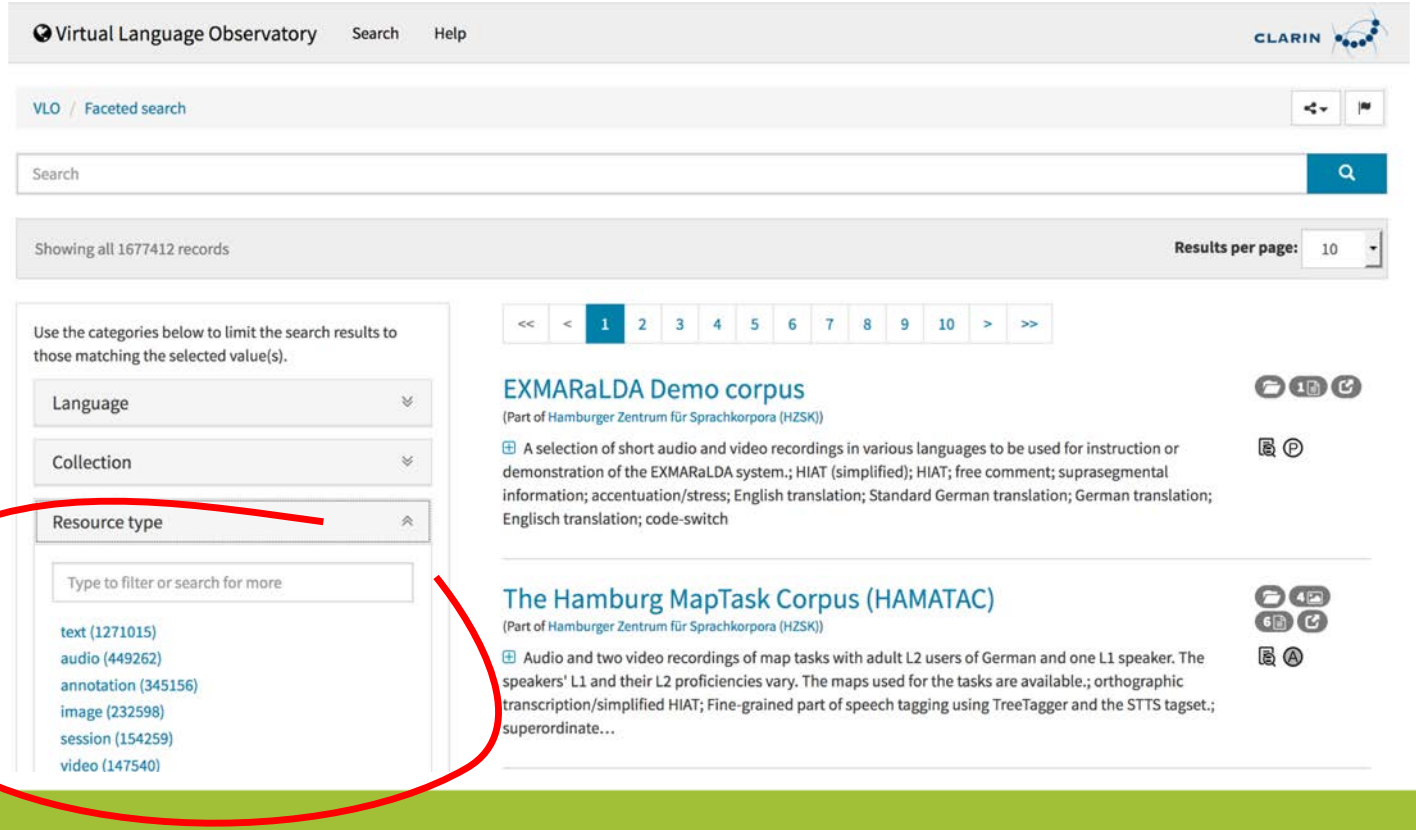
CLARIN Metadata Curation: Recent Developments

Twan Goosen, Menzo Windhouwer,
Susanne Haaf

CLARIN Centre Meeting
Utrecht, 4 June 2018

Why metadata curation?



Resource type facet: January 2018 

346 distinct values

530k records not covered

= 63% of CLARIN/Other harvest sets

Resource type facet: January 2018 ☃️

346 distinct values - a sample:

Boek

MovingImage

Boeken

Video

Book

LanguageDescription

Early printed book (1501-1800)

language_description

Nicht dokumentiert

Knjige

Electronic publications -- Great Britain -- 20th century text
Electronic publications -- Indonesia -- 20th century text
Electronic publications -- Japan -- 20th century
Electronic publications -- United States -- 20th century text
English drama -- Restoration, 1660-1700 text
English literature -- Middle English, 1100-1500 text
Philosophical texts -- Great Britain -- 18th century
Philosophical texts -- Great Britain -- 19th century

Resource type facet: June 2018  ??

