Flexible metadata schemes for research data repositories
The Common Framework in Dataverse and the CMDI use case

Jerry de Vries
DANS-KNAW
The Netherlands
jerry.de.vries@dans.knaw.nl

Vyacheslav Tykhonov
DANS-KNAW
The Netherlands
vyacheslav.tykhonov@dans.knaw.nl

Andrea Scharnhorst
DANS-KNAW
The Netherlands
andrea.scharnhorst@dans.knaw.nl

Eko Indarto
DANS-KNAW
The Netherlands
eko.indarto@dans.knaw.nl

Femmy Admiraal
DANS-KNAW
The Netherlands
femmy.admiraal@dans.knaw.nl

Mike Priddy
DANS-KNAW
The Netherlands
mike.priddy@dans.knaw.nl

Abstract

This paper presents how DANS, which participates in the CLARIAH+ project, works on a Common Framework which makes it possible to expose CMDI metadata via a DANS discovery service. The Common Framework refers to discussions in CLARIN about integrating standards in Dataverse. This paper informs CLARIAH+ about the explorations of the envisioned use of the Common Framework and reports about the possibilities and challenges of the interoperability of these metadata schemes. The challenges faced are: First, a proposal of a core set of CMDI metadata as recommendation. Second, the extraction of CMDI metadata and transform and load the metadata fields into the Dataverse core set of metadata. Third, a workflow for prediction and linking concepts from external controlled vocabularies to CMDI metadata values. Fourth, the extension of the Common Framework with support for FAIR controlled vocabularies to create FAIR metadata. Fifth, the extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format

1. Introduction

Research data repositories are increasingly expected to operate together. Standardization and alignment of metadata schemes used to describe datasets are a precondition for any platform to work (see as an example https://datacite.org). At the same time, data repositories usually serve specific knowledge domains, and have tailored their indexing practices towards those communities. In short, there is a tension between serving one or few communities in a very detailed manner and being integratable into a cross-domain platform. The Dataverse community responded to this natural tension by offering both a standard, common core set of metadata called Citation Block and the possibility to extend this core set with custom fields defined as a discipline specific metadata block.

This paper discusses in detail challenges and solutions when it comes to implement Common Framework principles into a very concrete Dataverse instance and a very concrete community.

(Conzett et al., 2020) More specifically, this paper reports how the Data Archive and Network Services institute (DANS), which participates in the CLARIAH+ project, works on a Common Framework which makes it possible to expose CMDI metadata via a DANS discovery service. With Common Framework we refer here to discussions about standards in CLARIN integrated in Dataverse. The envisioned use of the exploration reported in this paper is two-fold: primarily, it informs CLARIAH+ about possibilities and challenges when it comes to the interoperability of metadata schemes; secondly, it informs DANS as service provider of a long-term archive in its use of a technological backbone. DANS is currently migrating its research data archiving service from a Fedora-based platform (DANS-EASY) to Dataverse and introduced so-called DANS Datastations as specific Dataverse instances for designated communities. (Wals, 2021). The exploration described in this paper bases its analytic part on the current production system while at the same time, also informs the on-going migration process.

In the current DANS-EASY archive, CLARIN datasets are tagged as part of a specific collection containing 29 datasets. But, much more datasets use CMDI metadata, as search for 'CMDI' reveals (1096 datasets). The use of CMDI is often noted in either the Description metadata field or the Form metadata field of the Dublin Core Standard the current EASY is using. While the use of CMDI is noted, one cannot search in those CMDI notations. The CMDI based indexes are delivered as specific, additional files of the dataset, and hence not automatically searchable (in short called CMDI files). The core of the exploration of this paper is the development of a so-called Extract, Transform and Load-pipeline (ETL). This pipeline extracts all metadata fields from CMDI files archived in CLARIN datasets at DANS-EASY archive and automatically transforms this metadata to the defined core set of CMDI metadata and loads this metadata at a DANS Datastation. This results in findable and harvestable CLARIN metadata which is interoperable with CLARIN discovery services.

2. Five challenges

In this paper we detail challenges (and partly envisioned solutions) we are facing in our work on the Common Framework to expose CMDI metadata (ISO 24622-1:2015) (ISO 24622-2:2019) via a DANS discovery service and our work during the migration of the DANS archiving service to the DANS Datastation. Both are still ongoing processes.

2.1. Challenge 1: A proposal of a core set of CMDI metadata as recommendation

The first challenge is the fact that CMDI itself acts as a recommended standard, but that there is (not yet) a defined core set of CMDI metadata. (Goosen et al., 2014) The CLARIN taskforce CMDI, in which we are participating, is currently working on a proposal and acceptance of a core set of CMDI metadata as a recommendation for all CLARIN centers. This core set of metadata will be the basis for an extension of the Dataverse core set of metadata for describing CLARIN datasets when depositing them in the corresponding DANS Datastation

2.2. Challenge 2: Extraction of CMDI metadata and transform and load the metadata fields into the Dataverse core set of metadata

The second challenge concerns the Dataverse software itself, and how to best extend the core Dataverse metadata schema with a set of metadata whereby each metadata field should be part of some CMDI component and linked to the CMDI component registry. The creation of a pipeline to Extract, Transform and Load CMDI metadata fields into the Dataverse Core Set is part of this challenge. This ETL-pipeline uses the Dataverse DDI Converter tool which so far only supports

customized XSLT mappings for xml input. We are extending this functionality with Jinja2 templating in combination with key-value mapping for csv input. In this case the DDI Converter tool will not only be depending on xml input, to guarantee a broader use of the converter tool in the ETL-pipeline.

