

Version 2.0

nestor

vwh



nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion (http://nestor.sub.uni-goettingen.de/handbuch/) sowie über den Göttinger Universitätskatalog (http://www.sub.uni-goettingen.de) zur Verfügung.

www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

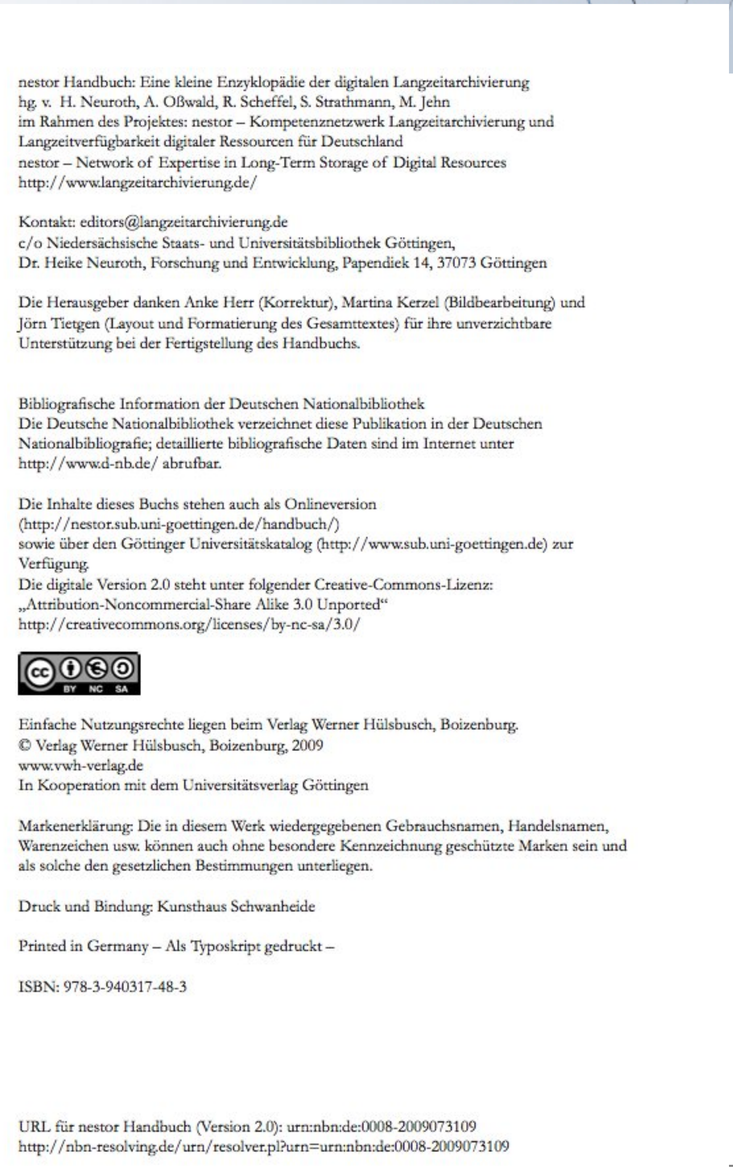
URL für nestor Handbuch (Version 2.0): urn:nbn:de:0008-2009073109
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2009073109

Was ist denn das?

- Beschreiben Sie dieses Buch unter dem Aspekt der
  - Auffindbarkeit
  - Des Inhalts

  Antwort: Die Bibliographie dieses Buches:
  Herausgeber: H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn,
  Titel=   nestor Handbuch: Eine kleine Enzyklopädie der
  digitalen Langzeitarchivierung
  Jahr =   2010,
  Identifikator= urn:nbn:de:0008-20100305186
  Farbe= blau, ...
  Schlagwörter= Archivierung, digitale Ressourcen, ...
- Betrachten Sie Ihre Beschreibung
  - Kommt Ihnen etwas bekannt vor?
  - Was?
  - Antwort: Das ist das, was man in einem Bibliothekskatalog finden würde.

H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn (Hrsg.)
nestor Handbuch
Eine kleine Enzyklopädie
der digitalen Langzeitarchivierung
Version 2.0
nestor
vwh


nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
hg. v. H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn
im Rahmen des Projektes: nestor – Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Korrektur), Martina Kerzel (Bildbearbeitung) und Jörn Tietgen (Layout und Formatierung des Gesamttextes) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte dieses Buchs stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/handbuch/)
sowie über den Göttinger Universitätskatalog (http://www.sub.uni-goettingen.de) zur Verfügung.
Die digitale Version 2.0 steht unter folgender Creative-Commons-Lizenz:
„Attribution-Noncommercial-Share Alike 3.0 Unported“
http://creativecommons.org/licenses/by-nc-sa/3.0/

BY NC SA

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2009
www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide

Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-940317-48-3

URL für nestor Handbuch (Version 2.0): urn:nbn:de:0008-2009073109
http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2009073109

Purpose of bibliographic data: description

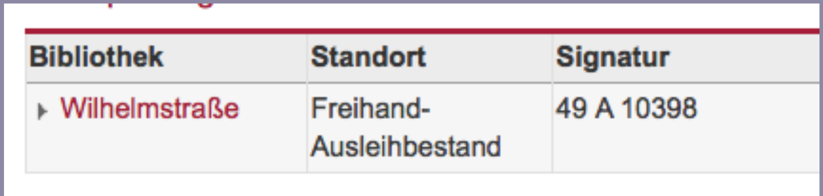
- ”external"
  - Title
  - Author
    - Authority file (e.g. VIAF)
  - publisher
  - year
  - Page number
  - cover
  - institution
    - Authority file again (GKD)

- ”internal"
  - topic
  - keyword
    - Standard keyword files (SND)
  - Position in classification system
    - Dewey Decimal Classification (DDC)
    - Universal Decimal Classification (UDC)

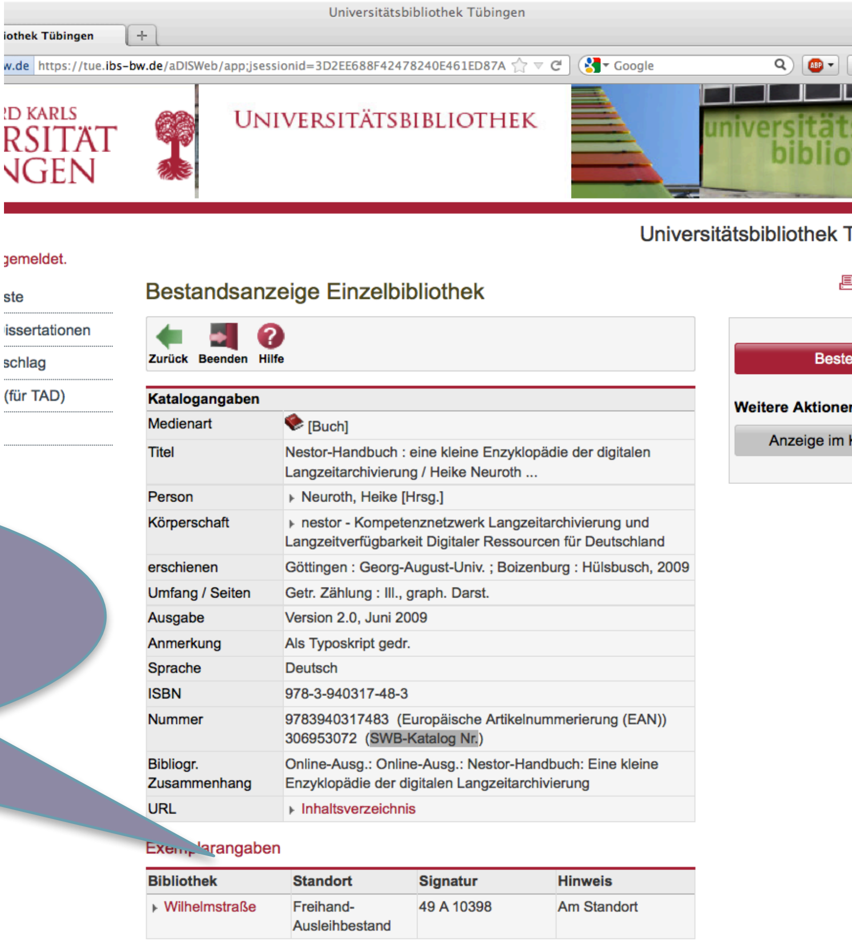
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Purpose of bibliographic data: locating

- Signature or location information
- Together with a library map



Bibliothek	Standort	Signatur
Wilhelmstraße	Freihand-Ausleihbestand	49 A 10398



Bestandsanzeige Einzelbibliothek

Katalogangaben	
Medienart	[Buch]
Titel	Nestor-Handbuch : eine kleine Enzyklopädie der digitalen Langzeitarchivierung / Heike Neuroth ...
Person	Neuroth, Heike [Hrsg.]
Körperschaft	nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland
erschienen	Göttingen : Georg-August-Univ. ; Boizenburg : Hülsbusch, 2009
Umfang / Seiten	Getr. Zählung : Ill., graph. Darst.
Ausgabe	Version 2.0, Juni 2009
Anmerkung	Als Typoskript gedr.
Sprache	Deutsch
ISBN	978-3-940317-48-3
Nummer	9783940317483 (Europäische Artikelnummerierung (EAN)) 306953072 (SWB-Katalog Nr.)
Bibliogr. Zusammenhang	Online-Ausg.: Online-Ausg.: Nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung
URL	Inhaltsverzeichnis

Exemplarangaben

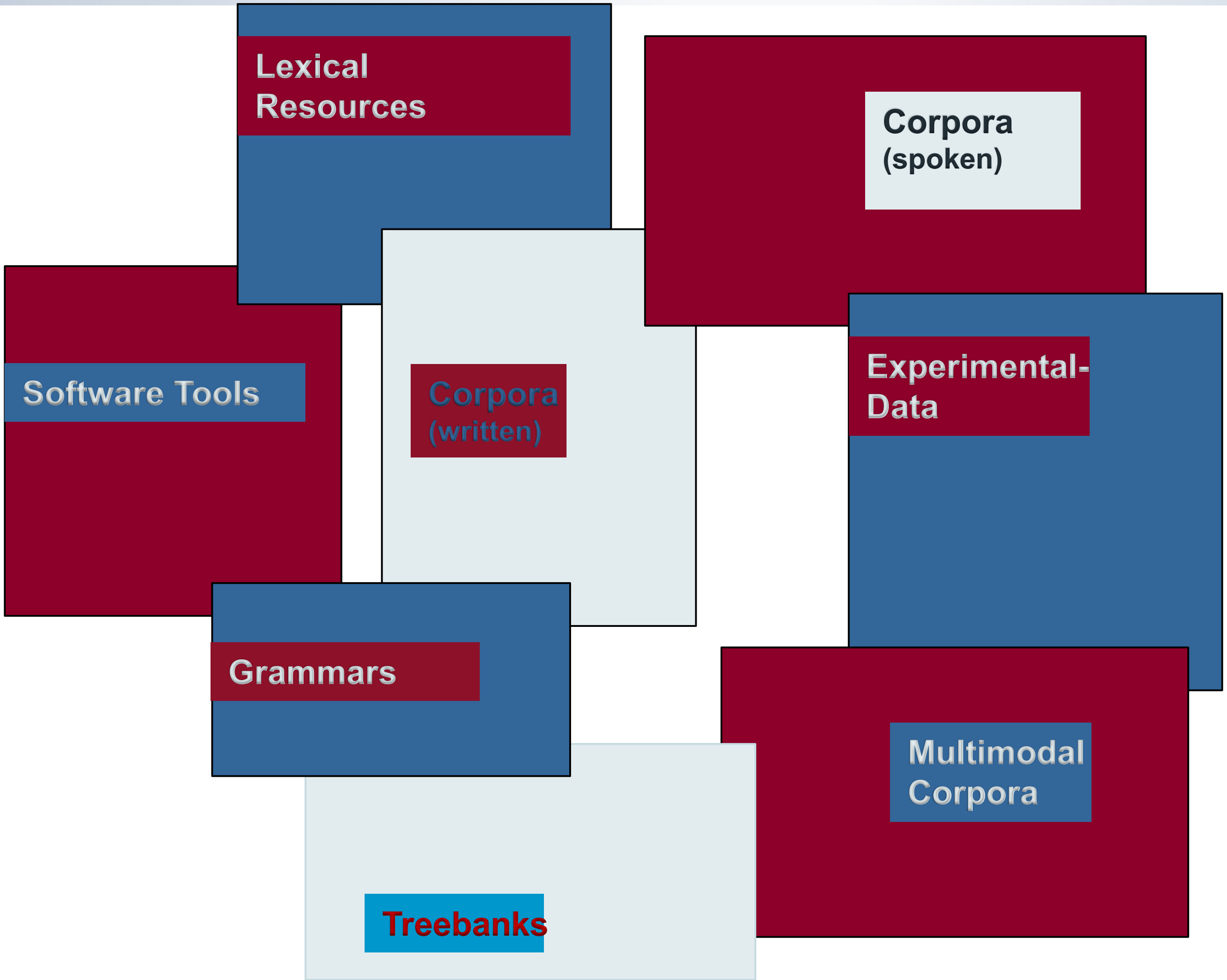
Bibliothek	Standort	Signatur	Hinweis
Wilhelmstraße	Freihand-Ausleihbestand	49 A 10398	Am Standort

Chapter 4

Example of linguistic resources

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Examples of linguistic resources



Lexical Resources

Corpora (spoken)

Software Tools

Corpora (written)

Experimental-Data

Grammars

Multimodal Corpora

Treebanks

Resource type: text corpus

Resource Name:	TüBa-D/Z
Resource Title:	Tübinger Baumbank des Deutschen/Zeitungskorpus
Resource Class:	Corpus
Corpus Type:	general corpus
Annotation Type:	POS, inflectional morphology, named entities, referential relations, grammatical functions
Annotation Mode:	semi-automatisch
Tagset:	Stuttgart-Tübingen Tagset (STTS)

Resource type: Tool

Software Tools

Resource Name:	Tree Tagger
Resource Title:	Tree Tagger: A Language - Independent POS Tagger
Resource Class:	Tool
Tool Type:	Annotation Tool, Lemmatizer, Chunker, POS Tagger
Input:	wortsegmentierter Text
Output:	Annotierter Text mit Informationen zu POS und Lemmata
Operating System:	Linux, Windows, Mac OS X, Solaris

Resource type: Experimental data

Resource Name:	Experiment 1
Resource Title:	Experiment on bridge verbs and V2 structures
Resource Class:	Experiment
Experiment Type:	Survey Data
Domain:	Linguistik
Elicitation Method:	Fragebogen
Elicitation Software:	WebExp
Sample Size:	23 Teilnehmer
Mean Age:	30,4 Jahre

Experimental-Data

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Drill

- What kinds of resources can you think of?
- What types of resources do you work with?
- What type of resources do you create?
- How would you describe them?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Übung

- Welche Ressourcen kennen Sie?
- Mit welchen Ressourcen arbeiten Sie?
- Was für Ressourcen erstellen Sie selber?
- Wie würden Sie die Ressourcen beschreiben?