?? distinct values

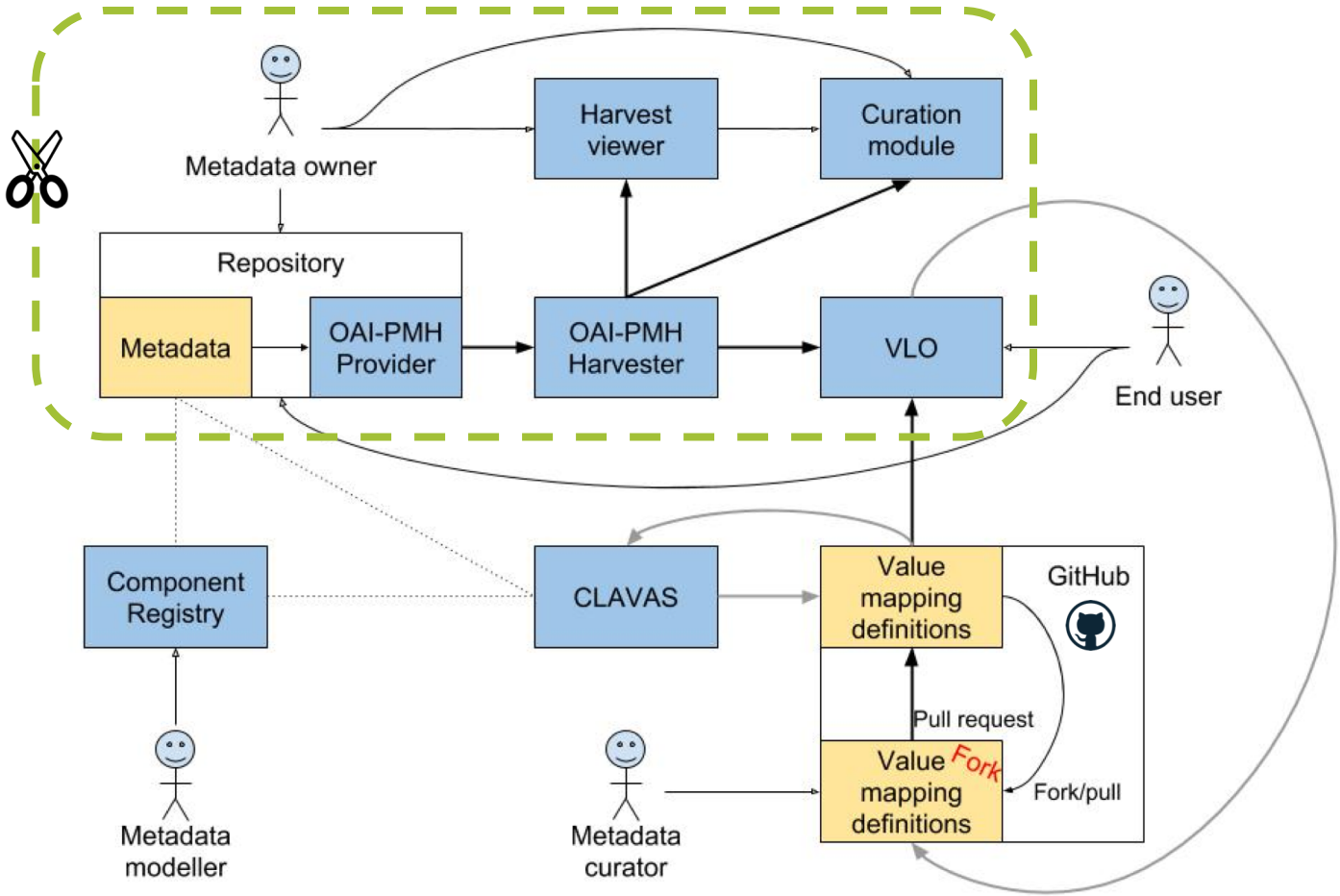
?? records not covered

= ??% of CLARIN/Other harvest sets

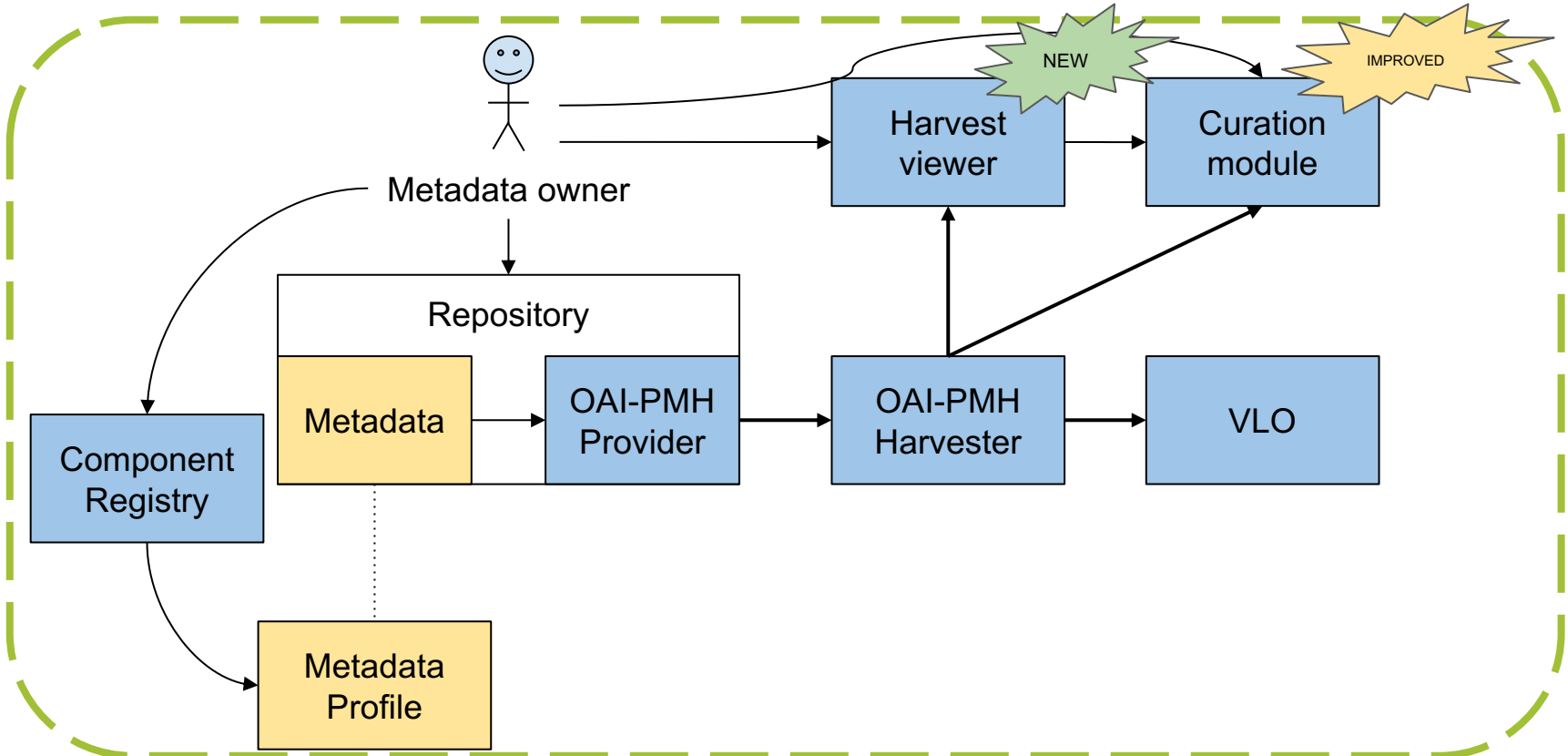
Outline

- Overview of CLARIN's Curation Ecosystem
- Curation Ecosystem for Metadata Owners
- Curation Ecosystem for Metadata Modellers
- Curation Ecosystem for Metadata Curators

Overview


Metadata owner
Harvest viewer
Curation module
Repository
Metadata
OAI-PMH Provider
OAI-PMH Harvester
VLO
End user
Component Registry
CLAVAS
Value mapping definitions
GitHub
Pull request
Value mapping definitions
Fork
Fork/pull
Metadata modeller
Metadata curator

