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
vyachesav.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

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

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

Abstract

In this paper we present an approach called Common Framework, which addresses issues of interoperability and flexibility of metadata schemes as developed by specific scientific communities, and as later supported by domain and cross-domain data repositories. The approach was triggered by a very concrete use case, namely the question how to expose Component Metadata Infrastructure (CMDI) metadata, stored in computational linguistics datasets in the DANS-EASY archive, for discovery services. The work in CLARIN to push further for the development of CMDI into a standard (ISO 24622-1:2015, ISO 24622-2:2019) forms part of the background of the use case. We used the Dataverse platform to deliver proof of concepts for various elements of the Common Framework, including the recommendation of standardised elements for Dataverse instances in CLARIN. At the core of the Common Framework is a design which envisions an interaction between different microservices, possibly also hosted by various service providers. Mechanisms of semantic mapping are used throughout a pipeline which starts at a set of existing metadata standards and values at a digital research data repository (Extraction) and their analysis. This leads to an alignment of these metadata standards with others standards (Transformation) and proposes enrichments to be used by other service providers but also to be imported back to the original source (Load). Some modules applied along this pipeline are discussed in detail, together with the challenges this specific use case entails. At the same time, we also stress generic aspects, as we are convinced that this approach can also be applied in other settings, other archival platforms and other domain specific metadata schemes. The high-level goal of this exploration is to explore ways to make research data collections FAIR (Findable, Accessible, Interoperable and Re-usable), and in particular interoperable and re-usable, while preserving the rigour of domain specific indexing practices.

Jerry de Vries, Vyacheslav Tykhonov, Andrea Scharnhorst, Eko Indarto, Mike Priddy and Femmy Admiraal 2022. Flexible Metadata Schemes for Research Data Repositories. The Common Framework in Dataverse and the CMDI Use Case. *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 168–180. DOI: https://doi.org/10.3384/9789179294441

1 Introduction

Research data repositories are increasingly expected to operate together. Standardisation and alignment of metadata schemes used to describe (index) datasets are a precondition for any platform to work (for example see 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 rigorous manner and being integratable into cross-domain platforms (see Figure 1).

The tension between specificity of metadata schemes and a genericity which enables interoperability is nothing new (e.g., Guéret et al., 2013). We see similar debates around the emergence of universal classifications in the bibliographic domain at the beginning of the twentieth century (e.g., Dewey and UDC) (McIlwaine, 2010); reinforced with the introduction of automatization in classification and indexing (Svenonius, 2000); and reappearing in a different shape with the emergence of web services. Currently, (traditional) phrases as *crosswalks, alignment, catalogues* mark the quest for interoperability in the growing universe of domain specific ontologies, classifications, thesauri which become semantic artefacts when living in the web (Hugo et al., 2020; European Commission, Directorate-General for Research and Innovation et al., 2021).

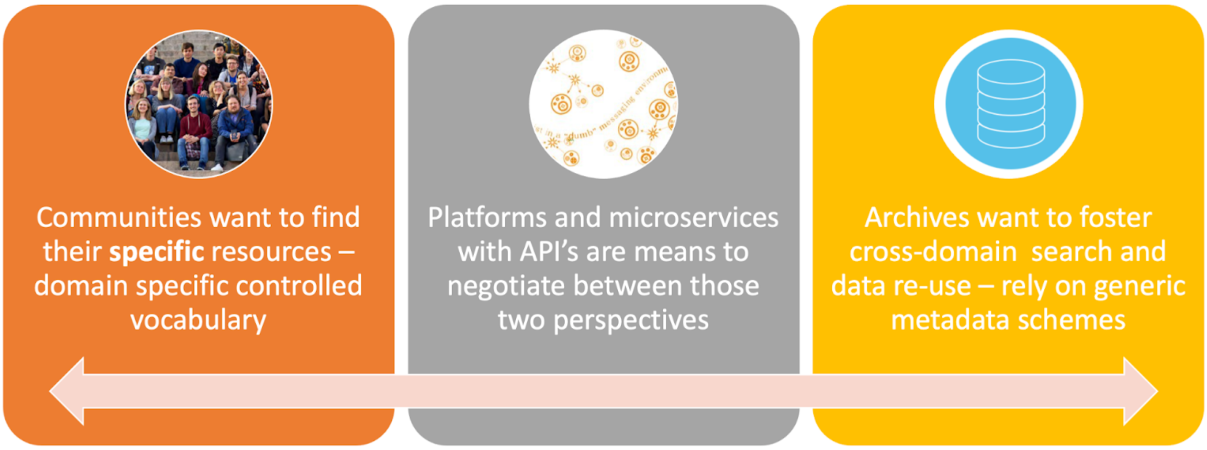

Communities want to find their **specific** resources – domain specific controlled vocabulary
Platforms and microservices with API's are means to negotiate between those two perspectives
Archives want to foster cross-domain search and data re-use – rely on generic metadata schemes

