



# Mapping METS and Dublin Core to CMDI: Making Textbooks available in the CLARIN VLO

***Francesca Fallucchi* and Ernesto William De Luca**

Georg-Eckert-Institute for International Textbook Research

University of Rome Guglielmo Marconi

# Outline

- Motivation
- Why CLARIN, why CMDI?
- CMDI process at GEI
  - TEI To CMDI
  - **METS To CMDI**
  - DC To CMDI
- Conclusion

# The Georg-Eckert-Institute for International Textbook Research

- In 1953 Georg Eckert (1912-1974) founded an „international textbook institute“
- In 1975 the Georg-Eckert-Institute was established in Brunswick, Germany
- It is member of the Leibniz Association since 2011
- The institute conducts applied and multidisciplinary research into textbooks and educational media related to textbooks, informed primarily by history and cultural studies
- Research, Transfer and Infrastructures are closely connected



# The World Views Project

<http://gei-worldviews.gei.de/>

World Views is an open edition featuring:

- Digital images from textbooks
- Full, TEI-XML annotated text
- Translations
- Comments and Essays on authors, educational systems, historic events featured etc.
- Metadata in different formats

Objective: Make data as open, visible and re-usable as possible

The screenshot displays the WorldViews website interface. At the top, the logo 'WORLD VIEWS' is on the left, and the title 'WORLDVIEWS. THE WORLD IN TEXTBOOKS' is in the center. A navigation menu includes 'HOME', 'SOURCES', 'COMMENTS AND ESSAYS', 'HISTORY OF EDUCATION', 'TOPICS', and 'COLLECTIONS'. On the right, there is a 'Reading list' icon. The main content area shows a digital edition of a textbook page titled 'SYSTEMS OF MEANING'. The page number '16' and the title 'GEOGRAFIA DO BRASIL' are visible. The text on the page discusses the influence of past generations on national life. Below the text, there is a section titled 'Intersection for global transportation routes, Brazil, 1972'. On the right side of the page, there is a sidebar with a search bar and a list of topics: 'Structures of Power, Dominance and Freedom', 'Transnational and International Relations and Organisations', 'Constructions of Space and Belonging', 'Violent Conflict and Peace', 'Models of Society and Visions of the Future', and 'Heterogeneity of Societies'. At the bottom of the page, there is a section titled 'THE PROJECT' with a brief description of the project's goals and scope.

# The GEI-Digital Project

<http://gei-worldviews.gei.de/>

Gei-Digital is an open edition featuring:

- Metadata in different formats
- The MIK-Center in Berlin is conducting the digitalisation of the content using the Goobi open source system.

Objective: Make data as open, visible and re-usable as possible

**gei.digital**  
The Digital Textbook Library

Home Search Browse News About the project Project partners Terms of Use DE / EN

**Last imports**

You are here: Home

27. September 2019  
Für Mittelklassen  
Deutsches Lesebuch für Volksschulen  
Lesebuch für gewerbliche Fortbildungsschulen  
Deutsches Lesebuch für die evangelischen Volksschulen des Regierungsbezirks Cassel  
[Band 2, [Schülerband]]

**GEI-Digital visualized**

**German Textbook Collection visualized**

GEI-Digital visualized provides interactive access to the collection of historic textbooks from Germany. It visualises temporal and spatial dimensions of the collection which supply insights into developments on the historic textbook market and its actors in Germany. As well as being a visualisation tool, it acts as an intuitive search instrument which allows users to search for digital German textbooks by school subject, educational level, publisher and publisher location.

**Collections**

*x*  
*26* f. v. l. b. d. s.

**Reading primers, imperial Germany (99)**

# Virtual Language Observatory (VLO)

<https://vlo.clarin.eu/>

The data becomes visible and searchable in the VLO

The screenshot shows the CLARIN Virtual Language Observatory (VLO) website. At the top, there is a navigation bar with "Virtual Language Observatory", "Search", and "Help" links, along with the CLARIN logo. The main heading is "CLARIN Virtual Language Observatory" with a sub-heading "Welcome to the VLO!". Below this, a text block instructs users to use the search bar to find resources or to browse everything and use facets. There are two buttons: "See all records" and "Learn more". A search bar is located below the text, with a search icon on the right. Below the search bar, it says "Showing all 1677412 records" and "Results per page: 10". On the left side, there are four facets: "Language", "Collection", "Resource type", and "Modality", each with a dropdown arrow. On the right side, there is a pagination bar showing "1" as the current page. Below the pagination, there are two search results. The first is "EXMARaLDA Demo corpus" (Part of Hamburger Zentrum für Sprachkorpora (HZSK)), with a description: "A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental information; accentuation/stress; English translation; Standard German translation; German translation; English translation; code-switch". The second result is "The Hamburg MapTask Corpus (HAMATAC)".