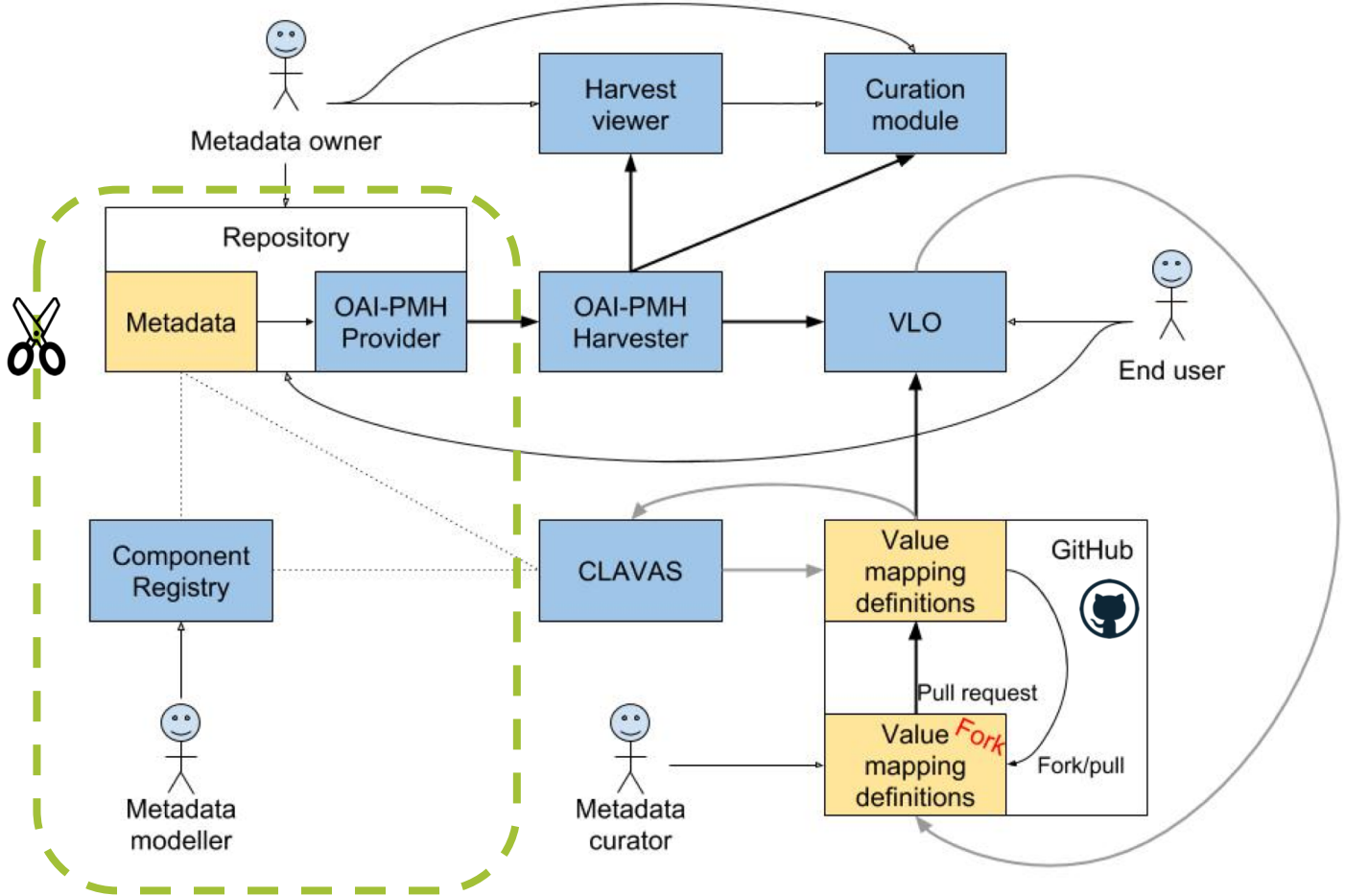
Curation Ecosystem for Metadata Owners



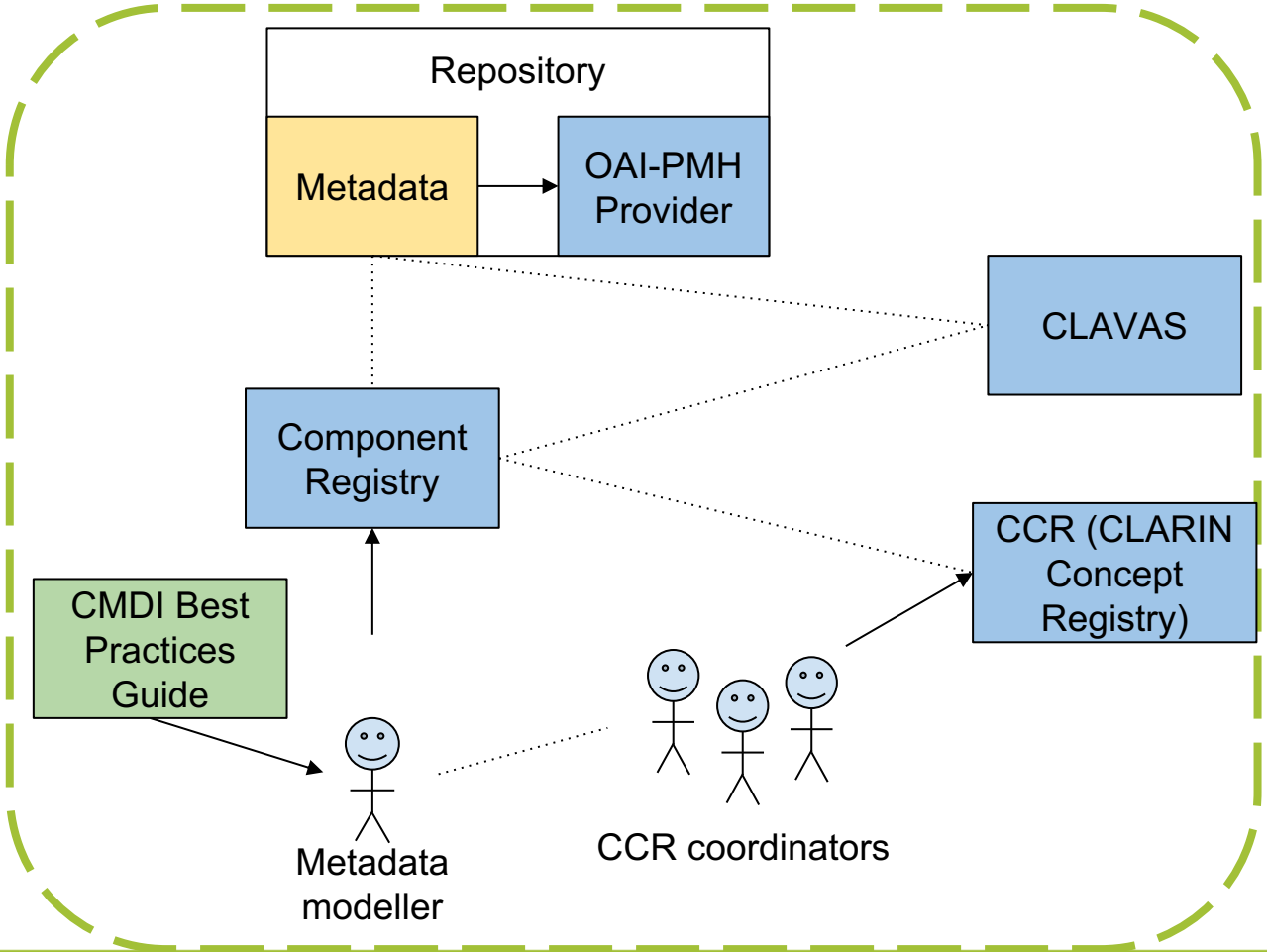
Developments

- **Harvest viewer** (just around the corner)
  - provides paged access to a harvest
  - provides access to endpoint specific harvest logs
  - provides a jump off point to the curation module
- **Curation Module** (clarin.oeaw.ac.at/curate/)
  - XML validation
  - URL checking

Overview


Metadata owner
Harvest viewer
Curation module
Repository
Metadata
OAI-PMH Provider
OAI-PMH Harvester
VLO
End user
Component Registry
CLAVAS
Value mapping definitions
GitHub
Pull request
Value mapping definitions
Fork
Fork/pull
Metadata modeller
Metadata curator

Curation Ecosystem for Metadata Modellers



Developments

CMDI Best Practices Guide

- Guide for modelling and creating (automatically or by hand) metadata
- Joint effort of Metadata Curation Taskforce and CMDI Taskforce
- Presented at CAC 2017
- Ongoing work on completing the guide, feedback is much appreciated!
- Recent version published here: clarin.eu/cmdi-best-practice-guide
- Where possible the usage of best practices can be checked by various validators
- Follow-up project evolved: Recommended Concepts & Components

Developments

CCR:

- Experiments around the use of generic or more specific concepts

Developments

CCR:

- Exper
- Conc

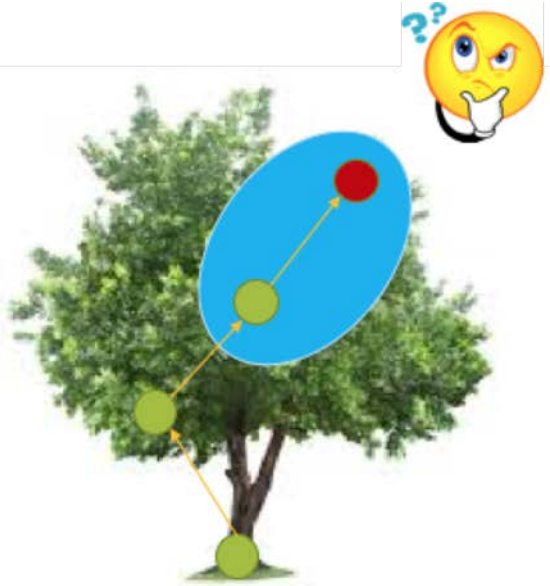
cepts

t values

Developments

CCR:

- Exper
- Conc
- Howe



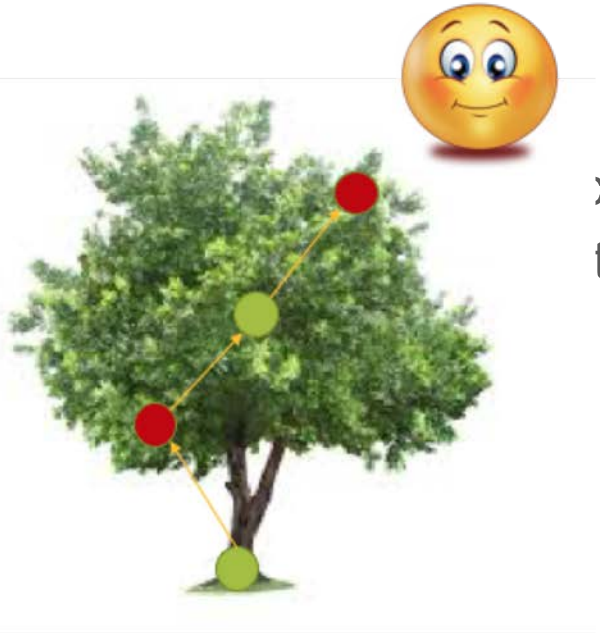
cepts

t values

Developments

CCR:

- Exper
- Conce
- Howe
- Creat
  the VI



cepts
t values
path for
les.

Developments

CCR:

- Experiments around the use of generic or more specific concepts
- ConceptLinks help the VLO to find the paths to relevant facet values
- However, the semantics of the concept might be too local!
- Create a set of generic concepts that can allow or disallow a path for the VLO to the target values, and can be used for many profiles.
- TODO
  - CCR: identify and specify the set of generic concepts
  - VLO: take more of the semantic context into account

