EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
HUMANITIES
Department of Linguistics
General and Computational Linguistics

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

Thorsten Trippel
Centre for Sustainability of Linguistic Data (NaLiDa)

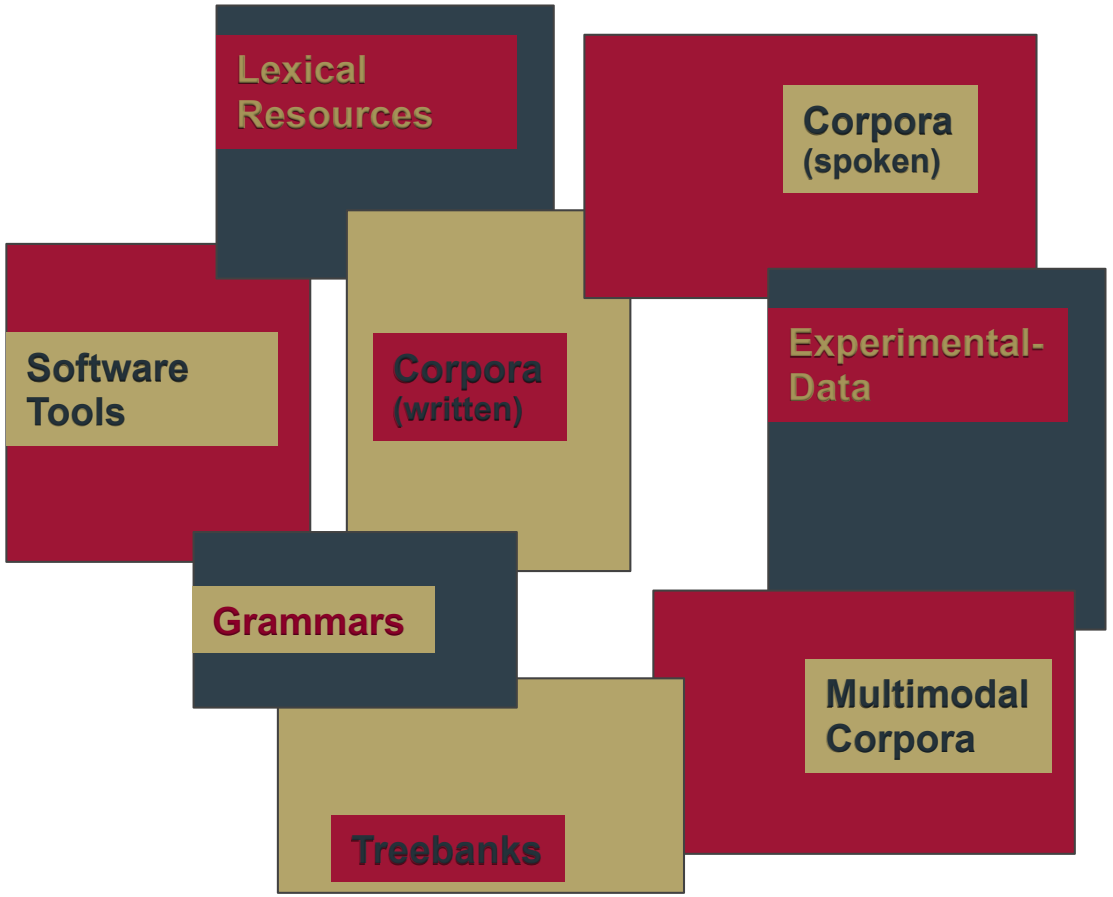
2012-06-05, Workshop on the Interoperability of Metadata, Oslo

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

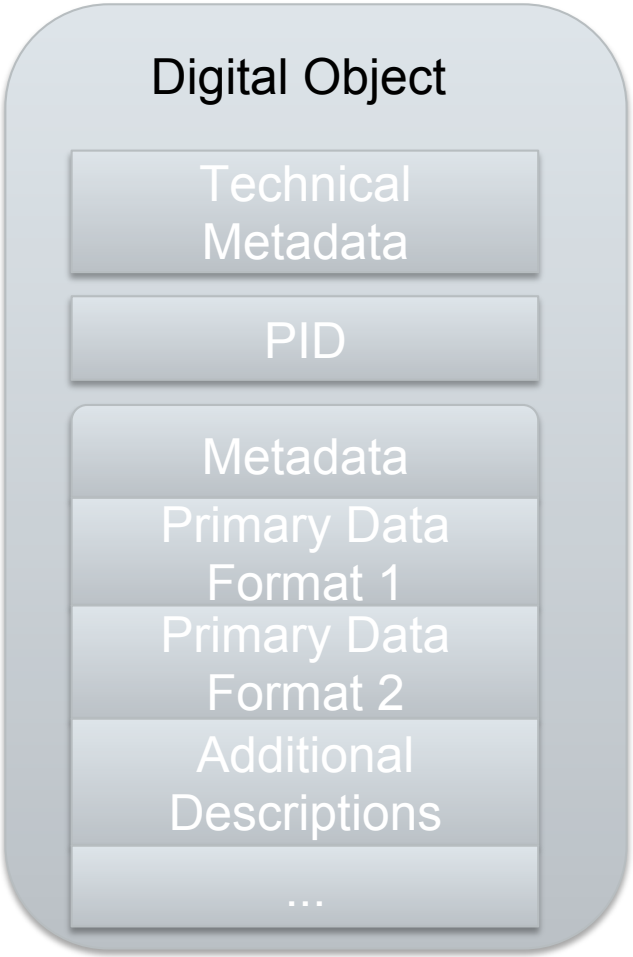
Extrinsic motivation for archiving: Funding

- 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

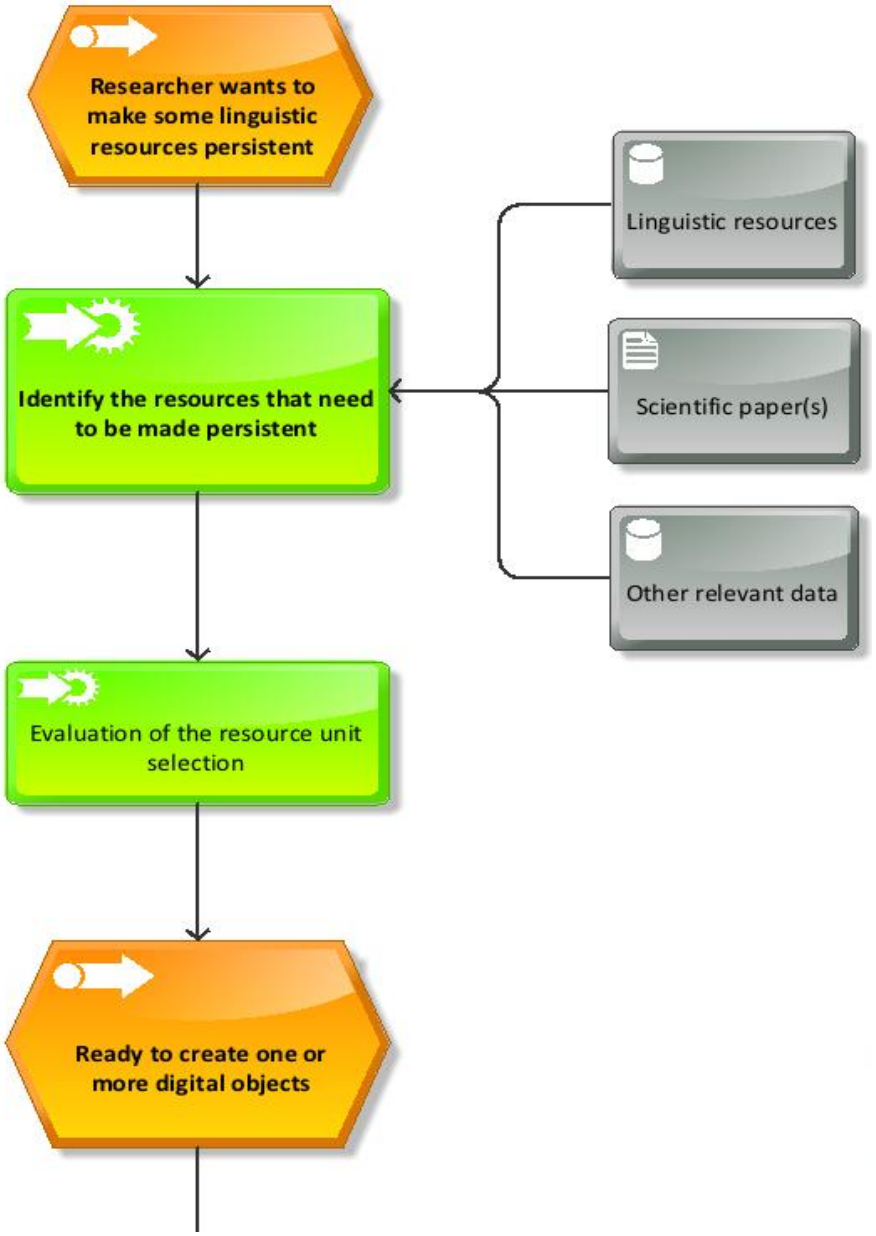


- Automatically created by archive systems
- Persistent identifier (PID), Uniform Resource Name (URN) of the DNB or DOI
- Descriptive metadata created by the researcher
- Primary data created by the researcher
- Descriptions, documents, publications, etc.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