Fallucchi Francesca

Mapping METS and Dublin Core to CMDI: Making Textbooks available in the CLARIN VLO



GEORG ECKERT  
INSTITUTE  
for International Textbook Research



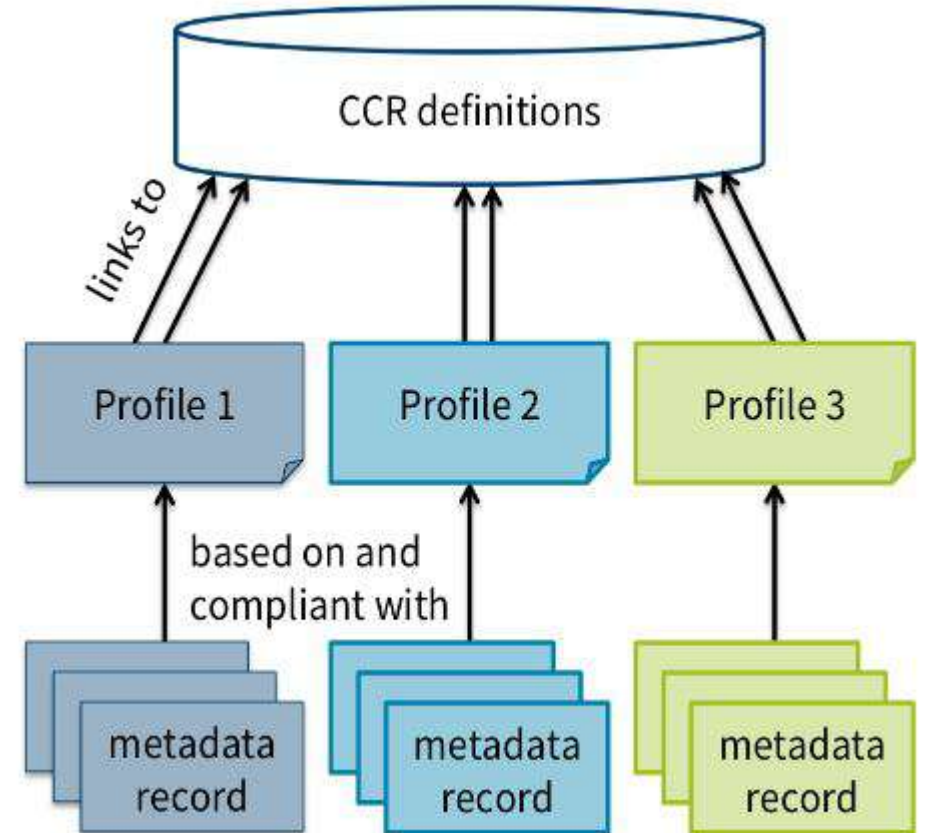
ANNUAL CONFERENCE 2019  
Leipzig, Germany

# Why CLARIN, why CMDI ?

- **Current standards:** CLARIN promotes the use of Component MetaData Infrastructure (CMDI).
- **Reduce dispersion of a multitude of formats:** using the CMDI suggested by CLARIN to overcome the dispersion produced by a multitude of formats of metadata for existing language resources and tools.
- **Interoperability and reusability of resources:** CMDI allows to overcome the problems that often descriptions like TEI headers for text or IMDI for multimedia collections have; such problems are due to the fact that these descriptions contain too specific information for any given research community.
- **Sharing resources:** CMDI offers component based metadata for harvesting resources via the Open Archives Initiative Metadata Harvesting Protocol (OAI-PMH)

# CMDI – Component Metadata Infrastructure

A metadata framework that is flexible enough to cover the different wishes from the various sub-disciplines and projects, but nevertheless has the expressive power to serve for the various functions.