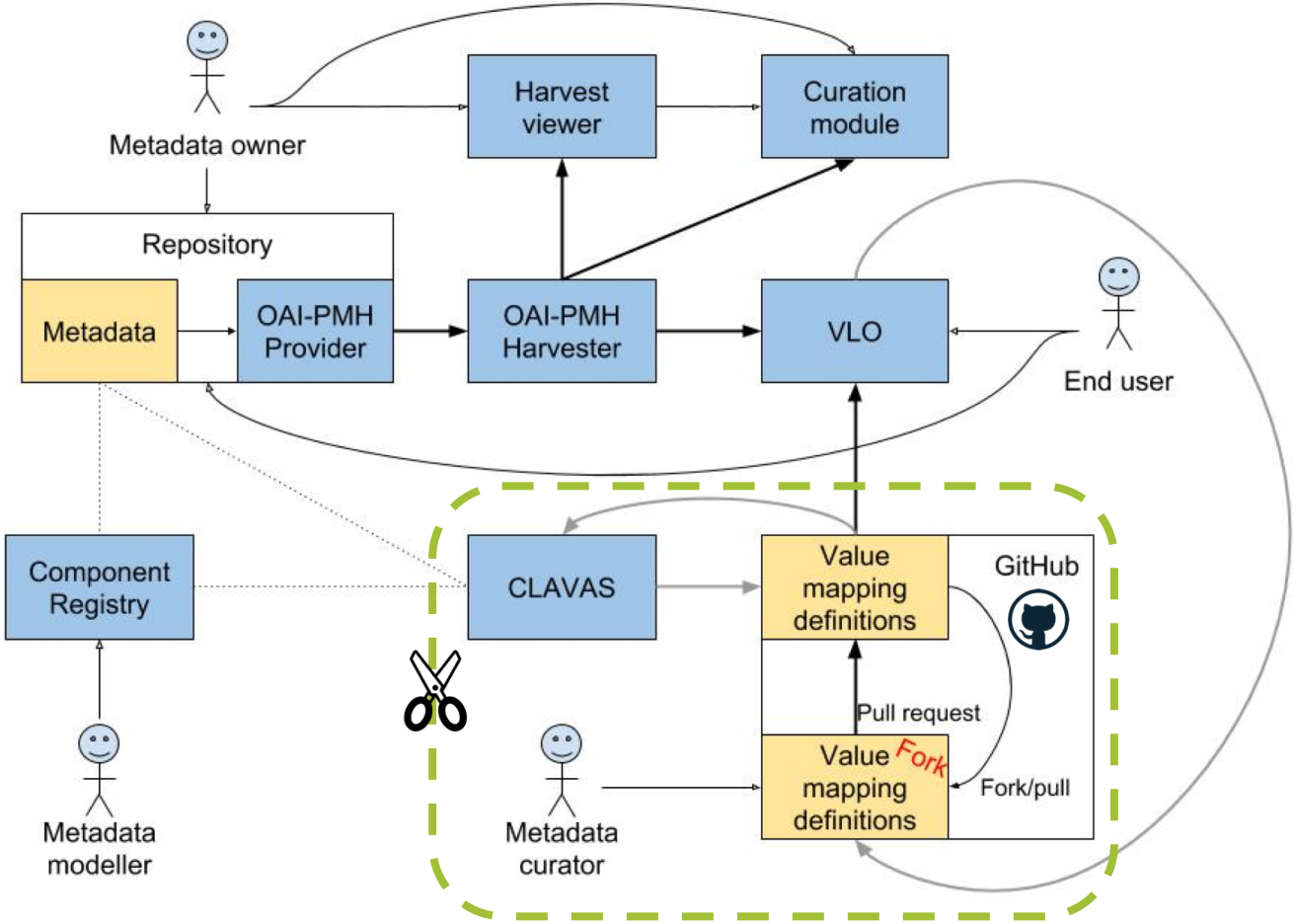
Developments

CLAVAS

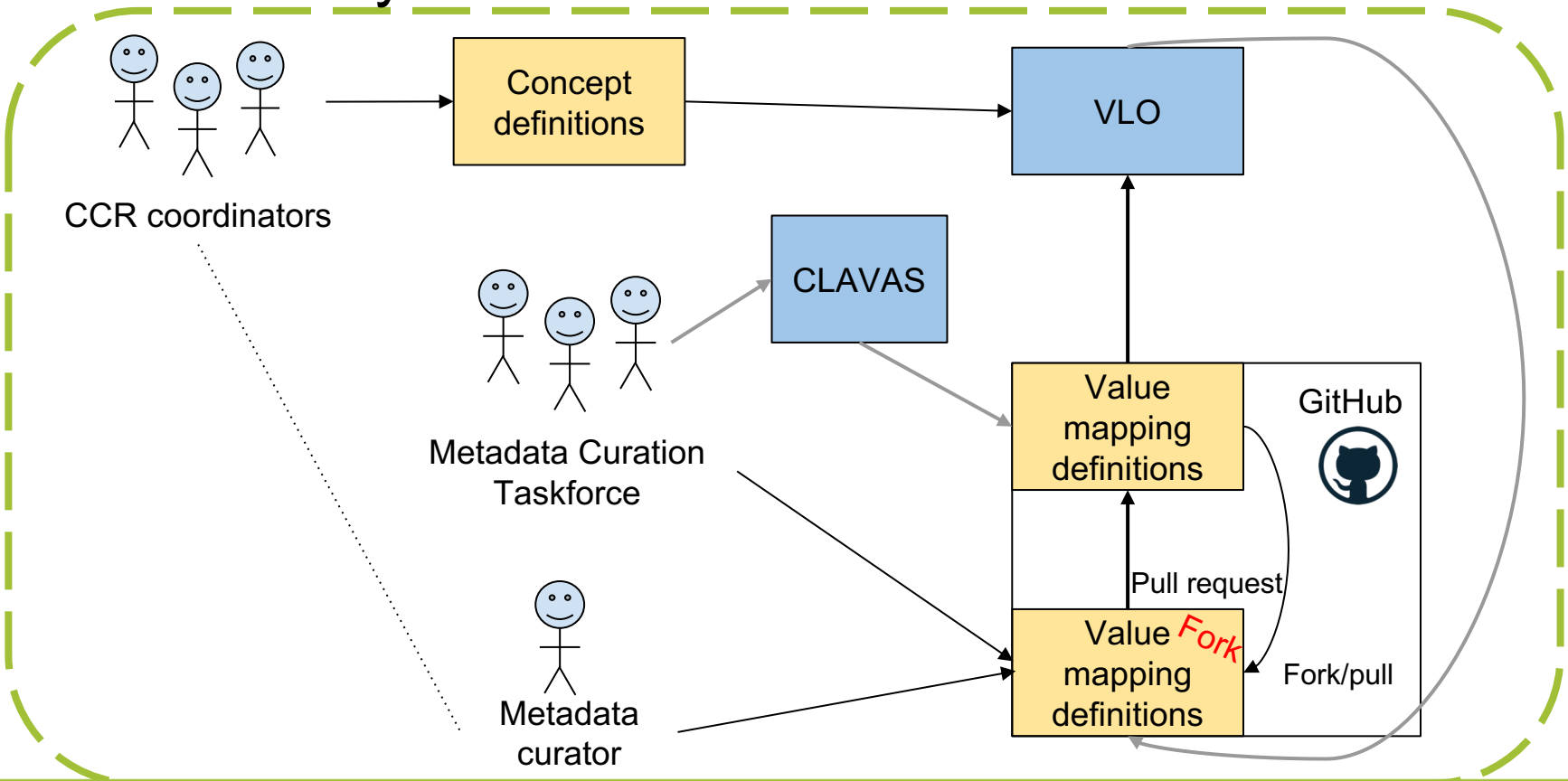
- To become the home of the vocabularies created and owned by the MD Curation Taskforce
  - The CLAVAS vocabulary can contain alternative and hidden labels, which are used in the generated mapping definitions
- In the ecosystem for metadata curators new vocabulary items, incl. labels of various kinds, can appear and mappings for them can be prototyped
  - On a regular basis the CLAVAS vocabulary can be updated to include them

- CLAVAS can also contain other vocabularies, i.e., not related to curation, but only if ownership and maintenance is clear!

