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

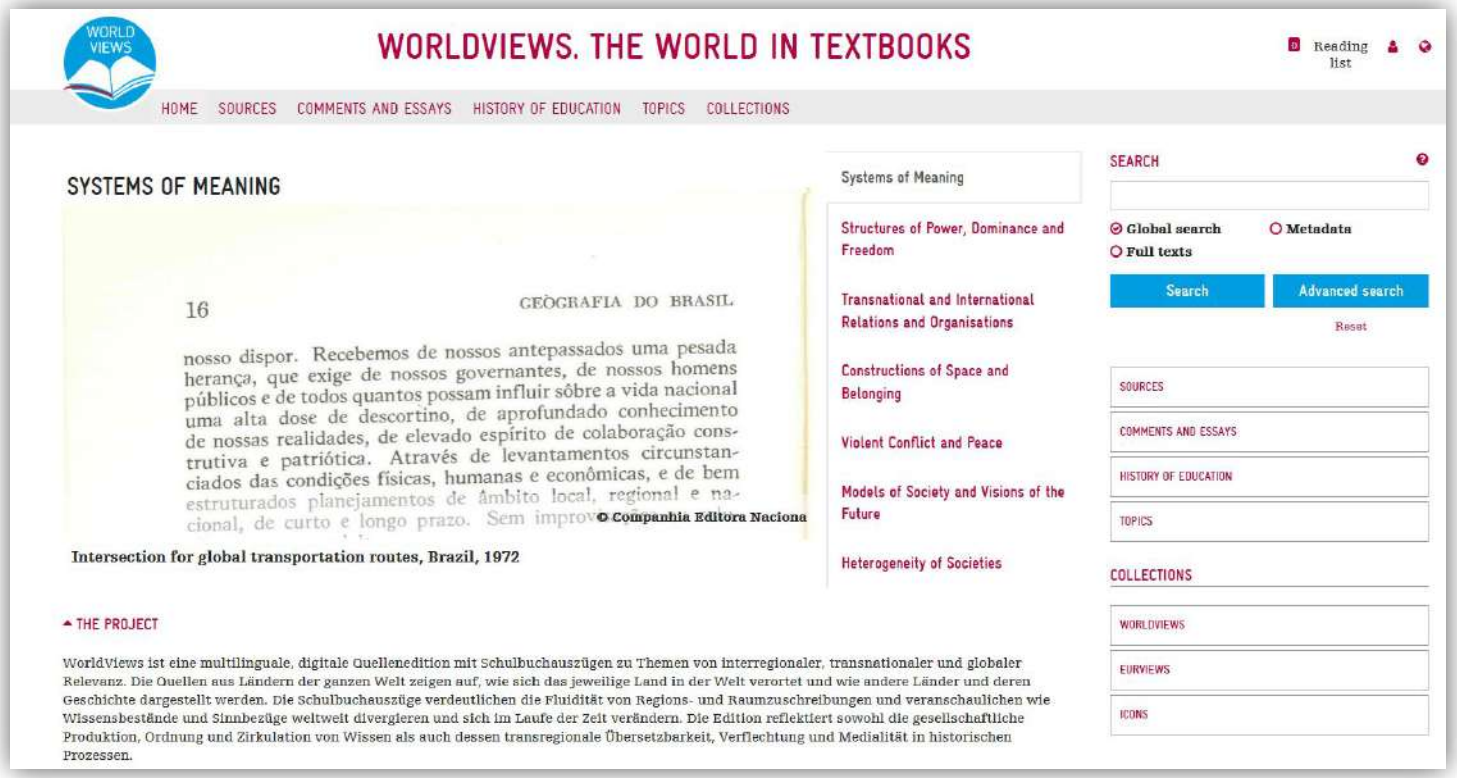
The World Views Project

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



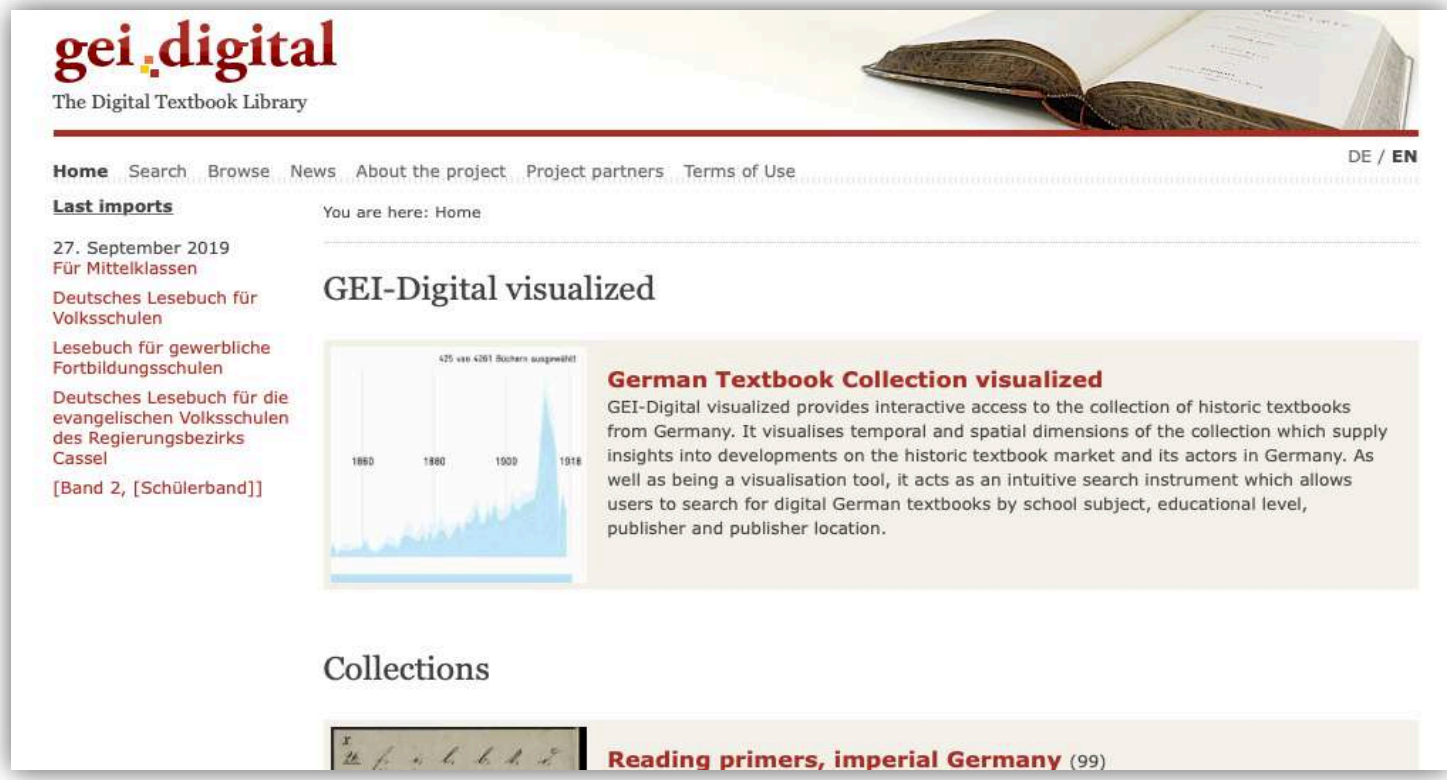
The GEI-Digital Project

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.