- Beispiele finden sich auf den vorherigen Folien

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 5

Data categories for describing resources

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Attribute-Value-Structures

- Pairs of "Name" und ”Value"
  - Name: Chomsky
  - First name: Noam
  - Name: Halle
  - First name: Morris
  - Title: The sound pattern of English
  - Year:1968

Name

Value

- Other terminology:
  - Data category value pair
  - Name-Value-Pair
  - Key-Value-Pair
  - AV-Structures
  - ...

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Attribute-Value-Structures

- Pairs of "Name" and ”Value"
  - Name: Chomsky
  - First name: Noam
  - Name: Halle
  - First name: Morris
  - Title: The sound pattern of English
  - Year:1968

- Who are the authors?
  - Hint:
    - Ambiguous structure
    - ”Order" is not meaningful in AV-Structures
  - There are two options
    - If everybody has one (and only one) name and first name
- How can you make something like that unique?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Data categories

- ”When creating metadata for a resource, data categories are utilized for providing categories for metadata. For example the data category *resource title* is filled with the title of a resource and *metadata creator* contains the name of the creator of the metadata."

  http://www.sfs.uni-tuebingen.de/nalida/de/doku/glossar.html
- Attributes in AV-Structures
- Terminology from data base development

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Example data categories

- Title
- Author
- Description
- Language
- Size
- Copyright
- Summary
- keyword(s)
- Address
- Institution
- Funder
- Subject field
- Type of resource
- Genre
- Programming language
- ...

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Drill

- Select a resource you know or one that you created. Which data categories would you use to describe them?
- Which data categories would be inadquate for this type of resource?
- Could you group the data categories? Which data categories should always be grouped together?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Übung

- Wählen Sie eine Ressource, die Sie kennen oder erstellt haben. Welche Datenkategorien würden Sie verwenden, um Ihre Ressource zu beschreiben?
  - Siehe vorherige Folien
- Welche Datenkategorien würden für diese Ressource nicht zutreffen?
  - Für ein Korpus ist "Programmiersprache" nicht relevant, für eine "Software" vielleicht das Genre, etc.
- Können Sie Gruppen von Datenkategorien bilden? Welche Datenkategorien sollten immer zusammen verwendet werden?
  - Straße und Hausnummer, PLZ und Ort, Name und Rolle (Funktion), ...

Chapter 6

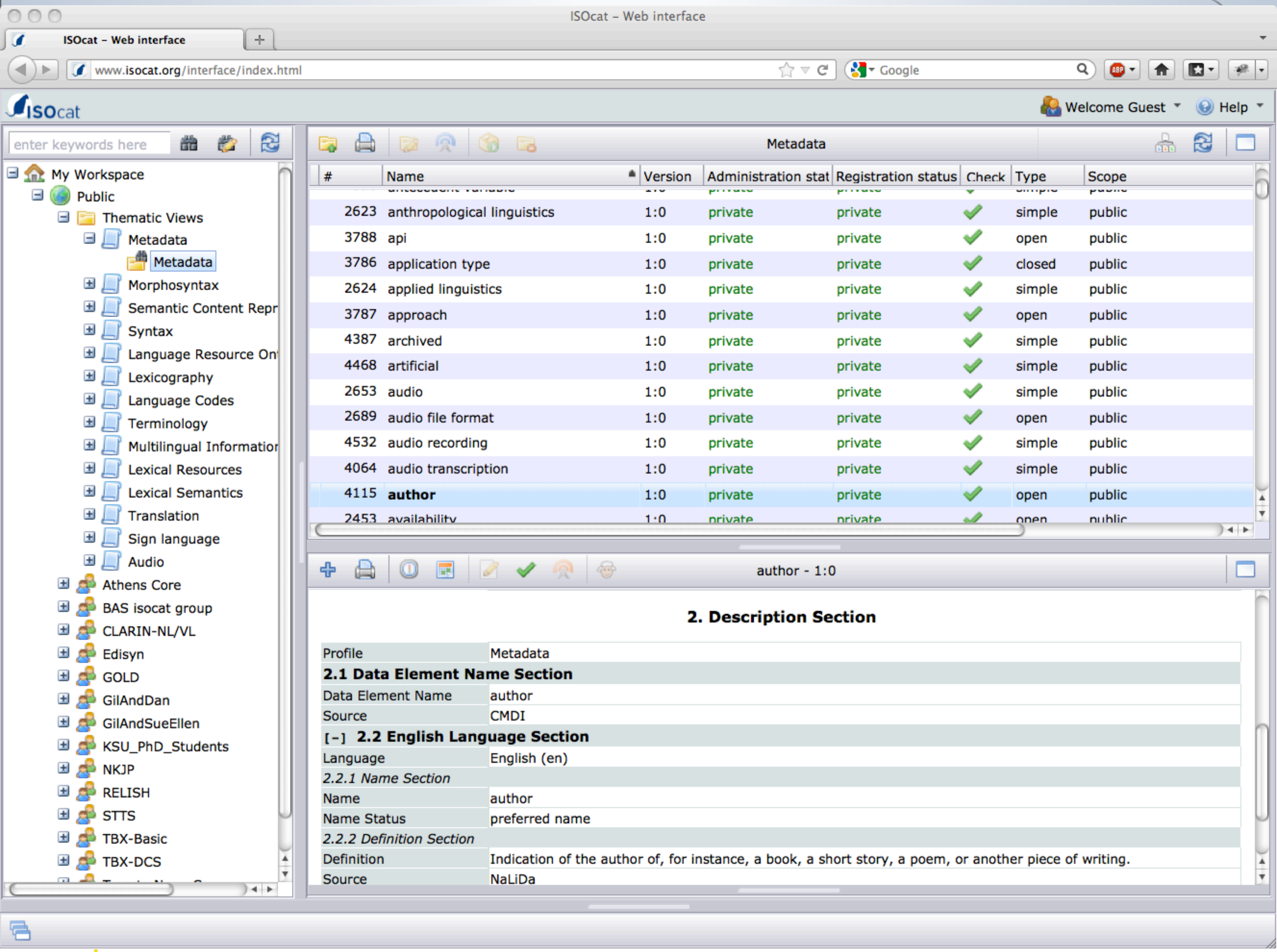
Central registry of data categories

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Central registry of data categories

- Standardising of terminology
- Avoiding *Tag Abuse* by common definitions
  - Unique references by identifiers
  - Partially controlled vocabulary
  - Semantic interoperability: Referring to the semantics of a category
- Candidates for resources: Data categories created by others could be relevant for your own metadata
- Sample registry: CLARIN Concept registry
  - Partly replaces isocat.org ISO 12620:2009



#	Name	Version	Administration stat	Registration status	Check	Type	Scope
2623	anthropological linguistics	1:0	private	private	✓	simple	public
3788	api	1:0	private	private	✓	open	public
3786	application type	1:0	private	private	✓	closed	public
2624	applied linguistics	1:0	private	private	✓	simple	public
3787	approach	1:0	private	private	✓	open	public
4387	archived	1:0	private	private	✓	simple	public
4468	artificial	1:0	private	private	✓	simple	public
2653	audio	1:0	private	private	✓	simple	public
2689	audio file format	1:0	private	private	✓	open	public
4532	audio recording	1:0	private	private	✓	simple	public
4064	audio transcription	1:0	private	private	✓	simple	public
4115	**author**	1:0	private	private	✓	open	public
2453	availability	1:0	private	private	✓	open	public

2. Description Section

Profile	Metadata
2.1 Data Element Name Section	
Data Element Name	author
Source	CMDI
[-] 2.2 English Language Section	
Language	English (en)
2.2.1 Name Section	
Name	author
Name Status	preferred name
2.2.2 Definition Section	
Definition	Indication of the author of, for instance, a book, a short story, a poem, or another piece of writing.
Source	NaLiDa

Drill

- Go to the CLARIN Concept
  - https://openskos.meertens.knaw.nl/ccr/browser/
- Select the metadata section
- Find 10 data categories unfamiliar to you and find out what they mean
- What kind of resource could make use of such a description?

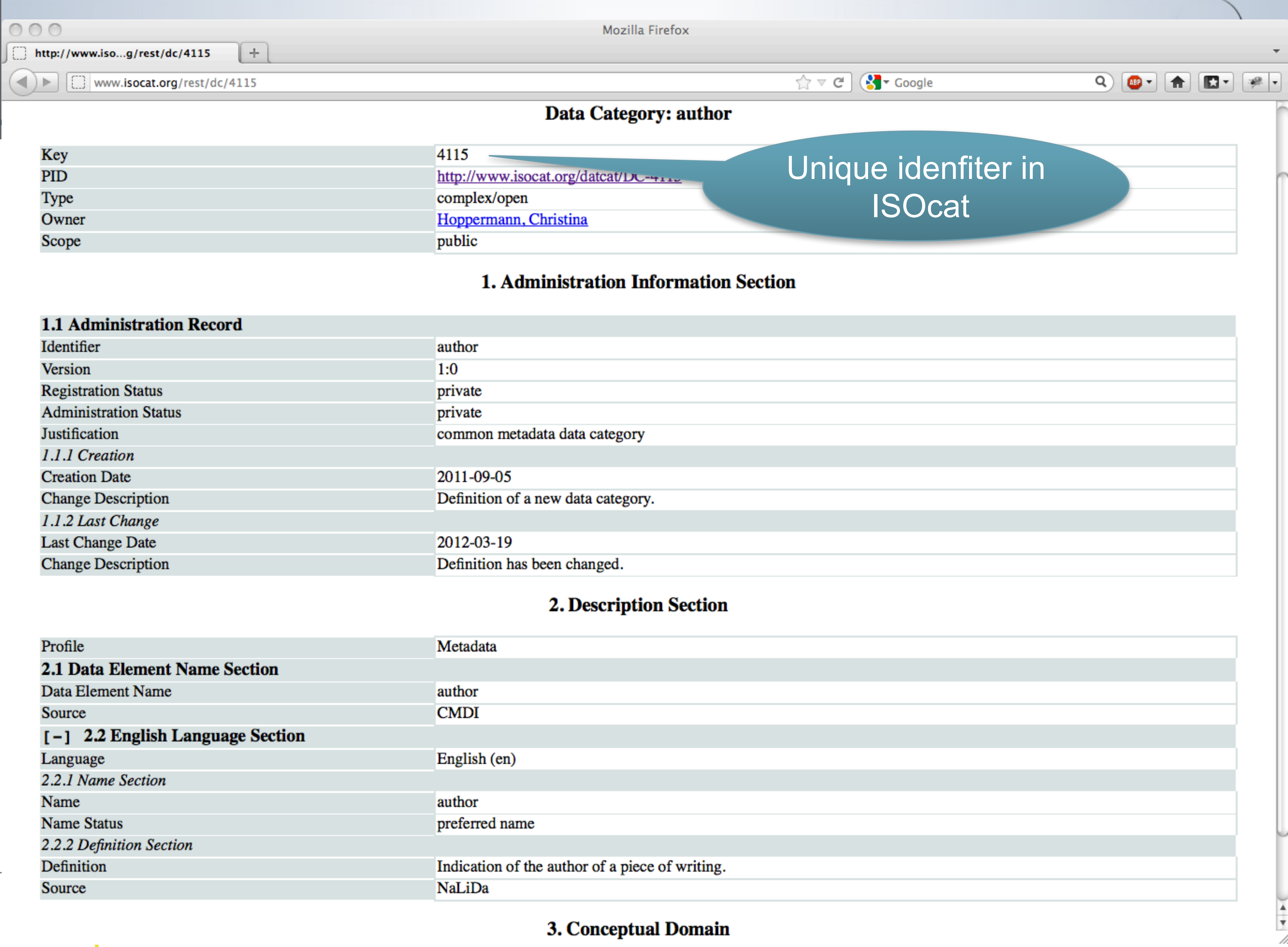
GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Übung

- Gehen Sie zum Datenkategorie Repository ISOcat
  - www.isocat.org
- Wählen Sie den Bereich Metadaten aus
- Finden Sie 10 Datenkategorien, die Ihnen nichts sagen und finden Sie heraus, was sich dahinter verbirgt
  - http://www.isocat.org/datcat/DC-2507 : Definition unter dem Link
  - http://www.isocat.org/datcat/DC-4507 ...

