Dataverse & CMDI

a proof of concept

Dataverse as a CLARIN compatible repository?

Menzo Windhouwer
menzo.windhouwer@di.huc.knaw.nl

Dataverse

- Open source research data repository software

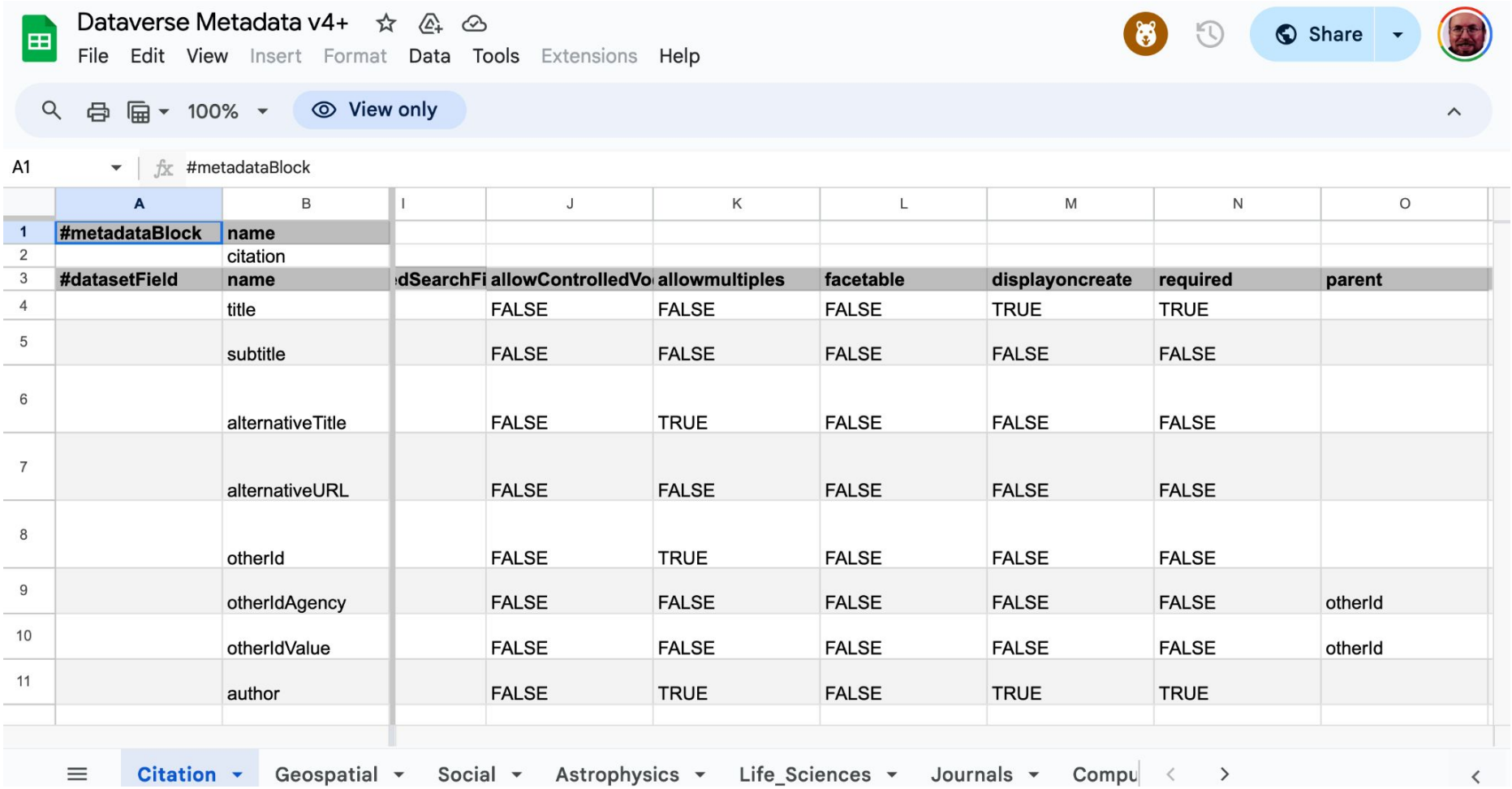
dataverse.org

- The Institute for Quantitative Social Science (IQSS) at Harvard
- Java-based
- One-stop shop a la DSpace
  - for changes you need to "break into" the codebase
    - get your changes into main is hard, see DANS
    - maintaining a fork is also hard, see CLARIN DSpace
  - vs modular/component setups a la Fedora Commons

