



Metadata in a project on sustainable linguistic data - creation, managing and use

Thorsten Trippel

Centre for Sustainability of Linguistic Data (NaLiDa)

2012-06-05, Workshop on the Interoperability of Metadata, Oslo

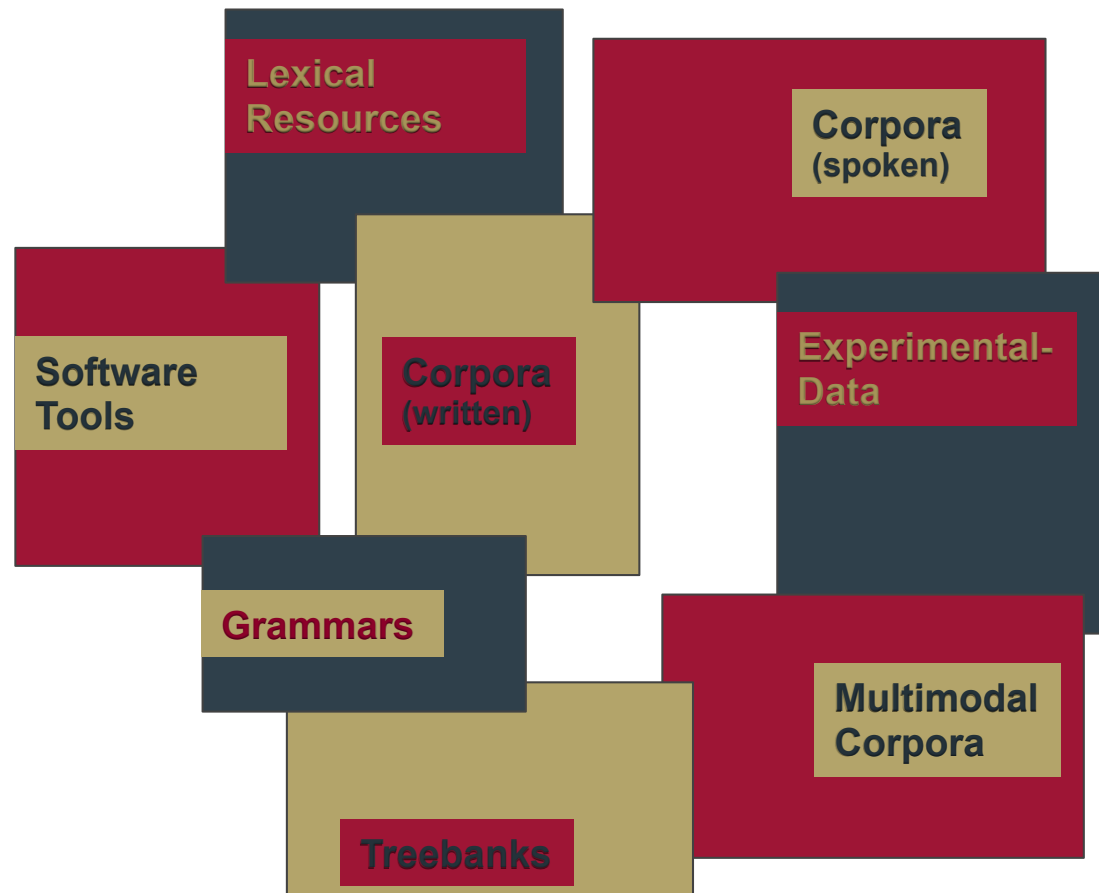


Structure

- Types of Linguistic Resources
- Researchers as customers
- The archiving workflow: pitfalls and obstacles
- University of Tübingen CMDI profiles
- Creating metadata



Background





The pit and the fall

- Huge variety
 - Dublin Core not sufficient for proper description
 - Description:= providing information to provide insights if a resource should be further investigated
 - Dublin core is not sufficient
- Different types: requiring different levels of description
- CMDI Metadata



Researchers as Customers

- Infrastructures need users of the infrastructure
- Researchers are supposed to be users
- Our customers: researcher



Why? Added value

- Integration into primary data repository
- Searchability of the resource
- Citability
- Interoperability
- Reusability of published data



Extrinsic motivation for archiving: Funding

- DFG (German Research Foundation):
 - min 10 years availability
 - Originating institution
- Wissenschaftsrat (Consulting council for the German Government):
 - In general publicly accessible
 - QA in research (anti plagiarism and fraud)
- Requests by other researchers

