Survey of Resources and Tools contributed by CLARIN ERIC members

Erhard Hinrichs
Thorsten Trippel

The final goal

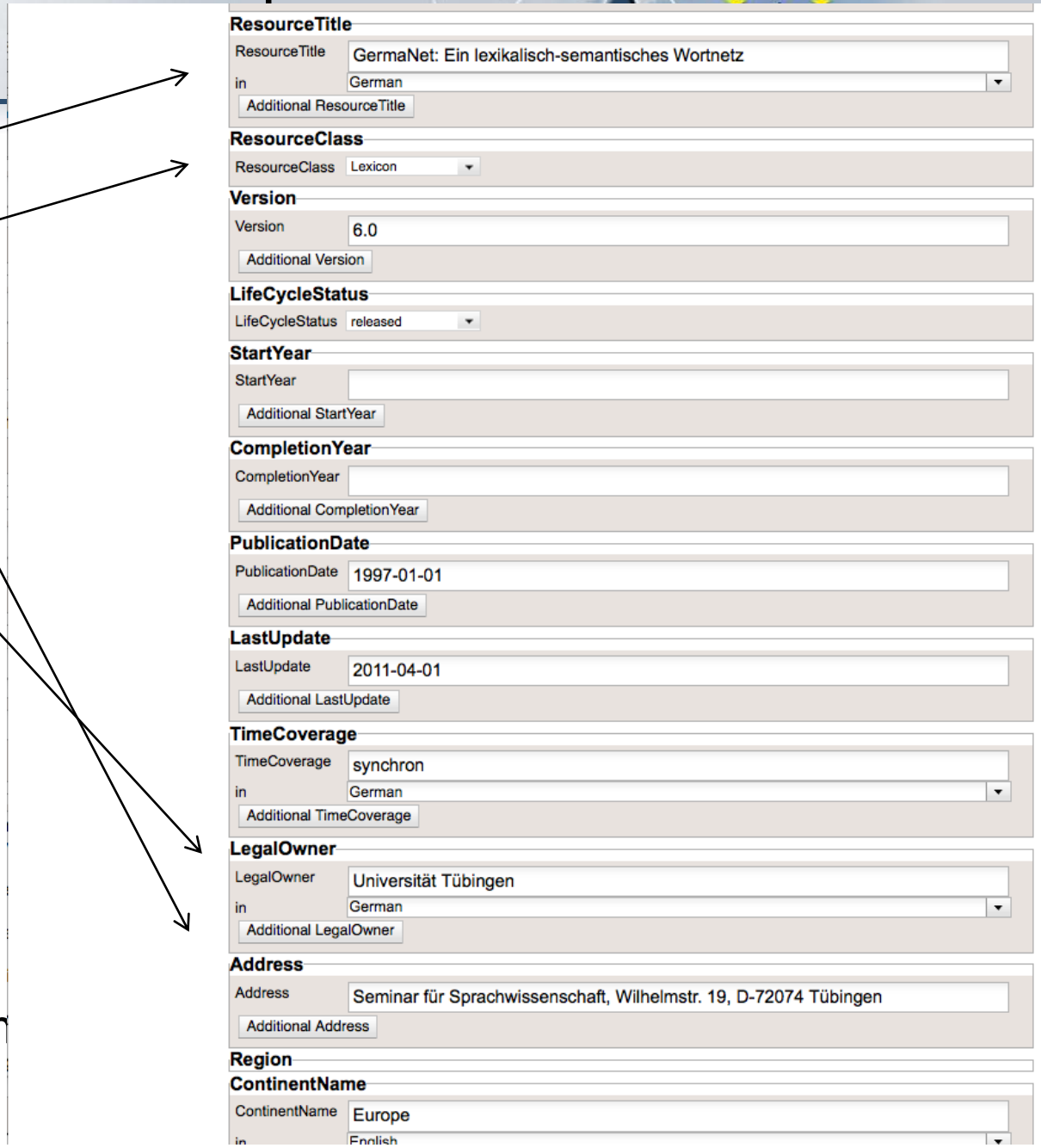
- CLARIN centers share their resource description for machine-to-machine communication
- Resource descriptions contain all information required by tools to reuse resources
- Descriptions are available in a common framework for interpretability
- Language Resources (LRs) and Tools (LRTs) are shared and reused according to researchers needs