# The World Views Project Use Case Scenario

TEI Document

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
reference document for sources
based on schema "tei_textbooks_wv.gei_schema_v9" and profi
last modified 01.08.2017
-->
<TEI version="5.0" xmlns="http://www.tei-c.org/ns/1.0" xml:id="
<teiHeader>
  <fileDesc>
    <titleStmt>
      <!-- "Titel (Originaltitel)" aus "Plugin: Que
<title level="a" xml:lang="langCode">[origina
      <!-- plus ggf.: -->
      <!-- "Übersetzung (de)" aus "Plugin: Quellenb
<title level="a" type="translated" xml:lang="
      <!-- "Übersetzung (en)" aus "Plugin: Quellenb
<title level="a" type="translated" xml:lang="
      <!-- "Person" aus "Plugin: Quellenbeschreibung
<author>
  <persName ref="GND-URI">
    <forename>[Vorname]</forename>
    <surname>[Nachname]</surname>
  </persName>
</author>
<!-- "Übersetzer" aus "Plugin: Quellenbeschre
<editor role="translator">
  <persName>[Name]</persName>
</editor>
</titleStmt>
<editionStmt>
  <!-- Hier soll die Versionsnummer der Quelle
<edition n="[Versionsnummer]">Version [Versio
</editionStmt>
<extent>
```

World View Service

WORLDVIEWS. THE WORLD IN TEXTBOOKS

HOME SOURCES COMMENTS AND ESSAYS HISTORY OF EDUCATION TOPICS COLLECTIONS

PEACE CONFERENCE IN MUNSTER 1648 VON HERFORDT, EWA