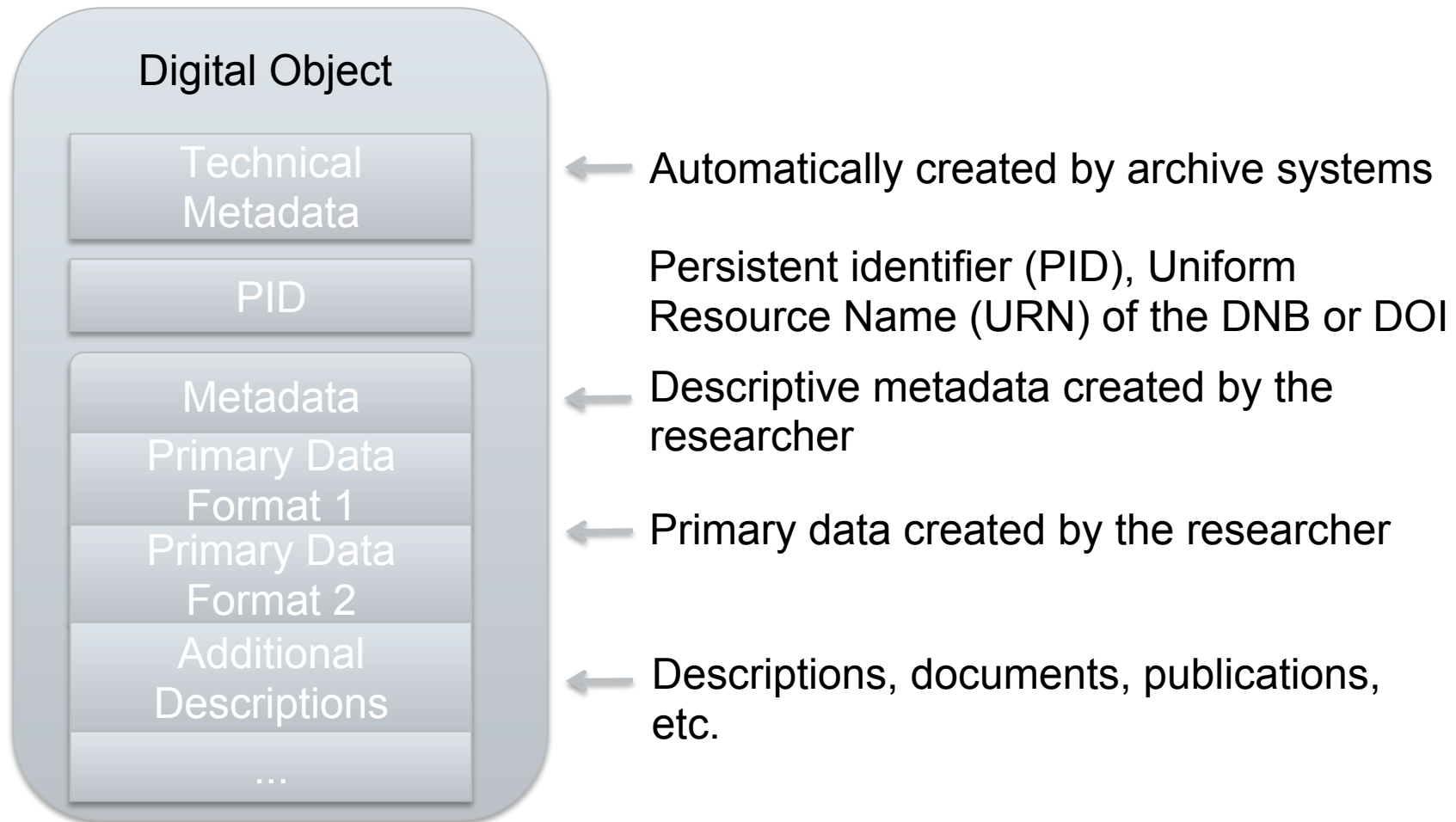


Archiving: How? Three phases:

- Preparing a resource for archiving
- Inserting the resource into the archive system
- Archive acceptance: Sanity check and PID assignment



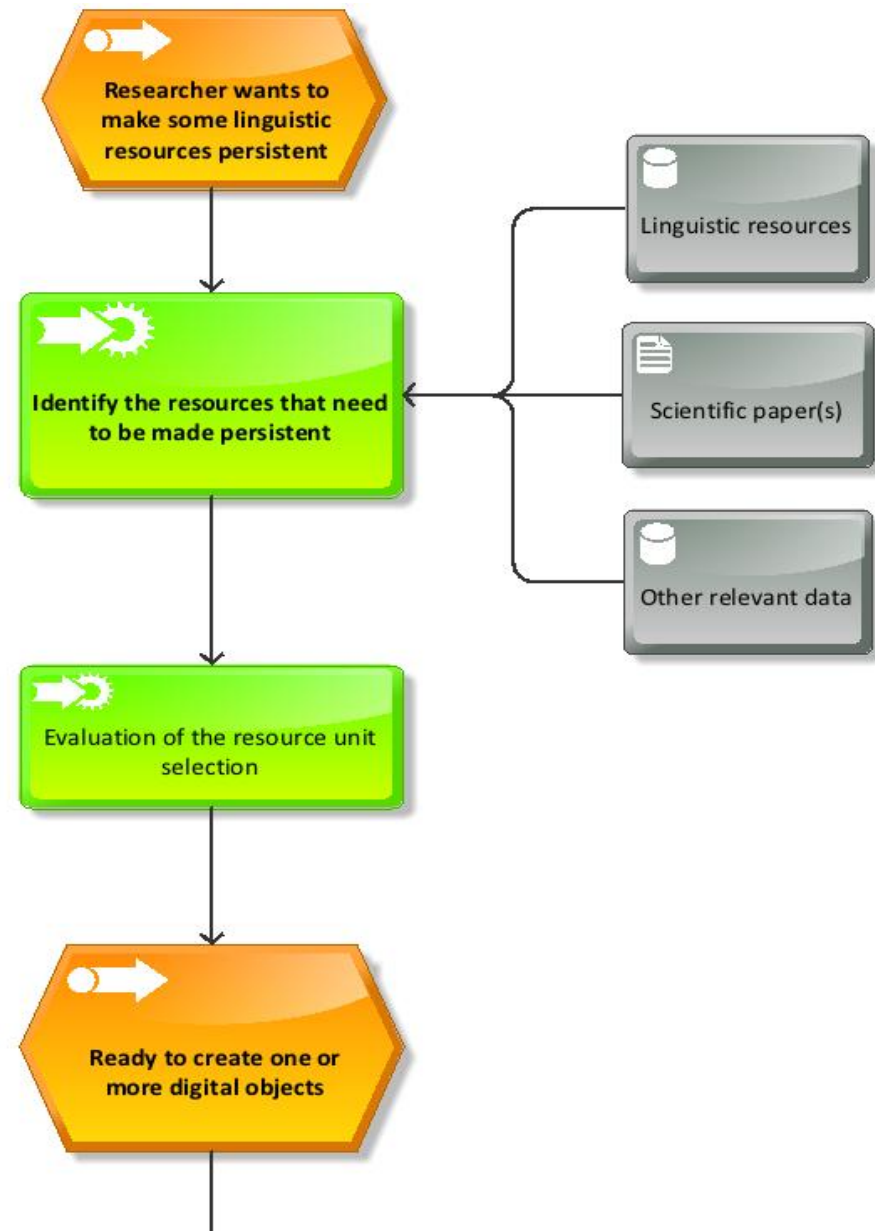
Technical Infrastructure: Fedora-Commons backend





Phase 1: Workflow

- Decide on the resource: What is one resource?
 - One experiment?
 - All parallel experiments in a study?
 - One lexicon?
 - One lexicon article?
 - One corpus?
 - One annotation layer?
- Rules:
 - One resource, one citable PID
 - A resource should "make sense" without other resources
 - See also ISO 24619:2011