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung



Data Category: author

Key	4115
PID	http://www.isocat.org/datcat/DC-4115
Type	complex/open
Owner	Hoppermann, Christina
Scope	public

Unique idenfiter in ISOcat

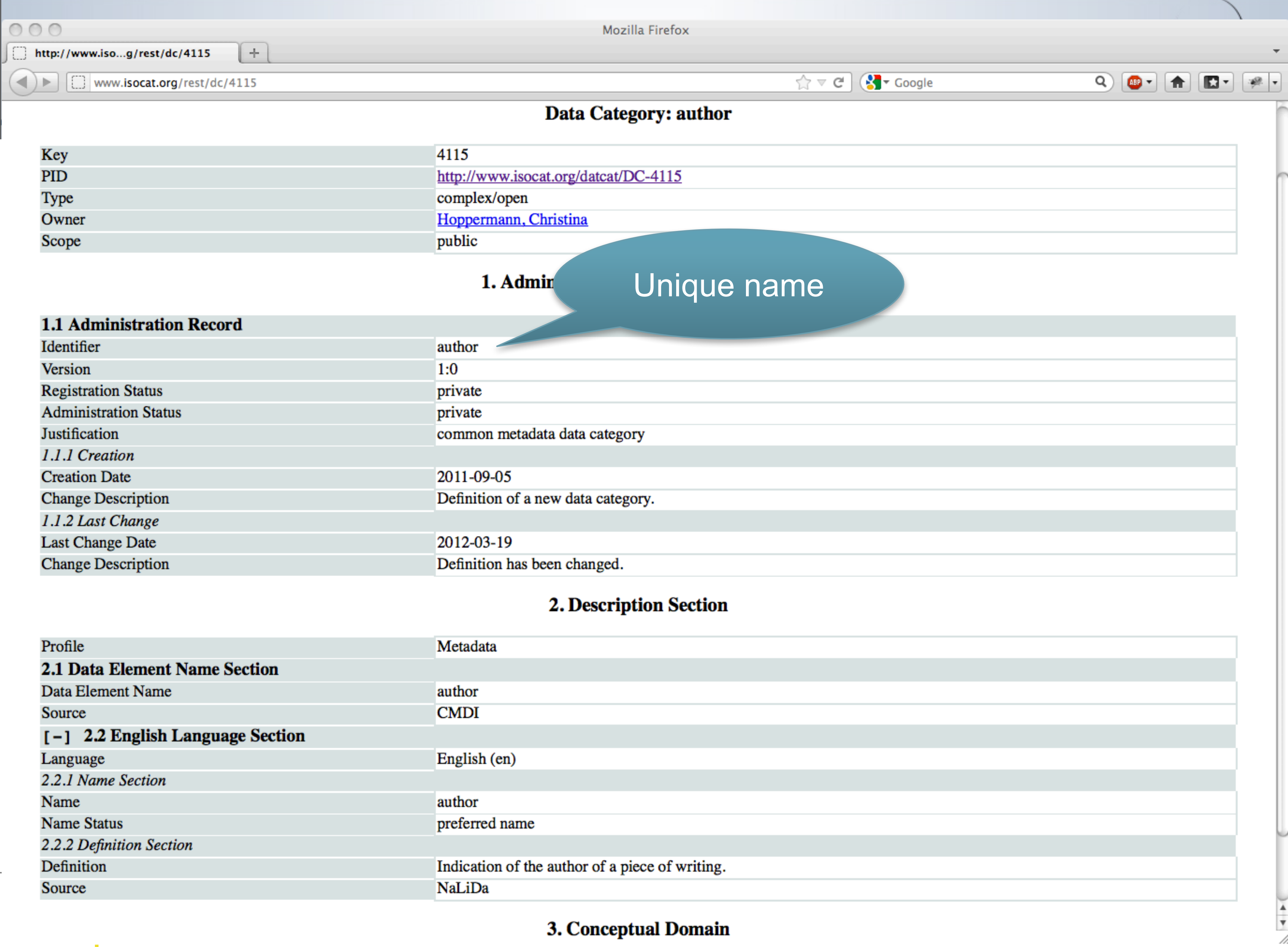
1. Administration Information Section

1.1 Administration Record	
Identifier	author
Version	1:0
Registration Status	private
Administration Status	private
Justification	common metadata data category
1.1.1 Creation	
Creation Date	2011-09-05
Change Description	Definition of a new data category.
1.1.2 Last Change	
Last Change Date	2012-03-19
Change Description	Definition has been changed.

2. Description Section

Profile	Metadata
2.1 Data Element Name Section	
Data Element Name	author
Source	CMDI
[–] 2.2 English Language Section	
Language	English (en)
2.2.1 Name Section	
Name	author
Name Status	preferred name
2.2.2 Definition Section	
Definition	Indication of the author of a piece of writing.
Source	NaLiDa

3. Conceptual Domain



Unique name

Data Category: author

Key	4115
PID	http://www.isocat.org/datcat/DC-4115
Type	complex/open
Owner	Hoppermann, Christina
Scope	public

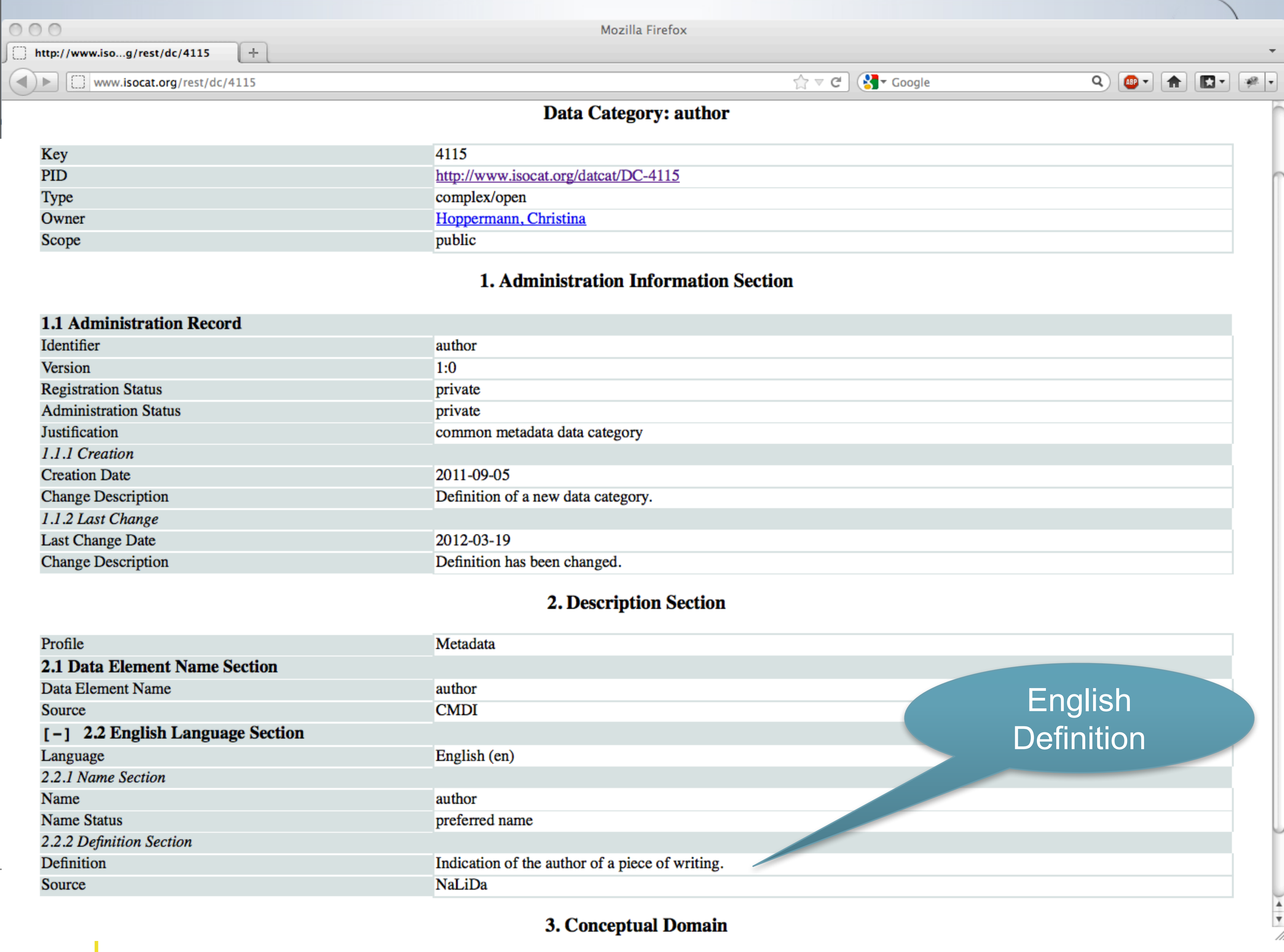
1. Admin

1.1 Administration Record	
Identifier	author
Version	1:0
Registration Status	private
Administration Status	private
Justification	common metadata data category
1.1.1 Creation	
Creation Date	2011-09-05
Change Description	Definition of a new data category.
1.1.2 Last Change	
Last Change Date	2012-03-19
Change Description	Definition has been changed.

2. Description Section

Profile	Metadata
2.1 Data Element Name Section	
Data Element Name	author
Source	CMDI
[–] 2.2 English Language Section	
Language	English (en)
2.2.1 Name Section	
Name	author
Name Status	preferred name
2.2.2 Definition Section	
Definition	Indication of the author of a piece of writing.
Source	NaLiDa

3. Conceptual Domain



Data Category: author

Key	4115
PID	http://www.isocat.org/datcat/DC-4115
Type	complex/open
Owner	Hoppermann, Christina
Scope	public

1. Administration Information Section

1.1 Administration Record	
Identifier	author
Version	1:0
Registration Status	private
Administration Status	private
Justification	common metadata data category
1.1.1 Creation	
Creation Date	2011-09-05
Change Description	Definition of a new data category.
1.1.2 Last Change	
Last Change Date	2012-03-19
Change Description	Definition has been changed.

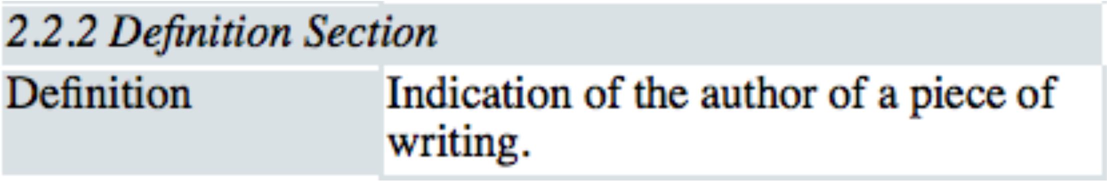
2. Description Section

Profile	Metadata
2.1 Data Element Name Section	
Data Element Name	author
Source	CMDI
[–] 2.2 English Language Section	
Language	English (en)
2.2.1 Name Section	
Name	author
Name Status	preferred name
2.2.2 Definition Section	
Definition	Indication of the author of a piece of writing.
Source	NaLiDa

3. Conceptual Domain

English Definition

The difference between an author and the DatCat "author"



2.2.2 Definition Section	
Definition	Indication of the author of a piece of writing.

- Definitions in CCR on metadata define the category
  - Which bits of information can you find in the field of a data category?
  - → ”Contains author"
  - Which value sets can you find there?
  - Where can you use this data category?
  - → Here: Example resource for which the data category is relevant
- Definitions **do not define** the semantics of a word
  - Here: they do not define what an author is

Definitions in the CCR

- Analytic definitions according to ISO 704
  - Consist of one fragment
  - Begin with the superordinate concept of the data category that is to be defined: direct superordinate term or above in a concept hierarchy
  - List essentially differentiating characteristics to related concepts
  - Do not contain a concept system
    - Concept-System and hierarchy are externally defined
    - Unification of definitions outside of the borders of a subject field
    - As general as possible

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

New DatCat or reusing an old one?

- Recycling of existing data categories if:
  - The concept in the CCR is the same as the needed one (the definition is sufficiently similar)
  - The definition is too narrow, but the creator or owner is willing to adjust the definition
  - The data category is from a different subject field but the meaning is obviously the desired one
  - The name of the data category is not the preferred variant
- Create a new Data Category if:
  - There is a different concepts/mental image/...
  - There is no similar term

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Drilling again?

- No, because
  - Persistency
  - Restricted creation
  - Only national content experts allowed

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 7

Generalizing groups of data categories

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Components

- Background
  - Use of similar data categories for different types of resources
  - Some data categories are typically used together
  - Selection of data categories is accelerated if they are grouped
- Groups of data categories: components
  - Cardinality: how often can or must a data category occur?
  - Value set: is there a closed vocabulary or a specific type of data?
- Components contain:
  - Data categories as elements
  - Other components
- Core model of the Component Metadata Infrastructure (CMDI, ISO 24622-1)
  - Co-developed by CLARIN
  - See: http://www.clarin.eu/cmdi

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Sample Components

- General Info Component:
  - Elements:
    - ResourceName
    - ResourceTitle
    - ResourceClass
    - Version
    - ...
  - Components:
    - Location
      - Address
      - Region
    - Descriptions
      - Description (in one language)

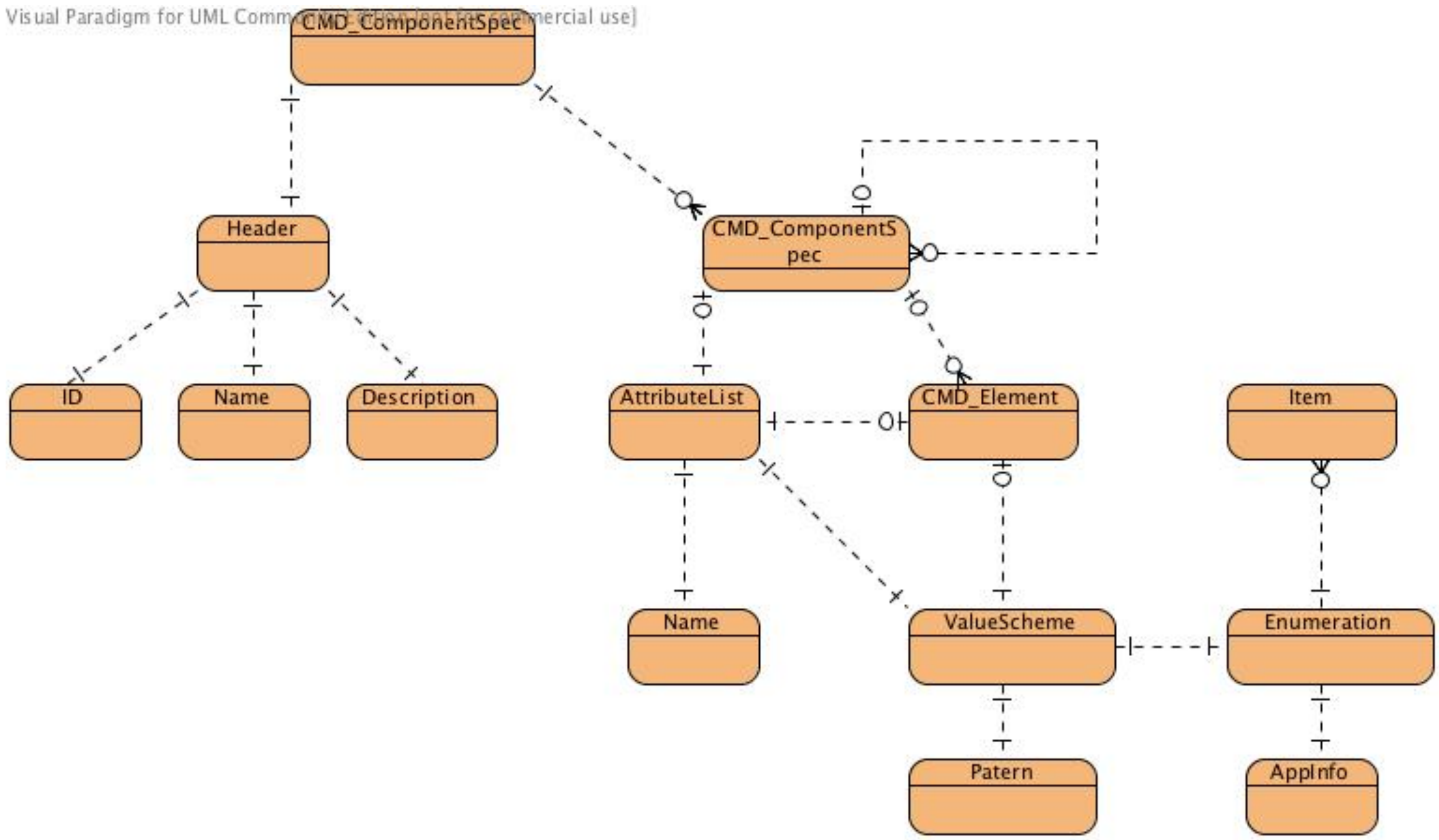
- Access Component:
  - Elements:
    - Availability
    - Distribution Medium
    - Catalogue Link
    - Price
  - Components:
    - DeploymentToolInfo
      - DeploymentTool
      - ToolType
      - ...
    - Contact
      - Person
      - Role
      - Address
      - ...