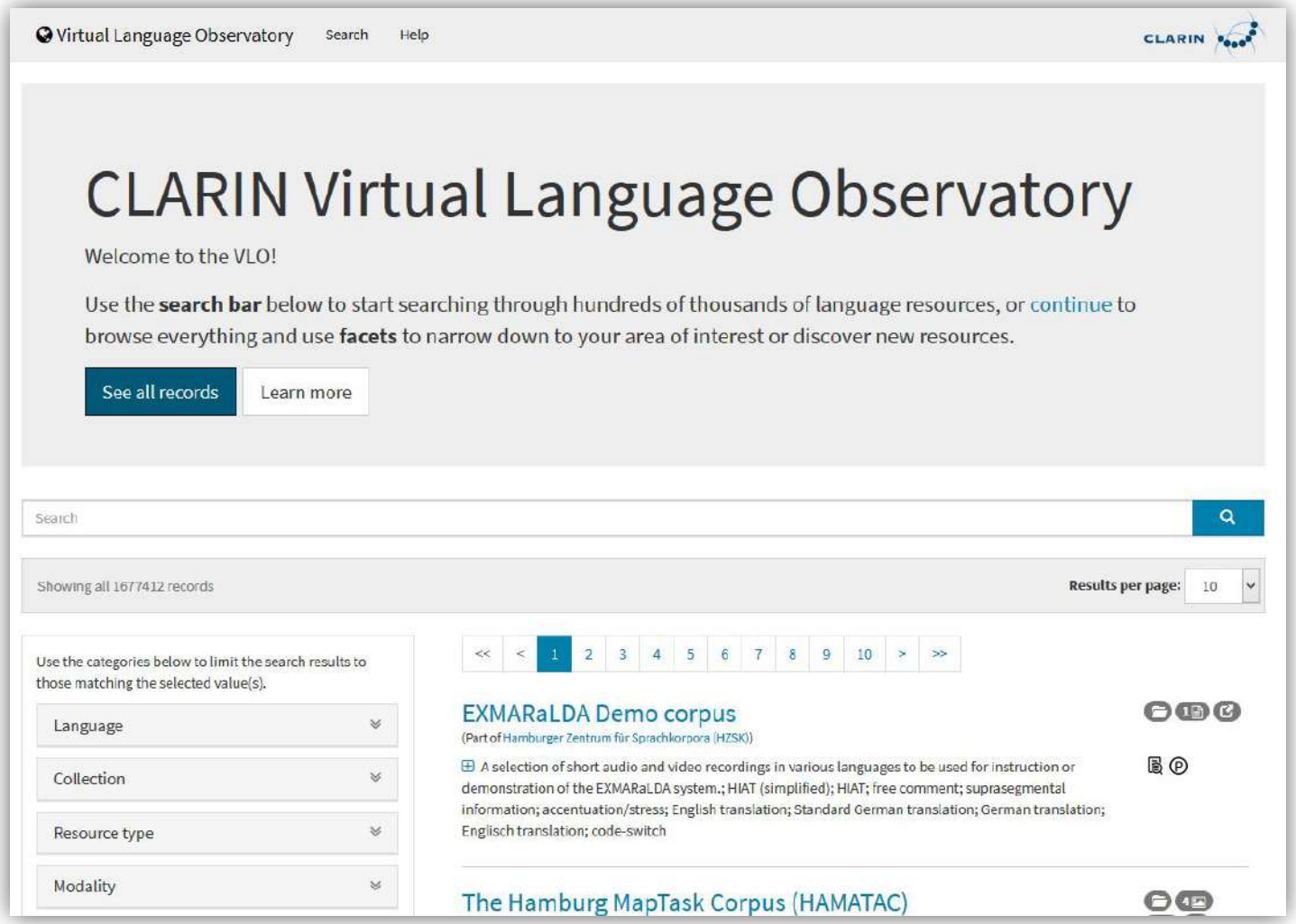
Objective: Make data as open, visible and re-usable as possible



Virtual Language Observatory (VLO)

https://vlo.clarin.eu/

The data becomes visible and searchable in the VLO

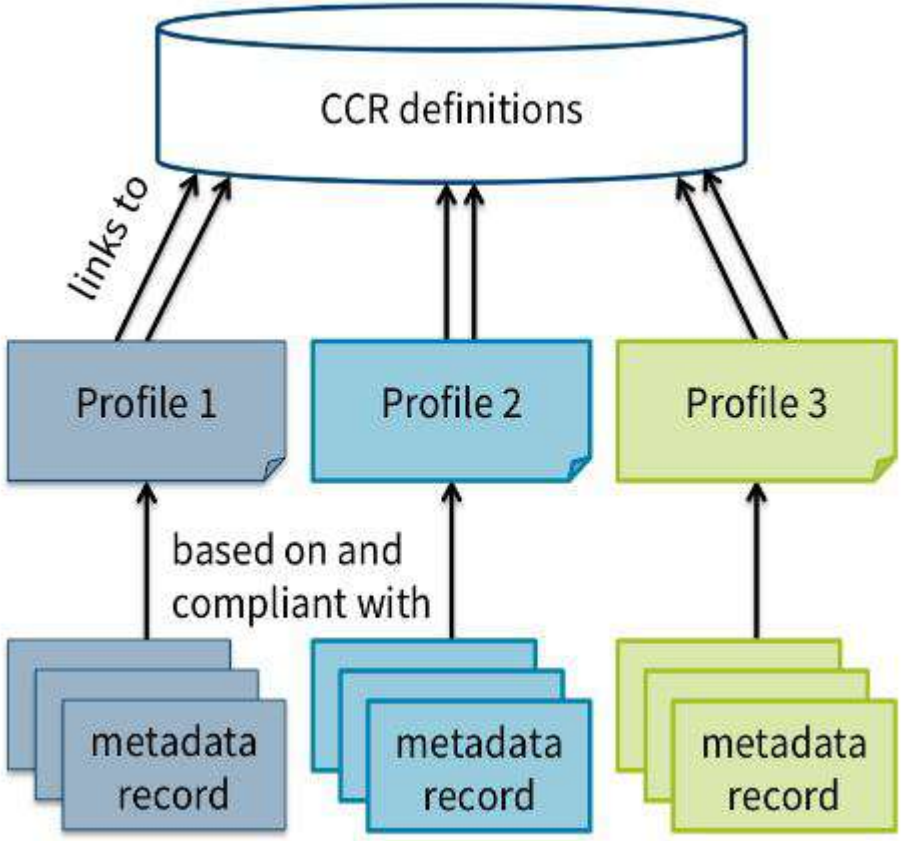


Why CLARIN, why CMDI ?

- **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.