Granularity according to ISO 24619:2011

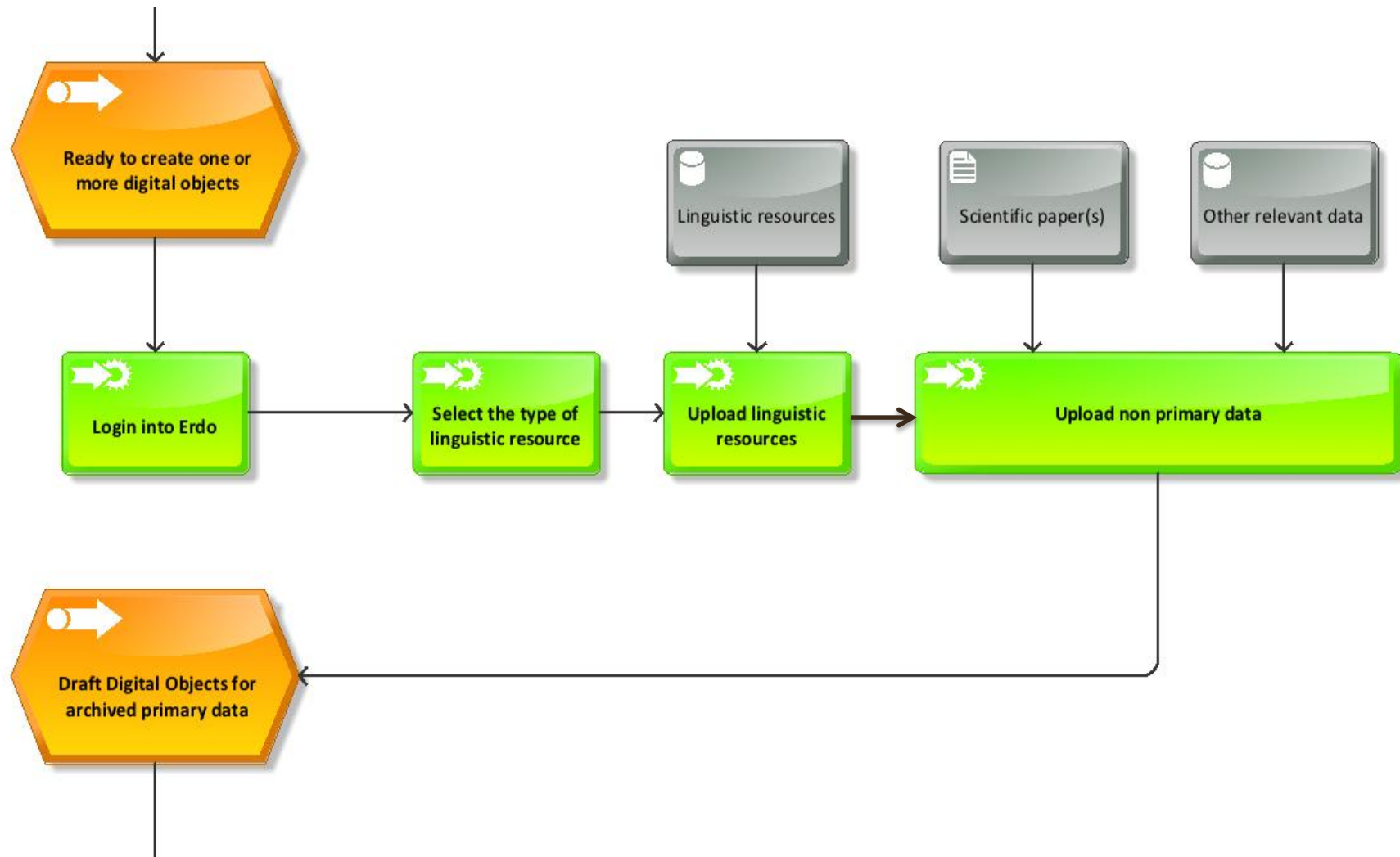
Individual IDs for resources in these cases:

- If there is an existing identifier scheme for a type of resources, for instance, ISBN, this level of granularity should be retained[...]
- If the resource is associated with the complete content of a digital file.
- If the resource is autonomous and exists outside a larger context.
- If a resource should be citable apart from any containing resource.

"Subject to the needs of resource creators with respect to the level of granularity they deem suitable to the specific resource environment."



Phase 2: Entering the archive (data upload)





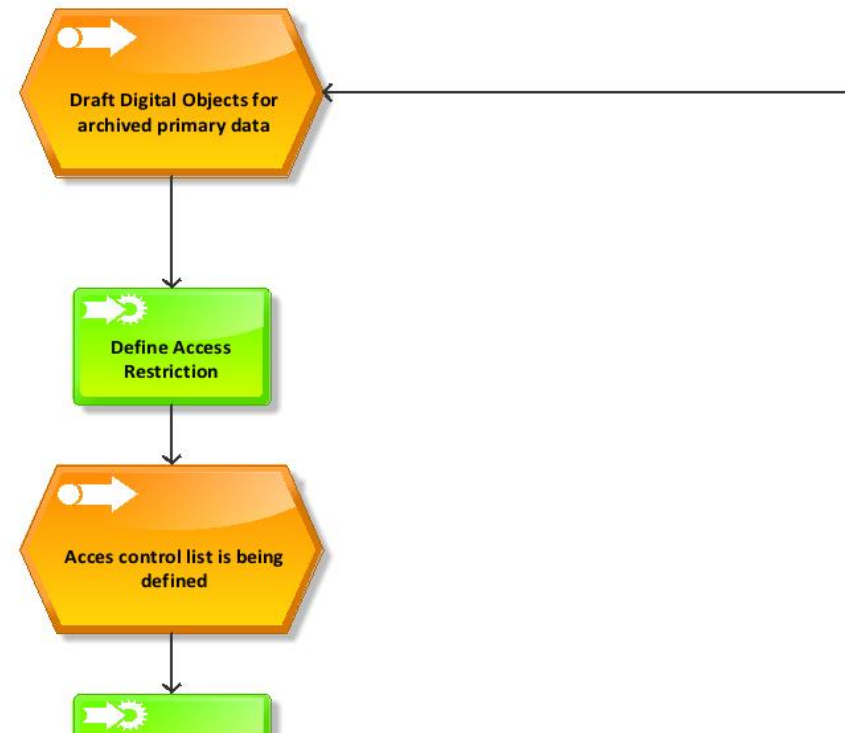
Documentation beyond the primary data

- Some primary research data: not self explanatory
- Existing documentation:
 - Publications
 - Technical documents
 - README-files
- Access restrictions to documentation
 - Publishers
 - "unfinished"
 - Internal information



Phase 2: Entering the archive (access control)

- Project internal: all read
- Allow other individuals to read
- Allow public to read



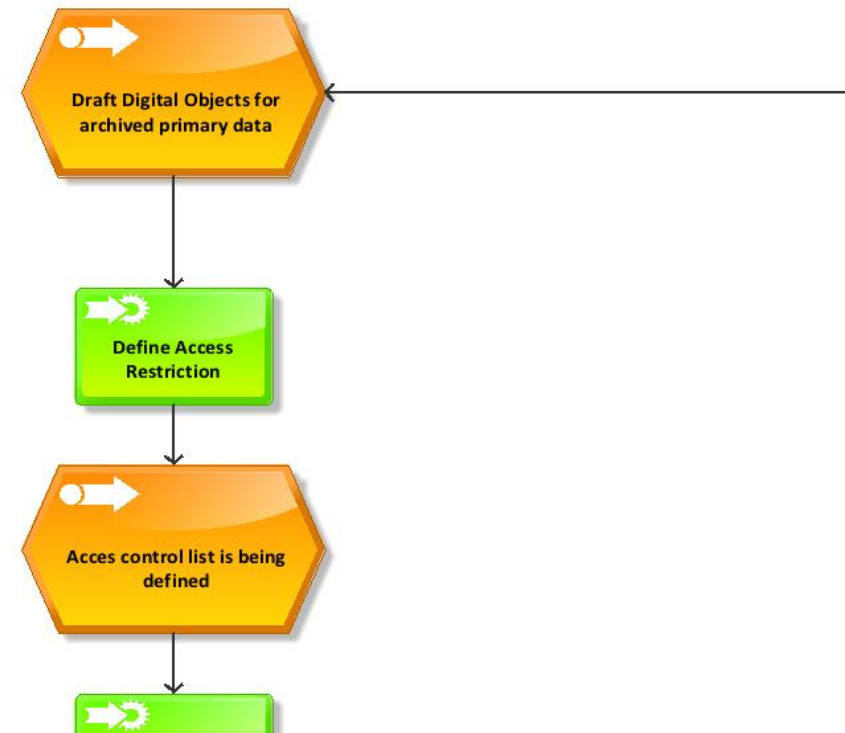


Phase 2: Entering the archive (access control)

- Project internal: all read
- Allow other individuals to read
- Allow public to read

The pay-off of persistency:

- Digital objects are not deletable
- Editing follows strict procedure
- Yes but no: workarounds