Integration: Milestones for metadata provisioning, metadata harvesting

- M2-1
  - List of different types of resources and tools
    - Offered by each member country
    - Contained in the Annexes of the CLARIN Agreements
  - List of existing metadata for LRTs.
- M2-2
  - Convert metadata into CMDI
  - Use relevant tools for metadata
- M2-3
  - Each country: Establish a repository for resources and tools (Type A/B centre)
  - Implement an OAI-PMH provider

Information on Resources required for Milestone 2-1

- Resource Title
- Resource Type
- Contact address
- Publication date
- Owner
- Description
- Size (with units!)
- Access information
  - Restrictions
  - Contact for getting access
- Additional information as relevant



Required: ∀LRTs and ∀countries ∃structured metadata

- Resource Title
- Resource Type
- Contact address
- Publication date
- Owner
- Description
- Size (with units!)
- Access information
  - Restrictions
  - Contact for getting access
- Additional information as relevant

Provided for each LRT by the data provider and compiled for each country by the national coordinators

Structured forms of metadata

- Dublin Core (DC)
- OLAC
- TEI
- IMDI
- Relational Database

Structured formats of metadata

- Dublin Core (DC)
- OLAC
- TEI
- IMDI
- Relational Database

Requirements in CLARIN: CMDI

Structured formats of metadata

- Dublin Core (DC)
- OLAC
- TEI
- IMDI
- Relational Database
- CMDI

Providing over OAI-PMH: CLARIN A/B Centers

Where we are: Milestone 2-1

- List of different types of resources and tools
  - Offered by each member country
  - Contained in the Annexes of the CLARIN Agreements
- List of existing metadata for LRTs.

Provided only by some members

Types of LRTs Contained in Annexes of CLARIN Agreements

- Annex 2 National contributions to be coordinated at the national level
  - Annex 2-A: Resources that are finished and usable by CLARIN partners
    - "National Work Programme, enhancements and improvements to existing CLARIN resources, tools and services."
  - Annex 2-B: Resources to be created within the CLARIN context or converted to CLARIN standards and afterwards available to
    - "National Work Programme, creation of new resources, tools and services for CLARIN."

Current Types of Resources

- Corpora (Text)
  - Treebanks
- Corpora (Audio)
- Corpora (Video)
- Lexical Resources
  - WordNet type
  - FrameNet type
  - Dictionaries
- Webservices
- Desktop- and Web applications

Example provision of resources

- Examples provided by various CLARIN partners
- ANNEX 2 A: finished, usable LRTs
- Large lists of LRs:
  - Provide examples (typical types)
  - Full lists if not available via OAI-PMH
  - (verbose) Description of the LRs collection
  - If available via OAI-PMH:
    - Examples and
    - OAI-PMH URL
- If LRs have a website: URL in the description of the LR

Resource: German newspaper corpus +CMDI +OAI-PMH +PID

OAI-PMH: http://clarinoai.informatik.uni-leipzig.de:8080/oaiprovider/?verb=ListRecords&metadataPrefix=cmdi

Pid: http://hdl.handle.net/11858/00-229C-0000-0003-1750-7

Name: Leipzig Corpora Collection

Url: http://corpora.informatik.uni-leipzig.de/

Pid: http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

CMDI: http://clarinws.informatik.uni-leipzig.de:8080/cmdi/http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

Name: German newspaper corpus - 10.000 sentences from 2008 (deu_news_2008_10K)

Resource: German newspaper corpus +CMDI +OAI-PMH +PID

OAI-PMH address

OAI-PMH: http://clarinoai.informatik.uni-leipzig.de:8080/oaiprovider/?verb=ListRecords&metadataPrefix=cmdi

Pid: http://hdl.handle.net/11858/00-229C-0000-0003-1750-7

Name: Leipzig Corpora Collection

Url: http://corpora.informatik.uni-leipzig.de/

Pid: http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

CMDI: http://clarinws.informatik.uni-leipzig.de:8080/cmdi/http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