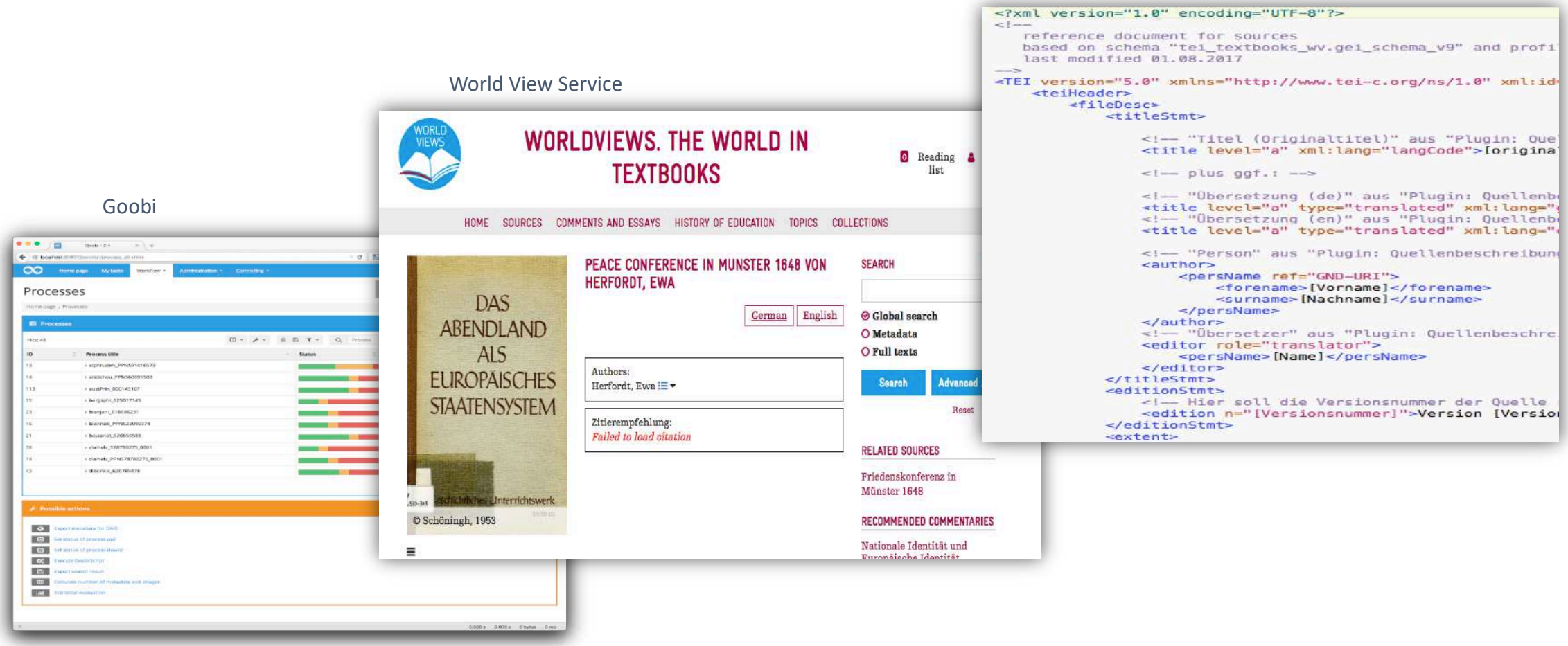


The World Views Project Use Case Scenario

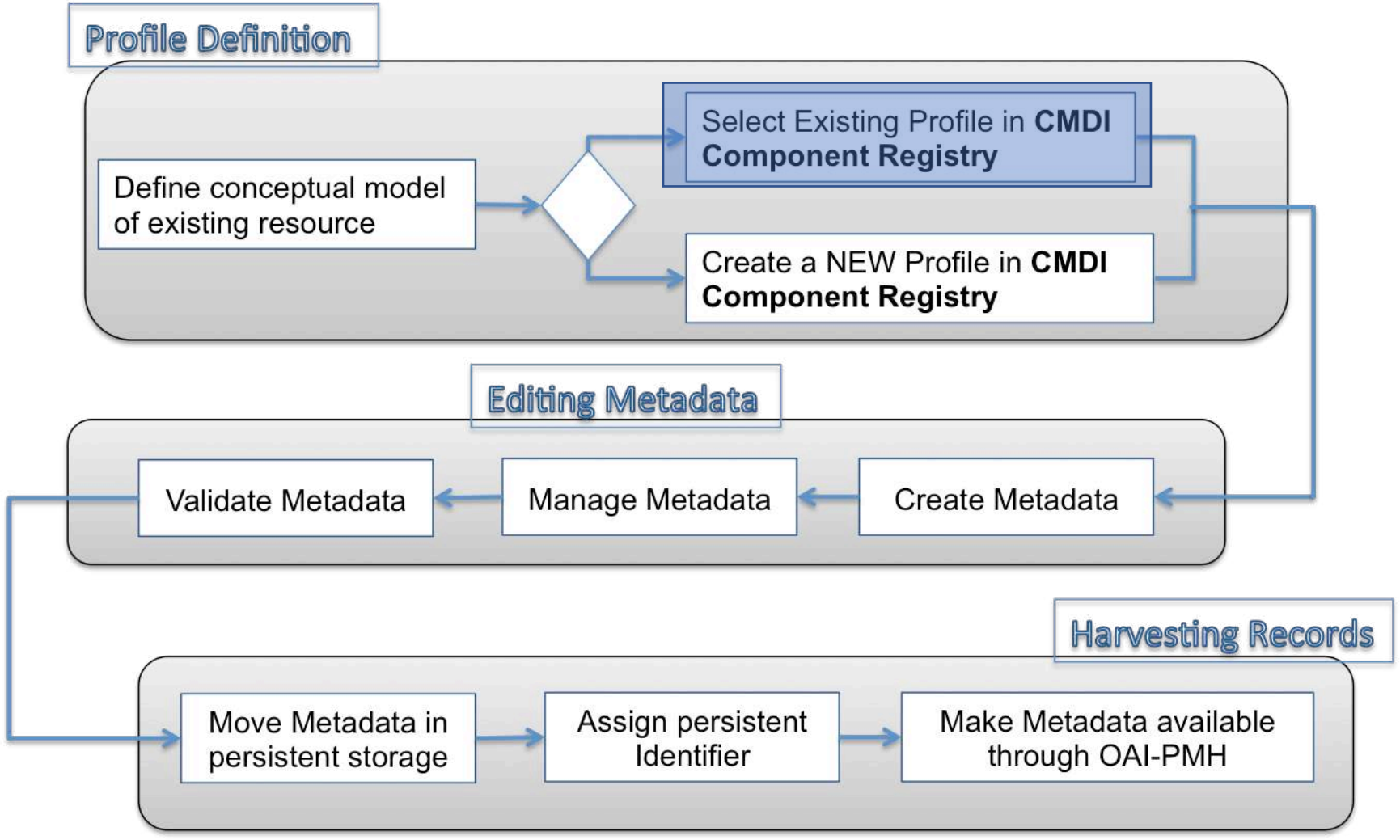
TEI Document

World View Service

Goobi



CMDI process at GEI

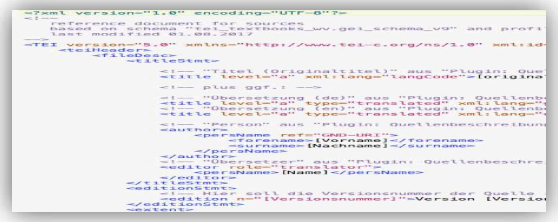


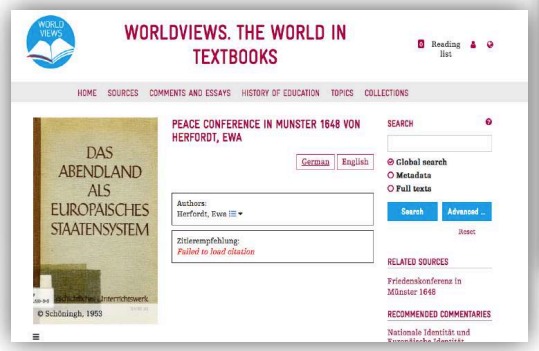
From GEI TEI Document To GEI CMDI

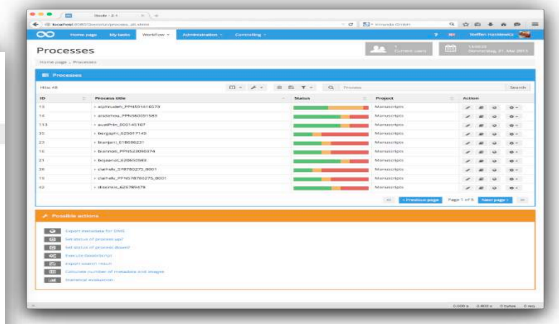
Goobi

World View Service

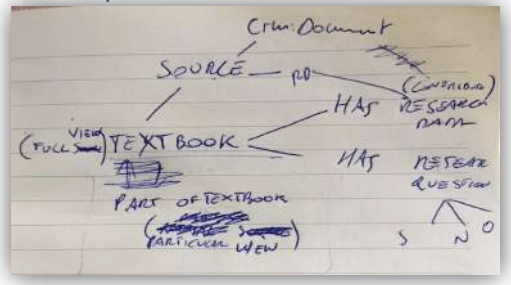