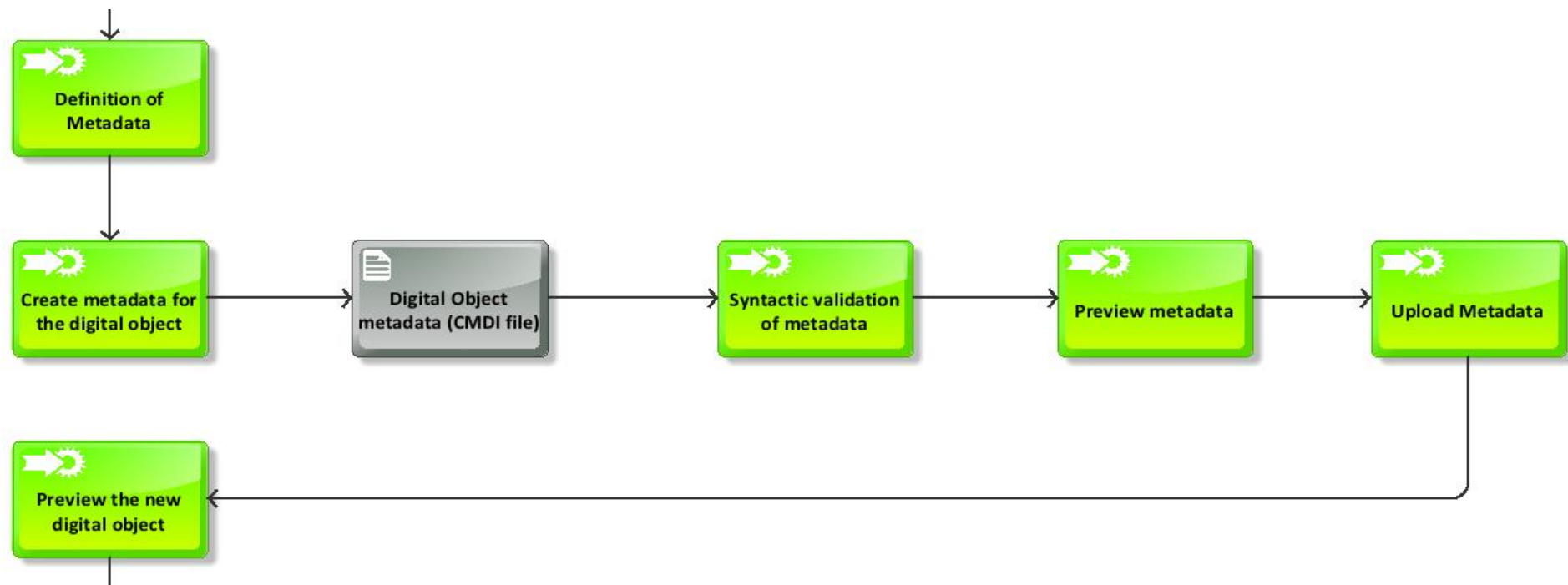
Editing and deleting persistent/archived data?

- User requirement
 - No finished LR
 - Versioning too techy
 - No changes, no data (!)
- Legal requirement
 - Cases of copyright infringement
 - Dichotomy: *delete at project end* vs. keep at least 10 years
- *GIF* vs. *JPEG* compared to *txt-*, vs. *txt+*,



Phase 2: Entering the archive (metadata creation)

- Metadata is structured data for describing and finding resources
- Essential for archives and catalogues





Metadata creation process

- Major concerns
 - Nobody wants to contribute metadata
 - Nobody wants to spend time on archiving
 - Tools are too cumbersome
 - Not all bits of information are available
- But:
 - Everybody wants their data too be found
 - Everybody wants to have the most correct representation of their data
 - Purpose dependent editors for the technophobe



Form-based CMDI-Editing

Resource:

GeneralInfo

ResourceName

ResourceName

ResourceTitle

ResourceTitle
 in

ResourceClass

ResourceClass

Version

Version

LifeCycleStatus

LifeCycleStatus

StartYear

StartYear

CompletionYear

CompletionYear

PublicationDate

PublicationDate

LastUpdate

LastUpdate

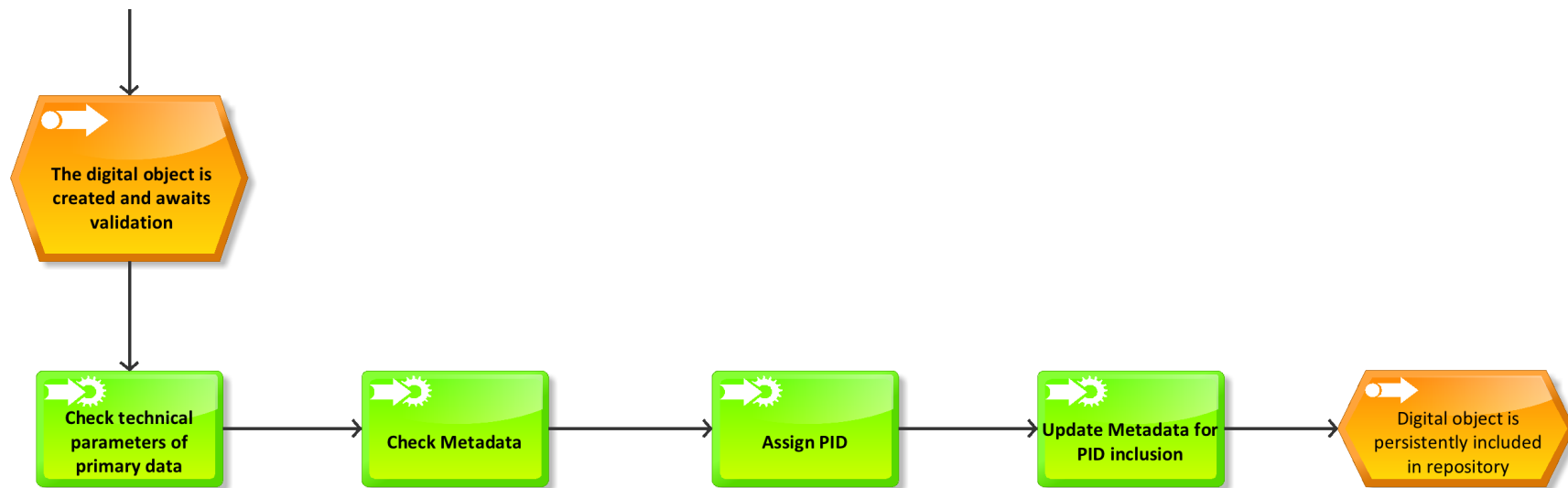
TimeCoverage

See <http://www.isocat.org>
 The definition of this data category is available via <http://www.isocat.org/datcat/DC-2544>





Phase 3: Archivists take over





Checking the data

- Access restrictions (!)
- But:
 - File size
 - Automated processes (XML syntax parsing)
 - Metadata checks
 - Number of files
 - Interrelation of files
 - Amount of filled in data