Metadata Blocks

KNAW Humanities Cluster /DI

- a set of metadata fields belonging together
- specified in a TSV file
  - google spreadsheet
  - documentation
  - fields can be grouped, e.g. author
- upload to the server
- enable it
- multiple blocks can be enabled and filled in
- the citation block is mandatory

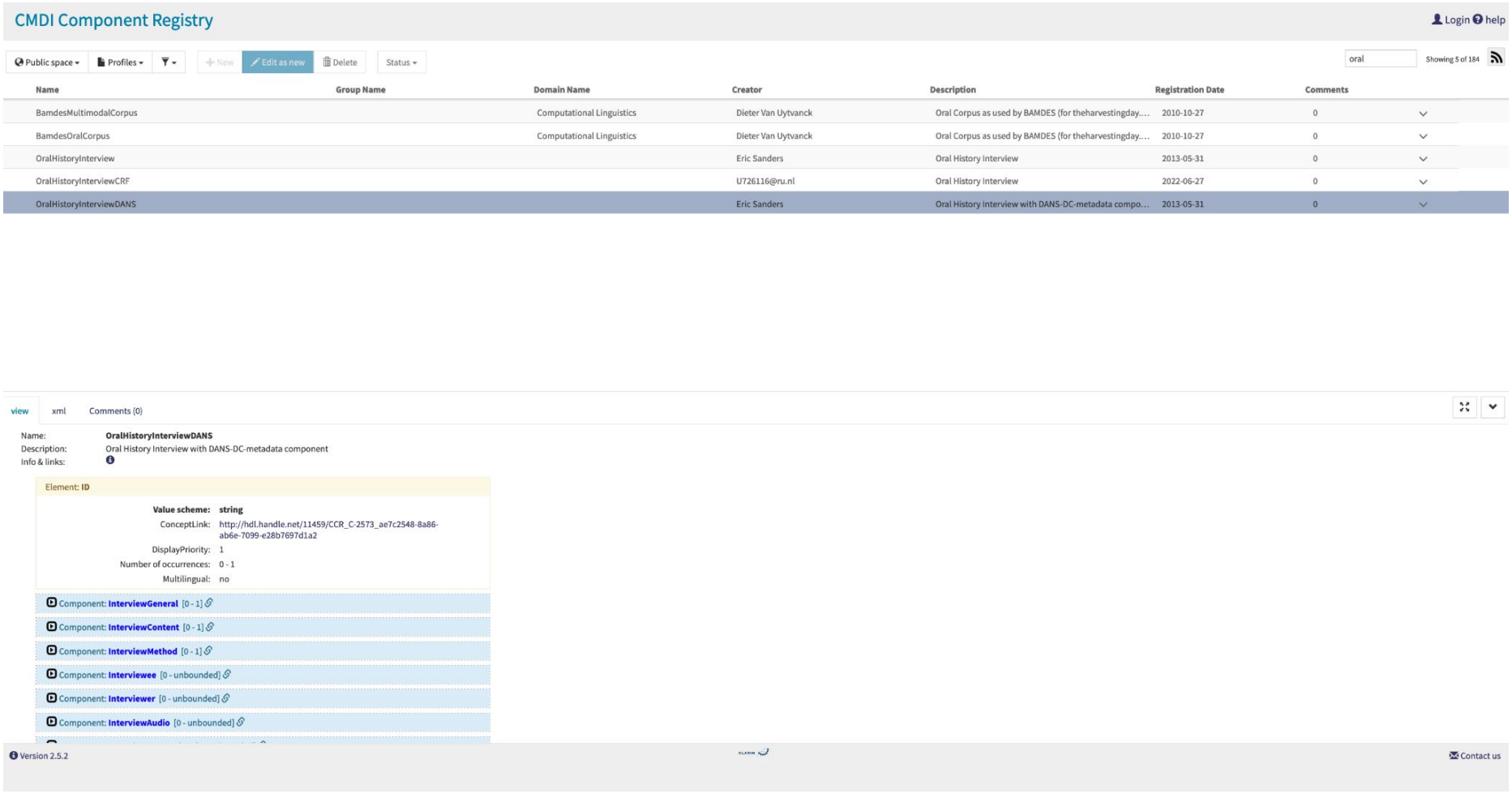


	A	B	I	J	K	L	M	N	O
1	**#metadataBlock**	**name**							
2		citation							
3	**#datasetField**	**name**	**dSearchFi**	**allowControlledVo**	**allowmultiples**	**facetable**	**displayoncreate**	**required**	**parent**
4		title		FALSE	FALSE	FALSE	TRUE	TRUE	
5		subtitle		FALSE	FALSE	FALSE	FALSE	FALSE	
6		alternativeTitle		FALSE	TRUE	FALSE	FALSE	FALSE	
7		alternativeURL		FALSE	FALSE	FALSE	FALSE	FALSE	
8		otherId		FALSE	TRUE	FALSE	FALSE	FALSE	
9		otherIdAgency		FALSE	FALSE	FALSE	FALSE	FALSE	otherId
10		otherIdValue		FALSE	FALSE	FALSE	FALSE	FALSE	otherId
11		author		FALSE	TRUE	FALSE	TRUE	TRUE	

Dataverse & CMDI: starting from CMD

KNAW Humanities Cluster /DI

- turn a CMD profile into a metadata block
- stylesheet
- example: OralHistoryInterviewDANS, record
  - various nesting levels
- the metadata block specification, i.e. the TSV, can handle it ✅
- however, the editor can't
  - deeper component nesting goes wrong