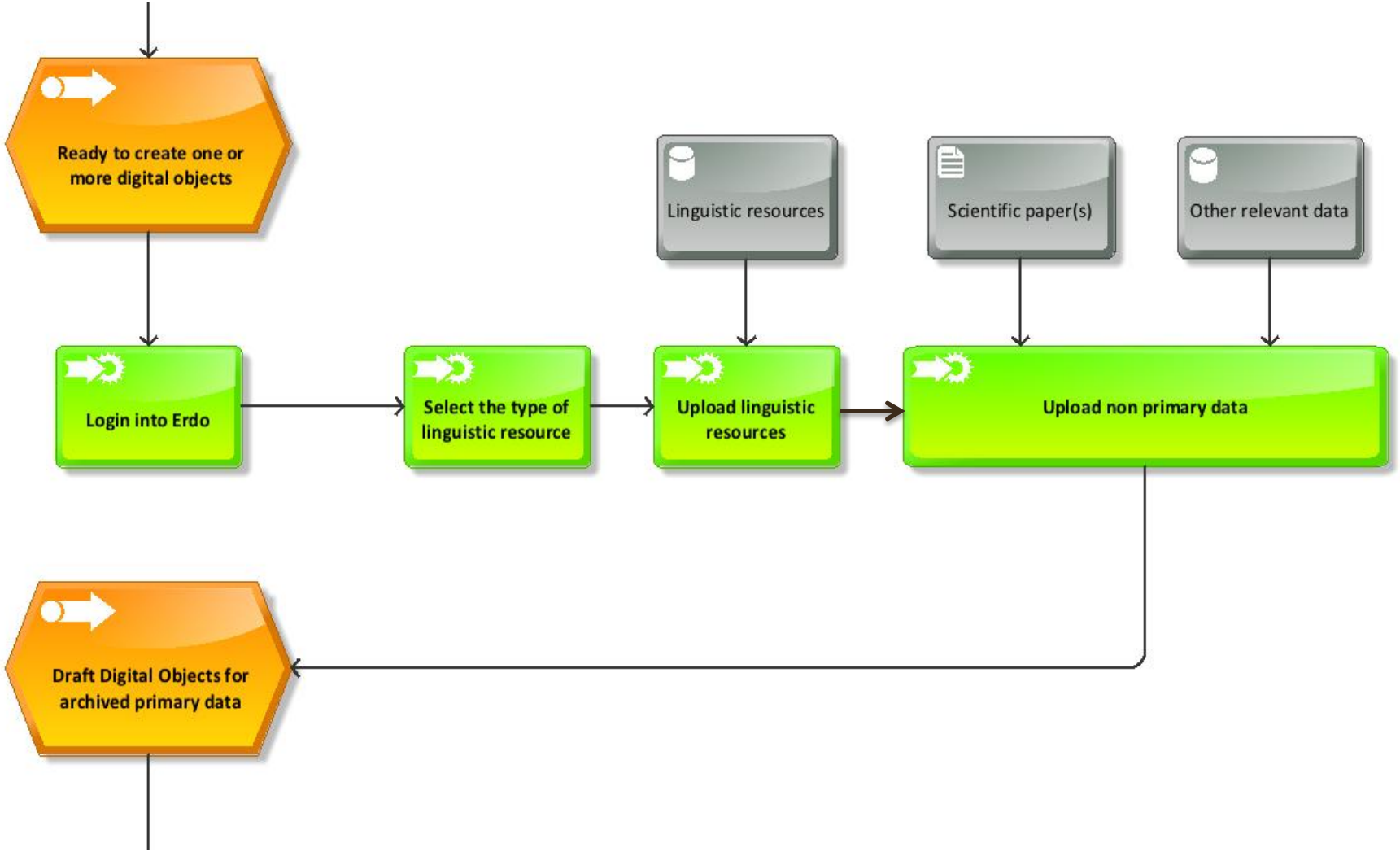
Individual IDs for resources in these cases:

- 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."

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Phase 2: Entering the archive (data upload)



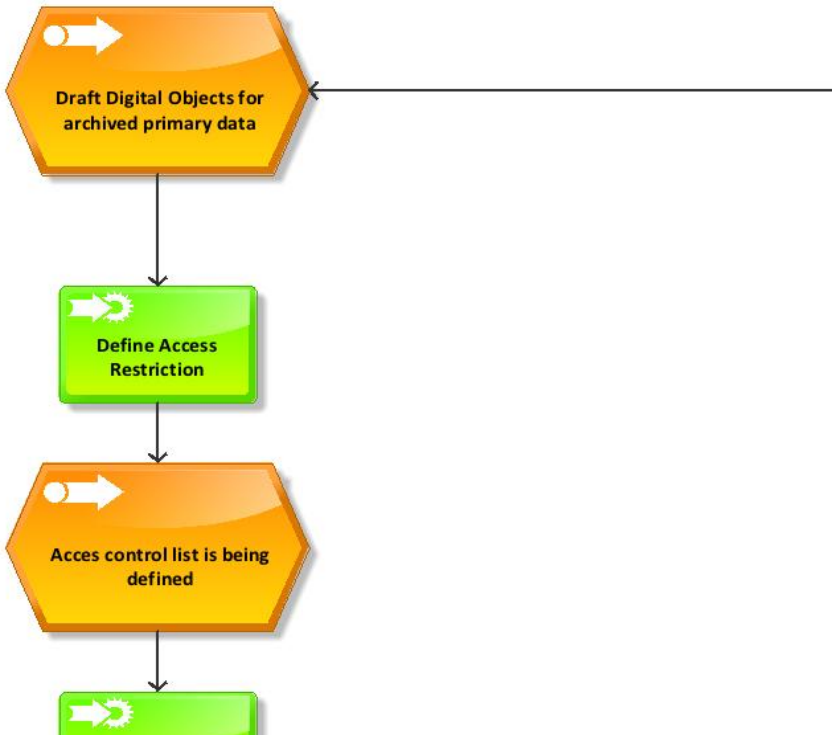
Documentation beyond the primary data

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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Phase 2: Entering the archive (access control)

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

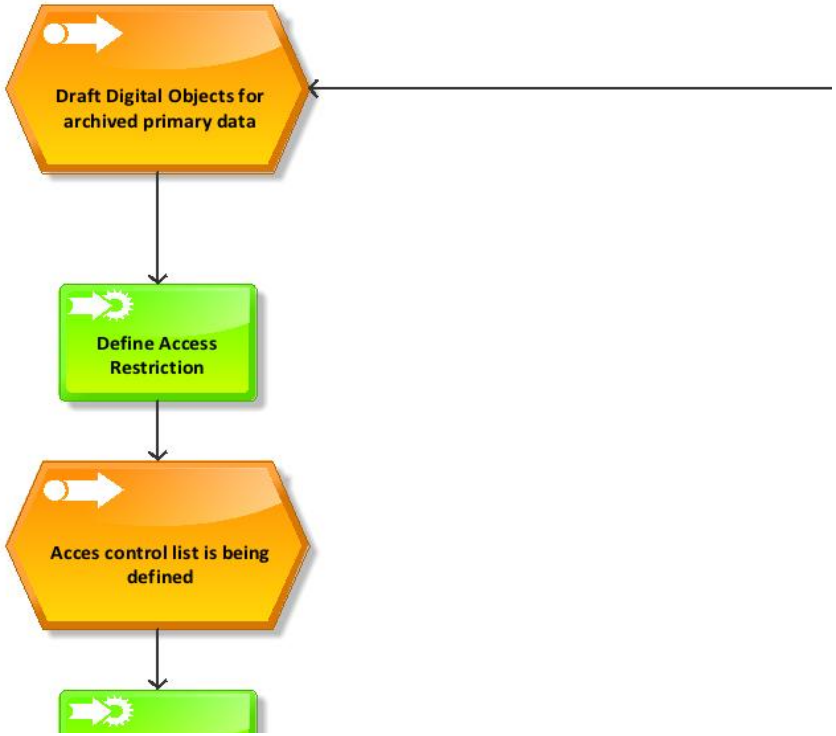


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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