Figure 1. Tension between specific resources and generic metadata schemes

While almost trivial, one cannot overemphasise the fact that the organisation of knowledge, and of systems and practices around it, is deeply contextualised, and depends on the concrete purpose for such knowledge orders to emerge (Smiraglia and Scharnhorst, 2021). So, it cannot be a surprise that each knowledge domain, each scientific field or speciality according to its epistemic frameworks and research perspectives develops its own specific organisation of knowledge. One could even say that research knowledge organisation systems have an element of being intrinsically not-interoperable. This is the curse to the virtue of deepening our knowledge in an ever more differentiated and specialised knowledge universe (Scharnhorst and Smiraglia 2021). Additionally, even in one domain the organisation of knowledge once achieved does not remain static (e.g., Tennis, 2018).

Formulated in terms of stability and volatility one could say that research itself naturally leads to changed knowledge. Volatility in this sense is what research is about. The organisation of new knowledge content needs to be a bit more stable to enable communication and knowledge transfer across all involved actors in a domain. But, with ever newly emerging knowledge also this domain organisation will need to change. If it comes to knowledge exchange across domains the reach/scale of interaction is bigger, and so stability is even more important, just to allow all parts of the information system to align with each other. One could also say the larger the (information) systems are, in which knowledge is produced and exchanged, the slower adaptation needs to take place to prevent a disconnect of parts of the system. However, at the end, even standardisation is relative, time-dependent and operates on different time scales.

It is these different timescales of change we refer to if we talk about flexible metadata schemes for repositories. In this paper, we look more closely into the negotiation between scientific communities and repositories, using one specific case: the DANS Long-term preservation archive EASY[1] and the CMDI metadata framework of the computational linguistics community (Goosen et al., 2015). But we developed our approach in a way that is also applicable for other metadata schemes: hence the choice of the name *Common Framework*.

A central aspect in our design is the exploration of moving from one application (in our case Fedora as the repository software behind EASY) to a modular approach of a coherent set of microservices working together, making the system more flexible towards the future. On a general level this is in line with thinking of infrastructures as consisting of networked 'microservices' (Wang, Y et al., 2021). We chose Dataverse as our main application to execute several workflows due to its active open development community[2], our own in-depth experiences with developing Dataverse microservices in various projects, the DANS experience as host of a Dataverse platform service for Dutch Higher Education repositories, and the fact that the CLARIN community has an instance of Dataverse in Norway with the colleagues of which we have already collaborated (Conzett et al., 2020). Moreover, Dataverse has already responded to the need of flexible metadata schemes by offering both a standard, common core set of metadata called Citation Block[3] and the possibility to extend this core set with custom fields defined as a discipline specific metadata block[4].

Our use case unfolds around a concrete pipeline - called the ETL pipeline (Extract-Transform-Load). We start with a CLARIN and Oral History collection, indexed on the dataset level with a specific metadata standard, and with more specific indexing information which can be found in a specific CMDI metadata file as part of the datasets in this collection. We call this phase *Extract*. We extract and analyse both schemas and values, with the aim to prepare alignments. These alignments are executed in the next phase (*Transform*). At the end, we discuss how enriched information can be feedback to the source repository as well as made available for other service providers (*Load*).

While we departed from DANS-EASY and the 'CMDI use case'[5], during our exploration it became obvious that various modules we developed as proof of concepts can also be applied to other settings: other metadata schemes, other problems of alignments and so one. This gradually led to the emergence of the *Common Framework*. The recommendation of standardised elements for Dataverse instances in CLARIN then becomes a special example of this *Common Framework*. At the core of the *Common Framework* is a design which envisions an interaction between different microservices, possibly also hosted by various service providers. Mechanisms of semantic mapping are the cornerstones of the framework.