  - e.g. Interview^{P} → Interviewer^{C} ✅ → Actor^{C} ❌


CMDI Component Registry
Login help
Public space
Profiles
New
Edit as new
Delete
Status
oral
Showing 5 of 184
Name
Group Name
Domain Name
Creator
Description
Registration Date
Comments
BamdesMultimodalCorpus
Computational Linguistics
Dieter Van Uytvanck
Oral Corpus as used by BAMDES (for theharvestingday....
2010-10-27
0
BamdesOralCorpus
Computational Linguistics
Dieter Van Uytvanck
Oral Corpus as used by BAMDES (for theharvestingday....
2010-10-27
0
OralHistoryInterview
Eric Sanders
Oral History Interview
2013-05-31
0
OralHistoryInterviewCRF
U726116@ru.nl
Oral History Interview
2022-06-27
0
OralHistoryInterviewDANS
Eric Sanders
Oral History Interview with DANS-DC-metadata compo...
2013-05-31
0
view
xml
Comments (0)
Name: OralHistoryInterviewDANS
Description: Oral History Interview with DANS-DC-metadata component
Info & links:
Element: ID
Value scheme: string
ConceptLink: http://hdl.handle.net/11459/CCR_C-2573_ae7c2548-8a86-ab6e-7099-e28b7697d1a2
DisplayPriority: 1
Number of occurrences: 0 - 1
Multilingual: no
Component: InterviewGeneral [0 - 1]
Component: InterviewContent [0 - 1]
Component: InterviewMethod [0 - 1]
Component: Interviewee [0 - unbounded]
Component: Interviewer [0 - unbounded]
Component: InterviewAudio [0 - unbounded]
Version 2.5.2
Contact us