TEI Document





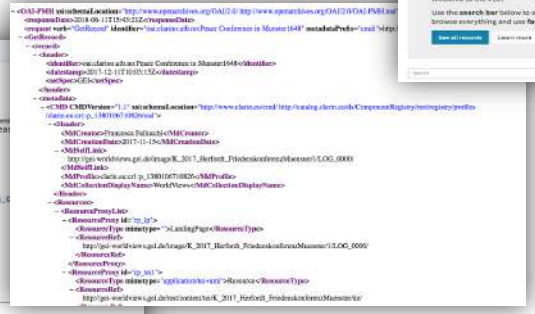


Conceptual Model



VLO

CMDI file



CMDI Metadata Editor

CMDI Component Registry

Fallucchi Francesca

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

Università degli Studi Guglielmo Marconi

GEORG ECKERT INSTITUTE for International Textbook Research

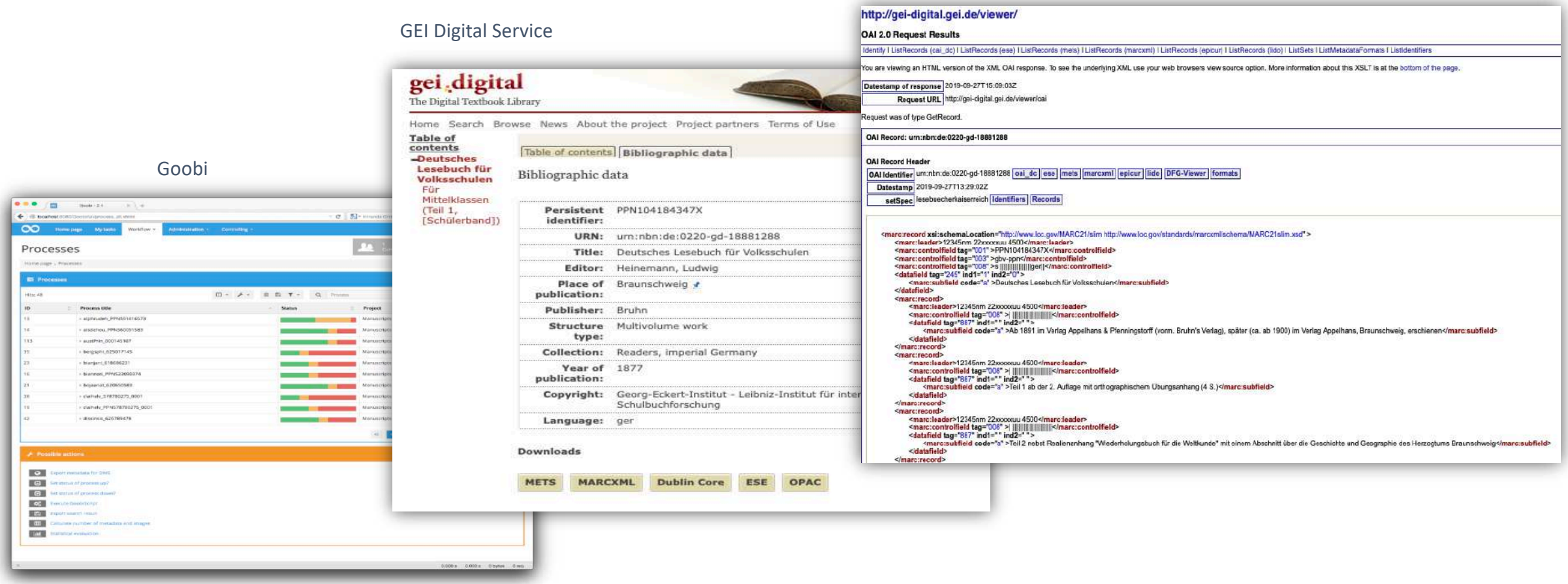
CLARIN | ANNUAL CONFERENCE 2019 Leipzig, Germany