In the next section, we unfold the steps which led to the *Common Framework*, the implementations and challenges we had to respond to.

2 Building a Common Framework

Figure 2 shows the major components upon which the *Common Framework* is built, along the pipeline we introduced above. In the course of the work, we identified two major linking tasks:

- Finding an appropriate ontology for the specific metadata fields (**red block**),
- The prediction and linkage of the appropriate concepts for their values from the list of available controlled vocabularies (**green block**).

The ultimate goal of the whole workflow lies within metadata enrichment, with adding Uniform Resource Identifiers (URIs) for both concepts and their values. For our case this means that we work on increasing the FAIR score of the datasets with associated CMDI metadata.

[1] https://easy.dans.knaw.nl/
[2] https://dataverse.org/
[3] https://guides.dataverse.org/en/latest/user/appendix.html
[4] https://guides.dataverse.org/en/4.20/admin/metadatacustomization.html
[5] https://github.com/CLARIAH/CLARIAH-plus/blob/main/use-cases/cases/DANS-cmdi.md

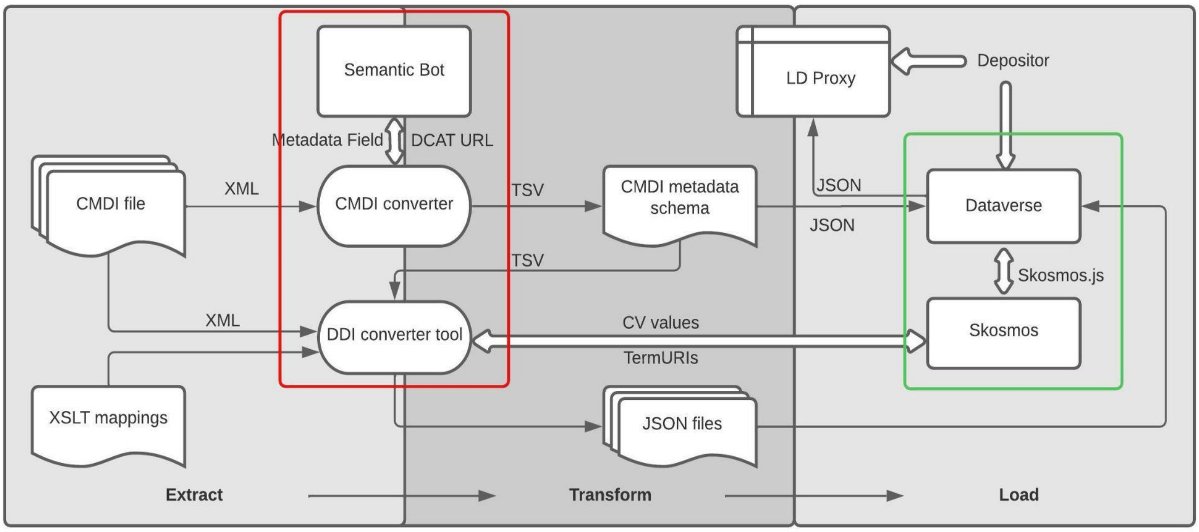


Figure 2. Schematic overview of the workflow (ETL pipeline)

The entire workflow consists of multiple microservices, indicated by blocks in the schematic description, which are separate components that we are building, reusing, or extending. Furthermore, this workflow can be extended with new open-source components (microservices) that already exist and are maintained by other institutions. This setup gives us the flexible opportunity to connect third-party services provided elsewhere. These components are represented as separate blocks in the workflow shown in Figure 2.

Component	Component provider	URL
CMDI converter	DANS-KNAW	https://github.com/DANS-labs/CLARIAH_CMDI
CMDI files	DANS-EASY archive	www.easy.dans.knaw.nl
Dataverse	DANS-KNAW	Internal DANS / Humanities Cluster[6] (HuC) development environment
DDI converter tool	DANS-KNAW	https://github.com/IQSS/dataverse-ddi-converter-tool
LD proxy	DANS-KNAW/ KNAW HuC	Resolver: https://github.com/DANS-labs/ld-server-less-resolver LD Proxy: https://github.com/KNAW-HUC/LDProxy
Semantic Bot	Open Data Soft	https://github.com/opendatasoft/semantic-bot
Skosmos	National Library of Finland	https://skosmos.org/
XSLT mappings	KNAW HuC	https://github.com/menzowindhouwer/dataverse2cmdi/blob/main/profile/prof2tsv.xsl