German English

Authors: Herfordt, Ewa

Zitierempfehlung: Failed to load citation

SEARCH

Global search Metadata Full texts

RELATED SOURCES

Friedenskonferenz in Münster 1648

RECOMMENDED COMMENTARIES

Nationale Identität und Europäische Identität

Goobi

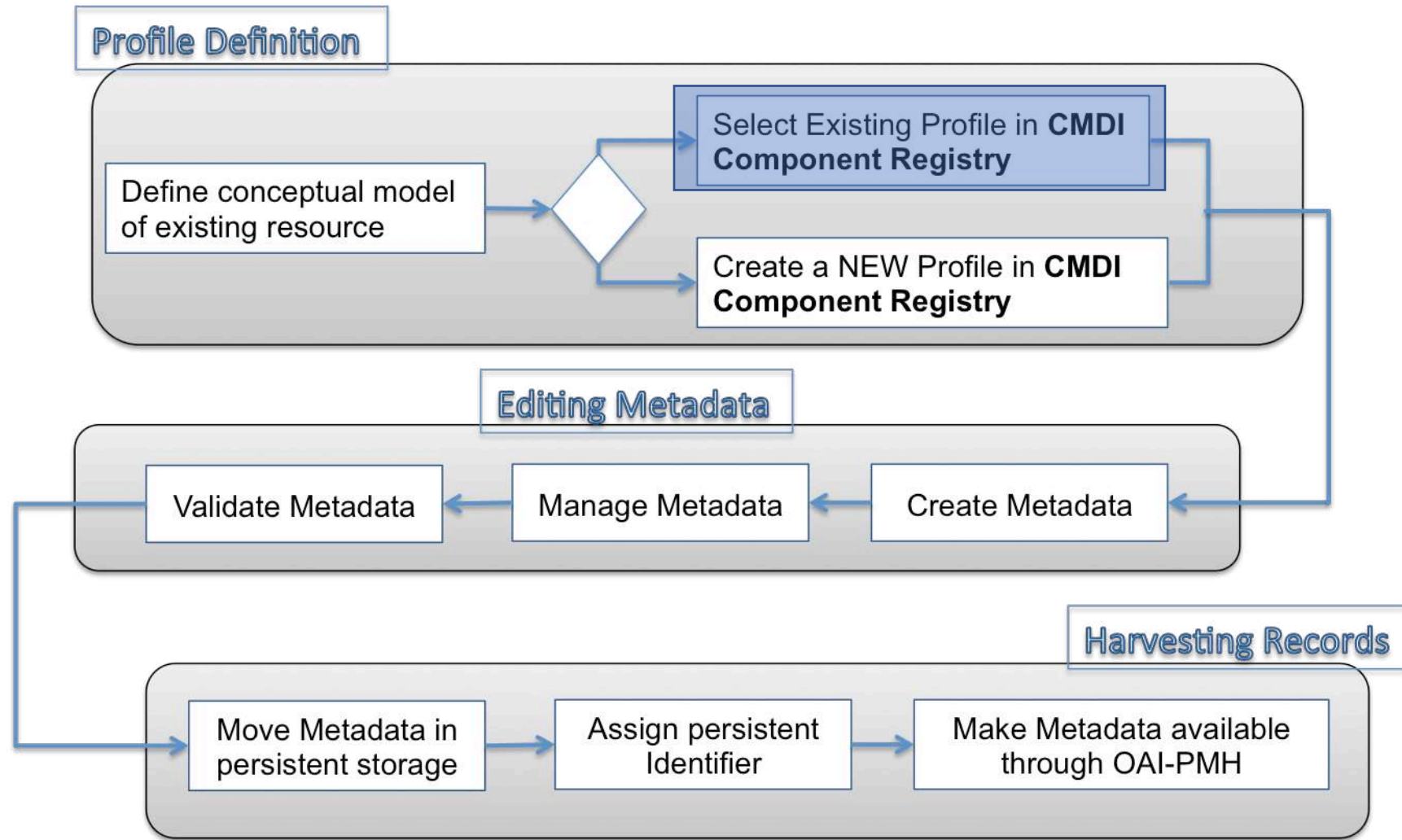
Processes

ID	Process title	Status
15	ajpvhvud_PPN601416579	
16	aridrhov_PPN60091589	
113	auzPha_000140107	
35	MargenL_02907145	
23	MargenL_01863621	
16	MargenL_PPN623000714	
21	MargenL_0060598	
38	stahak_319780272_0001	
19	stahak_PPN619810270_0001	
42	MargenL_026780478	

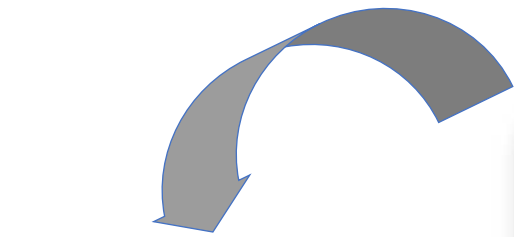
Possible actions:

- Export metadata for OAI
- Review of process app
- Get status of process-based
- Exclude description
- Export search result
- Calculate number of metadata and images
- Statistical evaluation

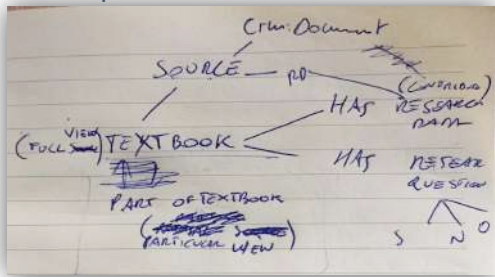
# CMDI process at GEI



# From GEI TEI Document To GEI CMDI



Conceptual Model

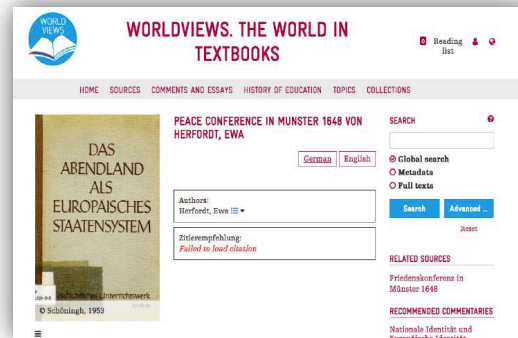


TEI Document

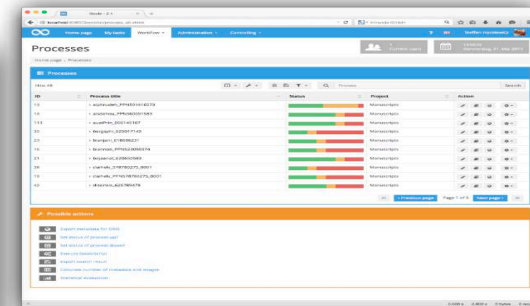
```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.tei-c.org/ns/1.0 http://www.tei-c.org/ns/1.0" type="text">
<text>
<p>
<span data-bbox="343 246 553 388">

```

World View Service



Goobi



VLO



CMDI file