The GEI-Digital Project Use Case Scenario

Goobi

GEI Digital Service

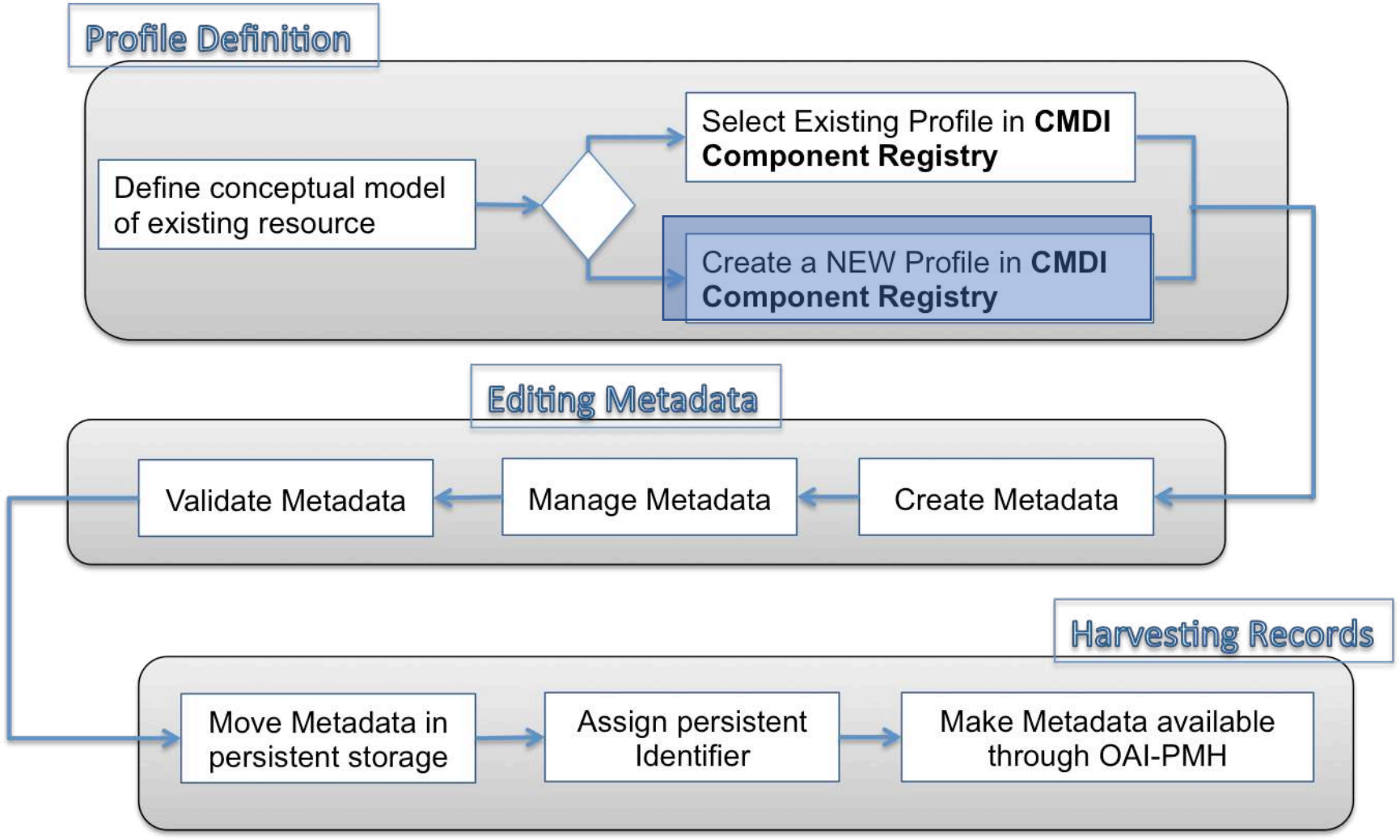
METS Document



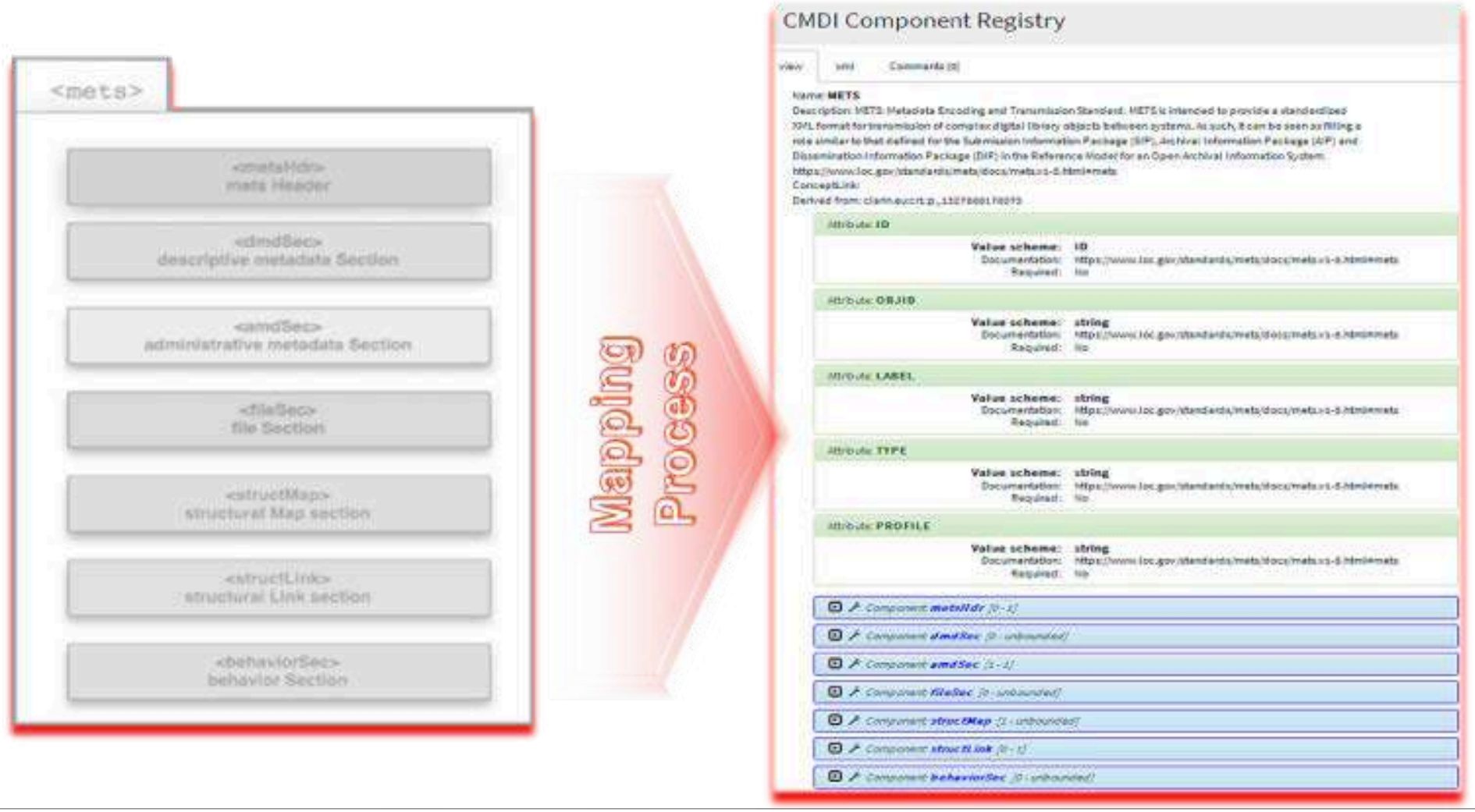
Fallucchi Francesca

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

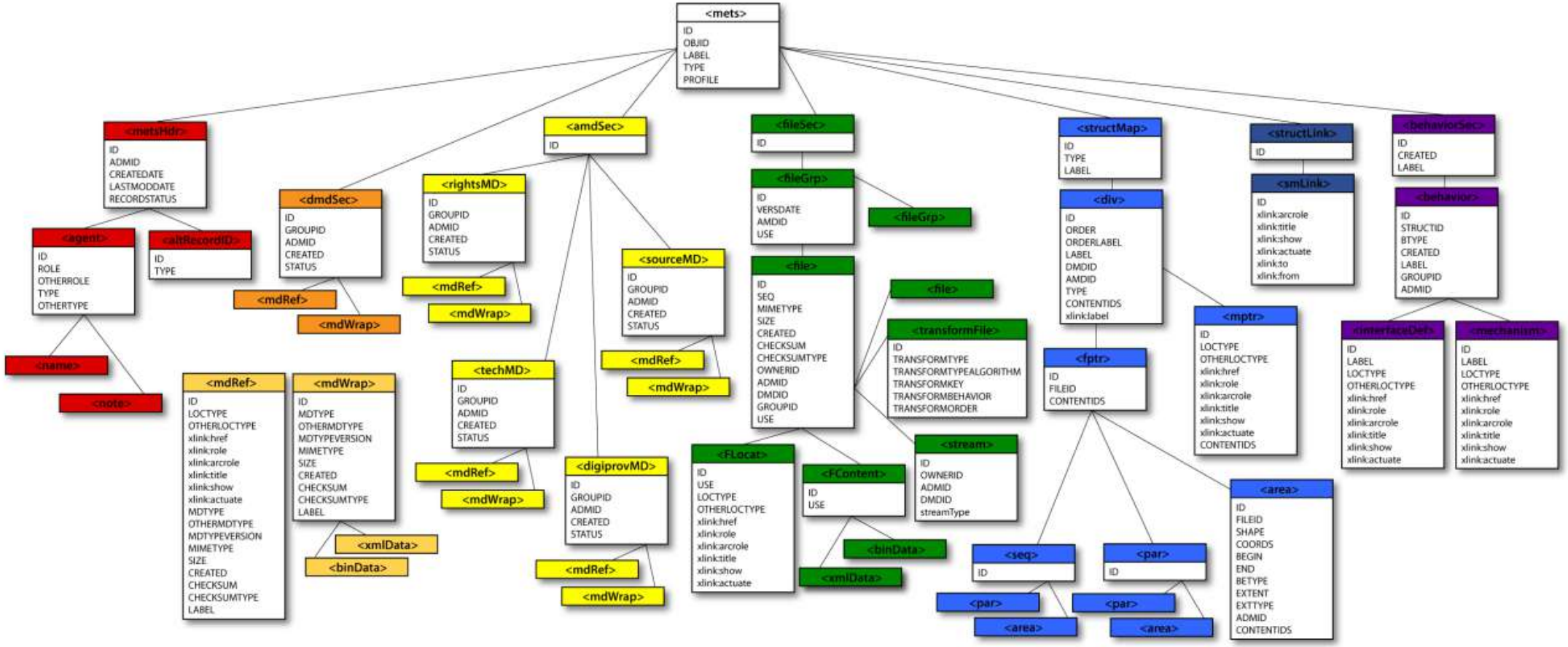
CMDI process at GEI



METS CMDI Profile

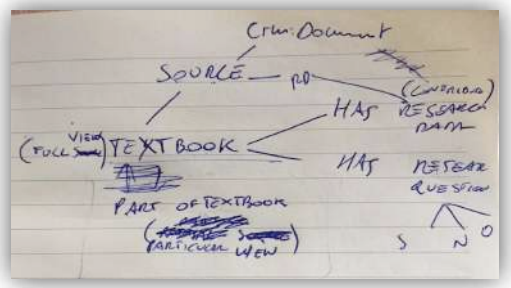


METS to CMDI Components, elements and attributes



From GEI METS Document To GEI CMDI