Table 1. Overview of components presented in the ETL pipeline

Examples from the workflow above and presented in Table 1 are: Skosmos as a service (for our proof of concept we have used the Skosmos instance of the National Library of Finland), a shared internal Dataverse instance, a local Semantic Bot installation and a DDI Converter tool. The arrows in the workflow (Figure 2) represent the interaction between the services. By reusing what is already available and eventually connecting it to our own DANS service infrastructure we hope to drive innovation. The integration of microservice-type components in existing workflows contributes, in principle, to their sustainability (sustainability by re-use and integration). In turn, the addition of microservices to existing

[6] https://huc.knaw.nl/

workflows contributes to the innovation of those workflows without the need to rebuild the entire workflow. Having said this, interoperability and maintenance are of course the underlying principles for this design to succeed. Ideally, the objective is to benefit from all the developments happening somewhere and matching these with our current demands and requirements. In the following we demonstrate this for the CMDI use case.

2.1 Describing CMDI use case

As part of the CLARIAH+[7] project DANS is working on the CMDI use case. The main goal of this use case is to identify all linguistic and oral history datasets containing a CMDI metadata file archived in the DANS-EASY repository and make these datasets visible and harvestable by CLARIN harvesters. What we call a dataset refers to one archival information package (AIP)[8] in the DANS-EASY archive which comes with a metadata file and can contain several folders or folder structures and/or individual files. The current DANS-EASY archive uses a subset of the Dublin Core Standards and DCMI Terms[9] as a metadata set to describe a dataset. The difficulty here is that the CMDI based metadata description of the dataset is attached to the dataset as a separate file, in either text or XML format, and thus not included in the EASY search index and also not harvested by CLARIN. The CMDI metadata fields are currently not visible for search interfaces, harvesters and metadata aggregators. The separate CMDI file forms part of the dissemination information package (DIP)[10] for harvesting or download by a researcher, thus it is possible for a user to view the CMDI metadata and potentially use it, however it is not manifest and prominent.

To make the CMDI metadata usable by metadata aggregators and for the discovery of datasets the first task was the identification of all AIPs which have been deposited with CMDI metadata in DANS-EASY. There are different ways to find them however, with the web interface, a general search through all metadata fields revealed (2020) 1096 datasets which use CMDI metadata, and it should be noted that its use is in either the Description metadata field or in the Form metadata field of the Dublin Core Standard. However, as noted above, while the use of CMDI is identified, the web interface cannot be used to automatically search in those CMDI notations as they are in a separate file and thus not indexed for search.

Once identified, the second task was to extract all CMDI specific metadata fields from the CMDI files and perform an analysis on the distribution of fields used and filled. Here, one needs to understand that a) CMDI is a standard, but does not have an obligatory set of core fields and b) the values of CMDI fields could be defined in a schema as free-text, closed vocabularies, specific data types or even regular expressions. We also see in the analysis of the CMDI metadata how this standard performs in the wild, and as expected we see quite some variation (Smiraglia et al., 2013; Odijk, 2016). An alternative to the tool we used could have been the SMC-Browser[11] (Durco et al., 2014) of which we were not aware at this time

The analysis was twofold. We first performed a frequency analysis of the CMDI metadata fields in our sample of 1097 records. As this sample was insufficient to draw conclusions on what might be a core set of metadata, we cooperated with CLARIN to do a second analysis executed on the Virtual Language Observatory[12] (VLO) that describes over a million datasets. CLARIN used our tool and performed a frequency and hierarchy analysis to see the distribution of metadata fields used in the CMDI descriptions of the VLO. We must point out here that the frequency analysis did not take into account the CMDI feature that allows different schemas to use different namespaces for the same element. Table 2 shows the result of the frequency analysis. The outcome of both analyses is the basis for further analyses and to identify core elements of CMDI.