Representing components: CMDI Component Specification Language (CCSL)

- Description language for components
- Defined in XML
- Descrives:
  - Elements of a component
    - cardinality
    - Reference to concept
    - Controlled vocabulary/data type
    - Multilinguality
    - Element name
  - <CMD_Element Multilingual="true“ CardinalityMax="1" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2453" name="Availability"/>

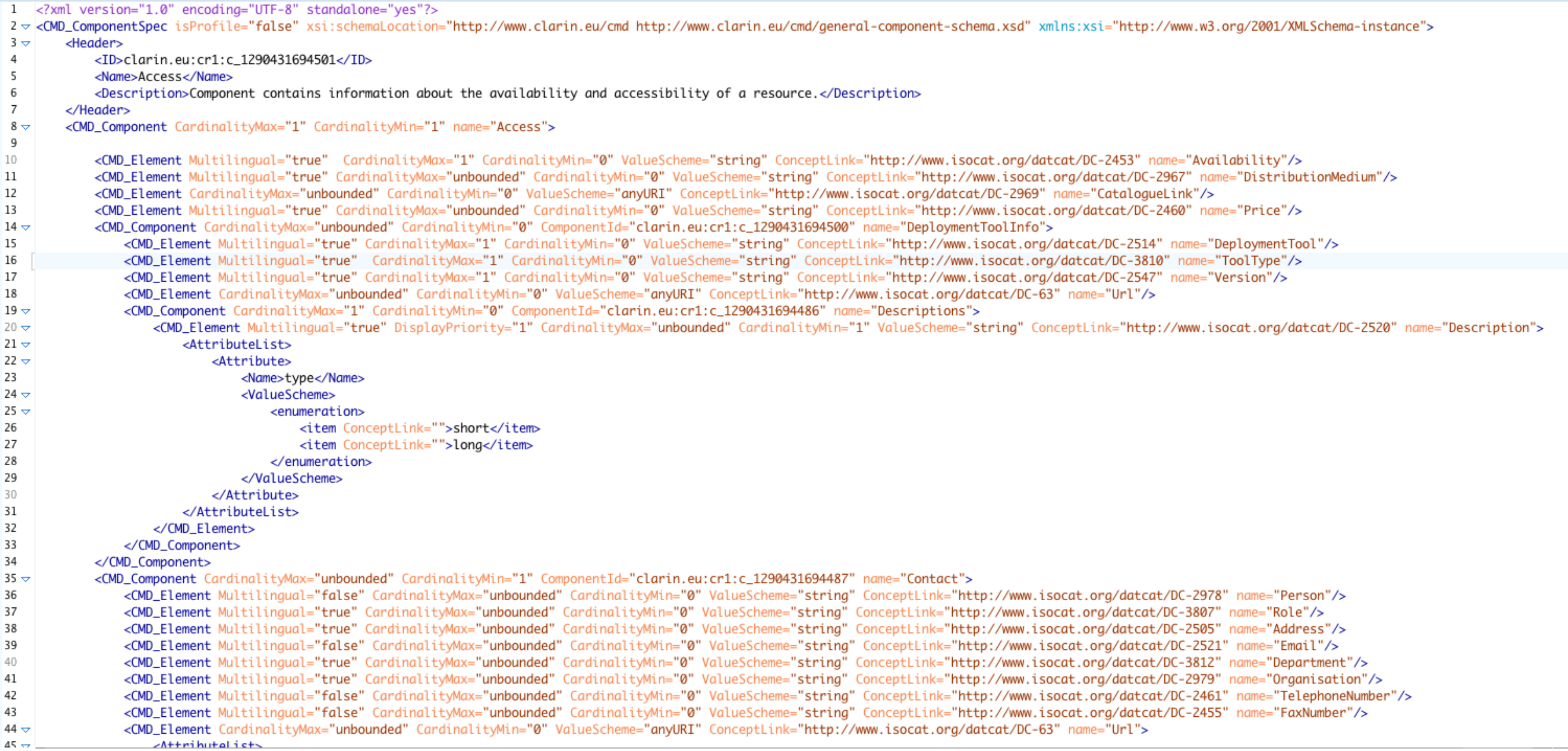
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

UML for CCSL



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Example CCSL in XML



```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<CMD_ComponentSpec isProfile="false" xsi:schemaLocation="http://www.clarin.eu/cmd http://www.clarin.eu/cmd/general-component-schema.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <Header>
        <ID>clarin.eu:cr1:c_1290431694501</ID>
        <Name>Access</Name>
        <Description>Component contains information about the availability and accessibility of a resource.</Description>
    </Header>
    <CMD_Component CardinalityMax="1" CardinalityMin="1" name="Access">

        <CMD_Element Multilingual="true"  CardinalityMax="1" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2453" name="Availability"/>
        <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2967" name="DistributionMedium"/>
        <CMD_Element CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="anyURI" ConceptLink="http://www.isocat.org/datcat/DC-2969" name="CatalogueLink"/>
        <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2460" name="Price"/>
        <CMD_Component CardinalityMax="unbounded" CardinalityMin="0" ComponentId="clarin.eu:cr1:c_1290431694500" name="DeploymentToolInfo">
            <CMD_Element Multilingual="true" CardinalityMax="1" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2514" name="DeploymentTool"/>
            <CMD_Element Multilingual="true"  CardinalityMax="1" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-3810" name="ToolType"/>
            <CMD_Element Multilingual="true" CardinalityMax="1" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2547" name="Version"/>
            <CMD_Element CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="anyURI" ConceptLink="http://www.isocat.org/datcat/DC-63" name="Url"/>
            <CMD_Component CardinalityMax="1" CardinalityMin="0" ComponentId="clarin.eu:cr1:c_1290431694486" name="Descriptions">
                <CMD_Element Multilingual="true" DisplayPriority="1" CardinalityMax="unbounded" CardinalityMin="1" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2520" name="Description">
                    <AttributeList>
                        <Attribute>
                            <Name>type</Name>
                            <ValueScheme>
                                <enumeration>
                                    <item ConceptLink="">short</item>
                                    <item ConceptLink="">long</item>
                                </enumeration>
                            </ValueScheme>
                        </Attribute>
                    </AttributeList>
                </CMD_Element>
            </CMD_Component>
        </CMD_Component>
        <CMD_Component CardinalityMax="unbounded" CardinalityMin="1" ComponentId="clarin.eu:cr1:c_1290431694487" name="Contact">
            <CMD_Element Multilingual="false" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2978" name="Person"/>
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-3807" name="Role"/>
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2505" name="Address"/>
            <CMD_Element Multilingual="false" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2521" name="Email"/>
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-3812" name="Department"/>
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2979" name="Organisation"/>
            <CMD_Element Multilingual="false" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2461" name="TelephoneNumber"/>
            <CMD_Element Multilingual="false" CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2455" name="FaxNumber"/>
            <CMD_Element CardinalityMax="unbounded" CardinalityMin="0" ValueScheme="anyURI" ConceptLink="http://www.isocat.org/datcat/DC-63" name="Url">
                <AttributeList>
```

Repetition

- What is a component
- What kind of information can you find in a component?
- How does the Concept Registry relate to components?
- Advanced issues:
  - What do you need to create components?

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Zur Wiederholung

- Was ist eine Komponente?
  - Eine Gruppierung von Datenkategorien
- Welche Informationen findet man in einer Komponente?
  - Welche Elemente und Komponenten Teil der Komponente sind
  - Häufigkeit
  - ISOcat-Referenz
  - Kontrolliertes Vokabular/Datentyp
  - Mehrsprachigkeit
  - Elementname
- Was hat ISOcat mit Komponenten zu tun?
  - Die Elementbeschreibungen kommen aus ISOcat
- Weiterführende Fragen:
  - Was brauchen Sie, um selbst eine Komponente zu erstellen?
    - Werkzeuge, die die Komplexität entfernen

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Kapitel 8

Generalising for types of resources

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Profiles

- Background
  - Each type of resource receives a metadata schema
  - Reuse of components
  - Different types of resources require different schemas
- One type of resource = one profile
  - Basis for generating an XSchema for validation
- Profile description contains
  - Data categories as elements
  - Additional components
- Core model of the Component Metadata Infrastructure (CMDI)
  - Developed together with the CLARIN
  - See: http://www.clarin.eu/cmdi

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Special profile: known metadata standards

- Dublin Core:
  - Dublin Core Metadata Initiative (DCMI)
  - Cataloguing of documents
  - Originally standardised DC: 15 core elements
  - qualified DC: further specification
- OLAC: Open Language Archives Community
  - Aiming at a library of language resources
  - Extending Dublin Core (DC)
  - Contains DC
  - More language related stuff
- IMDI: ISLE MetaData Initiative
  - Metadata for language resources
  - Especially spoken language
  - Very rich descriptions
  - Including speaker, roles, age, ….

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

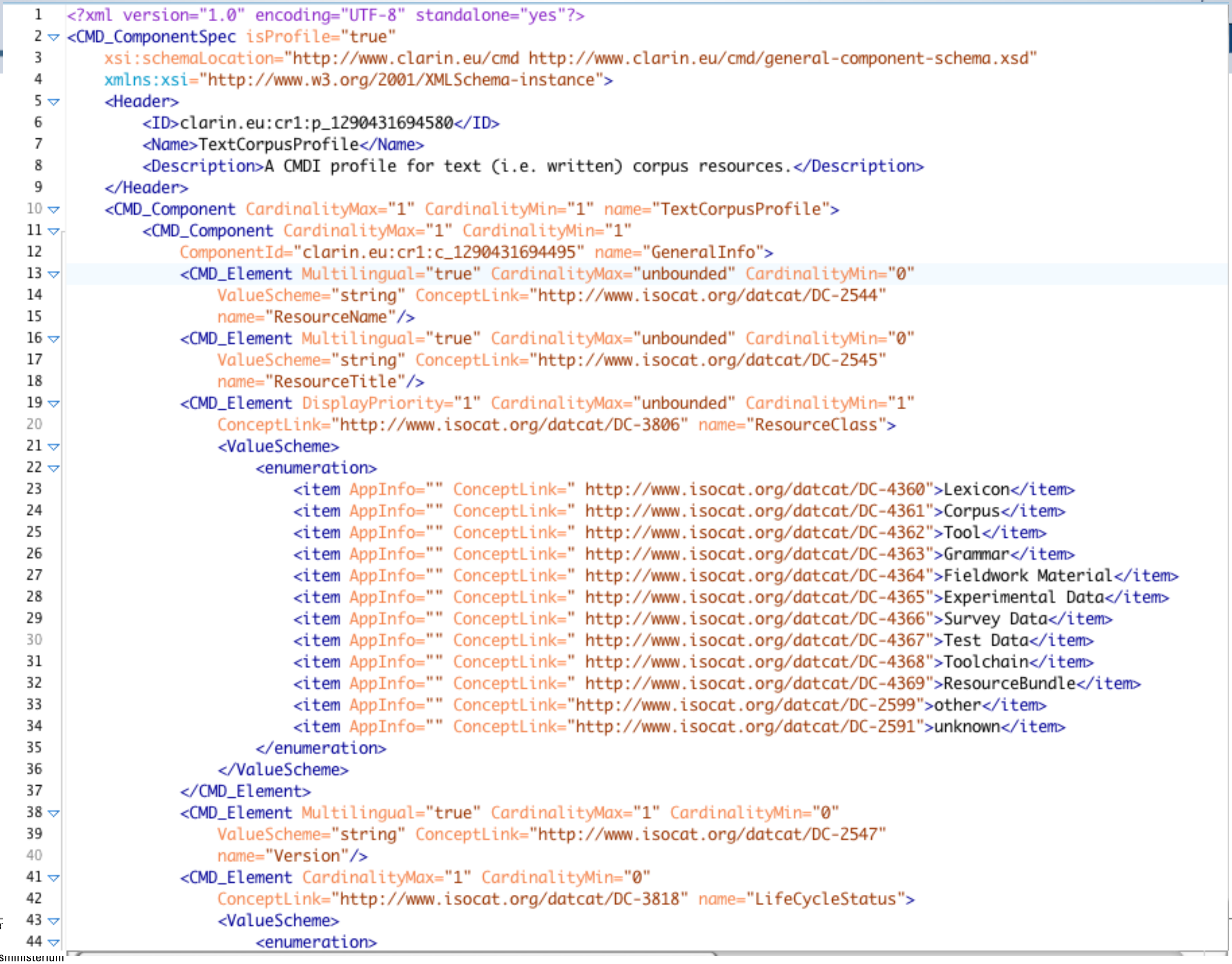
Sample profiles

- Experiments
  - Soziolinguistics
  - Psycholinguistics
- Lexical resourcen
- Diachrone textccorpora
- Corpora of spoken language
- Corpora of written language
- Software Tools
- Webservices
- Webservice-Toolchain

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Specification by CCSL

- Profiles defined as components
- Specified in CCSL: component type profile



```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<CMD_ComponentSpec isProfile="true"
    xsi:schemaLocation="http://www.clarin.eu/cmd http://www.clarin.eu/cmd/general-component-schema.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <Header>
        <ID>clarin.eu:cr1:p_1290431694580</ID>
        <Name>TextCorpusProfile</Name>
        <Description>A CMDI profile for text (i.e. written) corpus resources.</Description>
    </Header>
    <CMD_Component CardinalityMax="1" CardinalityMin="1" name="TextCorpusProfile">
        <CMD_Component CardinalityMax="1" CardinalityMin="1"
            ComponentId="clarin.eu:cr1:c_1290431694495" name="GeneralInfo">
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0"
                ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2544"
                name="ResourceName"/>
            <CMD_Element Multilingual="true" CardinalityMax="unbounded" CardinalityMin="0"
                ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2545"
                name="ResourceTitle"/>
            <CMD_Element DisplayPriority="1" CardinalityMax="unbounded" CardinalityMin="1"
                ConceptLink="http://www.isocat.org/datcat/DC-3806" name="ResourceClass">
                <ValueScheme>
                    <enumeration>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4360">Lexicon</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4361">Corpus</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4362">Tool</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4363">Grammar</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4364">Fieldwork Material</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4365">Experimental Data</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4366">Survey Data</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4367">Test Data</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4368">Toolchain</item>
                        <item AppInfo="" ConceptLink=" http://www.isocat.org/datcat/DC-4369">ResourceBundle</item>
                        <item AppInfo="" ConceptLink="http://www.isocat.org/datcat/DC-2599">other</item>
                        <item AppInfo="" ConceptLink="http://www.isocat.org/datcat/DC-2591">unknown</item>
                    </enumeration>
                </ValueScheme>
            </CMD_Element>
            <CMD_Element Multilingual="true" CardinalityMax="1" CardinalityMin="0"
                ValueScheme="string" ConceptLink="http://www.isocat.org/datcat/DC-2547"
                name="Version"/>
            <CMD_Element CardinalityMax="1" CardinalityMin="0"
                ConceptLink="http://www.isocat.org/datcat/DC-3818" name="LifeCycleStatus">
                <ValueScheme>
                    <enumeration>
```

Repetition

- What is a profile?
- What is in a profile?
- What is the connection between profiles and components?
- Advanced question:
  - What do you need to create a profile?