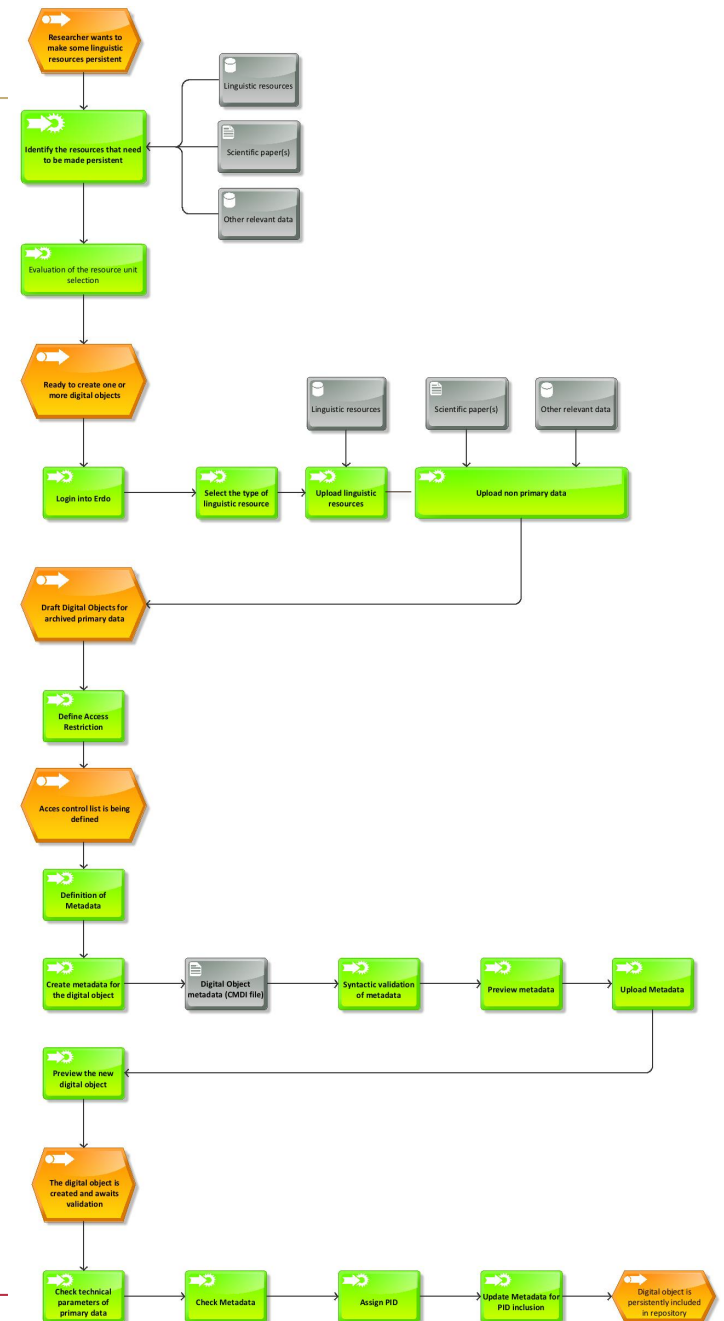
Which PID schema?

- DOI
 - → costs
 - → persistence (strong restrictions, problem for community)
- URN
 - → resolving service partly missing
 - → provider API (available?)
- Handle
 - → EPIC -- the European Persistent Identifier Consortium provides a Service for the European Research Community
- Self-Service PID
 - → Any plans for the weekend?



Whole workflow

No exceptions, revisions and options shown





Procedure for CMDI Component Creation

- Identifying type of resource and user group
- Reuse of components
- Recycling of components
- Creating new components
- Selecting, modifying data categories



Identifying Type of Resource and User Group

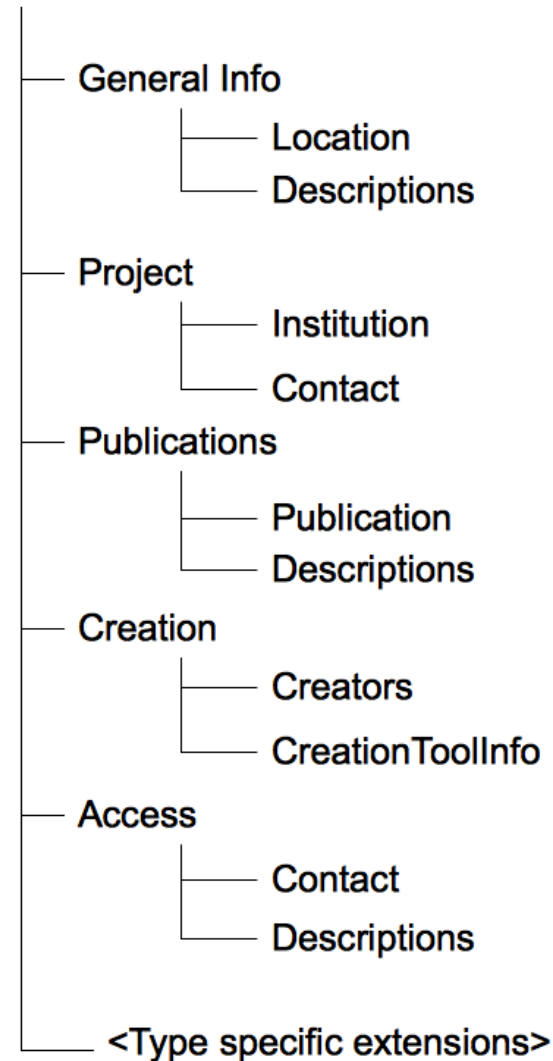
- One type of resource: one CMDI profile
 - Example: Different types of corpora require various profiles
 - Differences: spoken vs. written; field of study; used technical infrastructure; etc.
- Also depends on user group:
 - Technical background
 - Intended use in NLP environments vs. humanities computing



Reuse of Components

- Reuse of exiting components if possible
- General components often rather independent of resource type
- Type-specific extensions possible
- Advantage: partly reusing tools

Profile





Recycling of Components

- Reusing structures of existing components
 - Reuse existing components to create copy
 - Modification according to needs
- Frequent changes:
 - Cardinality of data categories
 - Allowing for multilinguality
 - Add (optional) additional data categories



Creating New Components

- When needed for new resource types
- Based on a collection of metadata categories
 - Forming sensible groups
 - Groups based on expected reusability
- Often reuses existing components as parts

Content:

- Profiles
- Components
- Data Categories

Profiles

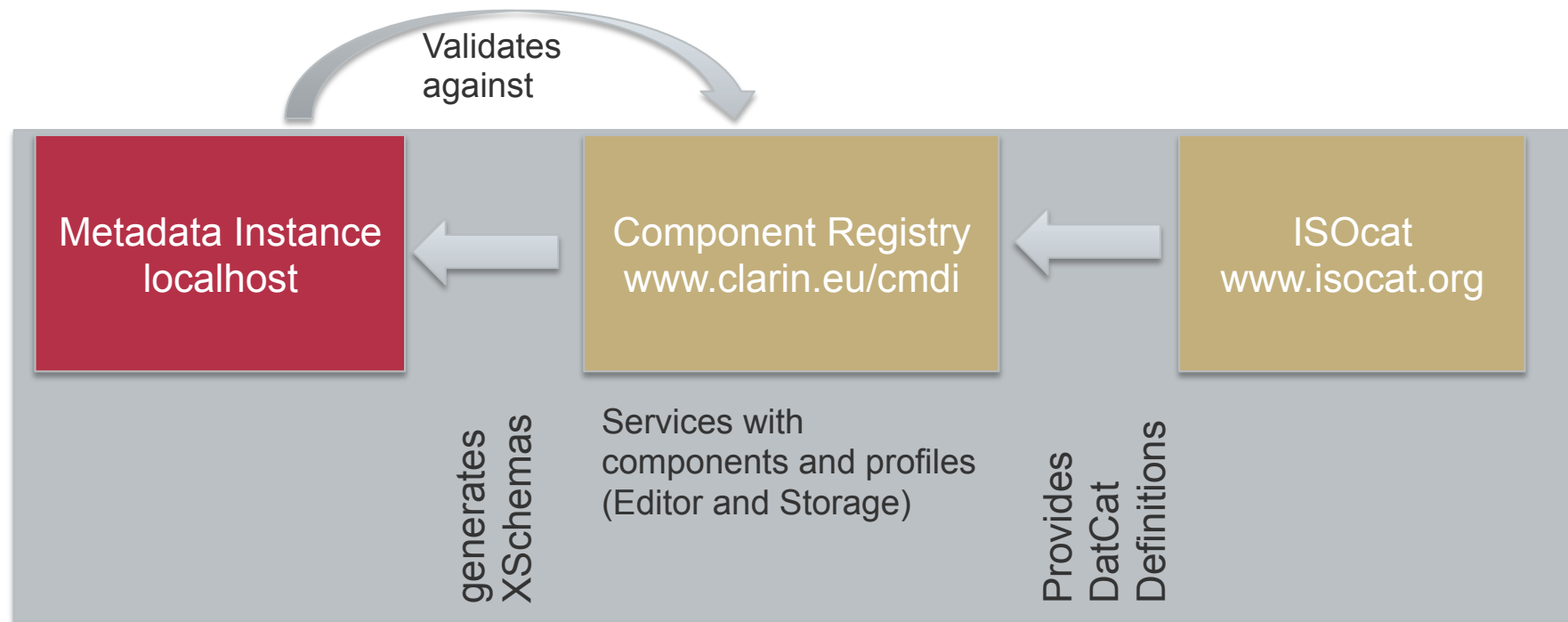
Profile Name	Resource Type	Description	Profile ID	Download Profile (XSD)
experiment profile	experimental data	This CMDI profile can be used for describing psychological studies, for example. The term "experiment" for the profile/components is generalised here so that it does not only refer to experimental designs but also to other types of study, such as test or survey data.	clarin.eu:cr1:p_1302702320451	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1302702320451/xsd
lexical resource profile	lexical resources: lexicons, dictionaries, wordnets, etc.	A CMDI profile for lexical resources (e.g. lexicons, wordnets, etc.).	clarin.eu:cr1:p_1290431694579	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694579/xsd
literary corpus profile	specialised corpus	A CMDI profile for text (i.e. written) corpus resources that are based on (digitised) literary works.	clarin.eu:cr1:p_1314870716423	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1314870716423/xsd
resource bundle	CMDI bundle	A CMDI profile for bundling arbitrary resources by URL and/or PID (PID preferred).	clarin.eu:cr1:p_1320657629649	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1320657629649/xsd
speech corpus profile	spoken language corpus	A CMDI profile for speech corpus resources. This profile is modified and resembles the IMDI metadata profile for spoken language resources.	clarin.eu:cr1:p_1302702320401	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1302702320401/xsd
text corpus profile	written corpus	A CMDI profile for text (i.e. written) corpus resources.	clarin.eu:cr1:p_1290431694580	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694580/xsd
tool profile	tool/software	A CMDI profile for tools/software.	clarin.eu:cr1:p_1290431694581	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694581/xsd
WebLicht web service	CMD core model for web services	This is a core model which can be used as a basis for web service registry specific models. In principle, this profile should not be instantiated. Instead, new profiles can be created that extend the structure built by the components in this profile. Valid instantiations of these extensions allow validation of the core elements against the schema of this profile. More detailed information, including facilities to validate compliance with this core model, can be found at: http://www.isocat.org/clarin/lws/cmd-core/	clarin.eu:cr1:p_1320657629644	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1320657629644/xsd
web service profile	web service	A CMDI profile for web services.	clarin.eu:cr1:p_1299509410083	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1299509410083/xsd
web tool chain	tool chain	A CMDI profile for describing the sequential application of web services or other tools in a tool chain.	clarin.eu:cr1:p_1320657629623	http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1320657629623/xsd



What is this all good for?