2.3. Challenge 3: Workflow for prediction and linking concepts from external controlled vocabularies to the CMDI metadata values

The third challenge consists in the extension of such a Common Framework for Dataverse beyond the CMDI case. Beyond the extension of the Citation Core set, what is also envisioned is to support a link between other 'indexing' metadata fields to the other Knowledge Organization System Providers. In particular, we think here of recommended FAIR controlled vocabularies and ontologies which potentially become part of the set of metadata fields. (Wilkinson et al., 2016) (Broeder et al., 2021), (Wang et al., 2021) Coming back to the CMDI case, this could lead to linking a or any CMDI metadata value to a recommended ontology or controlled vocabulary**.**

2.4. Challenge 4: Extension of the Common Framework with support for FAIR controlled vocabularies to create FAIR metadata

The fourth challenge is the support of FAIR controlled vocabularies. We use the SKOSMOS framework developed at the Finnish National Library. A semi-automatic workflow, which uses a SKOSMOS API, is developed to query any SKOSMOS representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint is used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields. These metadata fields link to the CMDI component registry in the CMDI metadata schema.

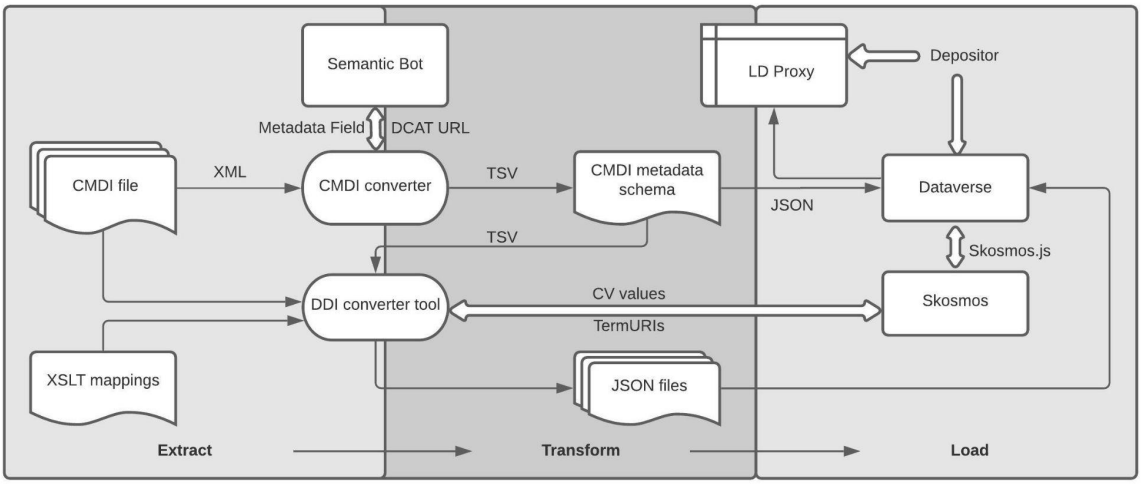
2.5. Challenge 5: Extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format

The fifth and last challenge is to extend the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format, following the used specification.

To support all these workflows in the Common Framework, the Apache Airflow is investigated as a proper solution to implement a reliable deposit pipeline. All insights and workflows will be shared with the CLARIN and CLARIAH community and we're looking for the collaboration on semantic mappings that should be used to get an appropriate ontology linkage not only on value level but also between fields available in CMDI Component Registry. Another task is to link CMDI component fields to the common ontologies like DCAT2 and Dublin Core using RDF Modelling Language (RML). The ultimate goal is to leverage the metadata with a search engine developed as a part of the ODISSEI portal that will allow users to search across linked concepts from the different disciplines available in the Linked Open Data Cloud (LoD) including linguistics sources (CLARIN). The same semantic mappings could be reused on a global scale to get CMDI datasets disseminated in the FAIR Data Points developed in the FAIRsFAIR project.

3. A CMDI pipeline

Describe the complete pipeline here and use image to show all components



4. From use case to general framework

4.1. From metadata to linked data

4.2. Flexible Semantic Mapping Framework (SEMAF)

A proposal for a Flexible Semantic Mapping Framework was conducted based on a number of interviews of people representing different research communities. (Broeder et al., 2021)

4.3. Knowledge graphs as fuel to power Open Science

4.4. Knowledge graphs for research data

4.5. Building a technology agnostic solution

4.6. Future developments in the distributed data management

4.7. FAIRness and FAIR Data Point

5. Future work

By supporting the enrichment of metadata, we help to make CLARIN datasets Findable and Accessible and make them Interoperable with other CLARIN datasets, and so ultimately also support Re-usability. FAIR compliance automatic of assessment tools, like F-UJI can be included in the Common Framework to evaluate the quality of the metadata. (Devaraju et al., 2020, 2021)

Future work will incorporate the application of Artificial Intelligence to get an automatic linkage of relations between CMDI values and relevant ontologies and controlled vocabularies, and create a semi-automatic mapping tool to generate RDF mappings for CMDI fields.