[7] https://www.clariah.nl/

[8] As described in the Reference Model for an Open Archival Information System (OAIS): https://public.ccsds.org/Pubs/650x0m2.pdf

[9] EASY metadata schema: https://easy.dans.knaw..nl/schemas/md/emd/emd.xsd

[10] As described in the Reference Model for an Open Archival Information System (OAIS): https://public.ccsds.org/Pubs/650x0m2.pdf

[11] https://clarin.oeaw.ac.at/smc-browser/index.html

[12] https://vlo.clarin.eu/ - Analysis executed 8th of April, 2020

cmdp:Description; 36774144	cmdp:Code; 18114816
cmdp:AnnotationType; 22291456	cmdp:MimeType; 16102720
cmdp:SizeUnit; 21238016	cmdp:LanguageName; 10574912
cmdp:Number; 21238016	cmdp:TotalSize; 10572928
cmdp:iso-639-3-code; 20817152	cmdp:Name; 10415552

Table 2. Most frequent used metadata fields based on VLO CMDI metadata

The variety in the use of a metadata standard, visible in our use case, has been long debated in the CLARIN community (Windhouwer et al., 2012). CMDI is a metadata framework primarily used for describing digital language resources. There is no obligatory subset of fields required for each CLARIN resource. Of course, this also poses a problem for the CLARIN-wide implementation of a federated search on the metadata in the Virtual Language Observatory, and the community itself is working on this intensively. Moreover, as for any indexing practice, metadata is not always complete or harmonised as the CLARIN community maintains limited mappings for VLO facets, and it is quite time consuming.

Despite these problems, the evaluation of the CMDI metadata from DANS-EASY and the Virtual Language Observatory led to a proposal for a core set of elements[13]. It remains a difficult process to eventually implement such a core set, to which all CLARIN data providers would need to agree to, and the proposal is still under discussion. However, such a draft proposed core set[14] was enough to start a Proof of Concept for the CMDI transformation.

We used the Dataverse platform to implement what became a CMDI metadata block. Firstly, a CMDI converter tool[15] was created and used to extract the CMDI fields from the XML files. Then to create a proposed CMDI metadata block in Dataverse these extracted fields were transformed to a CMDI metadata schema by using Tabs-Separated Values (TSV)[16] file. The Dataverse DDI converter tool[17] was used to convert the TSV files into a JSON file, which could be imported to Dataverse by using the Dataverse API to create the specific CMDI metadata block in Dataverse. With the CMDI metadata block in a Dataverse instance the following step is to load all the extracted CMDI metadata values to the corresponding metadata fields. As the original metadata files are present, a simple field value mapping using the JSON format did the job.

However, as we detail later, there is also a challenge of enriching this metadata by updating each field through linking the value to a corresponding term from a recommended controlled vocabulary. This last step is also part of the *Common Framework*, and thus could be applied to any metadata extraction and transformation. In our CMDI case, this last step (Load) helps us to make CMDI metadata more visible and findable.

2.2 From use case to general framework

The basis for the *Common Framework* has evolved from the desire to achieve a universal Federated Search across multiple data repositories. The current lack of crosswalks and mappings across different metadata schemes and the lack of enriched indexes with values from controlled vocabularies presents a challenge.

The exploration of agile solutions and proof of concepts, in principle of value for different communities, helped us in defining and understanding the problem domain and gave us a clear future perspective: the creation of FAIR metadata and related semantic services. Starting with a conceptual approach for semantic interoperability on the infrastructure level was the basis for finding a common, generic solution suitable for any metadata related use case. We had to make some critical changes in the Dataverse repository core software to implement this conceptual shift, which departs from the traditional

[13] Working document by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., https://docs.google.com/document/d/1sTgp_rdwE40tMqKqhuUQ2m-FJ74NURNQ1c2IAsC295E/edit

[14] Working document of mapping by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., https://docs.google.com/spreadsheets/d/1zKR5ErqL3wRX4tOL371l0-34jXVP0gNzgU2vFsLrbcI/edit

[15] https://github.com/DANS-labs/CLARIAH_CMDI

[16] https://guides.dataverse.org/en/latest/admin/metadatacustomization.html

[17] https://github.com/IQSS/dataverse-ddi-converter-tool

understanding of (meta)data management and leads to semantically driven services. In the following we describe the steps of these implementations and their conceptual importance.

The first conceptual step consisted of the leveraging of the Semantic Metadata API[18] being built for the Dataverse platform by the Global Dataverse Community Consortium[19] (GDCC) as a part of Dataverse's core. DANS has played a critical role in the testing and improvement of this new functionality that was introduced in version 5.6 of Dataverse.[20] The format, which follows the OAI-ORE[21] export recommendations, allows for a standardised transfer of metadata from, and to, external systems without knowledge of the Dataverse specifics, such as metadata block and field storage architecture. More importantly, the Semantic Metadata API allows for the update of metadata fields published in Dataverse, both on the level of the dataset, as well as on the level of the data in the dataset and therefore, could be widely used for metadata enhancement.