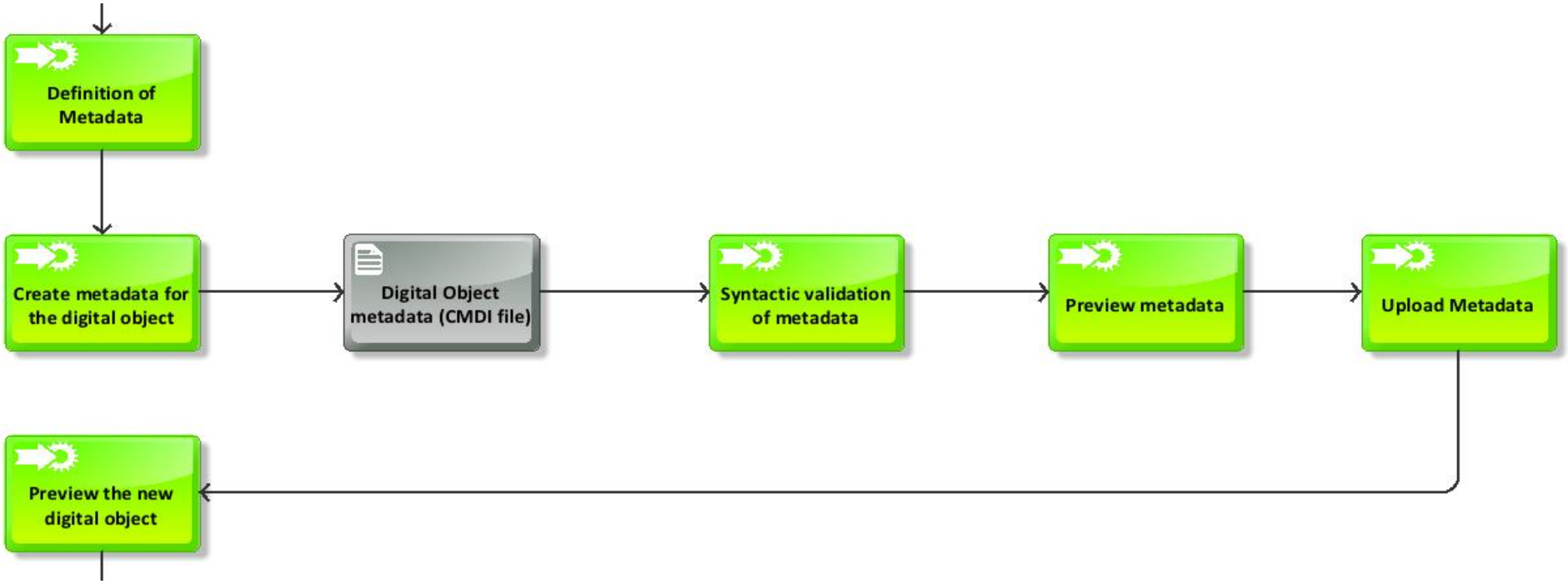
Editing and deleting persistent/archived data?

- 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+,*

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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 | Project | Publications | Creation | DocumentationLanguages | Access
LexiconContext | SizeInfo | TextTechnical | Save edited file

GeneralInfo

ResourceName
ResourceName

ResourceTitle
ResourceTitle in: GermaNet: Ein lexikalisch-semantisches Wortnetz (German)

ResourceClass
ResourceClass: Lexicon

Version
Version: 6.0

LifeCycleStatus
LifeCycleStatus: development

StartYear
StartYear: 1997

CompletionYear
CompletionYear: 2011

PublicationDate
PublicationDate: 1997-01-01

LastUpdate
LastUpdate: 2011-04-01

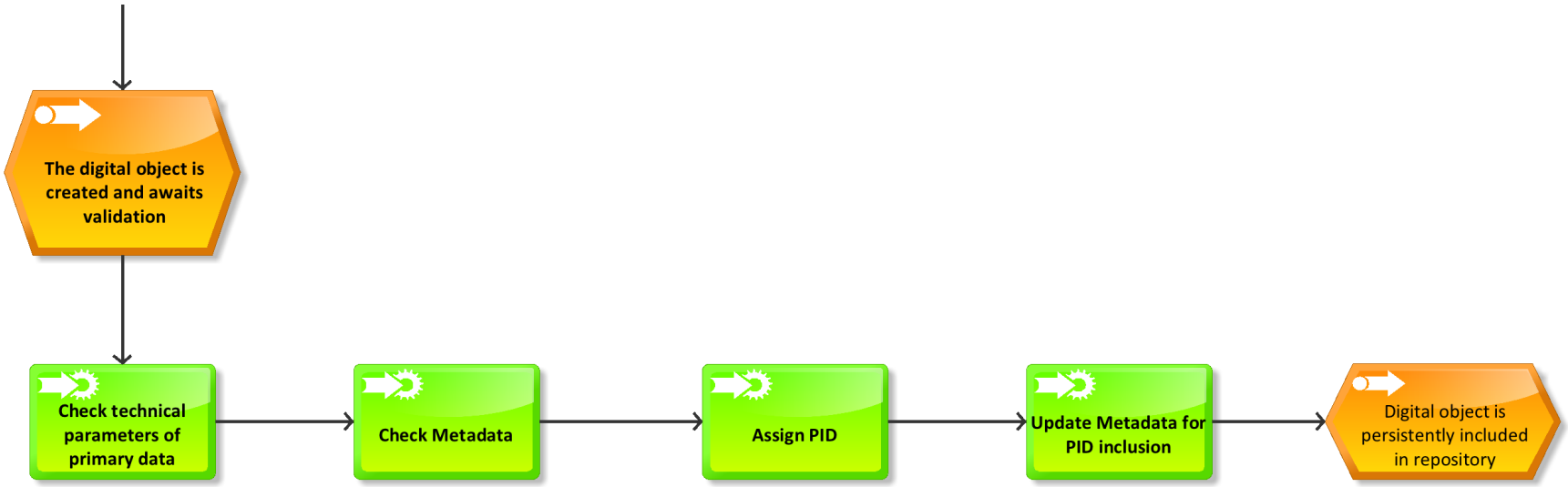
TimeCoverage

See http://www.isocat.org

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

http://www.sfs.uni-tuebingen.de/nalida/proforma/forms/ProFormA/15261.xhtml

Phase 3: Archivists take over



Checking the data

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

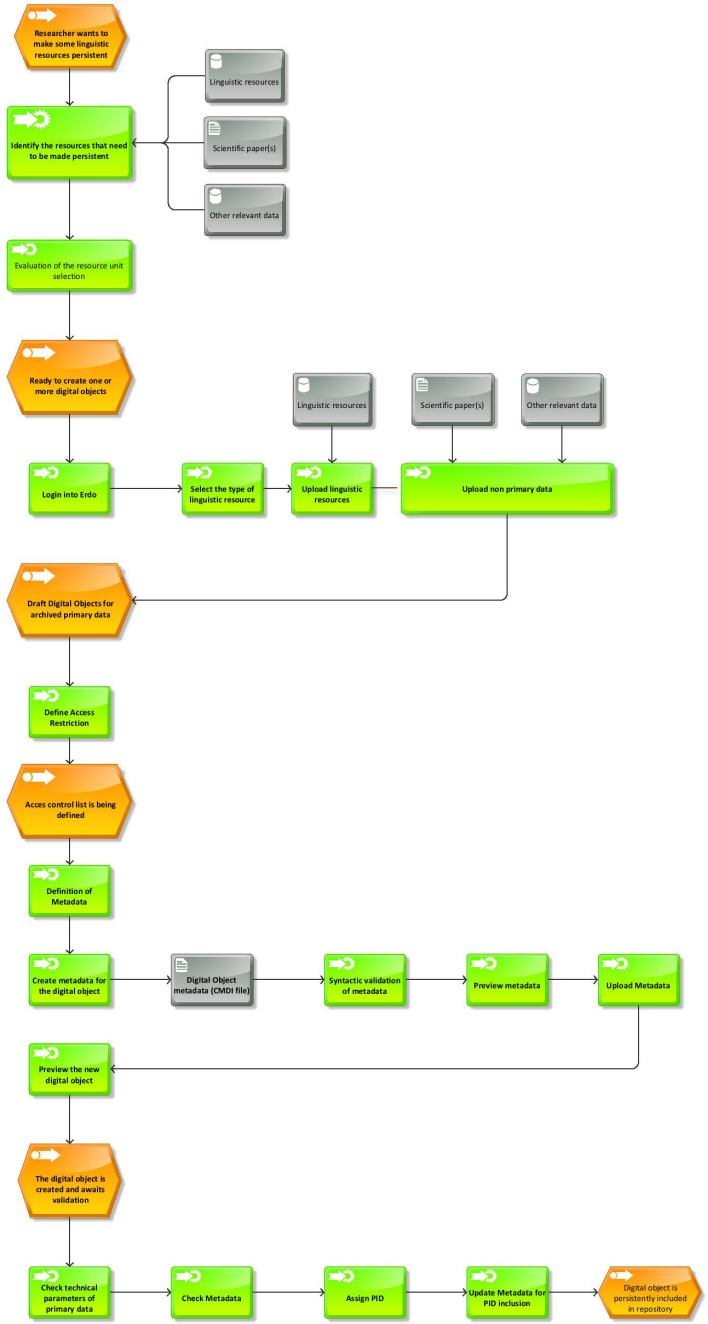
Which PID schema?