6. Conclusion

This paper elaborates on our (experimental) work of building a Common Framework to expose CMDI metadata via a DANS discovery service. This work relates to the migration of the DANS archive service to (a) newly to build DANS Datastation(s), which will serve as a basis for the discovery service.

However, the work is still ongoing and the challenges we reflect about when addressed unavoidably are leading to new challenges. For instance, we have been able to extend the Dataverse metadata model with a proposed core set of CMDI metadata which serves the needs of DANS as a basis for the discovery service. This resulted in a flexible solution which is easy to adjust in case the core set of CMDI metadata will be changed in the future. Its implementation in production services is still a challenge ahead.

To get to the proposed core set of CMDI metadata, we have analyzed all CMDI metadata stored in the DANS-EASY archive with the CMDI exploration tool. With the same tool we are able to transform each CMDI metadata file to the proposed core set.

To make the new metadata FAIR, we explored the possibilities of enriching the metadata with recommended external controlled vocabularies. This exploration has led to a flexible and generic solution to add custom external controlled vocabularies to Dataverse beyond the immediate CMDI case. A semi-automatic workflow, which uses a SKOSMOS API, is developed to query any SKOSMOS representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint is used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields.

To extend the semi-automatic workflow we have started to explore the possibilities of semantic gateway. We have started a proof of concept with a semantic gateway lookup API. This API is able to return a list of standardized concepts based on the selected vocabulary and a term. This will help to link each field in the proposed core set of metadata to the appropriate controlled vocabulary.

To make the circle round again we are in the phase of investigating the export of the Dataverse metadata back to the original CMDI format. The basic requirement for this should be that the Dataverse metadata schema must have CMDI metadata that can be extended with custom components which are used by the different CLARIN centers. Second, the original relationships between fields and concepts should be kept whereby the custom components should be added to a SKOS schema. If this will be possible, we should be able to reproduce the original CMDI metadata, which could be offered for download to any user without losing the quality of the original metadata.

This work has taught us that looking to the future and setting ourselves for big challenges is leading to new challenges. But these challenges are motivating us to build proper solutions with and for the community.

References

Broeder, D., Budrononi, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Weiland, C., Wittenberg, P. and Zwolf, C. M. 2021. SEMAF: *A Proposal for a Flexible Semantic Mapping Framework (version 1.0).* Zenodo. http://doi.org/10.5281/zenodo.4651421

Conzett, P., Goosen, T., Scharnhorst, A., Tykhonov, V., Van Uytvanck, D., de Vries, J. and Wittenberg, M. 2020, *How to weave domain specific information sources into a large, FAIR data fabric for the Digital Humanities? The use of the Dataverse platform.* In S Barbiers, A Fokkens & C Olesen (eds), Proceedings of the DH Benelux 2020. Zenodo. https://doi.org/10.5281/zenodo.3879031

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. and White, A. 2020. *FAIRsFAIR Data Object Assessment Metrics.* https://doi.org/10.5281/zenodo.4081213

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Åkerman, V., L'Hours, H., Davidson, J., and Diepenbroek, M. 2021. *From Conceptualization to Implementation: FAIR Assessment of Research Data Objects*. Data Science Journal, vol. 20, no. 1. https://doi.org/10.5334/dsj-2021-004

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Durco, M. and Schonefeld, O. 2015. *CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure.* in J Odijk (ed.), Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands, 116:004, Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköpings universitet, Linköping, pp. 36-53. http://www.ep.liu.se/ecp/116/004/ecp15116004.pdf

ISO 24622-1:2015. (2015). *Language resource management – Component metadata infrastructure (CMDI) – Part 1: The Component metadata model*. Standard, International Organization for Standardization, Geneva, CH

ISO 24622-2:2019. (2019). *Language resource management – Component metadata infrastructure (CMDI) – Part 2: The Component metadata specification language.* Standard, International Organization for Standardization, Geneva, CH

Wals, H. 2021. Focus on FAIR. DANS: Dutch national centre of expertise and repository for research data. DANS Strategy 2021-2025. The Hague, https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/dans-2021-2025/UK_DANS20212025.pdf

Wang, M., Qiu, L. and Wang, X. 2021, *A Survey on Knowledge Graph Embeddings for Link Prediction*. Symmetry, 13, 458. https://doi.org/10.3390/sym13030485

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Nature, 3, 160018. https://doi.org/10.1038/sdata.2016.18

https://pure.mpg.de/rest/items/item_1480943_7/component/file_1481203/content