Overview


Metadata owner
Harvest viewer
Curation module
Repository
Metadata
OAI-PMH Provider
OAI-PMH Harvester
VLO
End user
Component Registry
CLAVAS
Value mapping definitions
GitHub
Pull request
Value mapping definitions
Fork
Fork/pull
Metadata modeller
Metadata curator

Curation Ecosystem for Metadata Curators



Developments: **resource type mapping**

346 distinct values

Boek

Knjige

Boeken

Book

Early printed book (1501-1800)

Developments: **resource type mapping**

346 distinct values

MovingImage

Video

LanguageDescription

language_description

Developments: **resource type mapping**

346 distinct values

Electronic publications -- Great Britain -- 20th century text
Electronic publications -- Indonesia -- 20th century text
Electronic publications -- Japan -- 20th century
Electronic publications -- United States -- 20th century text
English drama -- Restoration, 1660-1700 text
English literature -- Middle English, 1100-1500 text
Philosophical texts -- Great Britain -- 18th century
Philosophical texts -- Great Britain -- 19th century

Developments: **resource type mapping**

346 distinct values

Nicht dokumentiert

Developments: **resource type mapping**

- Problem:
  - Resource Type values with similar semantics may be represented by different names
  - This led to a significant “expansion” of resource types

- Solution (by MD Curation Taskforce)
  - Define a limited vocabulary of resource types (not too specific, not too generic)
  - Map resource type names to terms in the vocabulary

Developments: **resource type mapping**

346 distinct values

530k records not covered

= 63% of CLARIN/Other harvest sets

Developments: **resource type mapping**

530k records not covered

A few **metadata profile names**:

- TextCorpusProfile
- LexicalResourceProfile
- TreebankProfile
- WebLichtWebService
- Dictionary
- SongAudio

Developments: **resource type mapping**

- Problem:
  - Profiles designed to describe a resource of a specific type do not explicitly encode this information

- Solution (by MD Curation Taskforce)
  - Map CMDI profiles without explicit resource type information to terms in the vocabulary

Resource type facet: January 2018 

346 distinct values

530k records not covered

= 63% of CLARIN/Other harvest sets

Resource type facet: June 2018 

53 distinct values

45k records not covered

= 5% of CLARIN/Other harvest sets

Resource type facet: June 2018 

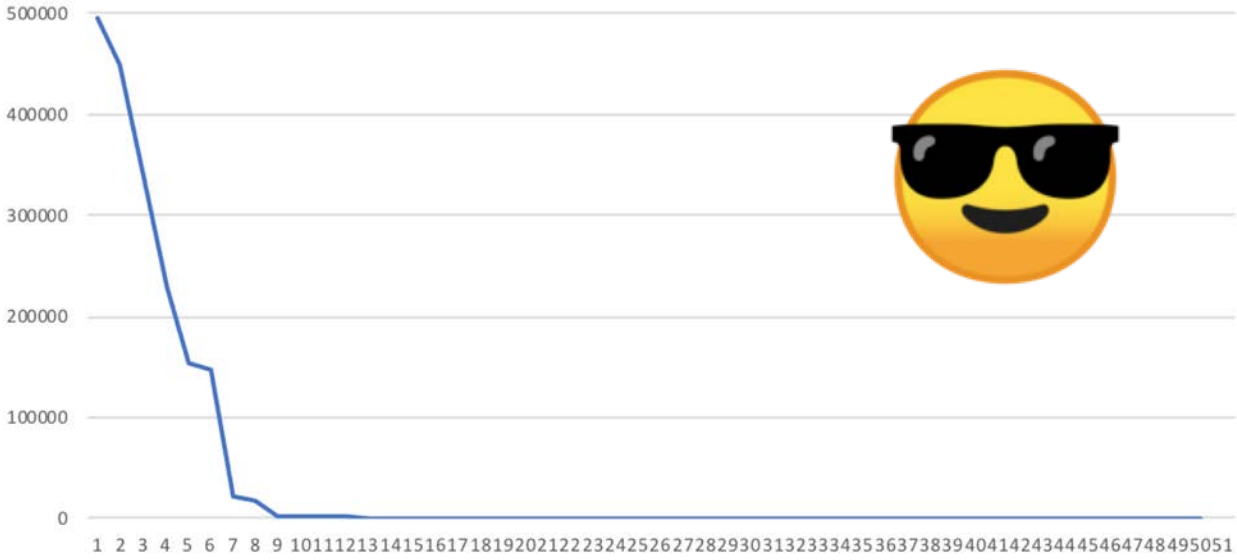
Top values

Text (494502)
Audio (449233)
Annotation (345157)
Image (231336)
Session (154259)
Video (147541)
Collection (21290)
Structured dataset (17757)
Corpus (1544)

Resource type facet: June 2018

Top values

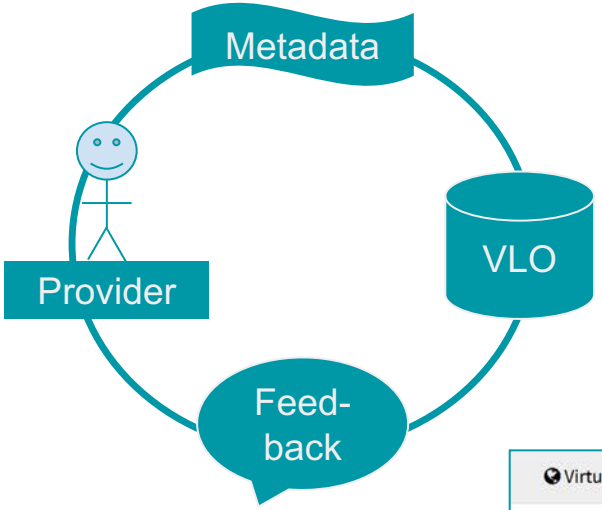
Text (494502)
Audio (449233)
Annotation (345157)
Image (231336)
Session (154259)
Video (147541)
Collection (21290)
Structured dataset (17757)
Corpus (1544)