The second conceptual step was the extension of the Dataverse API to fully support external controlled vocabularies (Tykhonov, 2021). This functionality was originally developed by DANS for work on a Skosmos framework[22] in the Social Science and Humanities Open Cloud[23] (SSHOC) project (Tykhonov et al., 2021) and extended by the GDCC to allow a more generic integration of Wikidata[24], ORCID[25], MeSH[26] and other controlled vocabularies (Tykhonov et al., 2021). A Skosmos implementation helps to get appropriate Simple Knowledge Organisation System (SKOS)[27] representations of the relevant controlled vocabularies and serves as a lookup service for metadata values and terms, returning the concept URI[28] that could be added to the metadata to enrich the dataset metadata. At the same time this process of metadata standardisation is extremely important for interoperability as the content of the selected concept is being cached in JSON[29] and indexed by Dataverse, and therefore, is available in the search interface. The utilisation of concept URIs facilitates users to find the dataset whilst querying in languages other than the original of the deposited dataset. For example, datasets with metadata described in Chinese, Russian or Arabic could be found with English search queries as soon as some of their terms are linked to external multilingual controlled vocabularies. It is important to understand that Dataverse is an open-source data repository and has a global community consortium, consequently all of its community members can potentially obtain and utilise this new semantic-based functionality after they upgrade their running data repository instance to version 5.7 or higher. When we worked on the consensus proposal, we also involved all community members in the process and accepted various comments, contributions, and feedback to make the collaborative solution as generic as possible. It was implemented in a fashion that it could be reused by any data repository system dealing with semantic artefacts (Hugo et al., 2020).

As a result of this work, users can export metadata from Dataverse repositories in the JSON, JSON-LD and other common formats and thus provide a more consistent way for the utilisation of this metadata by developers (and others) to create data-centric applications. Rich metadata descriptions will contain concept URIs with their cached records consisting of JSON export, this information could be indexed by various aggregators and used as a basis for the building of semantic search facilities, basically providing universal federated search across multiple data repositories.

2.3 Flexible Semantic Mapping Framework (SEMAF)

As indicated above, CMDI is a standard for which data providers usually define their own profiles, specifically tailored to their collections, consisting of various components of the CMDI framework. As a consequence, those self-defined metadata schemas in CMDI create complexity (Durco et al., 2018).

[18] https://guides.dataverse.org/en/latest/developers/dataset-semantic-metadata-api.html

[19] https://dataversecommunity.global/

[20] https://json-ld.org/

[21] https://www.openarchives.org/ore/

[22] https://github.com/SSHOC/Skosmos

[23] https://sshopencloud.eu/

[24] https://www.wikidata.org/

[25] https://orcid.org/

[26] https://www.ncbi.nlm.nih.gov/mesh/

[27] https://www.w3.org/2004/02/skos/

[28] https://www.wikidata.org/wiki/Wikidata:Data_access

[29] https://www.json.org/json-en.html

In turn, this has led to increased efforts on the maintenance of mappings for all CMDI fields[30]. Most of the CMDI mappings are done by the CLARIN community using XSLT transformations (Haaf et al., 2014). To improve the situation, Broeder et al. (2021) proposed the flexible Semantic Mapping Framework (SEMAF) to create, document and publish semantic mappings and cross-walks, linking different semantic artefacts within a particular scientific community and across scientific domains (Myers et al., 2021).

The aim is to keep all mappings in semantic form without taking into account the initial structure and hierarchy of the metadata records, and to reuse those mappings for CMDI, XML, CSV or any kind of input format.

We have developed a proof of concept using the Data Catalog Vocabulary[31] (DCAT) mappings for CMDI metadata fields to produce a standardised metadata schema where every CMDI metadata field is mapped to a DCAT URL. As soon as this link is established, another process is used to extract all the values from appropriate fields and to update metadata records with the corresponding URIs, helped by the Skosmos look-up service (see Figure 2, phase Load). SKOS is applied here to model thesauri-like resources with simple `skos:broader`, `skos:narrower` and `skos:related` properties.