Name: German newspaper corpus - 10.000 sentences from 2008 (deu_news_2008_10K)

Resource: German newspaper corpus +CMDI +OAI-PMH +PID

Resource
PID

OAI-PMH: http://clarinoai.informatik.uni-leipzig.de:8080/oaiprovider/?verb=ListRecords&metadataPrefix=cmdi

Pid: http://hdl.handle.net/11858/00-229C-0000-0003-1750-7

Name: Leipzig Corpora Collection

Url: http://corpora.informatik.uni-leipzig.de/

Pid: http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

CMDI: http://clarinws.informatik.uni-leipzig.de:8080/cmdi/http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

Name: German newspaper corpus - 10.000 sentences from 2008 (deu_news_2008_10K)

Resource: German newspaper corp +CMDI +OAI-PMH +PID

Name of the Resource

OAI-PMH: http://clarinoai.informatik.uni-leipzig.de:8080/oaiprovider/?verb=ListRecords&metadataPrefix=cmdi

Pid: http://hdl.handle.net/11858/00-229C-0000-0003-1750-7

Name: Leipzig Corpora Collection

Url: http://corpora.informatik.uni-leipzig.de/

Pid: http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

CMDI: http://clarinws.informatik.uni-leipzig.de:8080/cmdi/http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

Name: German newspaper corpus - 10.000 sentences from 2008 (deu_news_2008_10K)

Resource: German +CMDI +OAI-PMH -

URL for further information (human readable)

OAI-PMH: http://clarinoai.in[illegible].uni-leipzig.de:8080/oai[illegible]der/?verb=ListRecords&metadataPrefix=cmdi

Pid: http://hdl.handle.net/11858/00-229C-0000-0003-1750-7

Name: Leipzig Corpora Collection

Url: http://corpora.informatik.uni-leipzig.de/

Pid: http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

CMDI: http://clarinws.informatik.uni-leipzig.de:8080/cmdi/http://hdl.handle.net/11858/00-229C-0000-0001-B06F-3

Name: German newspaper corpus - 10.000 sentences from 2008 (deu_news_2008_10K)

Resource: FOLK –CMDI –OAI-PMH + Website

Name: Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)
Resource type: spoken language corpus

Resource info: 52 recordings (1d 10h 24'), 50 utterances, 82 transcriptions, 124 speakers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

Resource: FOLK –CMDI –OAI-PMH + Web

Name of the resource

Name: Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)

Resource type: spoken language corpus

Resource info: 52 recordings (1d 10h 24'), 50 utterances, 82 transcriptions, 124 speakers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

Resource: FOLK –CMDI –OAI-PMH + Web

Type of resource

Name: Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK)

Resource type: spoken language corpus

Resource info: 52 recordings (1d 10h 24'), 50 utterances, 82 transcriptions, 124 speakers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

Resource: FOLK –CMDI –OAI-PMH + Web

Description of the resource

Name: Forschungs- und Lehrk… gesprochenes Deutsch (FOLK)

Resource type: spok… language corpus

Resource info: 52 recordings (1d 10h 24'), 50 utterances, 82 transcriptions, 124 speakers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

Resource: FOLK –CMDI –OAI-PMH + Web

URL with further description

Name: Forschungs- und Lehrkor… …prochenes Deutsch (FOLK)

Resource type: spoken la… …ge corpus

Resource info: 52 r… …ordings (1d 10h 24'), 50 utterances, 82 transcriptions, 124 speakers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

Resource: FOLK –CMDI –OAI-PMH + Web

Contact information

Name: Forschungs- und Lehrkorp… …chenes Deutsch (FOLK)

Resource type: spoken langua… …pus

Resource info: 52 recordin… …d 10h 24'), 50 utterances, 82 transcriptions, 124 sp…kers

URL: http://agd.ids-mannheim.de/html/folk.shtml

Contact person(s): Thomas Schmidt < thomas.schmidt@ids-mannheim.de>

List of resource and tool information in Annex 2A and B

- List Resources with
  - Title
  - Contact
  - Owner
  - Type of resource
  - Description
  - Available at which conditions
  - Address of contact
  - Additional information according to the type of resource
- CMDI files have to contain all of this
  - If CMDI files already exist, name of the resource, contact information and URL of CMDI file is sufficient