```
<?xml version="1.0" encoding="UTF-8"?>
<CMDI xmlns="http://www.gei-c.org/ns/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.gei-c.org/ns/1.0 http://www.gei-c.org/ns/1.0" type="text">
<text>
<p>

```

CMDI Metadata Editor



CMDI Component Registry



# The GEI-Digital Project Use Case Scenario

Goobi

The screenshot shows the Goobi web interface. At the top, there are navigation tabs: Home page, My tasks, Workflow, Administration, and Controlling. Below this is a 'Processes' section with a table listing various processes. The table has columns for ID, Process title, Status, and Project. Below the table, there are 'Possible actions' such as 'Export metadata for DIME', 'Get status of process app?', 'Get status of process download', 'Export search result', 'Calculate number of metadata and images', and 'Statistical evaluation'.

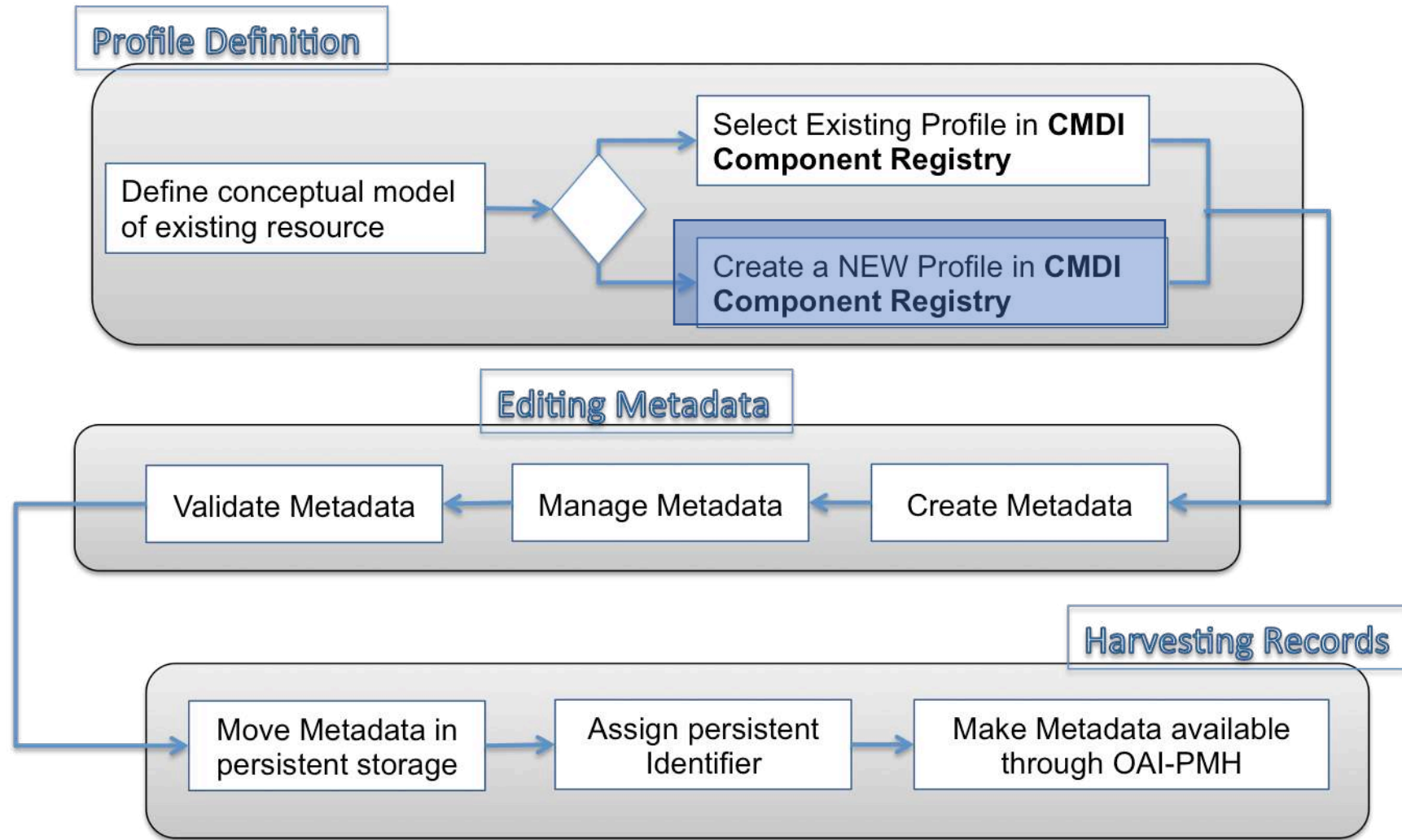
GEI Digital Service

The screenshot shows the GEI Digital Service website. The header includes the logo 'gei digital' and the tagline 'The Digital Textbook Library'. Below the header, there are navigation links: Home, Search, Browse, News, About the project, Project partners, Terms of Use. The main content area is titled 'Bibliographic data' and shows details for a specific record. The record is for 'Deutsches Lesebuch für Volksschulen' by Heinemann, Ludwig, published in Braunschweig by Bruhn in 1877. The language is German. There are also links for 'Downloads' in METS, MARCXML, Dublin Core, ESE, and OPAC formats.

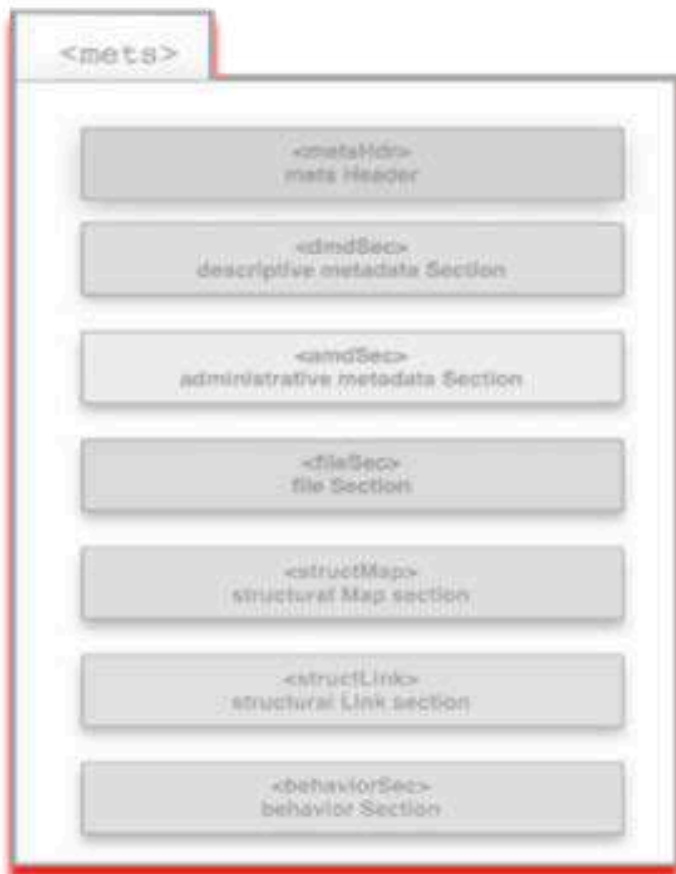
METS Document

The screenshot shows a web browser displaying the METS Document viewer. The URL is 'http://gei-digital.gei.de/viewer/'. The page title is 'OAI 2.0 Request Results'. Below the title, there is a list of identifiers: Identify | ListRecords (oai\_dc) | ListRecords (eas) | ListRecords (mets) | ListRecords (marcxml) | ListRecords (epicur) | ListRecords (lido) | ListSets | ListMetadataFormats | ListIdentifiers. The page content includes a 'Datestamp of response' (2019-09-27T15:09:03Z) and