This approach allows us to load all elements, such as properties and attributes, from CMDI records and build a knowledge graph from them. This solution is suitable for any metadata enhancement task where dataset metadata is being enriched with concept URIs linked from various controlled vocabularies such as Skosmos, Wikidata and others, linking to the appropriate nodes in the knowledge graph. In turn, this knowledge graph could be serialised in formats suitable for the integration with different systems. For example, JSON-LD serialisation works well for Dataverse, TURTLE and RDF-XML serialisations for Apache Jena Fuseki triple stores[32] (Tykhonov et al., 2021).

2.4 Using Machine Learning for metadata enrichment

While testing the workflow (schematically depicted in Figure 2) we discovered that the quality of the linking approach, based on the Deterministic (Exact) Matching Method (Shlomo, 2019), and currently available in Skosmos and Wikidata services, was rather poor. In some cases, this lookup process returns a lot of irrelevant candidates as it does not take into account the ambiguity based on the applied context, with the associated possibility of selecting an inappropriate concept URI and creating a false data linkage and an incorrect assertion.

We began to experiment with Machine Learning (ML) in order to add context to improve this workflow and the first results are very promising. We ran a Doccano annotation tool[33], which facilitated collaborative labelling of concepts, over the text of CMDI records and received results with recognised concepts and entities, delivered through our SpaCy[34] based Machine Learning pipeline (Figure 3). All annotations are shared between all users as a part of a collaborative effort as, in principle, the labelling of concepts could be improved. Users can also create new labels, highlight them in the text and enrich annotation with comments. After the annotation is complete the ML model is retrained.

Such a collaborative approach enables an increase in the quality of the concepts detected and provides more accurate information about types or classes of concepts, for example, for persons (PERSON), organisations (ORG), dates (DATE), Geo-Political Entity (GPE). It may guide the linking process to create appropriate links between concepts and controlled vocabularies, however, human oversight is required to review ambiguities and changes, for example, in names of persons and places. Future work will incorporate experimenting with the association of concepts, for example, with PERSON to lookup in the ORCID registry (if living), GPE in the Geonames service[35], ORG in the Global Research Identifier Database[36] (GRID), etc. One should emphasise that this experiment did not check for the ambiguity of, for instance, multiple individuals or places having the same name.

[30] https://www.clarin.eu/content/component-metadata

[31] https://www.w3.org/TR/vocab-dcat-2/

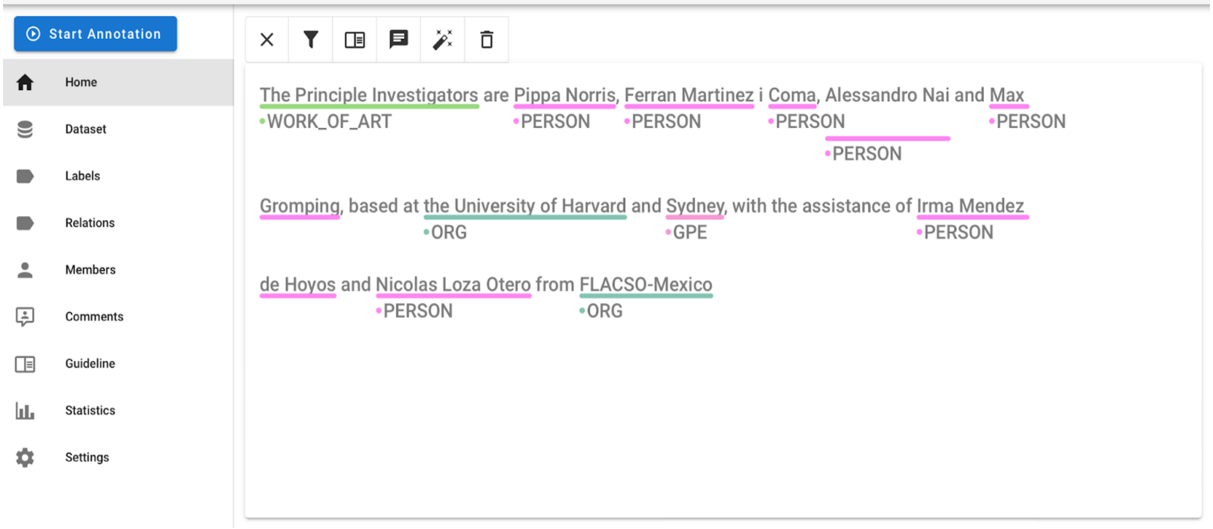
[32] https://jena.apache.org/documentation/fuseki2/