Zur Wiederholung

- Was ist ein Profil?
  - Eine besondere Komponente für einen bestimmten Typ von Ressource
- Welche Informationen findet man in einem Profil?
  - Alles, was man auch in einer Komponente findet plus die Information, dass es sich um ein Profil handelt
- Was ist der Zusammenhang von Profil und Komponente?
  - Profile sind Komponenten und enthalten in der Regel Komponenten und Elemente
- Weiterführende Fragen:
  - Was brauchen Sie, um selbst ein Profil zu erstellen?
    - Immer noch Werkzeuge mit geringerer Komplexität

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 9

CMDI Component Registry

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Erläuterung für Lehrende

- Die Component Registry ist eine Webapplikation, mit der man sich vor dem Kurs vertraut machen sollte.
- Zugang: http://catalog.clarin.eu/ds/ComponentRegistry/#
- Zum Erstellen von Profilen und Komponenten ist ein CLARIN-Webseiten-Account erforderlich. Wenn Sie nicht Partner von CLARIN sind, beachten Sie bitte die Bedingungen, unter denen Sie Zugang erhalten, siehe http://www.clarin.eu/user
- Die Erfahrung zeigt, dass schwer vorherzusagen ist, wie lange Nutzer brauchen, um sich mit neuen Applikationen vertraut zu machen. Möglicherweise ist es notwendig, dass Sie diese Schritte vorführen.

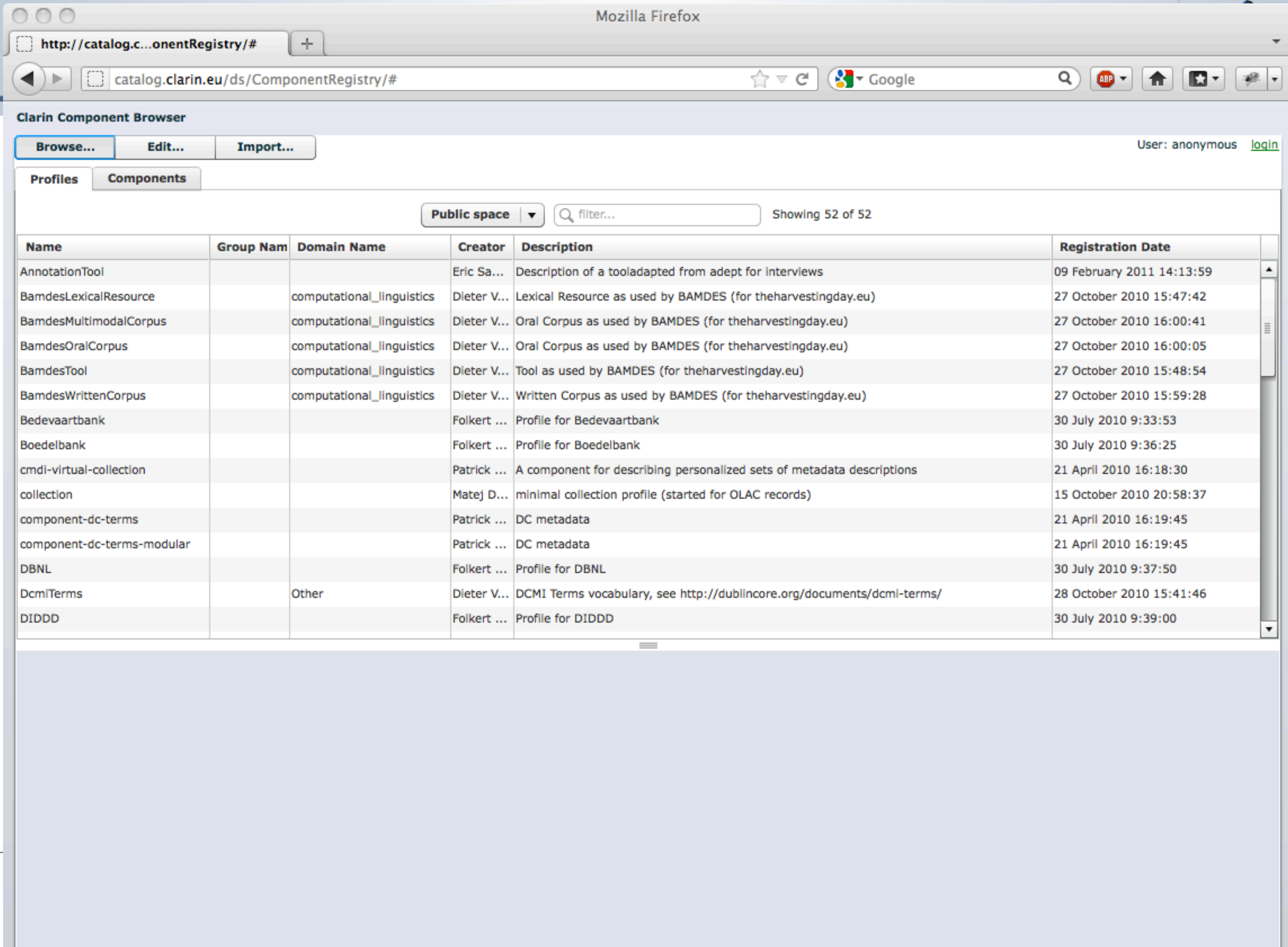
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Motivation for the Component Registry

- High complexity of component model
- Tools not part of the models
  - Creation of components
  - Connecting components with components
  - Include data categories
- Reuse of components requires central registry
  - Unique reference
  - Easily accessible

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

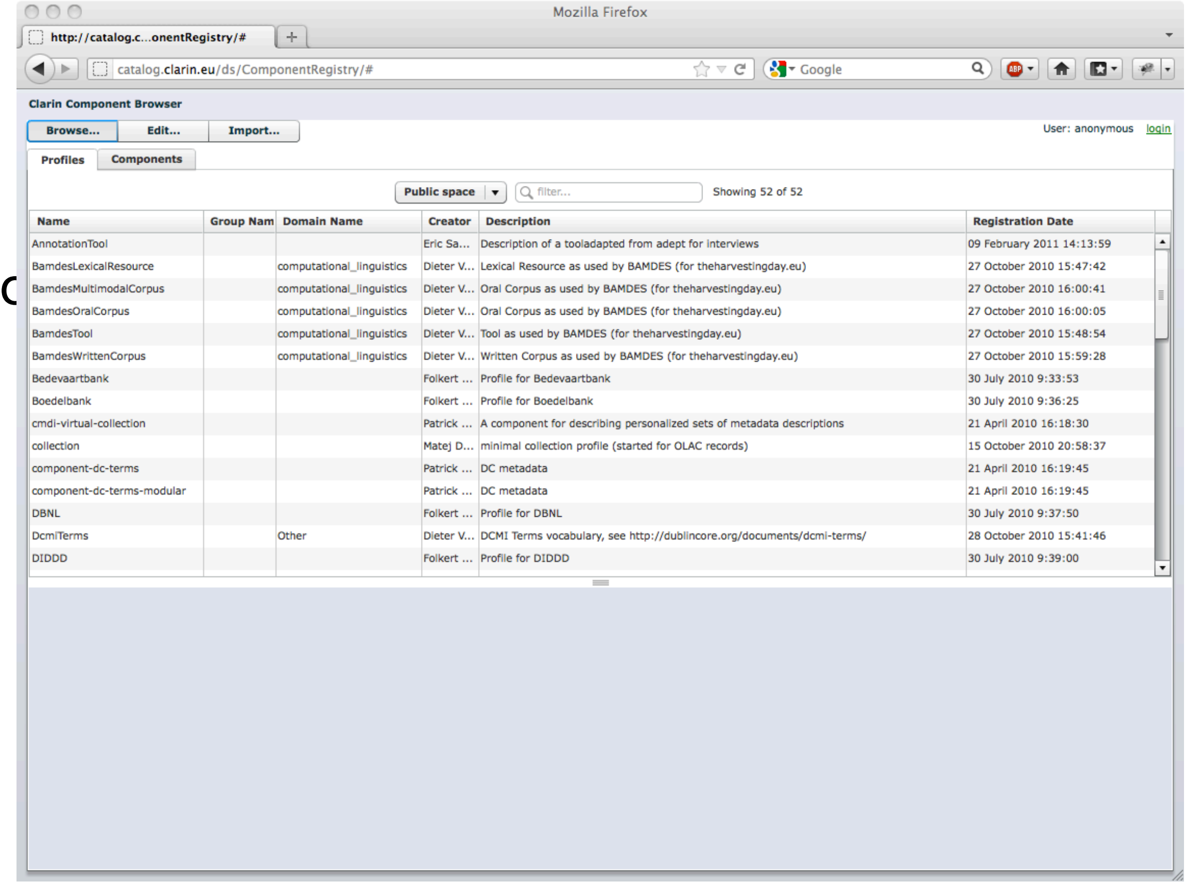


Showing 52 of 52

Name	Group Nam	Domain Name	Creator	Description	Registration Date
AnnotationTool			Eric Sa...	Description of a tooladapted from adept for interviews	09 February 2011 14:13:59
BamdesLexicalResource		computational_linguistics	Dieter V...	Lexical Resource as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:47:42
BamdesMultimodalCorpus		computational_linguistics	Dieter V...	Oral Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 16:00:41
BamdesOralCorpus		computational_linguistics	Dieter V...	Oral Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 16:00:05
BamdesTool		computational_linguistics	Dieter V...	Tool as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:48:54
BamdesWrittenCorpus		computational_linguistics	Dieter V...	Written Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:59:28
Bedevaartbank			Folkert ...	Profile for Bedevaartbank	30 July 2010 9:33:53
Boedelbank			Folkert ...	Profile for Boedelbank	30 July 2010 9:36:25
cmdi-virtual-collection			Patrick ...	A component for describing personalized sets of metadata descriptions	21 April 2010 16:18:30
collection			Matej D...	minimal collection profile (started for OLAC records)	15 October 2010 20:58:37
component-dc-terms			Patrick ...	DC metadata	21 April 2010 16:19:45
component-dc-terms-modular			Patrick ...	DC metadata	21 April 2010 16:19:45
DBNL			Folkert ...	Profile for DBNL	30 July 2010 9:37:50
DcmiTerms		Other	Dieter V...	DCMI Terms vocabulary, see http://dublincore.org/documents/dcmi-terms/	28 October 2010 15:41:46
DIDDD			Folkert ...	Profile for DIDDD	30 July 2010 9:39:00

Clarin Component Browser: Overview of existing public components and profiles

- Profile name
- Description
- Date
- Creator
- Group
- Profile-URL→ Show Info

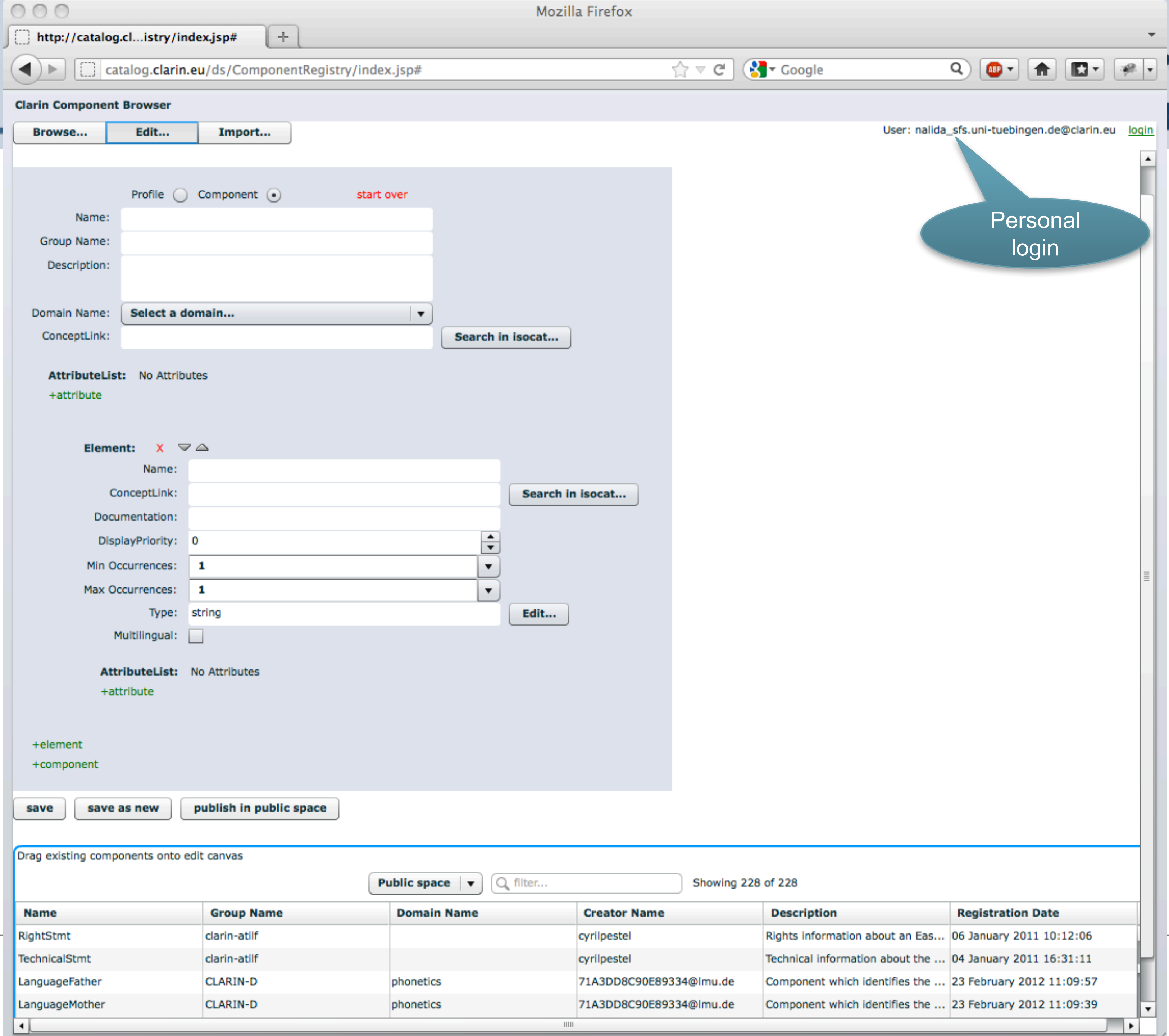


Mozilla Firefox
http://catalog.c...onentRegistry/#
catalog.clarin.eu/ds/ComponentRegistry/#
Google