Copy of Dataverse Metadata v4+

File Edit View Insert Format Data Tools Extensions Help

Share

A1 #metadataBlock

	A	B	C	D	E	F	G	H	I	J
1	#metadataBlock	name	dataverseAlias	displayName	CMDProfile					
2		OralHistoryInterv		OralHistoryInterv	clarin.eu:cr1:p_1369752611610					
3	#datasetField	name	title	description	watermark	fieldType	displayOrder	displayFormat	advancedSearch	allowControlledV
4		ID	ID			text	1		FALSE	FALSE
5		InterviewGenera	InterviewGenera			none	2		FALSE	FALSE
6		NumberOfSpeak	NumberOfSpeak			text	1		FALSE	FALSE
7		CreationDate	CreationDate			text	2		FALSE	FALSE
8		PublicationDate	PublicationDate			text	3		FALSE	FALSE
9		Duration	Duration			text	4		FALSE	FALSE
10		Owner	Owner			text	5		FALSE	FALSE
11		Genre	Genre			text	6		FALSE	TRUE
12		Modality	Modality			none	7		FALSE	FALSE
13		Modality	Modality			text	1		FALSE	TRUE
14		Description	Description			none	2		FALSE	FALSE
15		Description	Description			text	1		FALSE	FALSE
16		Multilinguality	Multilinguality			none	8		FALSE	FALSE
17		Multilinguality	Multilinguality			text	1		FALSE	TRUE
18		Access	Access			none	9		FALSE	FALSE

MRA (custom) Digaai (custom) CHIA (custom) PSRI (custom) OralHistoryInterview

OralHistoryInterviewDANSMetadata

ID

InterviewGeneral

One or more of these fields may become required if you add to one or more of these optional fields.

NumberOfSpeakers
CreationDate
PublicationDate
Duration
Owner
Genre
Select...
Access
Creators
Location

InterviewContent

InterviewKeyWords
InterviewSummary
Mission

InterviewMethod

ParticipantName
InterviewSort
RecruitmentMethod
PreInterviewInformation
DataCollectionMethod
TopicList

Interviewee

Interviewer

One or more of these fields may become required if you add to one or more of these optional fields.

RelationToInterviewee
RelationToProject
BirthPlace
ResidencePlace
Actor

Dataverse & CMDI: starting from DV

KNAW Humanities Cluster /DI

- create a working Dataverse metadata block
- convert that to a CMD profile/component
  - stylesheet 1 & 2
  - example: citationProfile
  - needs admin rights to import easily into the ComponentRegistry
    i. run stylesheet on your TSV
    ii. create empty component/profile
    iii. mail XML & id to cmdi@clarin.eu
    iv. complete the profile/components
      - concept links (VLO mapping)
      - replace inline components with external components

CMDI Component Registry

menzo.windhouwer@di.huc.knaw.nl settings help admin

dataverse | Profiles | New | Edit | Move to team | Publish | Delete | Status | Type to filter... | Showing 1 of 1

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
citationProfile			Menzo Windhouwer	4Dataverse	2022-02-27	0

view | xml | Comments (0)

Name: **citationProfile**
Description: 4Dataverse
Info & links:

Element: **title**

Value scheme: **string**
DisplayPriority: 1
Number of occurrences: 1 - 1
Multilingual: no

Element: **subtitle**

Value scheme: **string**
Number of occurrences: 0 - 1
Multilingual: no

Element: **alternativeTitle**

Value scheme: **string**
Number of occurrences: 0 - 1
Multilingual: no

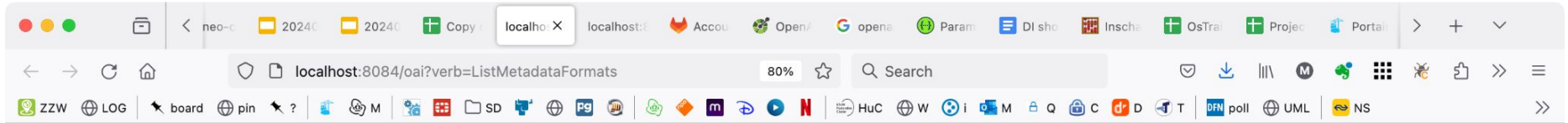
Element: **subject**

Dataverse & CMDI: OAI

- extended the OAI provider with support for CMDI
  - metadataFormat
  - records
    - stylesheet
    - needs a config to map the block to a profile
      - unfortunately we couldn't extend the TSV
      - path is passed on via an environment variable