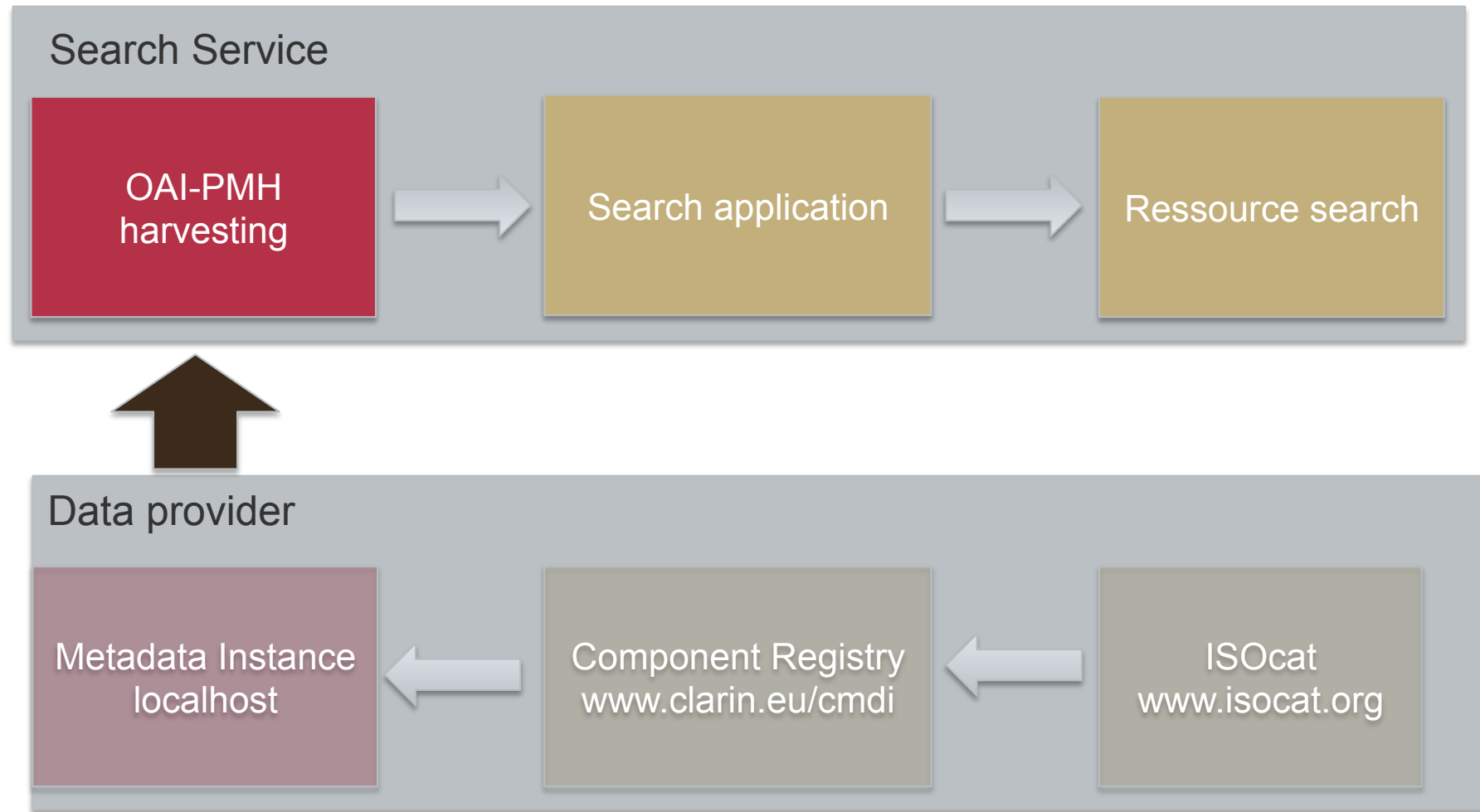


Semantic interoperability and consistent Syntax of Metadata Schemas: Repositories in CMDI /simdi/





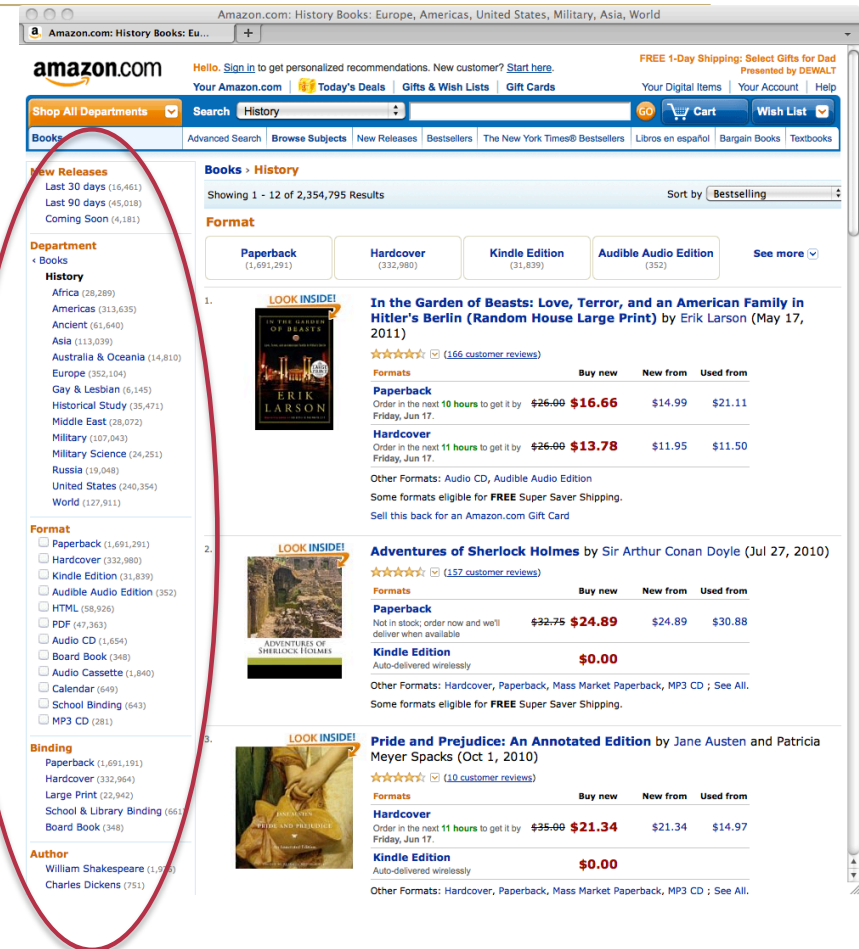
CMDI to search





Faceted Search for sustainable resources

- Method used in e-commerce applications
 - Often top level not sophisticated
 - Combined with full text search
- Unconditional facets
 - For any type of resource
 - Values in the facets respect previous selections
 - Automatic update
- Conditional facets
 - For specific resource types only
 - Else the same as for unconditional facets





Faceted Browser mit bedingten und unbedingten Facetten

Facet: modality (6)		Facet: language (60)	
modality	Occurrences	language	Occurrences
Pointing gestures	448	Albanian	1
Signs	77	Bosnian	3
Speech	9339	British Sign Language	23
Unspecified	13	Bulgarian	1
verbal and non-verbal interaction	766	Chinese	1
Writing	98	Croatian	4
		dk	1
		Dutch	5
		Dutch Sign Language	29
Facet: resourceclass (8)		Facet: country (5)	
resourceclass	Occurrences	country	Occurrences
general corpus	1	France	93
learner corpus	2	Germany	10188
LexicalResource	1	Netherlands	21
Lexicon	4	Sweden	8
other	64	United Kingdom	21
Resource	3618		
Tool	277		
WrittenCorpus	15		
Facet: organisation (9)			
organisation	Occurrences		
Magdeburg-Stendal University of Applied Sciences	10		
Max Planck Institute for Psycholinguistics	708		
Max-Planck-Institut für Bildungsforschung	1443		
Radboud University Nijmegen	67		
SFB 441	34		
SFB 632	27		
Universität Tübingen	8		
University of Leipzig	168		
University of Stuttgart	32		



Faceted Browser mit bedingten und unbedingten Facetten

- Unconditional facets
 - Modality
 - Language
 - Resourceclass
 - Country
 - Organisation
- Special: Full text search

Facet: modality (6)		Facet: language (60)	
modality	Occurrences	language	Occurrences
Pointing gestures	448	Albanian	1
Signs	77	Bosnian	3
Speech	9339	British Sign Language	23
Unspecified	13	Bulgarian	1
verbal and non-verbal interaction	766	Chinese	1
Writing	98	Croatian	4
		dk	1
		Dutch	5
		Dutch Sign Language	29
Facet: resourceclass (8)		Facet: country (5)	
resourceclass	Occurrences	country	Occurrences
general corpus	1	France	93
learner corpus	2	Germany	10188
LexicalResource	1	Netherlands	21
Lexicon	4	Sweden	8
other	64	United Kingdom	21
Resource	3618		
Tool	277		
WrittenCorpus	15		
Facet: organisation (9)			
organisation	Occurrences		
Magdeburg University of Applied Sciences	10		
Max Planck Institute for Psycholinguistics	708		
Max-Planck-Institut für Bildungsforschung	1443		
Radboud University Nijmegen	67		
SFB 441	34		
SFB 632	27		
Universität Tübingen	8		
University of Leipzig	168		
University of Stuttgart	32		



Conditional Facets

- Tools
 - Tool type
 - Input type
 - Output type
 - Application type
- Unconditional facets minimized
- Other resource types: other conditional facets