Clarin Component Browser

Showing 52 of 52

Name	Group Nam	Domain Name	Creator	Description	Registration Date
AnnotationTool			Eric Sa...	Description of a tooladapted from adept for interviews	09 February 2011 14:13:59
BamdesLexicalResource		computational_linguistics	Dieter V...	Lexical Resource as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:47:42
BamdesMultimodalCorpus		computational_linguistics	Dieter V...	Oral Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 16:00:41
BamdesOralCorpus		computational_linguistics	Dieter V...	Oral Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 16:00:05
BamdesTool		computational_linguistics	Dieter V...	Tool as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:48:54
BamdesWrittenCorpus		computational_linguistics	Dieter V...	Written Corpus as used by BAMDES (for theharvestingday.eu)	27 October 2010 15:59:28
Bedevaartbank			Folkert ...	Profile for Bedevaartbank	30 July 2010 9:33:53
Boedelbank			Folkert ...	Profile for Boedelbank	30 July 2010 9:36:25
cmdi-virtual-collection			Patrick ...	A component for describing personalized sets of metadata descriptions	21 April 2010 16:18:30
collection			Matej D...	minimal collection profile (started for OLAC records)	15 October 2010 20:58:37
component-dc-terms			Patrick ...	DC metadata	21 April 2010 16:19:45
component-dc-terms-modular			Patrick ...	DC metadata	21 April 2010 16:19:45
DBNL			Folkert ...	Profile for DBNL	30 July 2010 9:37:50
DcmiTerms		Other	Dieter V...	DCMI Terms vocabulary, see http://dublincore.org/documents/dcmi-terms/	28 October 2010 15:41:46
DIDDD			Folkert ...	Profile for DIDDD	30 July 2010 9:39:00



Mozilla Firefox
http://catalog.cl...istry/index.jsp#
catalog.clarin.eu/ds/ComponentRegistry/index.jsp#
Google

Clarin Component Browser

Browse...
Edit...
Import...

User: nalida_sfs.uni-tuebingen.de@clarin.eu login

Personal login

Profile ○ Component ● start over

Name:

Group Name:

Description:

Domain Name: Select a domain...

ConceptLink: Search in isocat...

AttributeList: No Attributes

+attribute

Element: X

Name:

ConceptLink: Search in isocat...

Documentation:

DisplayPriority: 0

Min Occurrences: 1

Max Occurrences: 1

Type: string Edit...

Multilingual: □

AttributeList: No Attributes

+attribute

+element

+component

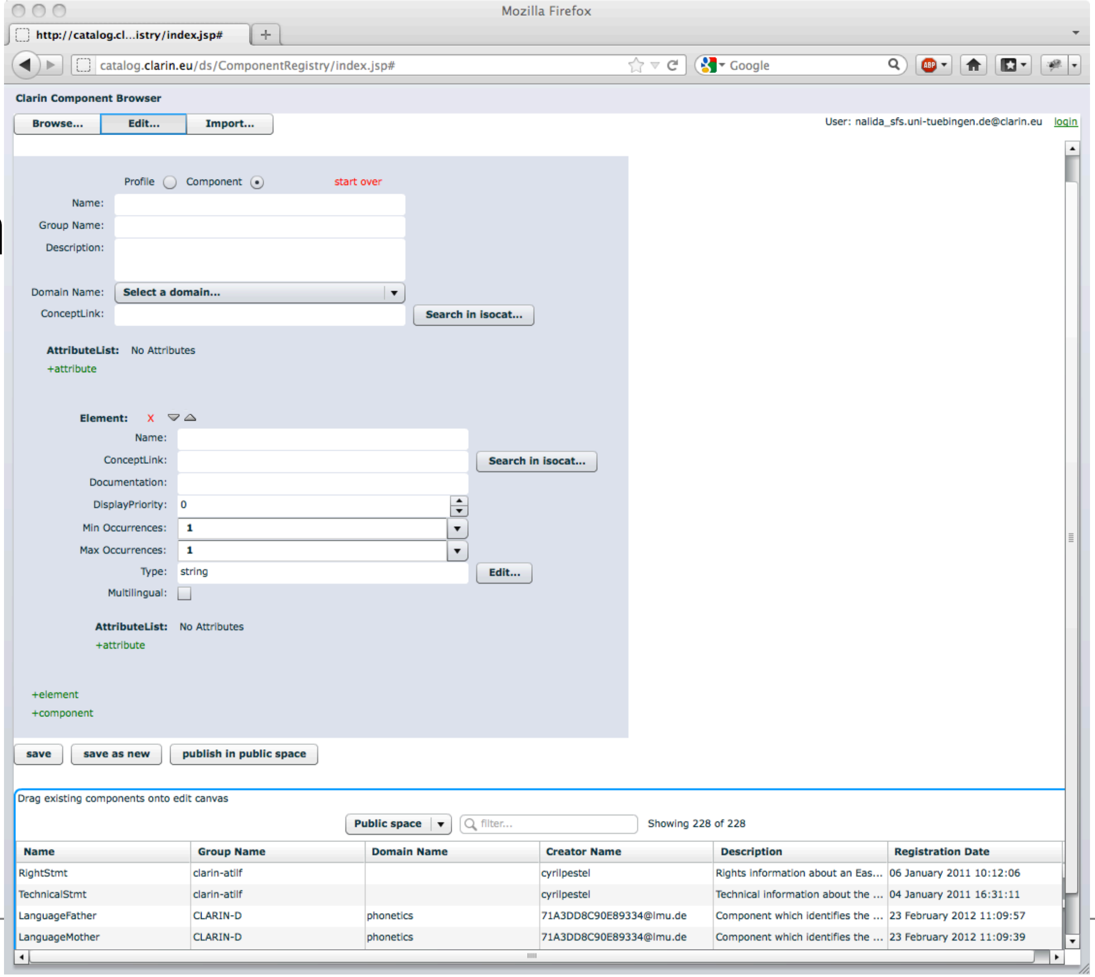
Drag existing components onto edit canvas

Public space filter... Showing 228 of 228

Name	Group Name	Domain Name	Creator Name	Description	Registration Date
RightStmt	clarin-atilf		cyrilpestel	Rights information about an Eas...	06 January 2011 10:12:06
TechnicalStmt	clarin-atilf		cyrilpestel	Technical information about the ...	04 January 2011 16:31:11
LanguageFather	CLARIN-D	phonetics	71A3DD8C90E89334@lmu.de	Component which identifies the ...	23 February 2012 11:09:57
LanguageMother	CLARIN-D	phonetics	71A3DD8C90E89334@lmu.de	Component which identifies the ...	23 February 2012 11:09:39

Creating new profiles and components

- Elements referring to Concept Registry
- Selecting compone
- Cardinality
- Data types for elem



Clarin Component Browser

User: nalida_sfs.uni-tuebingen.de@clarin.eu

Profile ○ Component ●

Name:

Group Name:

Description:

Domain Name: Select a domain...

ConceptLink:

AttributeList: No Attributes

Element:

Name:

ConceptLink:

Documentation:

DisplayPriority: 0

Min Occurrences: 1

Max Occurrences: 1

Type: string

Multilingual: □

AttributeList: No Attributes

Drag existing components onto edit canvas

Showing 228 of 228

Name	Group Name	Domain Name	Creator Name	Description	Registration Date
RightStmt	clarin-atilf		cyrilpestel	Rights information about an Eas...	06 January 2011 10:12:06
TechnicalStmt	clarin-atilf		cyrilpestel	Technical information about the ...	04 January 2011 16:31:11
LanguageFather	CLARIN-D	phonetics	71A3DD8C90E89334@lmu.de	Component which identifies the ...	23 February 2012 11:09:57
LanguageMother	CLARIN-D	phonetics	71A3DD8C90E89334@lmu.de	Component which identifies the ...	23 February 2012 11:09:39

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Practical

- Go to http://catalog.clarin.eu/ds/ComponentRegistry/
  - Click "login"
  - Login using Shibboleth
- Explore a couple of profiles
- Click edit
- Create some profile and components; do NOT publish them
- Tutorial http://www.sfs.uni-tuebingen.de/nalida/en/docu/tutorials/component-registry.html

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

New or used: Reusing components

- Advantage reuse:
  - Quickly create components and profiles
  - Components are “better” testet and probably more complete
  - Reuse of tools
    - XPaths with a couple of wildcards
    - No need for individual applications
- Advantage new creation
  - More of a control
- Recommendation: Reuse whenever possible or create new ones on the basis of existing ones

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

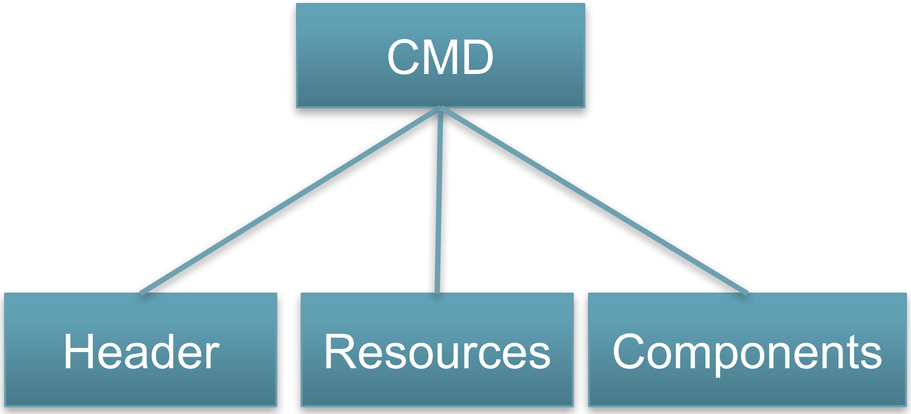
Chapter 10

Creating CMDI instances

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Structure of a CMDI instance

- Three sections
  - Header: Description of the metadata
  - Resources: reference to the resources
- Components: according to the type of resource



GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CMDI Header

- Information:
  - Name of metadata creator: MdCreator
  - Date of metadata creation: MdCreationDate
  - URL of metadata file: MdSelfLink
  - Identifier of the CMDI Profile: MdProfile
  - Name of the archive: MdCollectionDisplayName
- Creation:
  - Archiving process
  - URL of the metadaten file persistent
    - Name of the creator as known by the (Login)
    - CMDI profile known by selection
    - Name of Archive known by archive
    - Creation without archiving process: manual or by defaults

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CMDI Resources

- Information
  - ResourceProxyList: List of PIDs of primary data JournalFileProxyList: List of files with provenance metadata
  - ResourceRelationList: Relation to other resources
  - IsPartOfList: if the resource is part of a larger unit
- Creation:
  - Archiving process:
    - ResourceProxyList: PIDs created during archiving
  - Other lists as needed

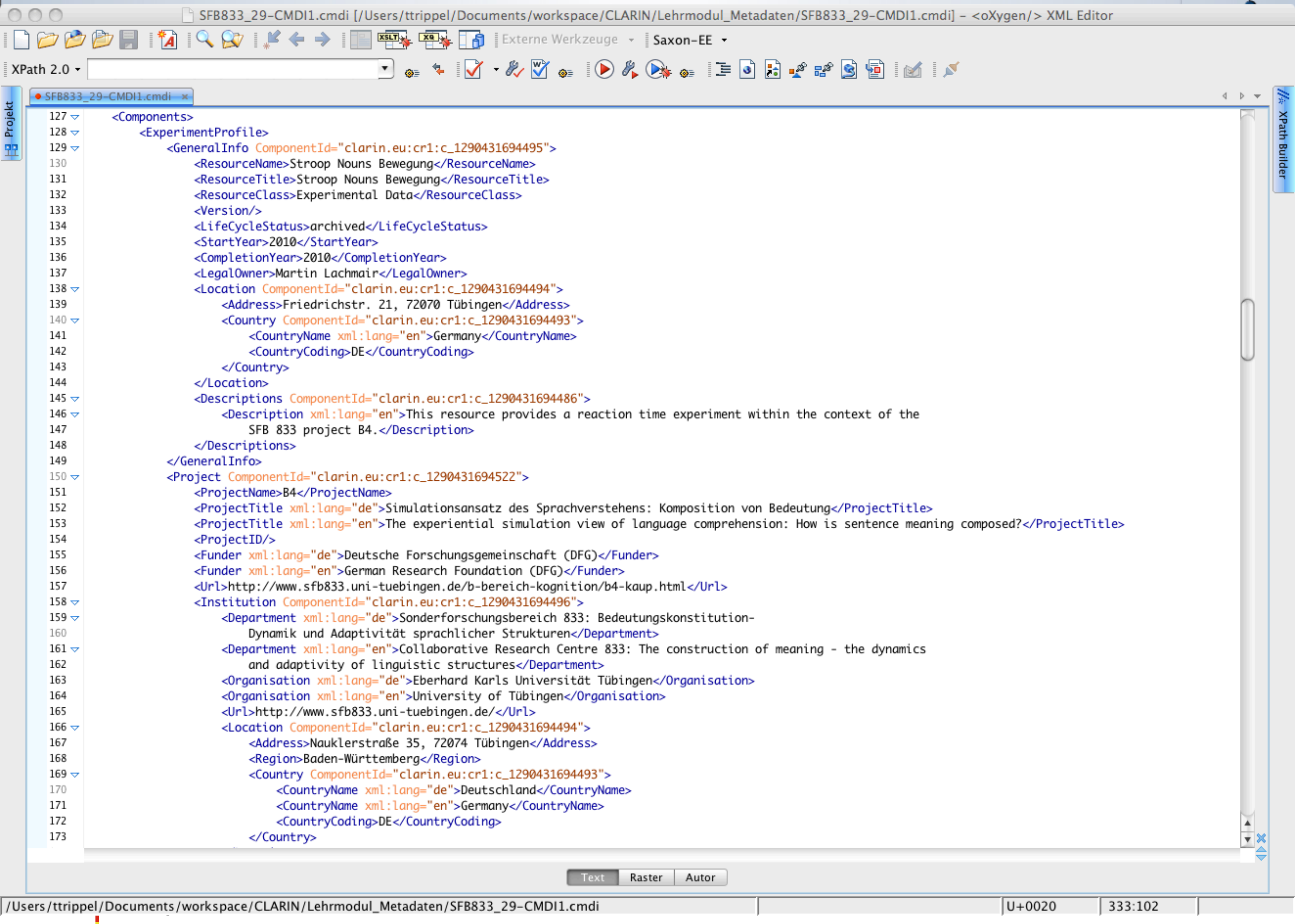
GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