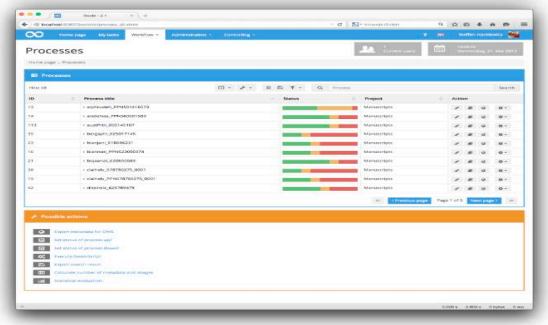
Conceptual Model



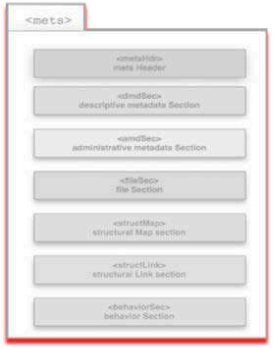
METS Document

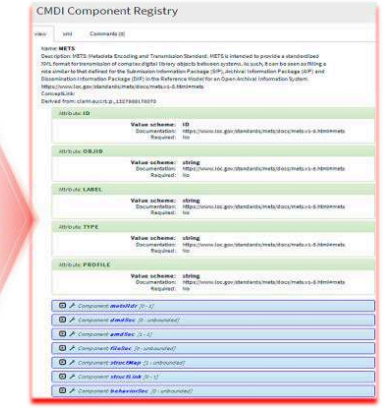
GEI Digital Service

Goobi



METS CMDI Profile

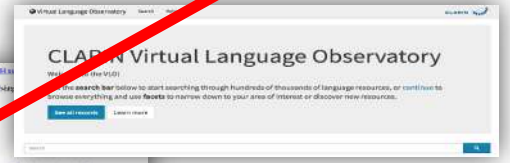




CMDI Metadata Editor

CMDI file

VLO



Fallucchi Francesca

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

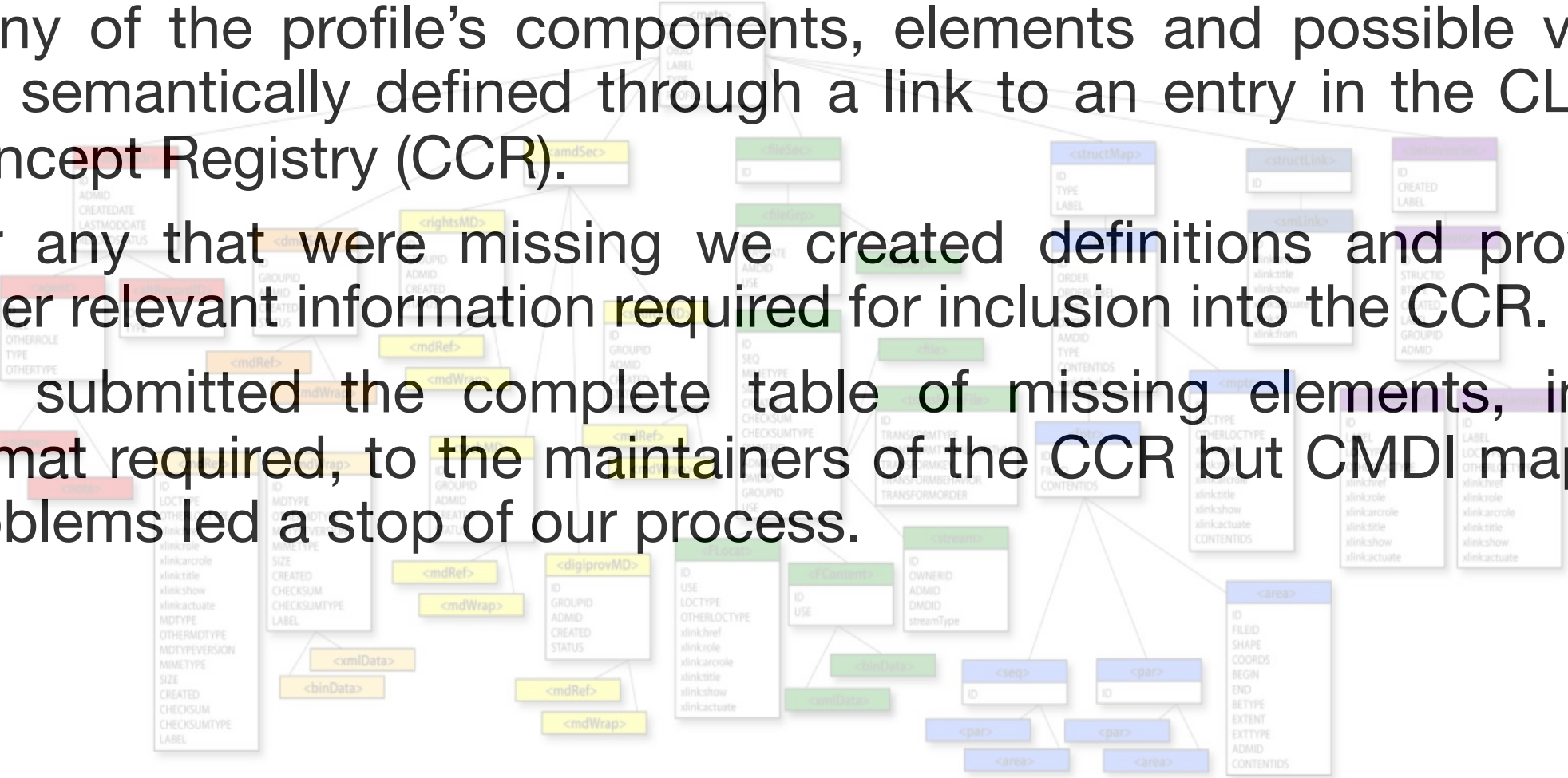
Università degli Studi Guglielmo Marconi