The screenshot displays the NALIDA Faceted Browsing interface. The left sidebar contains a list of facets, with several highlighted by red circles:

- Facet: organisation (1)**
- Facet: tooltype (19)**
- Facet: inputtype (9)**
- Facet: outputtype (9)**
- Facet: applicationtype (1)**

The main content area shows a table of occurrences for the selected facets. The table has two columns: the facet name and the number of occurrences. The facets are organized into groups, with the first group containing 'Facet: modality (1)', 'Facet: language (3)', 'Facet: country (1)', 'Facet: organisation (1)', and 'Facet: tooltype (19)'. The second group contains 'Facet: inputtype (9)' and 'Facet: outputtype (9)'. The third group contains 'Facet: applicationtype (1)'. The table lists various tools and their occurrences, such as 'Lemmatizer' (1), 'syntactic tagging' (2), 'clustering' (2), 'stemming/lemmatization' (1), 'keyword extraction' (1), 'webservice' (1), 'Web service' (120), 'Morphological Analyser' (1), 'Annotated corpus extractor' (1), 'raw text' (1), 'word-segmented text' (2), 'tokenized text with sentence and document boundaries' (1), 'query' (2), 'transformer' (3), 'textcorpus.0.4' (17), 'textcorpus.0.3' (52), 'descriptive morphological lexicon' (22), 'binary' (25), 'transformer' (1), 'textcorpus.0.4' (26), 'textcorpus.0.3' (70), 'morphologically analysed word forms' (1), 'structured corpus' (1), 'dspin.pid.wrapper.dspinlexicon' (22), 'binary' (2), 'Annotated Text with part-of-speech and lemma information' (1), 'structured corpus' (1), 'web application' (2), and 'local/desktop' (2).

The right sidebar shows the 'resourceclass: Tool' section, which lists documents (277) and a list of documents with their IDs and titles. The documents are listed in a table with columns for ID and title. The titles include 'NaLiDaTestRepository oai:ut:NaLiDa:sfb833 Tool A3 PyCWB', 'NaLiDaTestRepository oai:ut:NaLiDa:sfb833 Tool A4 WELCOME', 'NaLiDaTestRepository oai:ut:NaLiDa:smor nalida', 'NaLiDaTestRepository oai:ut:NaLiDa:treeTagger nalida', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 102', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 103', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 108', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 109', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 110', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 116', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 117', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 129', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 132', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 133', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 136', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 137', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 139', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 140', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 145', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 146', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 148', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 150', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 151', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 152', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 154', 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 155', and 'NaLiDaTestRepository oai:ut:NaLiDa:weblicht 165'.

At the bottom right, there is a section for 'full text search results' with a table showing 'id' and 'score'.



- Previously selected facets with value
- List of result set

NALIDA Faceted Browsing

Facet: modality (1) **resourceclass: Tool**

Facet: language (3)

Facet: country (1)

Facet: organisation (3)

Facet: tooltype (19)

tooltype	Occurrences
Lemmatizer	1
syntactic tagging	2
clustering	2
stemming/lemmatization	1
keyword extraction	1
webservice	1
Webservice	120
Morphological Analyser	1
distributed corpus collector	1

Facet: Inputtype (9)

inputtype	Occurrences
raw text	1
word-segmented text	2
tokenized text with sentence and document boundaries	1
query	2
transformer	3
textcorpus_0.4	17
textcorpus_0.3	52
dspin_pid_wrapper.dspinlexicon	22
binary	25

Facet: outputtype (9)

outputtype	Occurrences
transformer	1
textcorpus_0.4	26
textcorpus_0.3	70
morphologically analysed word forms	1
structured corpus	1
dspin_pid_wrapper.dspinlexicon	22
binary	2
Annotated Text with part-of-speech and lemma information	1
structured corpus	1

Facet: applicationtype (3)

applicationtype	Occurrences
web application	2
local/desktop	2

Documents (277)

documents

NaLiDaTestRepository oai:ut:NaLiDa:sfb833 Tool A3 PyCWB

NaLiDaTestRepository oai:ut:NaLiDa:sfb833 Tool A4 WELCOME

NaLiDaTestRepository oai:ut:NaLiDa:smor nalida

NaLiDaTestRepository oai:ut:NaLiDa:treeTagger nalida

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 102

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 103

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 108

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 109

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 110

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 116

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 117

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 129

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 132

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 133

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 136

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 137

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 139

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 140

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 145

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 146

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 148

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 150

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 151

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 152

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 154

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 155

NaLiDaTestRepository oai:ut:NaLiDa:weblicht 165

Page 1 of 6

full text search results

id	score
----	-------



Faceted Browser mit bedingten und unbedingten Facetten

After the
selection of
facets

The screenshot shows a web interface for a faceted browser. On the left, a facet titled 'Facet: language ()' is expanded, showing a list of languages. The 'language' facet is selected, and the 'Occurrences' tab is active. On the right, a list of facets is displayed, including 'country: Germany', 'modality: Writing', 'organisation: University of Stuttgart', 'resourceclass: Tool', and 'tooltype: Lemmatizer'. Below these facets, a section titled 'Documents (1)' is shown, containing a single document entry: 'TreeTagger - a language independent part-of-speech tagger'. This document entry is circled in red.



Resulting Info

- Summary of metadata
- Structured
- Providing required information for accessing

TreeTagger - a language independent part-of-speech tagger

TreeTagger - a language indep... +

Resource: TreeTagger - a language independent part-of-speech tagger

General Info Project Creation Access Copyright Tool context

Resource Context About...

General Info

Resource Name	TreeTagger
Resource Title	TreeTagger - a language independent part-of-speech tagger
Resource Class	Tool
Version	3.2
Publication Date	1994
Location	Institut für Maschinelle Sprachverarbeitung (IMS), Azenbergstraße 12, D-70174 Stuttgart
Description	The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese, Swahili, Northern Sotho, Dzongkha and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available. The TreeTagger can also be used as a chunker for English, German, and French.



Making document view editable

- Easy to process for customers
- Help if required
- Generated for different types of resources
- Assistance in filling in
 - Data type
 - Picklist
 - Required
 - ...
- Special purpose editors



Form-based CMDI-Editing

Resource:

GeneralInfo

ResourceName

ResourceName

ResourceTitle

ResourceTitle
 in

ResourceClass

ResourceClass

Version

Version

LifeCycleStatus

LifeCycleStatus

StartYear

StartYear

CompletionYear

CompletionYear

PublicationDate

PublicationDate

LastUpdate

LastUpdate

TimeCoverage

See <http://www.isocat.org>
 The definition of this data category is available via <http://www.isocat.org/datcat/DC-2544>





Summary and Outlook

- General archiving workflow
- Challenges embedded in user community
 - Non-archivists
 - Some checking by archivists required
 - Privacy concerns
- Tool support essential
- Pitfalls in the archiving details



Thank you.

Contact

Centre for Sustainability of Linguistic Data (NaLiDa)

University of Tübingen, Department of Linguistics

Wilhelmstraße 19

72074 Tübingen · Germany

nalida@sfs.uni-tuebingen.de

<http://www.sfs.uni-tuebingen.de/nalida/>