CMDI Components

- Information:
  - Components
  - Data categories from Concept Registry
- Creation:
  - In an editor
    - XML editor
    - Text editor
    - Metadata editor
  - By transformation
    - From metadata in other formats
    - From relational databases

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

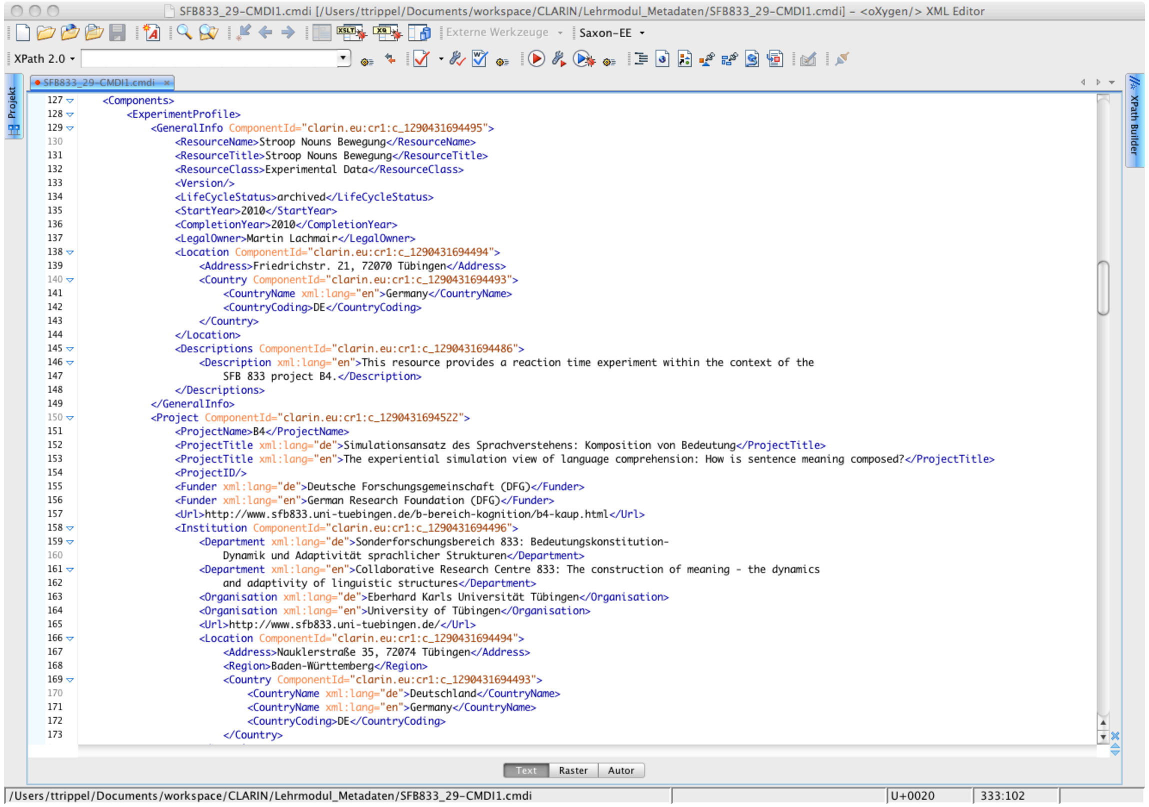


SFB833_29-CMDI1.cmdi [/Users/ttrippel/Documents/workspace/CLARIN/Lehrmodul_Metadaten/SFB833_29-CMDI1.cmdi] - <oXygen/> XML Editor
Externe Werkzeuge
Saxon-EE
XPath 2.0
SFB833_29-CMDI1.cmdi
Projekt
XPath Builder
Text
Raster
Autor
/Users/ttrippel/Documents/workspace/CLARIN/Lehrmodul_Metadaten/SFB833_29-CMDI1.cmdi
U+0020
333:102

```
<Components>
    <ExperimentProfile>
        <GeneralInfo ComponentId="clarin.eu:cr1:c_1290431694495">
            <ResourceName>Stroop Nouns Bewegung</ResourceName>
            <ResourceTitle>Stroop Nouns Bewegung</ResourceTitle>
            <ResourceClass>Experimental Data</ResourceClass>
            <Version/>
            <LifeCycleStatus>archived</LifeCycleStatus>
            <StartYear>2010</StartYear>
            <CompletionYear>2010</CompletionYear>
            <LegalOwner>Martin Lachmair</LegalOwner>
            <Location ComponentId="clarin.eu:cr1:c_1290431694494">
                <Address>Friedrichstr. 21, 72070 Tübingen</Address>
                <Country ComponentId="clarin.eu:cr1:c_1290431694493">
                    <CountryName xml:lang="en">Germany</CountryName>
                    <CountryCoding>DE</CountryCoding>
                </Country>
            </Location>
            <Descriptions ComponentId="clarin.eu:cr1:c_1290431694486">
                <Description xml:lang="en">This resource provides a reaction time experiment within the context of the
                    SFB 833 project B4.</Description>
            </Descriptions>
        </GeneralInfo>
        <Project ComponentId="clarin.eu:cr1:c_1290431694522">
            <ProjectName>B4</ProjectName>
            <ProjectTitle xml:lang="de">Simulationsansatz des Sprachverstehens: Komposition von Bedeutung</ProjectTitle>
            <ProjectTitle xml:lang="en">The experiential simulation view of language comprehension: How is sentence meaning composed?</ProjectTitle>
            <ProjectID/>
            <Funder xml:lang="de">Deutsche Forschungsgemeinschaft (DFG)</Funder>
            <Funder xml:lang="en">German Research Foundation (DFG)</Funder>
            <Url>http://www.sfb833.uni-tuebingen.de/b-bereich-kognition/b4-kaup.html</Url>
            <Institution ComponentId="clarin.eu:cr1:c_1290431694496">
                <Department xml:lang="de">Sonderforschungsbereich 833: Bedeutungskonstitution-
                    Dynamik und Adaptivität sprachlicher Strukturen</Department>
                <Department xml:lang="en">Collaborative Research Centre 833: The construction of meaning - the dynamics
                    and adaptivity of linguistic structures</Department>
                <Organisation xml:lang="de">Eberhard Karls Universität Tübingen</Organisation>
                <Organisation xml:lang="en">University of Tübingen</Organisation>
                <Url>http://www.sfb833.uni-tuebingen.de/</Url>
                <Location ComponentId="clarin.eu:cr1:c_1290431694494">
                    <Address>Nauklerstraße 35, 72074 Tübingen</Address>
                    <Region>Baden-Württemberg</Region>
                    <Country ComponentId="clarin.eu:cr1:c_1290431694493">
                        <CountryName xml:lang="de">Deutschland</CountryName>
                        <CountryName xml:lang="en">Germany</CountryName>
                        <CountryCoding>DE</CountryCoding>
                    </Country>
```

Creation of a CMDI instance in an XML-Editor

- Edit in XML code
  - Syntax highlighting
  - automatic validation
- disadvantage:
  - complexity
  - Learning curve of archivist



SFB833_29-CMDI1.cmdi [/Users/ttrippel/Documents/workspace/CLARIN/Lehrmodul_Metadaten/SFB833_29-CMDI1.cmdi] – <oXygen/> XML Editor
Externe Werkzeuge
Saxon-EE
XPath 2.0
SFB833_29-CMDI1.cmdi
Projekt
XPath Builder
Text
Raster
Autor
/Users/ttrippel/Documents/workspace/CLARIN/Lehrmodul_Metadaten/SFB833_29-CMDI1.cmdi
U+0020
333:102

```xml
    <Components>
        <ExperimentProfile>
            <GeneralInfo ComponentId="clarin.eu:cr1:c_1290431694495">
                <ResourceName>Stroop Nouns Bewegung</ResourceName>
                <ResourceTitle>Stroop Nouns Bewegung</ResourceTitle>
                <ResourceClass>Experimental Data</ResourceClass>
                <Version/>
                <LifeCycleStatus>archived</LifeCycleStatus>
                <StartYear>2010</StartYear>
                <CompletionYear>2010</CompletionYear>
                <LegalOwner>Martin Lachmair</LegalOwner>
                <Location ComponentId="clarin.eu:cr1:c_1290431694494">
                    <Address>Friedrichstr. 21, 72070 Tübingen</Address>
                    <Country ComponentId="clarin.eu:cr1:c_1290431694493">
                        <CountryName xml:lang="en">Germany</CountryName>
                        <CountryCoding>DE</CountryCoding>
                    </Country>
                </Location>
                <Descriptions ComponentId="clarin.eu:cr1:c_1290431694486">
                    <Description xml:lang="en">This resource provides a reaction time experiment within the context of the
                        SFB 833 project B4.</Description>
                </Descriptions>
            </GeneralInfo>
            <Project ComponentId="clarin.eu:cr1:c_1290431694522">
                <ProjectName>B4</ProjectName>
                <ProjectTitle xml:lang="de">Simulationsansatz des Sprachverstehens: Komposition von Bedeutung</ProjectTitle>
                <ProjectTitle xml:lang="en">The experiential simulation view of language comprehension: How is sentence meaning composed?</ProjectTitle>
                <ProjectID/>
                <Funder xml:lang="de">Deutsche Forschungsgemeinschaft (DFG)</Funder>
                <Funder xml:lang="en">German Research Foundation (DFG)</Funder>
                <Url>http://www.sfb833.uni-tuebingen.de/b-bereich-kognition/b4-kaup.html</Url>
                <Institution ComponentId="clarin.eu:cr1:c_1290431694496">
                    <Department xml:lang="de">Sonderforschungsbereich 833: Bedeutungskonstitution-
                        Dynamik und Adaptivität sprachlicher Strukturen</Department>
                    <Department xml:lang="en">Collaborative Research Centre 833: The construction of meaning - the dynamics
                        and adaptivity of linguistic structures</Department>
                    <Organisation xml:lang="de">Eberhard Karls Universität Tübingen</Organisation>
                    <Organisation xml:lang="en">University of Tübingen</Organisation>
                    <Url>http://www.sfb833.uni-tuebingen.de/</Url>
                    <Location ComponentId="clarin.eu:cr1:c_1290431694494">
                        <Address>Nauklerstraße 35, 72074 Tübingen</Address>
                        <Region>Baden-Württemberg</Region>
                        <Country ComponentId="clarin.eu:cr1:c_1290431694493">
                            <CountryName xml:lang="de">Deutschland</CountryName>
                            <CountryName xml:lang="en">Germany</CountryName>
                            <CountryCoding>DE</CountryCoding>
                        </Country>
```

Comedi

- http://clarino.uib.no/comedi/metadata-editor

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 11

Using metadata

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Exchanging metadata

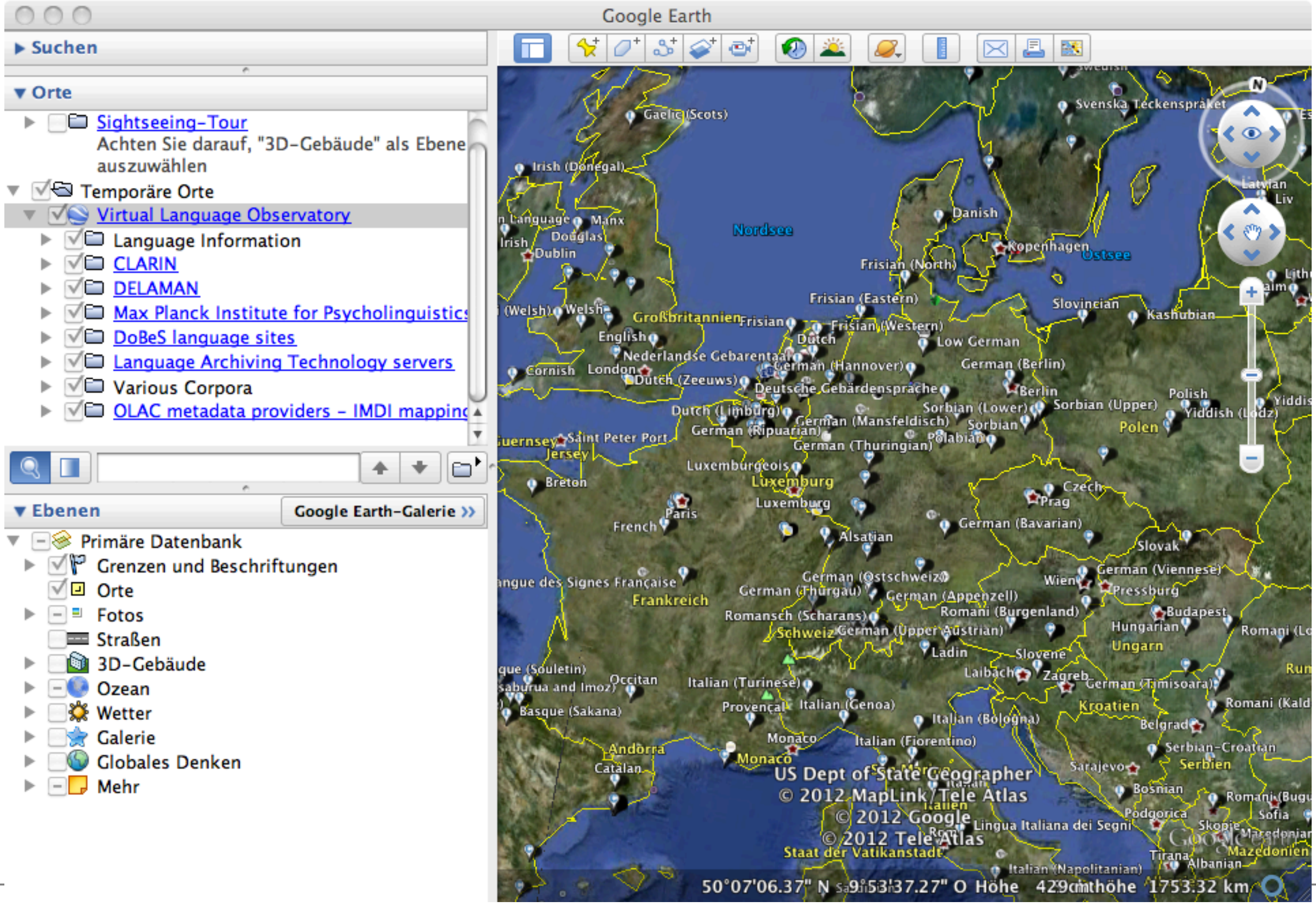
- Spreading the word for visibility
- Application/search engine for resources
- Resource as such can stay at repository
- Integration into other data sources
  - Library catalogues
  - Websites
- Protocol: OAI-PMH
  - Open Archives Initiative – Protocol for Metadata Harvesting
  - Metadata on OAI-PMH Server
  - Clients harvest changed data
- Most important:
  - Search