- 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?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Whole workflow

No exceptions, revisions and options shown



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Procedure for CMDI Component Creation

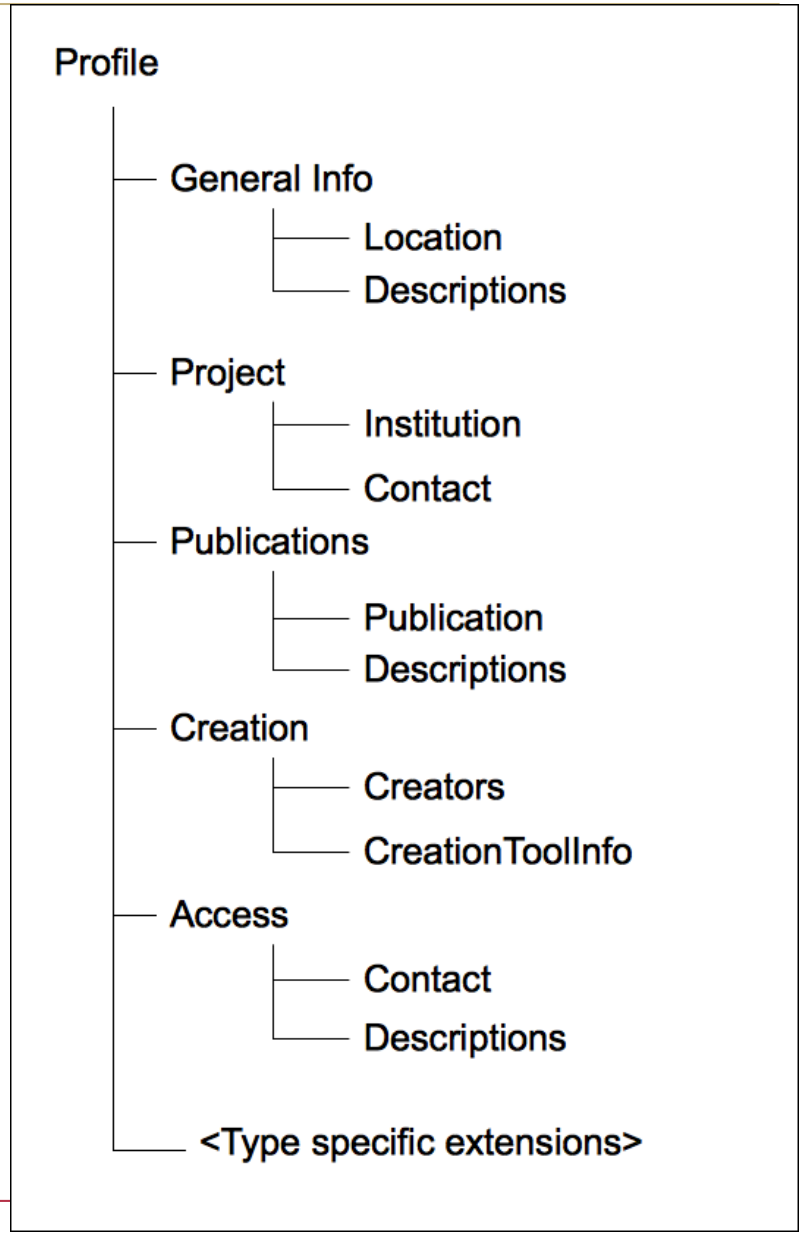
- Identifying type of resource and user group
- Reuse of components
- Recycling of components
- Creating new components
- Selecting, modifying data categories