a 'Request URL' (http://gei-digital.gei.de/viewer/oai). The main content is the OAI Record Header for the record 'urn:nbn:de:0220-gd-18881288'. The header includes the OAI Identifier, Datestamp, and setSpec. Below the header, there is a large block of XML metadata, including MARC records for the book 'Deutsches Lesebuch für Volksschulen'.

# CMDI process at GEI



# METS CMDI Profile



Mapping  
Process

### CMDI Component Registry

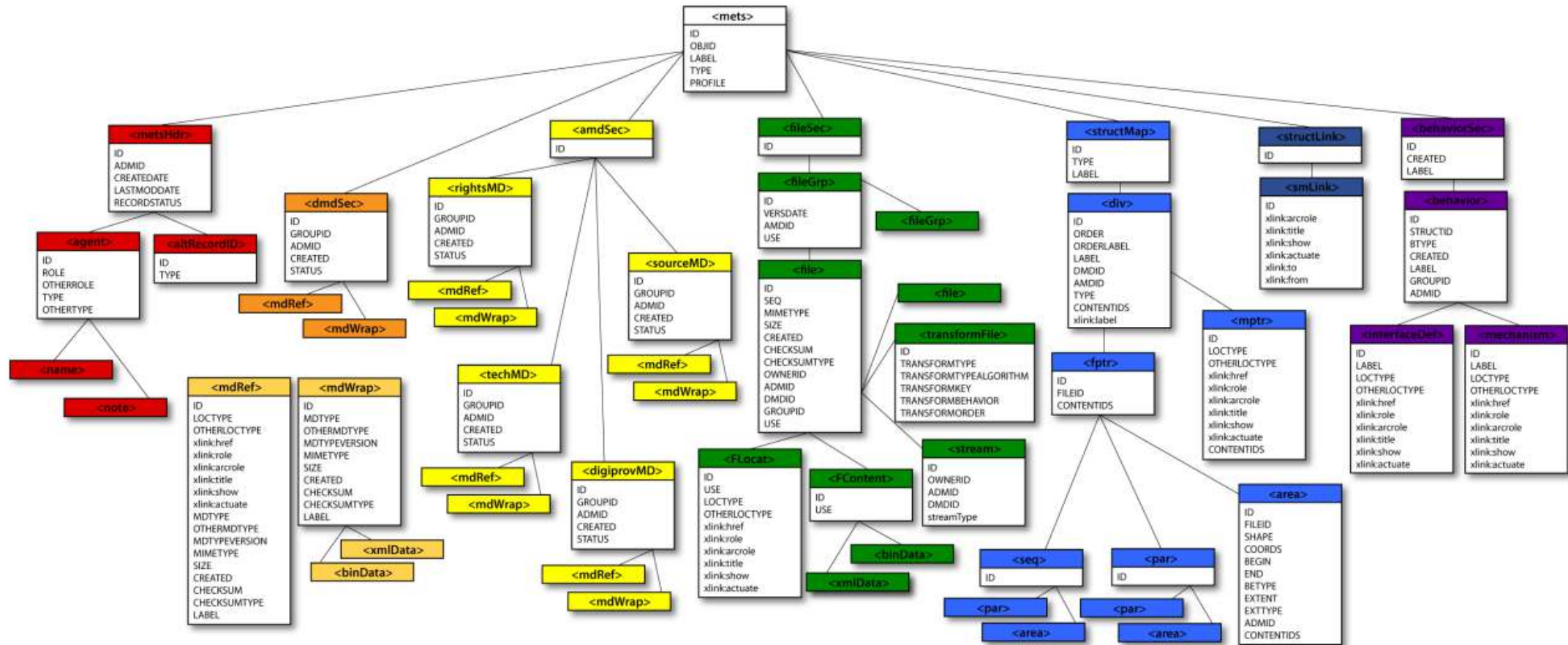
view | xml | Comments (0)

**Name:** METS  
**Description:** METS: Metadata Encoding and Transmission Standard. METS is intended to provide a standardized XML format for transmission of complex digital library objects between systems. As such, it can be seen as filling a role similar to that defined for the Submission Information Package (SIP), the Ingest Information Package (IIP) and Dissemination Information Package (DIP) in the Reference Model for an Open Archival Information System.  
<https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets>  
**Concept Link:**  
Derived from: [clarin.eucrt.p\\_1107800170070](http://clarin.eucrt.p_1107800170070)

Attribute ID	Value scheme:	Documentation:	Required:
	ID	<a href="https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets">https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets</a>	Yes
Attribute OBJID			
	string	<a href="https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets">https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets</a>	Yes
Attribute LABEL			
	string	<a href="https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets">https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets</a>	Yes
Attribute TYPE			
	string	<a href="https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets">https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets</a>	Yes
Attribute PROFILE			
	string	<a href="https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets">https://www.loc.gov/standards/mets/docs/mets-s-5.html#mets</a>	Yes