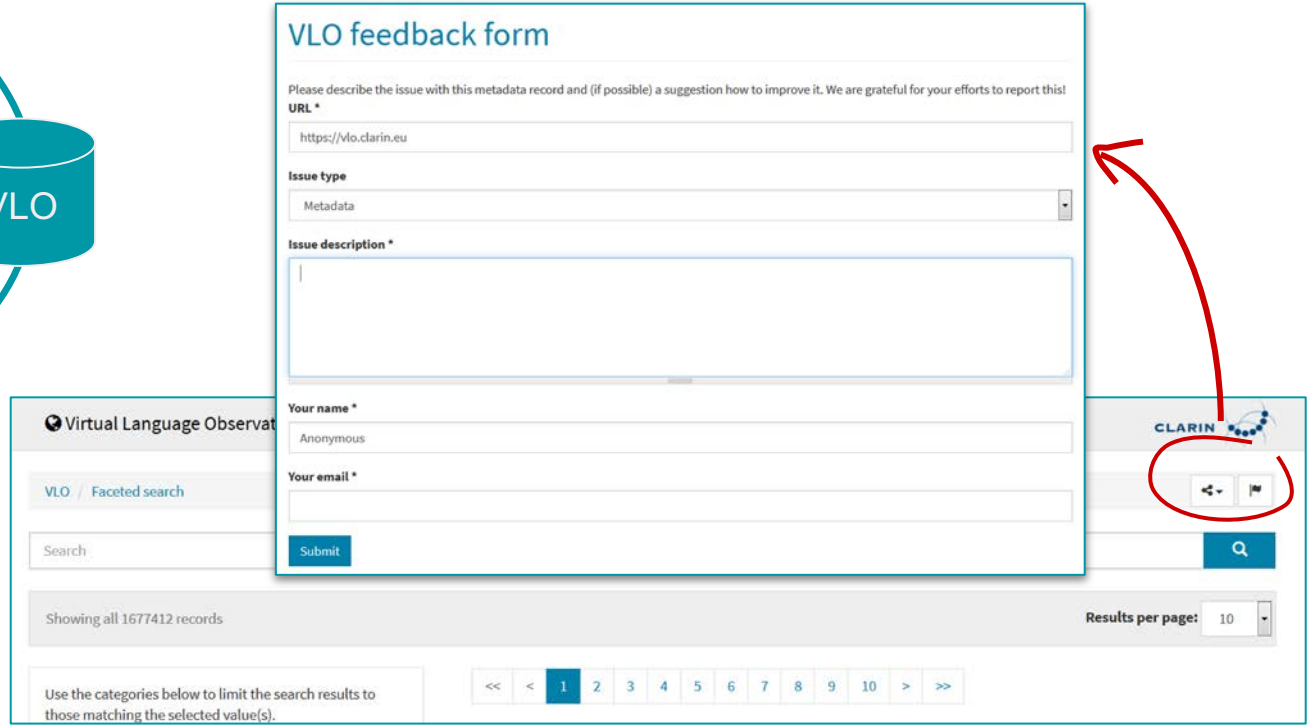


Collaborative Workflow

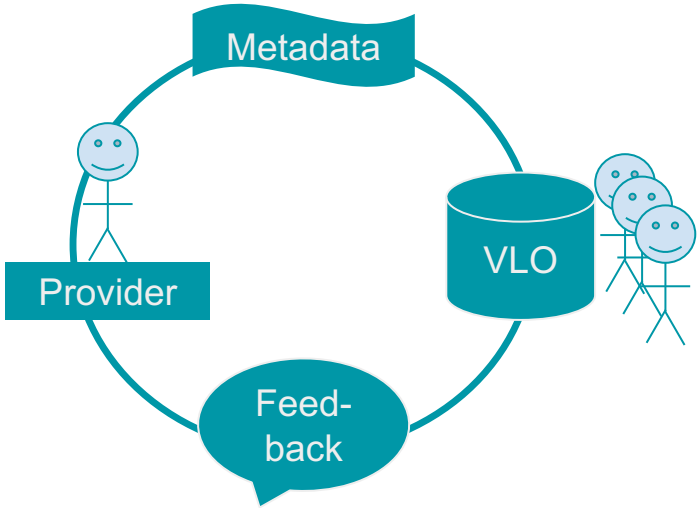
- **MD curation** in CLARIN is a collaborative task
- Involves various groups of specialists
- Has to be attended to at different points in the workflow
- There are several aspects to it, e.g.
  - VLO: Fixing the concept→facet mapping
  - Component Registry: Fixing semantic mappings in metadata components (concept links)
  - VLO/Helpdesk: Reporting quality issues in metadata records and working with metadata owners to resolve these
  - CLAVAS: Defining and refining vocabularies
  - VLO: Defining post-hoc value mapping rules
- Can be done actively (by actually fixing problems) or as a reporter

Collaborative Workflow: VLO Feedback loop





Collaborative Workflow: Curation TF Feedback loop



Reminder:

Tomorrow 9:00
Hands-on curation session for metadata providers

Afternoon 15:30
Taskforce meeting for (aspiring) curators

Thank you!