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

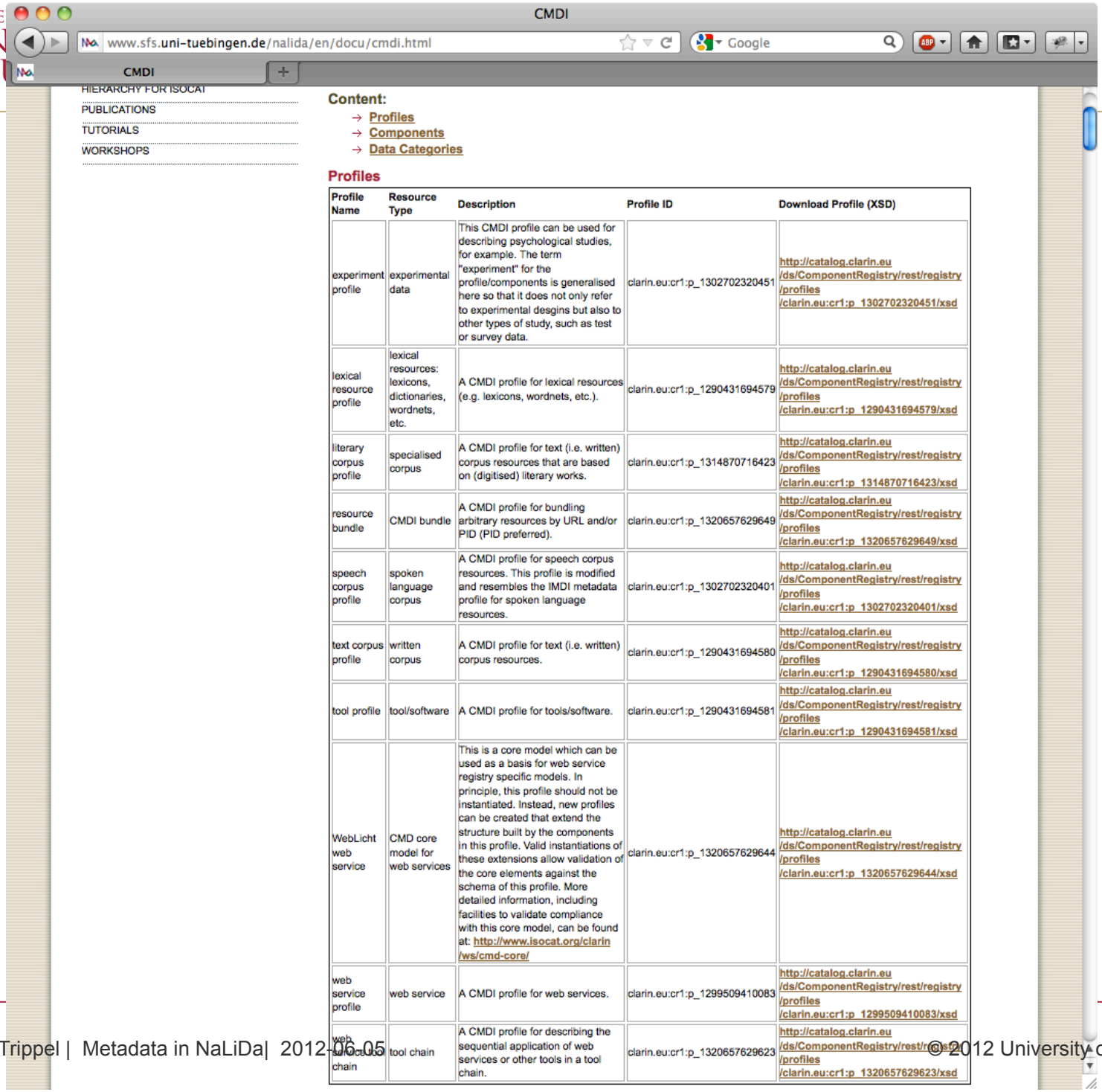


Recycling of Components

- 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

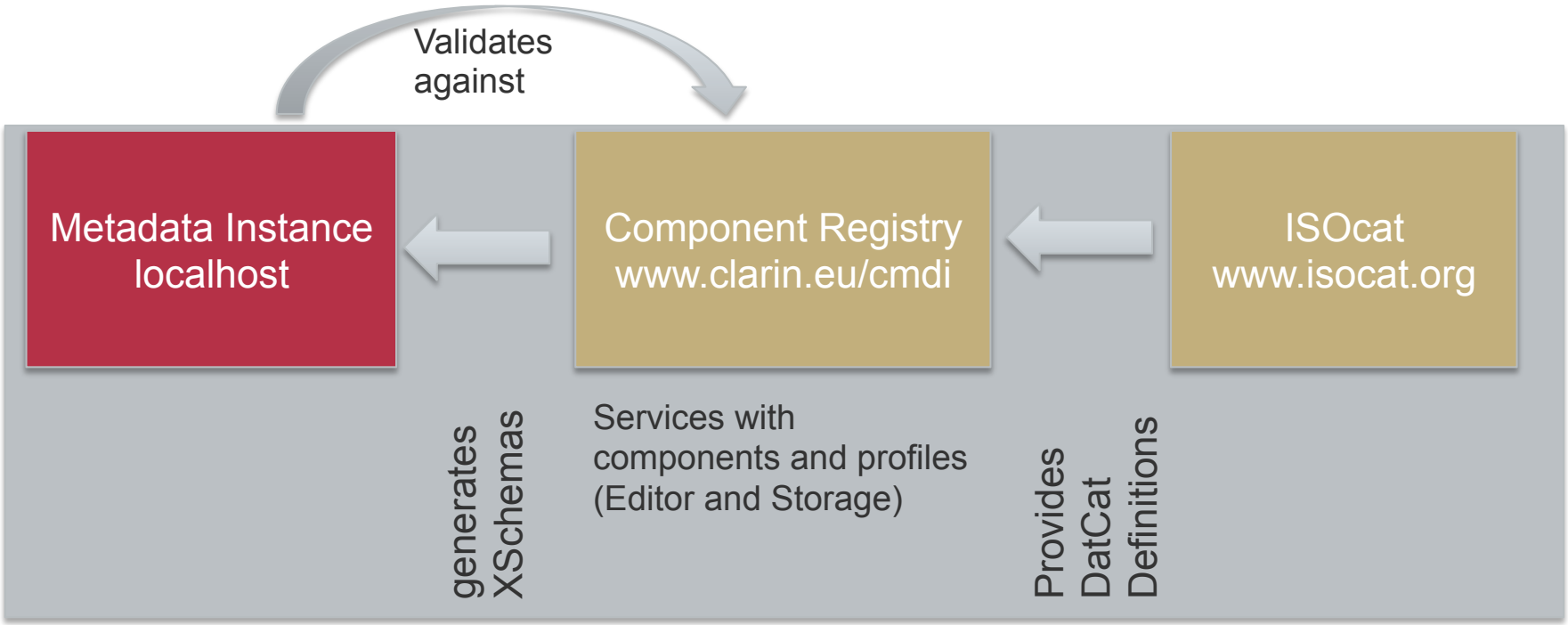


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

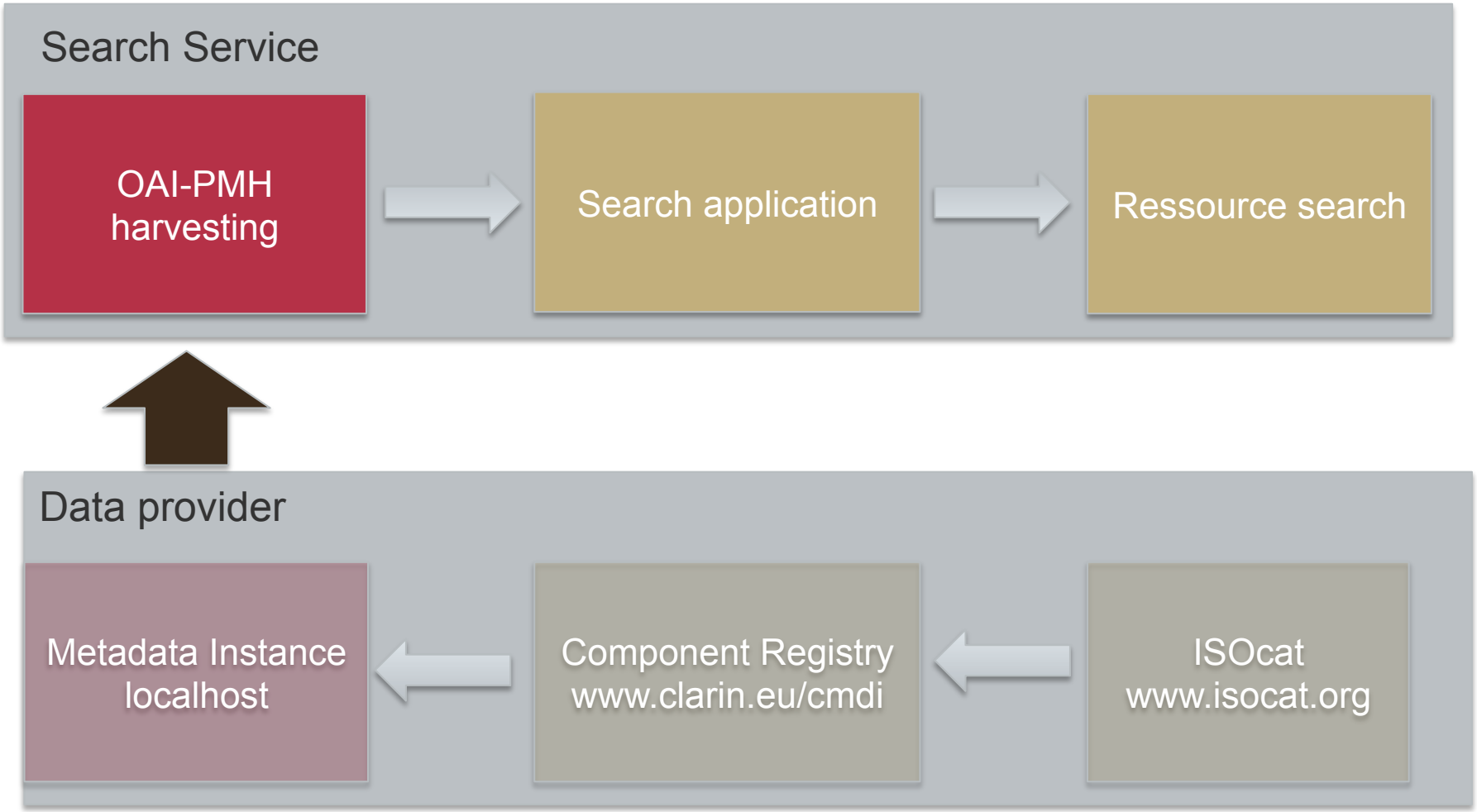
What is this all good for?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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