What is next: Integration Milestone 2-2

- Convert metadata into CMDI
- Use relevant tools for metadata

Convert Metadata into CMDI…

```
<?xml version="1.0" encoding="UTF-8"?>
<CMD xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.clarin.eu/cmd/" CMDVersion="1.1"
xsi:schemaLocation="http://www.clarin.eu/cmd/
http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1290431694579/xsd">
  <Header>
    <MdCreator>Reinhild Barkey</MdCreator>
    <MdCreationDate>2010-11-11</MdCreationDate>
    <MdSelfLink>http://hdl.handle.net/11858/00-1778-0000-0005-896E-B</MdSelfLink>
    <MdProfile>clarin.eu:cr1:p_1290431694579</MdProfile>
    <MdCollectionDisplayName>Tübingen Language Resources</MdCollectionDisplayName>
  </Header>
  <Resources>
    <ResourceProxyList>
      <ResourceProxy id="resprox1">
        <ResourceType>Resource</ResourceType>
        <ResourceRef>http://www.sfs.uni-tuebingen.de/GermaNet</ResourceRef>
      </ResourceProxy>
    </ResourceProxyList>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <LexicalResourceProfile>
      <GeneralInfo>
        <ResourceName>GermaNet</ResourceName>
        <ResourceTitle xml:lang="de">GermaNet: Ein lexikalisch-semantisches Wortnetz</ResourceTitle>
        <ResourceClass>Lexicon</ResourceClass>
        <Version>6.0</Version>
        <LifeCycleStatus>released</LifeCycleStatus>
    ….
```

…using appropriate tools

```
<?xml version="1.0" encoding="UTF-8"?>
<CMD xmlns:xsi="http://www.w3.org/2001/XM
xsi:schemaLocation="http://www.clarin.eu/cmd
http://catalog.clarin.eu/ds/ComponentRegistry/
  <Header>
    <MdCreator>Reinhild Barkey</MdCreator>
    <MdCreationDate>2010-11-11</MdCreatio
    <MdSelfLink>http://hdl.handle.net/11858/(
    <MdProfile>clarin.eu:cr1:p_129043169457
    <MdCollectionDisplayName>Tübingen Lang
  </Header>
  <Resources>
    <ResourceProxyList>
      <ResourceProxy id="resprox1">
        <ResourceType>Resource</ResourceTy
        <ResourceRef>http://www.sfs.uni-tueb
      </ResourceProxy>
    </ResourceProxyList>
    <JournalFileProxyList/>
    <ResourceRelationList/>
  </Resources>
  <Components>
    <LexicalResourceProfile>
      <GeneralInfo>
        <ResourceName>GermaNet</Resource
        <ResourceTitle xml:lang="de">GermaN
        <ResourceClass>Lexicon</ResourceCla
        <Version>6.0</Version>
        <LifeCycleStatus>released</LifeCycleSt
    ....
```

ResourceTitle
ResourceTitle: GermaNet: Ein lexikalisch-semantisches Wortnetz
in: German
Additional ResourceTitle