GEORG ECKERT INSTITUTE for International Textbook Research

CLARIN | ANNUAL CONFERENCE 2019 Leipzig, Germany

Mapping problems

- Many of the profile's components, elements and possible values are semantically defined through a link to an entry in the CLARIN Concept Registry (CCR).
- For any that were missing we created definitions and provided other relevant information required for inclusion into the CCR.
- We submitted the complete table of missing elements, in the format required, to the maintainers of the CCR but CMDI mapping problems led a stop of our process.



Mapping problems

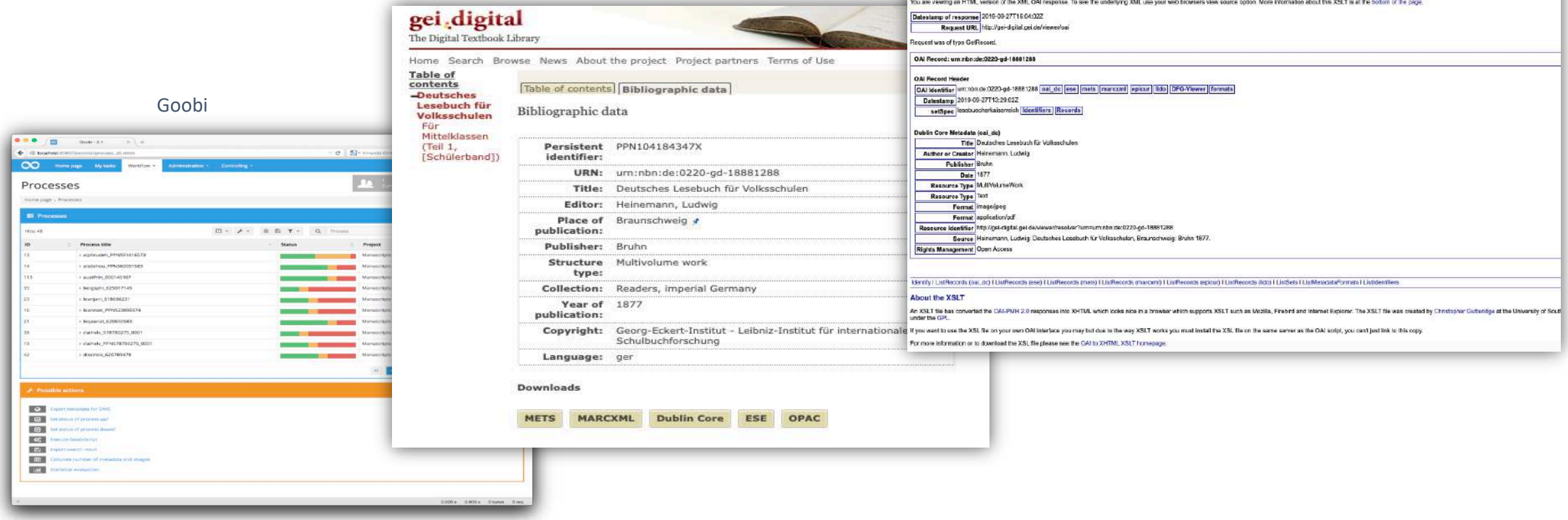
- CMDI cannot manage recursive structures such as those based on the *<mets:div>* components which are mandatory for mapping the physical layer of the digital objects.
  - METS semantics states that: "The structural divisions of the hierarchical organization provided by a *<structMap>* are represented by division 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.

The GEI-Digital Project Use Case Scenario

DC Document

GEI Digital Service

Goobi



Fallucchi Francesca

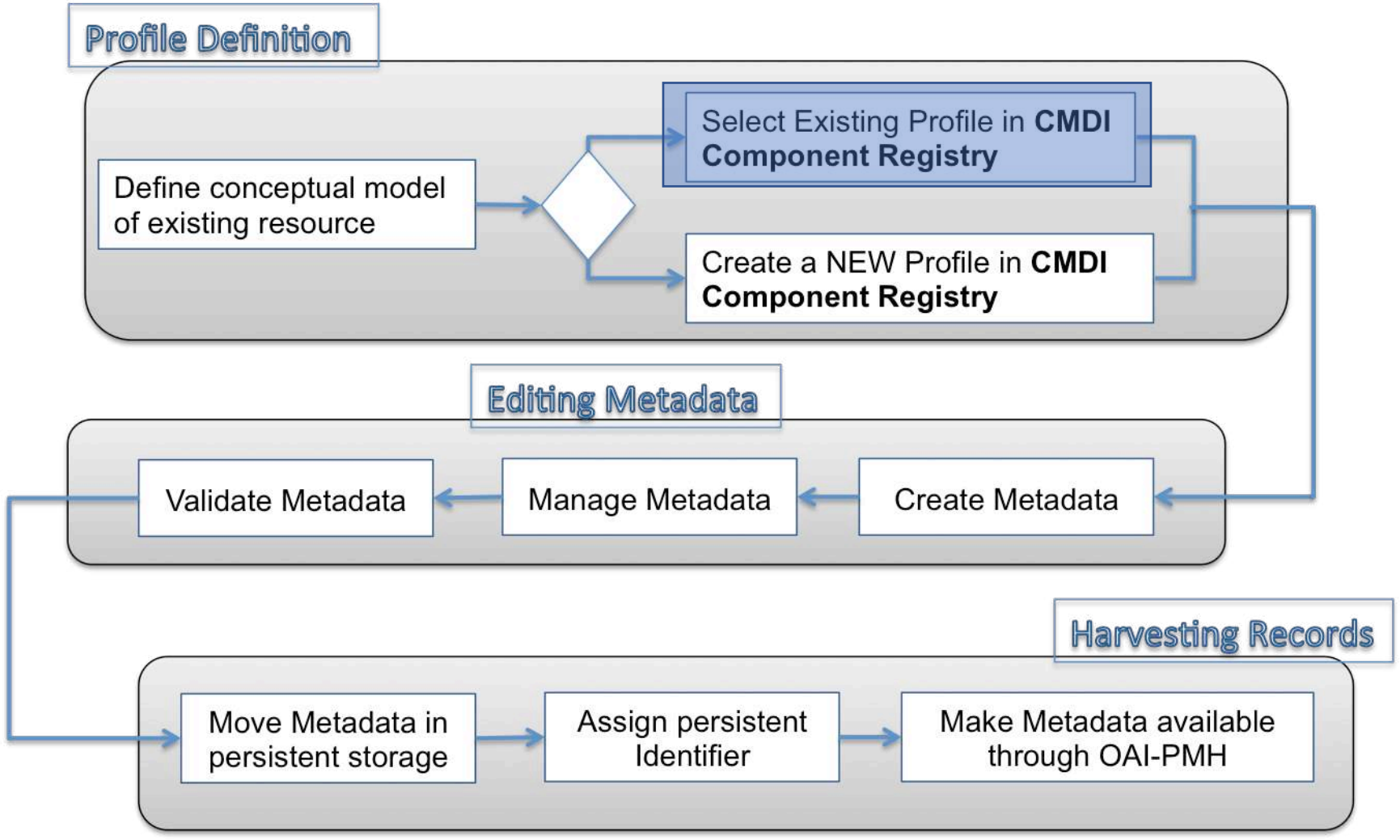
Mapping METS and Dublin Core to CMDI: Making Textbooks available in the CLARIN VLO