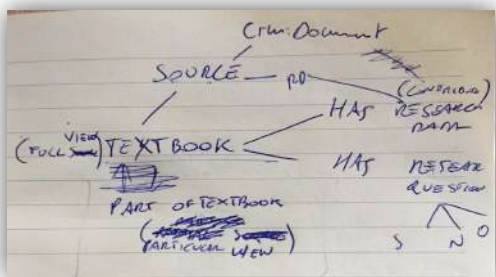
- Component: [metaHdr](#) [0-1]
- Component: [amdSec](#) [0-unbounded]
- Component: [amdSec](#) [5-1]
- Component: [fileSec](#) [0-unbounded]
- Component: [structMap](#) [1-unbounded]
- Component: [structLink](#) [0-1]
- Component: [behaviorSec](#) [0-unbounded]

# METS to CMDI Components, elements and attributes

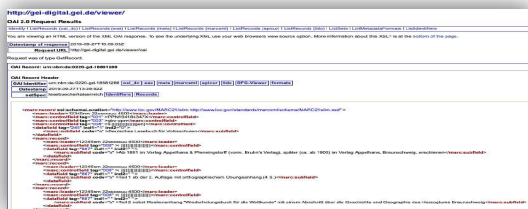


# From GEI METS Document To GEI CMDI

Conceptual Model



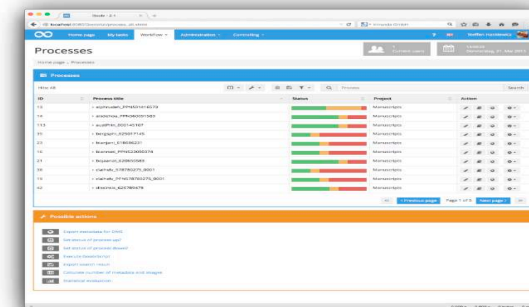
METS Document



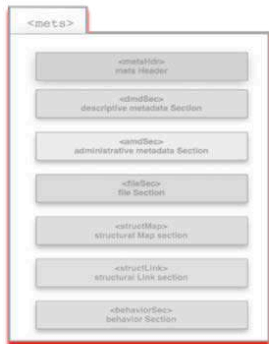
GEI Digital Service



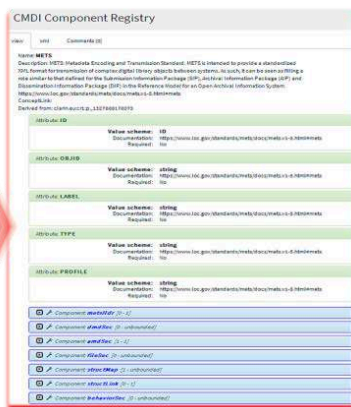
Goobi



METS CMDI Profile



Mapping Process



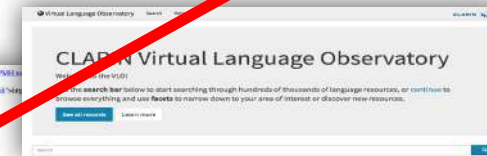
CMDI Metadata Editor



CMDI file



VLO



Fallucchi Francesca

Mapping METS and Dublin Core to CMDI: Making Textbooks available in the CLARIN VLO



GEORG ECKERT INSTITUTE For International Textbook Research



ANNUAL CONFERENCE 2019 Leipzig, Germany



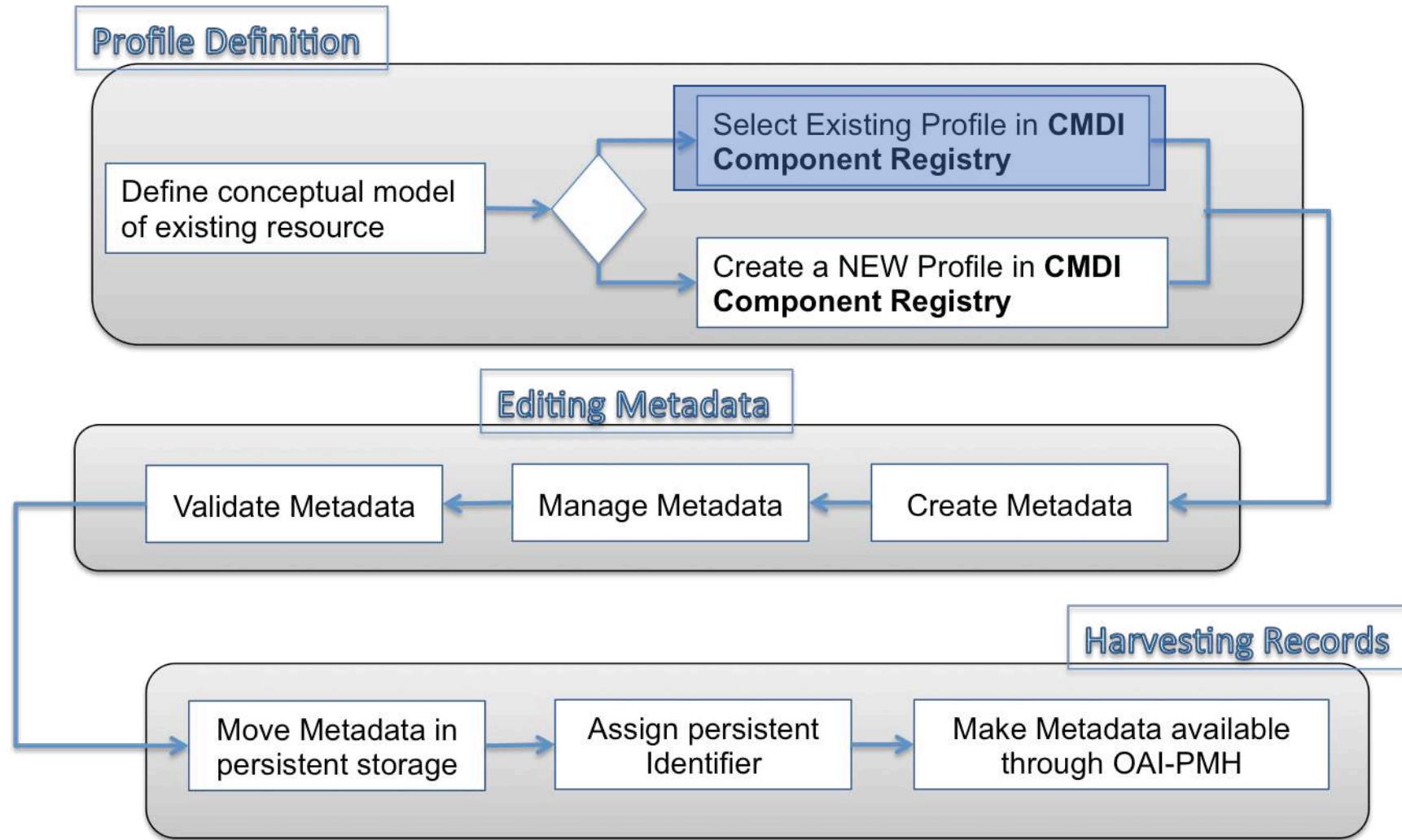


# Mapping problems

- CMDI cannot manage recursive structures such as those based on the `< mets:div >` components which are mandatory for mapping the physical layer of the digital objects.
  - METS semantics states that: "The structural divisions of the hierarchical organization provided by a `< structMap >` are represented by division `< div >` elements, which can be nested to any depth."
- CMDI cannot manage cross-reference links such as `< mets:smLink >` components, which can be found in any METS structure.
- CMDI cannot manage different concepts tied to the same label.
  - In METS such cases are disambiguated by their position in the XML structure.



# CMDI process at GEI





# CMDI process at GEI Results

By using CMDI, full-text resources can immediately be indexed by CLARIN's Virtual Language Observatory and can be analyzed using its various tools and services such as Weblicht.

The CMDI description of GEI resources allows for internally standardized search and retrieval operations in federated search scenarios.