ResourceClass
ResourceClass: Lexicon

Version
Version: 6.0
Additional Version

LifeCycleStatus
LifeCycleStatus: released

StartYear
StartYear:
Additional StartYear

CompletionYear
CompletionYear:
Additional CompletionYear

PublicationDate
PublicationDate: 1997-01-01
Additional PublicationDate

LastUpdate
LastUpdate: 2011-04-01
Additional LastUpdate

TimeCoverage
TimeCoverage: synchron
in: German
Additional TimeCoverage

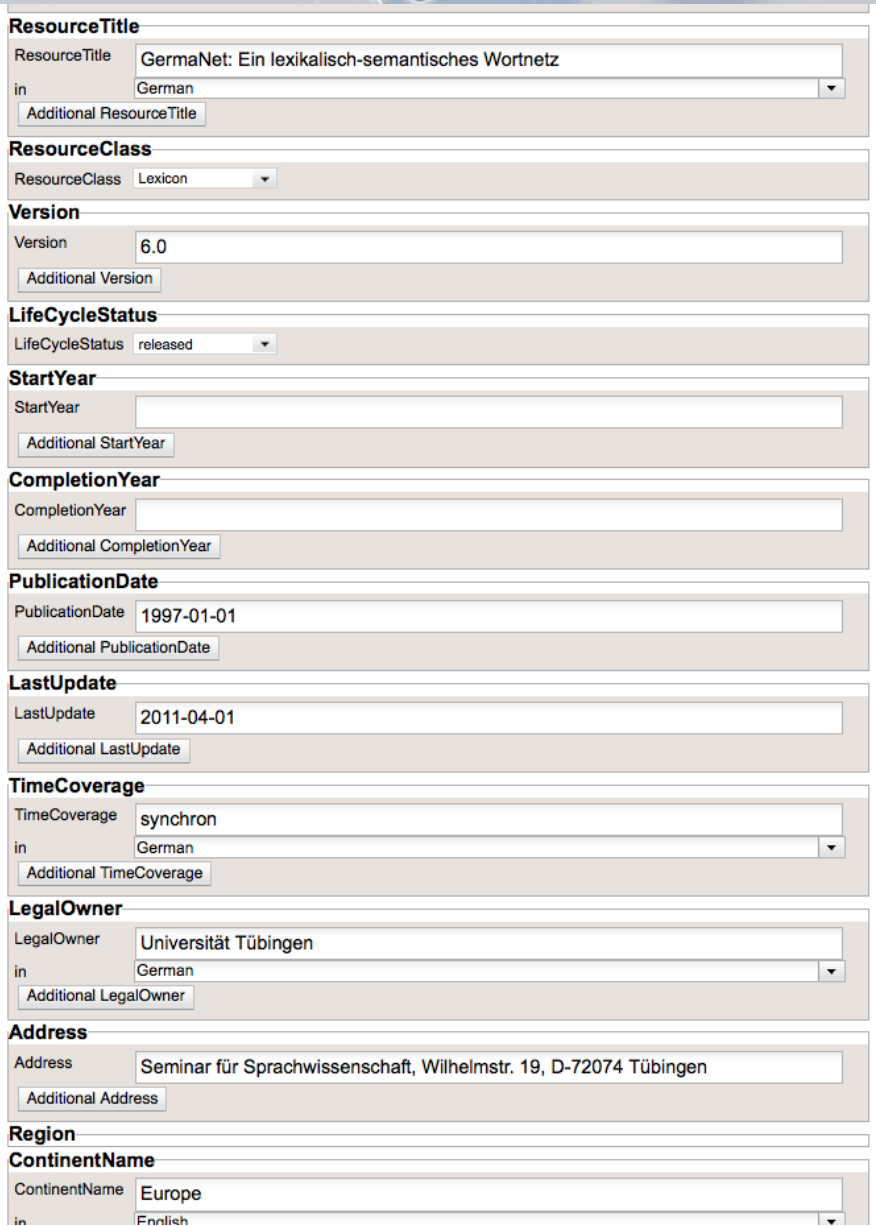
LegalOwner
LegalOwner: Universität Tübingen
in: German
Additional LegalOwner

Address
Address: Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen
Additional Address