Università degli Studi Guglielmo Marconi

GEORG ECKERT INSTITUTE for International Textbook Research

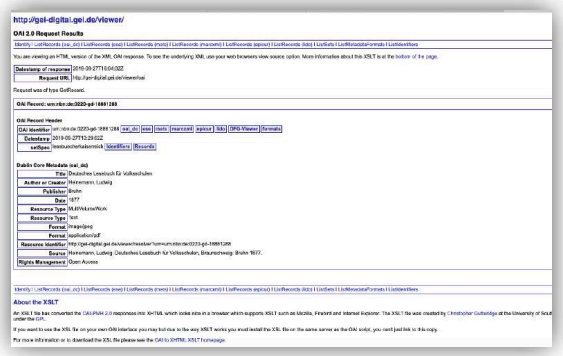
CLARIN ANNUAL CONFERENCE 2019 Leipzig, Germany

CMDI process at GEI



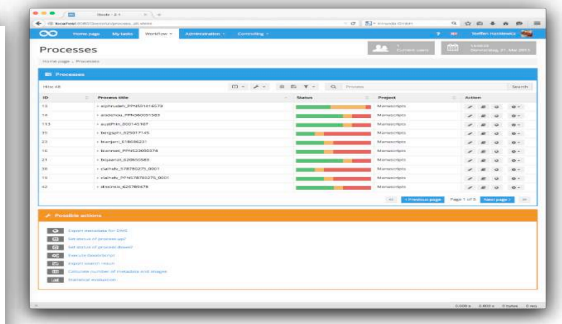
From GEI DC Document To GEI CMDI

DC Document

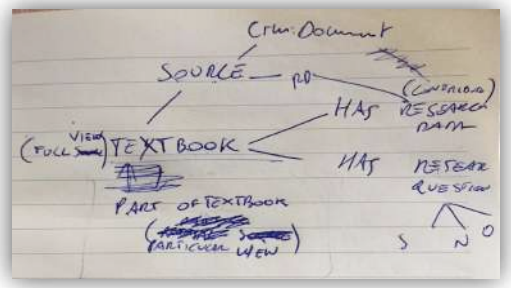


GEI Digital Service

Goobi



Conceptual Model

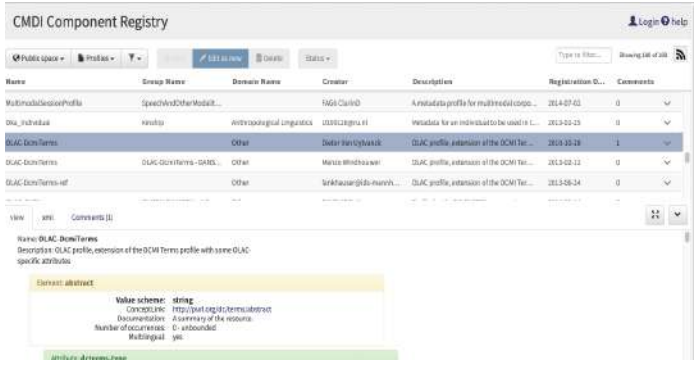


VLO

CMDI File

Dcto CMDI Converter

CMDI Component Registry



Fallucchi Francesca

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

Università degli Studi Guglielmo Marconi

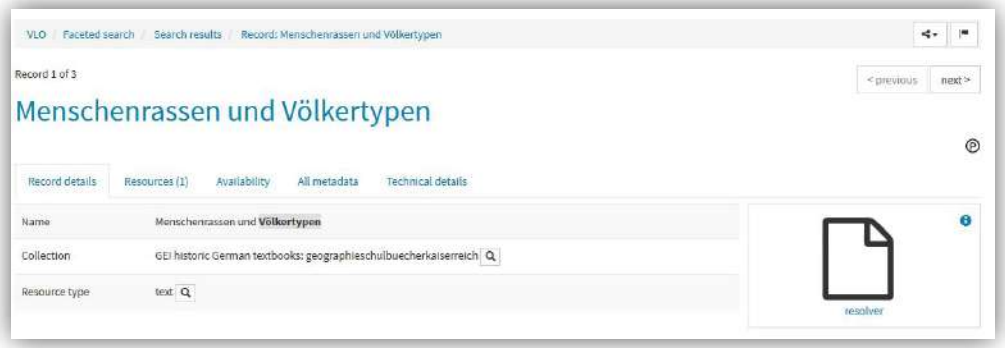
GEORG ECKERT INSTITUTE for International Textbook Research

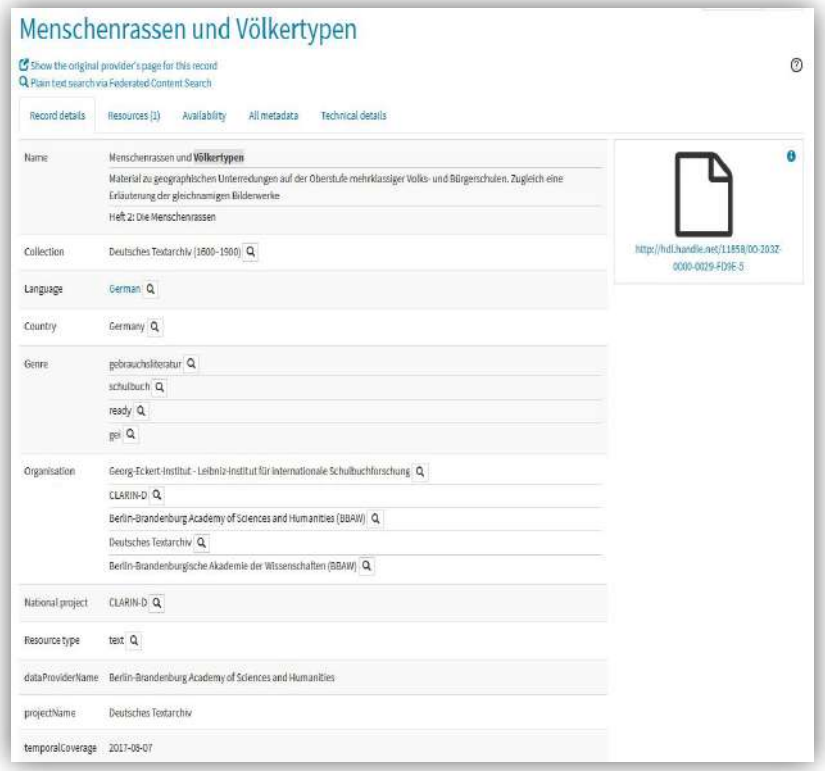
CLARIN | ANNUAL CONFERENCE 2019 Leipzig, Germany

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.





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?