**Menschenrassen und Völkertypen**

Show the original provider's page for this record  
Plan text search via Federated Content Search

Record details Resources (1) Availability All metadata Technical details

Name Menschenrassen und Völkertypen  
Material zu geographischen Untersuchungen auf der Oberstufe mehrklassiger Völk- und Bürgerschulen. Zugleich eine Erläuterung der gleichnamigen Bildwerke.  
Heft 2: Die Menschenrassen

Collection Deutsches Testarchiv (1800-1900)

Language German

Country Germany

Genre gebrauchsliteratur  
schulbuch  
ready  
pe

Organisation Geog-Eckert-Institut - Leibniz-Institut für Internationale Schulbuchforschung  
CLARIN-D  
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)  
Deutsches Testarchiv  
Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

National project CLARIN-D

Resource type text

dataProviderName Berlin-Brandenburg Academy of Sciences and Humanities

projectName Deutsches Testarchiv

temporalCoverage 2017-08-07

<http://hdl.handle.net/11858/00-2032-0000-0026-F0E-5>

VLO Faceted search Search results Record: Menschenrassen und Völkertypen

Record 1 of 3

**Menschenrassen und Völkertypen**

Record details Resources (1) Availability All metadata Technical details

Name Menschenrassen und Völkertypen

Collection GEI historic German textbooks: geographischeschulbuecherkaiserreich

Resource type text

resolver

# Conclusion

- We describe a mapping process to established standards and provide APIs for other services, in order to integrate our resources and tools into the CLARIN infrastructure and make them discoverable in VLO.
- We focus on CMDI as a unique metadata descriptive standard for encoding administrative and structural metadata of the resources of GEI-Digital repository.
- We create the METS CMDI profile but the CMDIfication process is stopped because for mapping problems, the METS to CMDI mapping revealed itself impracticable.
- Our library team likes and used METS because they want to record more than is possible with DC.

# Thanks for your Attention!



# Questions?