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Visualising resources in Google-Earth

- See http://www.clarin.eu/vlo/



Google Earth

▶ Suchen

▼ Orte

▶ □ Sightseeing-Tour
Achten Sie darauf, "3D-Gebäude" als Ebene auszuwählen

▼ ☑ Temporäre Orte

▼ ☑ Virtual Language Observatory

▶ ☑ Language Information

▶ ☑ CLARIN

▶ ☑ DELAMAN

▶ ☑ Max Planck Institute for Psycholinguistics

▶ ☑ DoBeS language sites

▶ ☑ Language Archiving Technology servers

▶ ☑ Various Corpora

▶ ☑ OLAC metadata providers - IMDI mappinc

▼ Ebenen Google Earth-Galerie »

▼ ⊟ Primäre Datenbank

▶ ☑ Grenzen und Beschriftungen

☑ Orte

▶ ⊟ Fotos

□ Straßen

▶ □ 3D-Gebäude

▶ ⊟ Ozean

▶ □ Wetter

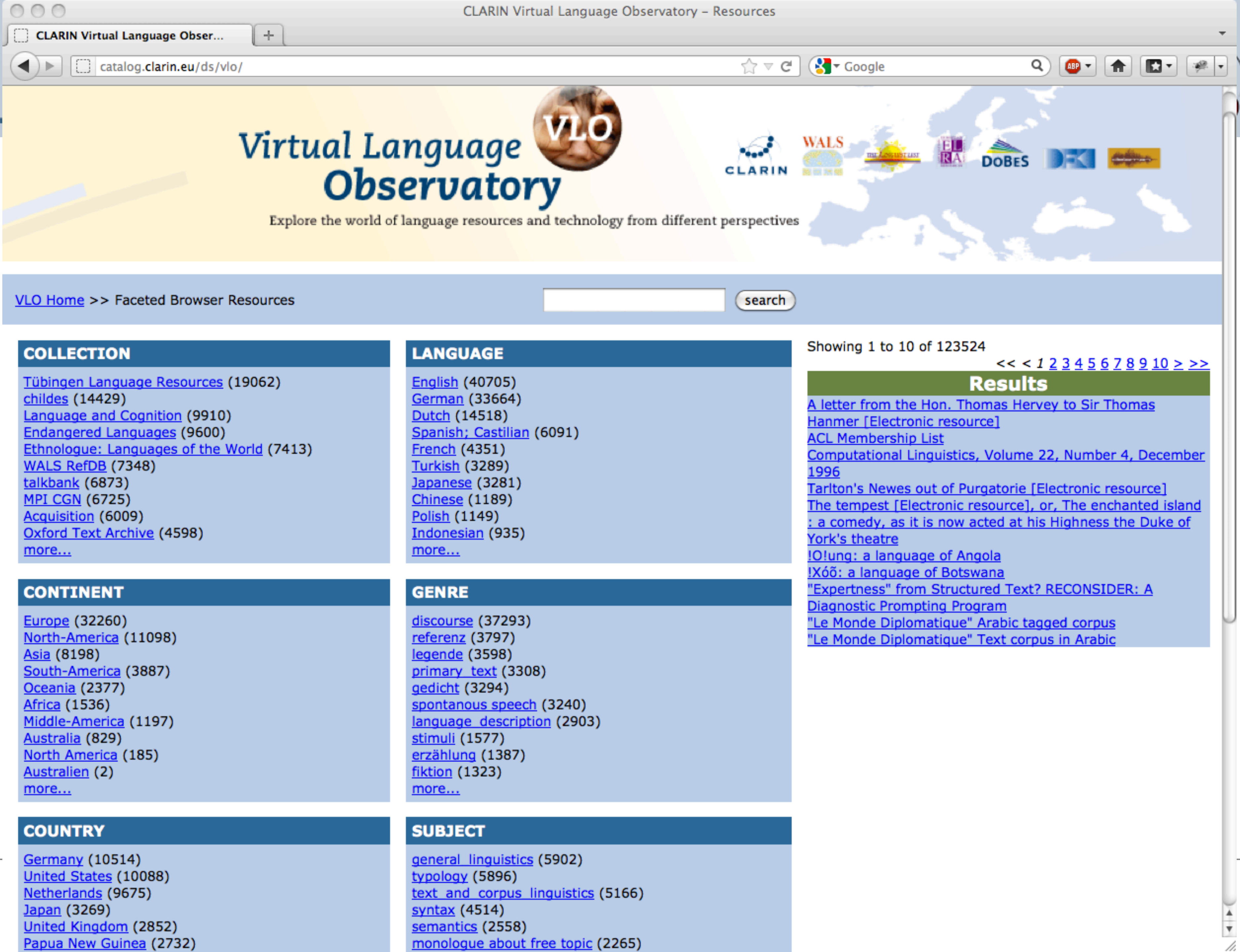
▶ □ Galerie

▶ □ Globales Denken

▶ ⊟ Mehr

US Dept of State Geographer
© 2012 MapLink/Tele Atlas
© 2012 Google
© 2012 Tele Atlas

50°07'06.37" N 9°53'37.27" O Höhe 429 m Sichthöhe 1753.32 km


CLARIN Virtual Language Observatory – Resources
CLARIN Virtual Language Obser...
catalog.clarin.eu/ds/vlo/
Google
VLO
CLARIN
WALS
DOBES
search

Virtual Language Observatory

Explore the world of language resources and technology from different perspectives

VLO Home >> Faceted Browser Resources

COLLECTION

Tübingen Language Resources (19062)
childes (14429)
Language and Cognition (9910)
Endangered Languages (9600)
Ethnologue: Languages of the World (7413)
WALS RefDB (7348)
talkbank (6873)
MPI CGN (6725)
Acquisition (6009)
Oxford Text Archive (4598)
more...

CONTINENT

Europe (32260)
North-America (11098)
Asia (8198)
South-America (3887)
Oceania (2377)
Africa (1536)
Middle-America (1197)
Australia (829)
North America (185)
Australien (2)
more...

COUNTRY

Germany (10514)
United States (10088)
Netherlands (9675)
Japan (3269)
United Kingdom (2852)
Papua New Guinea (2732)

LANGUAGE

English (40705)
German (33664)
Dutch (14518)
Spanish; Castilian (6091)
French (4351)
Turkish (3289)
Japanese (3281)
Chinese (1189)
Polish (1149)
Indonesian (935)
more...

GENRE

discourse (37293)
referenz (3797)
legende (3598)
primary_text (3308)
gedicht (3294)
spontanous speech (3240)
language_description (2903)
stimuli (1577)
erzählung (1387)
fiktion (1323)
more...

SUBJECT

general_linguistics (5902)
typology (5896)
text_and_corpus_linguistics (5166)
syntax (4514)
semantics (2558)
monologue about free topic (2265)

Showing 1 to 10 of 123524

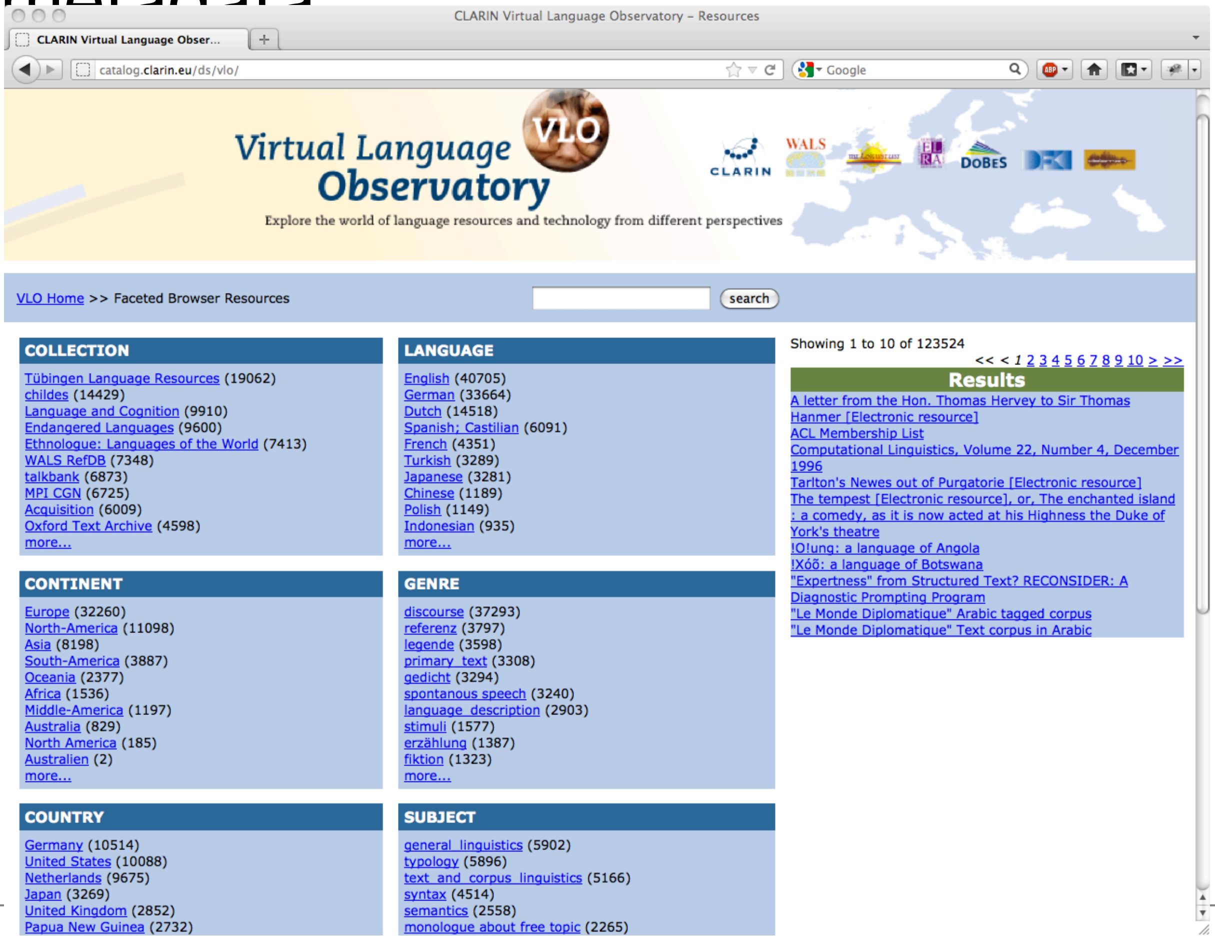
<< < 1 2 3 4 5 6 7 8 9 10 > >>

Results

A letter from the Hon. Thomas Hervey to Sir Thomas Hanmer [Electronic resource]
ACL Membership List
Computational Linguistics, Volume 22, Number 4, December 1996
Tarlton's Newes out of Purgatorie [Electronic resource]
The tempest [Electronic resource], or, The enchanted island : a comedy, as it is now acted at his Highness the Duke of York's theatre
!O!ung: a language of Angola
!Xóõ: a language of Botswana
"Expertness" from Structured Text? RECONSIDER: A Diagnostic Prompting Program
"Le Monde Diplomatique" Arabic tagged corpus
"Le Monde Diplomatique" Text corpus in Arabic

CLARIN Virtual Language Observatory

- Search by data categories
- Based on CMDI metadata


CLARIN Virtual Language Observatory – Resources
CLARIN Virtual Language Obser...
catalog.clarin.eu/ds/vlo/
Google
Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DoBeS

VLO Home >> Faceted Browser Resources

search

COLLECTION

Tübingen Language Resources (19062)
childes (14429)
Language and Cognition (9910)
Endangered Languages (9600)
Ethnologue: Languages of the World (7413)
WALS RefDB (7348)
talkbank (6873)
MPI CGN (6725)
Acquisition (6009)
Oxford Text Archive (4598)
more...

CONTINENT

Europe (32260)
North-America (11098)
Asia (8198)
South-America (3887)
Oceania (2377)
Africa (1536)
Middle-America (1197)
Australia (829)
North America (185)
Australien (2)
more...

COUNTRY

Germany (10514)
United States (10088)
Netherlands (9675)
Japan (3269)
United Kingdom (2852)
Papua New Guinea (2732)

LANGUAGE

English (40705)
German (33664)
Dutch (14518)
Spanish; Castilian (6091)
French (4351)
Turkish (3289)
Japanese (3281)
Chinese (1189)
Polish (1149)
Indonesian (935)
more...

GENRE

discourse (37293)
referenz (3797)
legende (3598)
primary_text (3308)
gedicht (3294)
spontanous speech (3240)
language_description (2903)
stimuli (1577)
erzählung (1387)
fiktion (1323)
more...

SUBJECT

general_linguistics (5902)
typology (5896)
text_and_corpus_linguistics (5166)
syntax (4514)
semantics (2558)
monologue about free topic (2265)

Showing 1 to 10 of 123524

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Results

A letter from the Hon. Thomas Hervey to Sir Thomas Hanmer [Electronic resource]
ACL Membership List
Computational Linguistics, Volume 22, Number 4, December 1996
Tarlton's Newes out of Purgatorie [Electronic resource]
The tempest [Electronic resource], or, The enchanted island : a comedy, as it is now acted at his Highness the Duke of York's theatre
!O!ung: a language of Angola
!Xóõ: a language of Botswana
"Expertness" from Structured Text? RECONSIDER: A Diagnostic Prompting Program
"Le Monde Diplomatique" Arabic tagged corpus
"Le Monde Diplomatique" Text corpus in Arabic

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Chapter 12

Summary

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Metadaten in general

- Description of resources
- Purpose: Overview and locating resources
- Different metadata formats
- Different resource types = different metadata schemas
- Integration of metadata in archive systems/repositories/catalogues

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Metadaten for resources

- Description by filling in data categories
- Definition of data categories: Concept Registry
- Grouping of data categories: components
- Components combined to describe one type of resource: profile
- Tools:
  - Concept Registry: Data categories
  - Component Registry: creation and central storage of components
  - ARBIL and Comedi: CMDI editors

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

TEST

- What is the connection between metadata, data categories, components and profiles?
- Why is it necessary to create metadata?
- Create metadata for a book using CMDI!
  - Select data categories
  - Create a component
  - Create an instance matching the component

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Contact: Seminar für Sprachwissenschaft

Dr. Thorsten Trippel

Wilhelmstr. 19, 72074 Tübingen

Telefon: +49 7071 29-77352

thorsten.trippel@uni-tuebingen.de