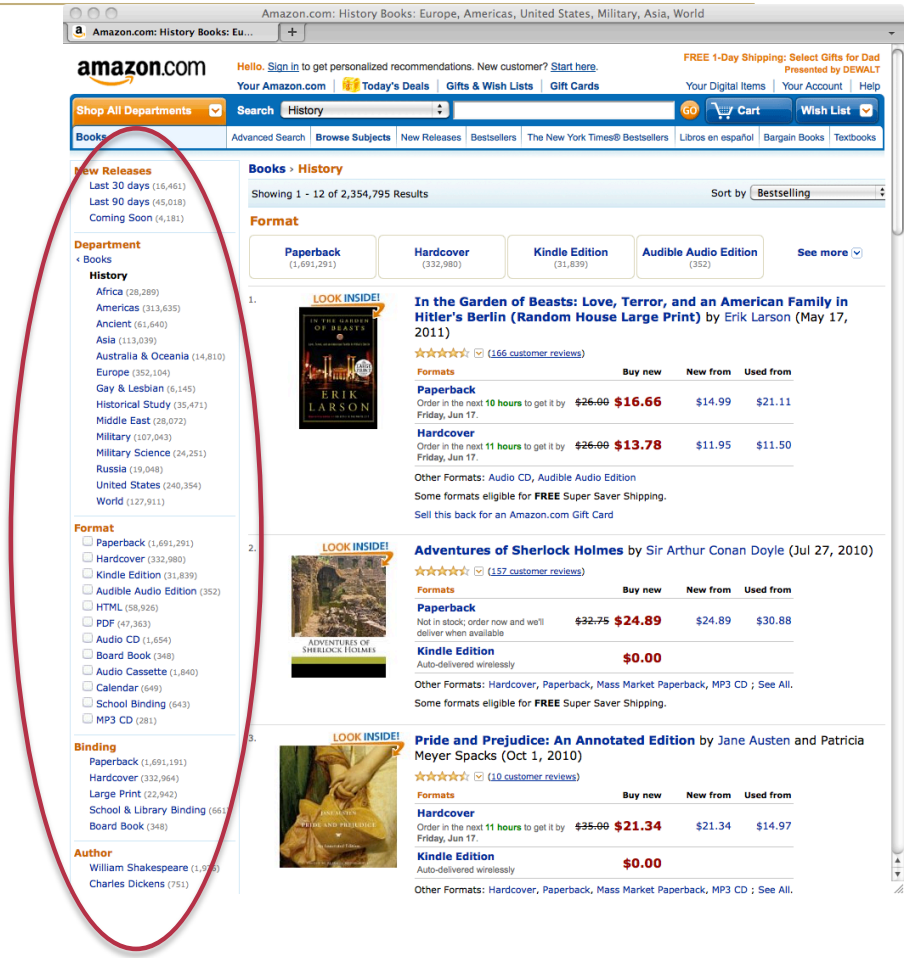
CMDI to search



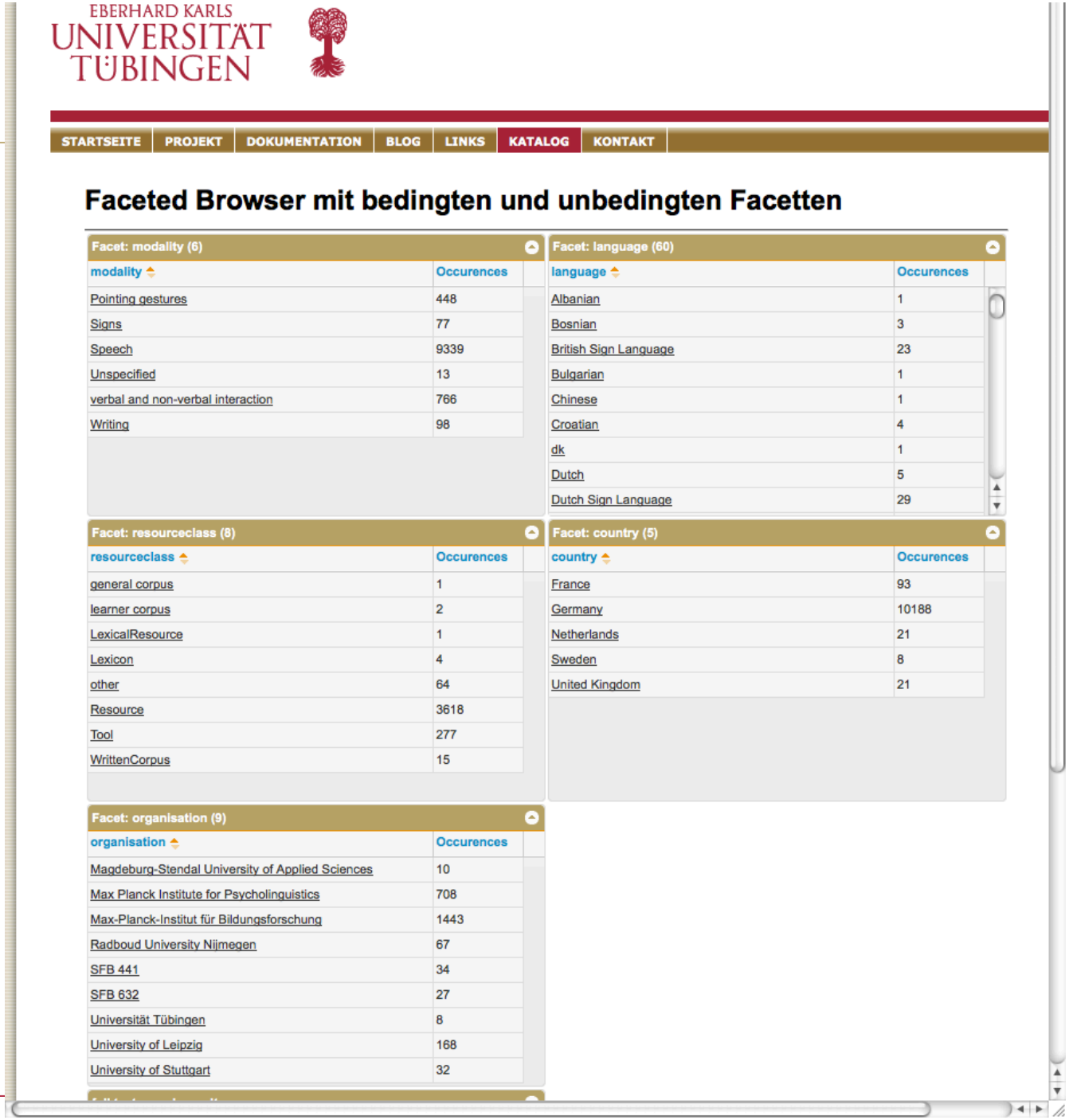
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