Region
ContinentName
ContinentName: Europe
in: English

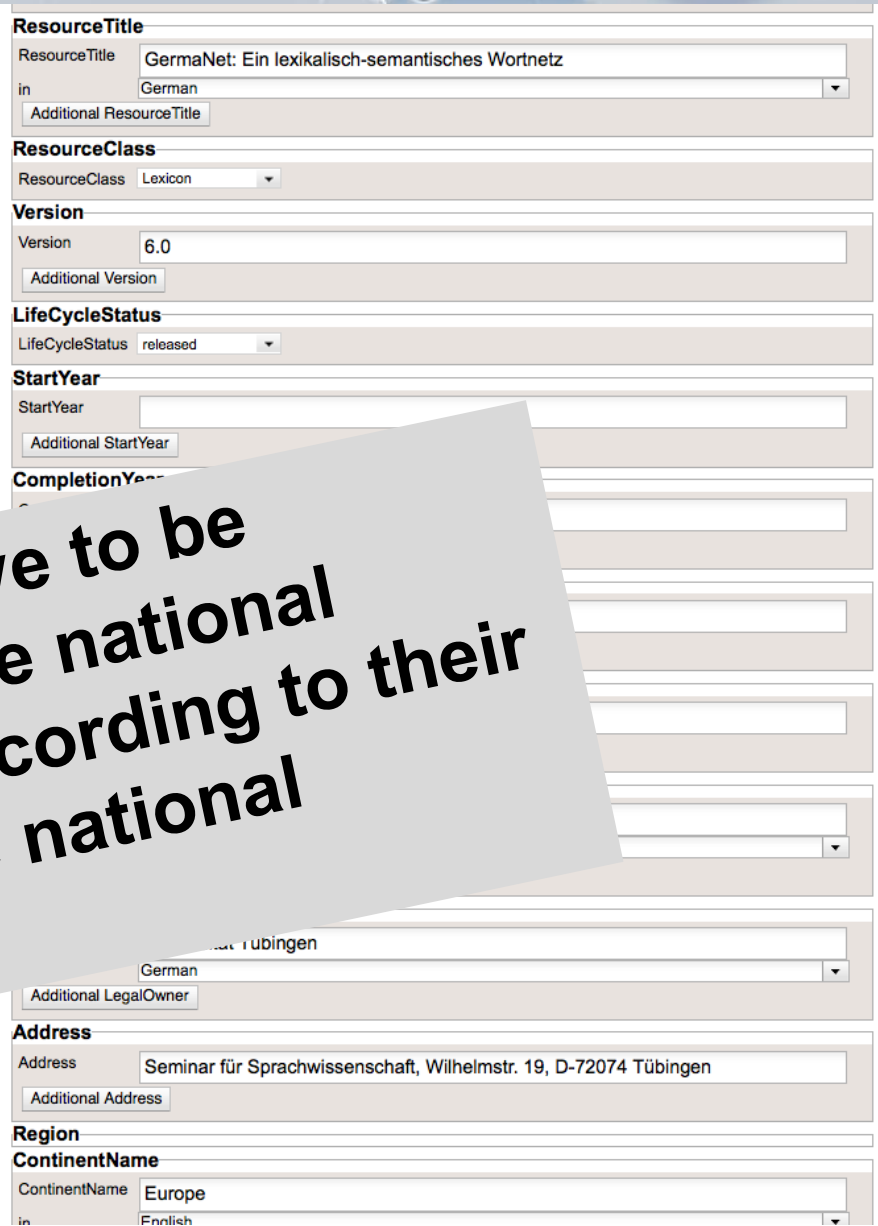
…using appropriate tools

- Creation of new metadata: Arbil, ProFormA, …
- Converting legacy metadata
  - Excel2CMDI
  - XSLT transformation OLAC, IMDI, …
- Export filter for relational databases



…using appropriate tools

- Creation of new metadata: Arbil, ProFormA, …
- Converting legacy metadata
  - Excel2CMDI
  - XSLT transfo
-



Some tools will still have to be adjusted/created for the national infrastructures and according to their requirements – by the national CLARIN centers

The future: Integration Milestone 2-3

- Each country: Establish a repository for resources and tools (Type A/B center)
- Implement an OAI-PMH provider
  - Provides access to the CMDI metadata
  - Resources with PID
- Requires:
  - CMDI metadata for language resources by the data providers
  - PID framework to be used for resources

Summary

- Final goal of integration with metadata:
  - CMDI via OAI-PMH
  - Milestones for metadata creation and provision
  - Starting point: information in CLARIN-Agreements on the resources
  - All data countries: provide LRT information in structured form for the agreements
- Current state with Milestone 2-1
  - List of some resource types
  - Example resources from agreements
- Showed the roadmap towards
  - CMDI for each LRT to be created by resource providers (M 2-2)
  - OAI-PMH integration for CLARIN A/B centers (M 2-3)

THANK YOU