- caveats
  - Dataverse 5.9 (6.x is the current)
  - currently only does one block to a profile, should become multiple blocks mapped to components in a profile

https://github.com/menzowindhouwer/dataverse/blob/develop%2Bct/CMDI.md



This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2024-06-11T07:48:08Z</responseDate>
  <request verb="ListMetadataFormats">http://localhost:8084/oai</request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>CMDI</metadataPrefix>
      <schema>
        https://infra.clarin.eu/CMDI/1.x/xsd/cmd-envelop.xsd
      </schema>
      <metadataNamespace>http://www.clarin.eu/cmd/1</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>Datacite</metadataPrefix>
      <schema>
        http://datacite.org/schema/kernel-3 http://schema.datacite.org/meta/kernel-3/metadata.xsd
      </schema>
      <metadataNamespace>http://datacite.org/schema/kernel-3</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>oai_ddi</metadataPrefix>
      <schema>
        https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd
      </schema>
      <metadataNamespace>ddi:codebook:2_5</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>oai_datacite</metadataPrefix>
      <schema>
        http://schema.datacite.org/meta/kernel-4.1/metadata.xsd
      </schema>
      <metadataNamespace>http://datacite.org/schema/kernel-4</metadataNamespace>
    </metadataFormat>
    <metadataFormat>
      <metadataPrefix>dataverse_json</metadataPrefix>
      <schema>JSON schema pending</schema>
      <metadataNamespace>
        Custom Dataverse metadata in JSON format (Dataverse4 to Dataverse4 harvesting only)
      </metadataNamespace>
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
−<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2024-06-12T12:33:12Z</responseDate>
  <request verb="ListRecords" metadataPrefix="CMDI">http://localhost:8084/oai</request>
 −<ListRecords>
  −<record>
   −<header>
     <identifier>doi:10.5072/FK2/FGSWBF</identifier>
     <datestamp>2024-06-07T15:07:21Z</datestamp>
    </header>
   −<metadata>
    −<cmd:CMD xsi:schemaLocation="http://www.clarin.eu/cmd/1 https://infra.clarin.eu/CMDI/1.x/xsd/cmd-envelop.xsd http://
     www.clarin.eu/cmd/1/profiles/clarin.eu:cr1:p_1639731773881 https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/
     clarin.eu:cr1:p_1639731773881/xsd" CMDVersion="1.2">
     −<cmd:Header>
       <cmd:MdCreator>Root</cmd:MdCreator>
       <cmd:MdCreationDate>2024-06-07</cmd:MdCreationDate>
       <cmd:MdSelfLink>https://doi.org/10.5072/FK2/FGSWBF</cmd:MdSelfLink>
       <cmd:MdProfile>clarin.eu:cr1:p_1639731773881</cmd:MdProfile>
      </cmd:Header>
     −<cmd:Resources>
      −<cmd:ResourceProxyList>
       −<cmd:ResourceProxy id="lp">
```

Dataverse & CMDI: mix & match

KNAW Humanities Cluster /DI

1. create a "flat" profile
2. to TSV
3. test as metadataBlock
4. (OAI) test in VLO
5. if needed update profile, and back to 2.

(doesn't needs admin rights)

Conclusion

KNAW Humanities Cluster /DI

Dataverse as a CLARIN compatible repository?

- yes, that's possible 🤓
  - but with a limited use of the powers of CMDI
- as a fork or as part of the core?

Harvesting from Dataverse?

- yes, CMDI using 👆
- or other XML-based formats
  - from DANS we (can) get Datacite/DC via OAI
- or we might also "harvest" the JSON
  - CLARIN's harvester has been extended in the dutch CLARIAH+ project to use other "protocols" than OAI

KNAW Humanities Cluster /DI

Special thanks!

Vic Ding
Slava Tykhonov
Eko Indarto

Speak to Slava for a (CMDI) KG on the side/top of Dataverse & (general) AI good/badness!

CLARIN