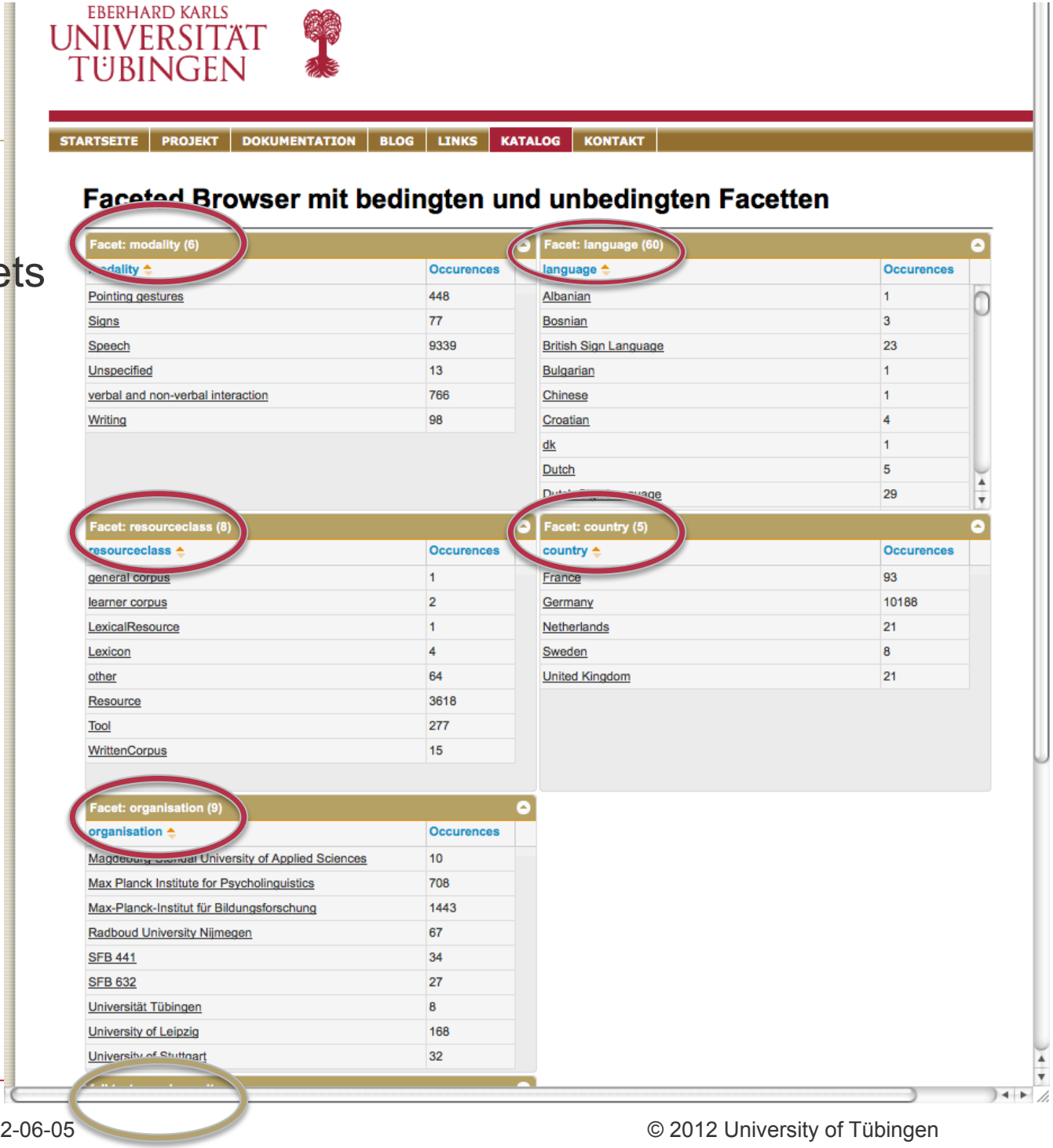


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

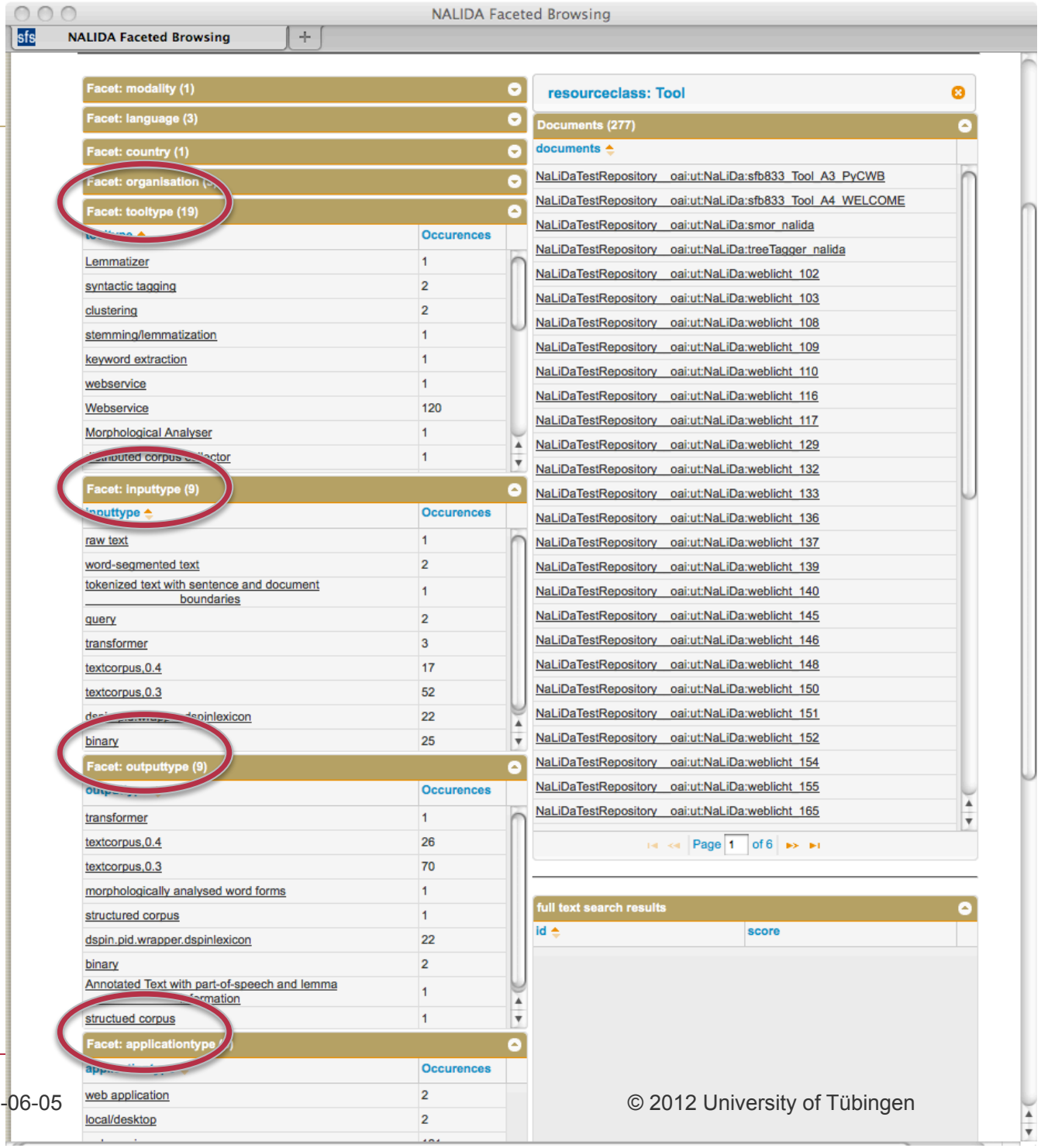
- Unconditional facets
  - Modality
  - Language
  - Resourceclass
  - Country
  - Organisation
- Special: Full text search



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

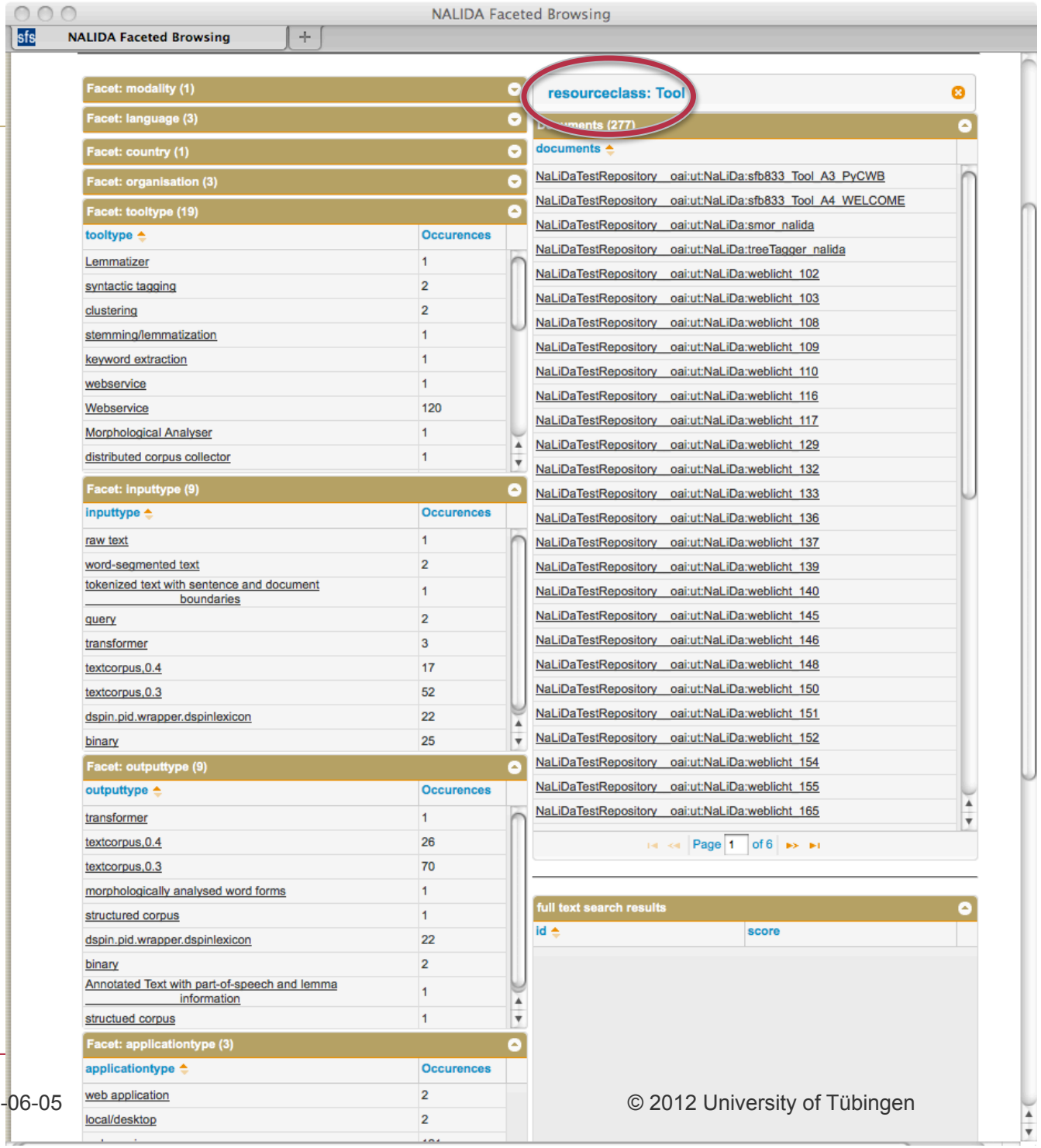
Conditional Facets

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



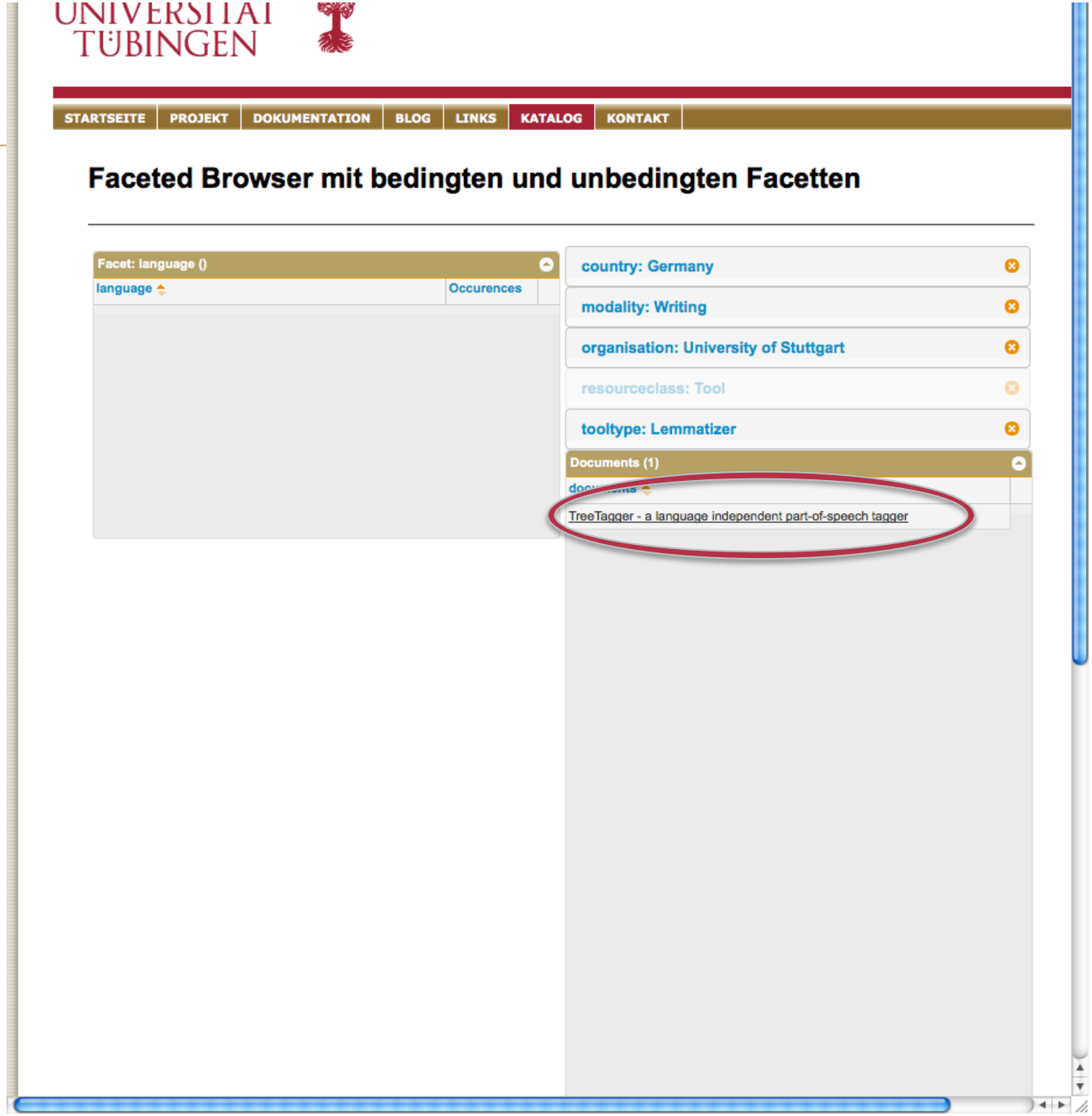
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Previously selected facets with value
- List of result set



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

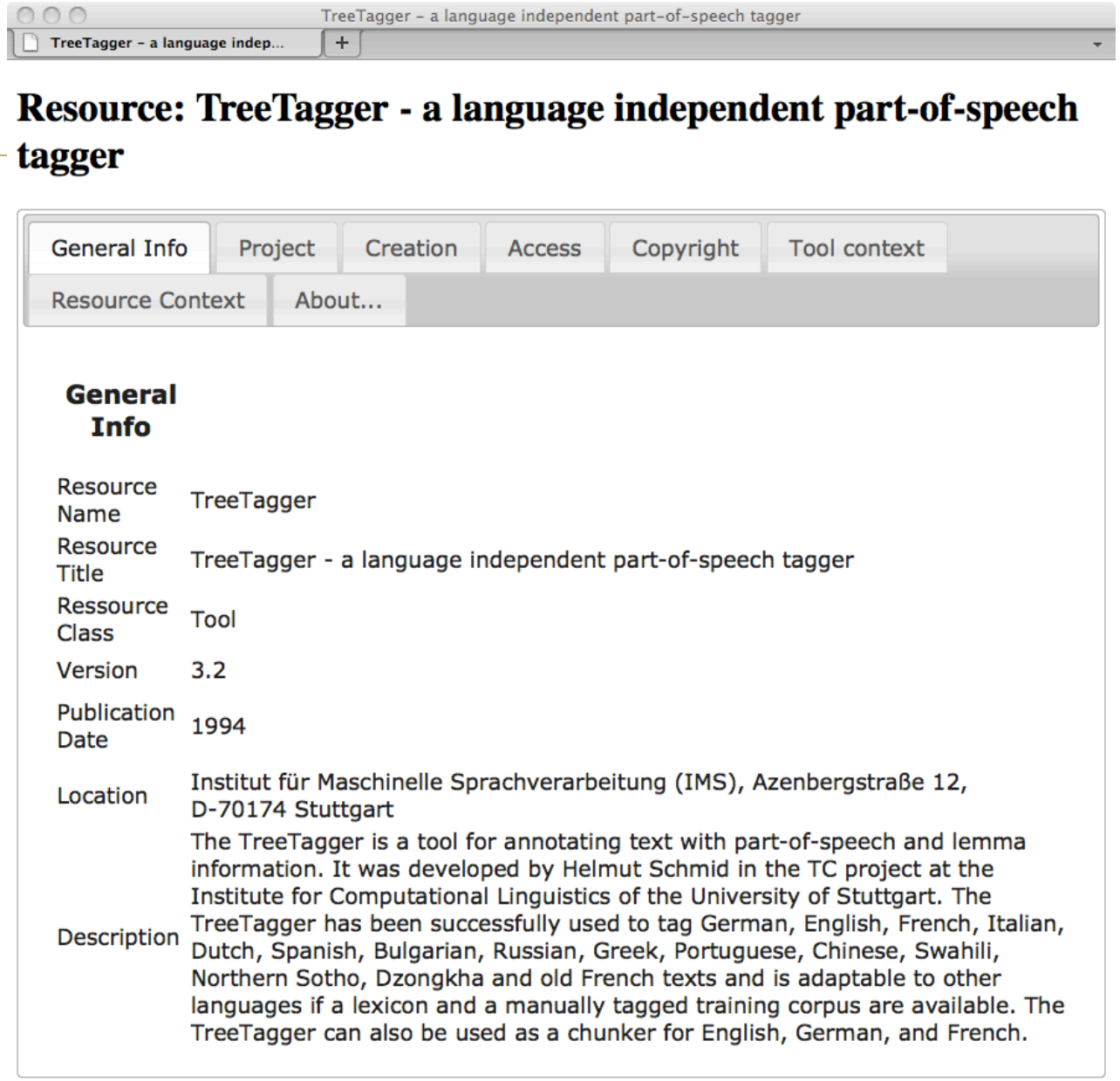
After the selection of facets



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Resulting Info

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



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 | Project | Publications | Creation | DocumentationLanguages | Access
LexiconContext | SizeInfo | TextTechnical | Save edited file

GeneralInfo

ResourceName
ResourceName

ResourceTitle
ResourceTitle in: GermaNet: Ein lexikalisch-semantisches Wortnetz (German)

ResourceClass
ResourceClass: Lexicon

Version
Version: 6.0

LifeCycleStatus
LifeCycleStatus: development

StartYear
StartYear: 1997

CompletionYear
CompletionYear: 2011

PublicationDate
PublicationDate: 1997-01-01

LastUpdate
LastUpdate: 2011-04-01

TimeCoverage

See http://www.isocat.org

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

http://www.sfs.uni-tuebingen.de/nalida/proforma/forms/ProFormA/15261.xhtml

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

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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/