[33] https://github.com/doccano/doccano

[34] https://github.com/DANS-labs/spacy-DANS

[35] http://www.geonames.org/

[36] https://www.grid.ac/



The Principle Investigators are Pippa Norris, Ferran Martinez i Coma, Alessandro Nai and Max Gromping, based at the University of Harvard and Sydney, with the assistance of Irma Mendez de Hoyos and Nicolas Loza Otero from FLACSO-Mexico

Figure 3. Automatic annotation in the Doccano text annotation tool

The utilisation of ML tools appears promising for automated metadata enrichment, but it is quite resource consuming if we wish to achieve the highest possible quality. It is manifest that only collaboration, additionally on an international level, will turn these explorations into standardised methods.

2.5 Beyond the CMDI use case

The CLARIAH+ project and the CMDI use case gave us the opportunity to explore and comprehend the first two parts of the Common Framework pipeline and helped us to demonstrate and test the complete pipeline: metadata extraction, metadata transformation and loading (archiving) of the enhanced metadata, in an instance of Dataverse.

Throughout this work we have taken the opportunity to collaboration with various other research groups and organisations:

- Extraction and evaluation of CMDI metadata fields, based on all CMDI metadata archived in DANS-EASY, in collaboration with the ODISSEI[37] project
- Definition of a core set of CMDI metadata fields in the cooperation with the CLARIN community[38]
- Creation of a workflow for the prediction and linking of concepts from external controlled vocabularies to the CMDI metadata values (metadata enrichment), in collaboration with the CESSDA[39] community
- Extension of the *Common Framework* with the support for controlled vocabularies to create metadata that is available in a FAIR way, joining forces with Netwerk Digitaal Erfgoed[40] (NDE) team
- Extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format, together with KNAW Humanities Cluster (HuC)

The envisioned use of the *Common Framework* workflow as shown in explorations 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, about a portfolio of registered microservices which form a generic and extensible pipeline. DANS is currently migrating its research data archiving service from a Fedora-based platform (DANS-

[37] https://odissei-data.nl/en/

[38] Working document of mapping by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., https://docs.google.com/spreadsheets/d/1zKR5ErqL3wRX4tOL371l0-34jXVP0gNzgU2vFsLrbcI/edit

[39] https://www.cessda.eu/

[40] https://netwerkdigitaalerfgoed.nl/

EASY) to other platforms including introducing Dataverse to function as the DANS Data Stations as a specific repository service for designated communities (Wals, 2021). The designated communities will be served with larger metadata aggregations that include references to data not curated and/or hosted by DANS. The exploration described in this paper bases its analytic part on the current production system while, at the same time, informs the on-going migration process.

3 Future work

A substantial amount of work has been completed, but we are not finished yet. From the CMDI use case we discovered that CMDI as a standard is lacking a defined core set of CMDI metadata (Goosen et al., 2014). We remain in close cooperation with the CLARIN CMDI taskforce working on a proposal for, and acceptance of, a core set of CMDI metadata as a recommendation for all CLARIN centres.

The CMDI use case gave us the opportunity to prove the *Common Framework* approach. The following steps are to extend this Framework and to implement it for other cases. Beyond the extension of the Citation Core set of Dataverse, it is envisioned to support a link between other 'indexing' metadata fields to the other Knowledge Organisation Systems providers. In particular, we think here of recommended FAIR controlled vocabularies and ontologies which potentially may become part of the set of metadata fields (Wilkinson et al., 2016; Broeder et al., 2021; Wang, M. et al., 2021). Coming back to the CMDI case, this could lead to linkages of recognized, or any, CMDI metadata values to a recommended ontology or controlled vocabulary with the aim to produce '5-star Linked Open Data'[41].

To contribute further to the FAIRification of controlled vocabularies and other KOS or Semantic artefacts we wish to experiment further with the creation of a semi-automatic workflow, using a Skosmos API, to query Skosmos representations of recommended controlled vocabularies. Therefore, explorations of the NDE's Network of Terms[42] GraphQL[43] endpoint will be continued to create links between appropriate controlled vocabularies for the terms extracted from the CMDI fields. These metadata fields will link to the CMDI component registry in the CMDI metadata schema.

Within the ODISSEI project DANS is going to work further on the creation of a production implementation of the microservices infrastructure. In the recently granted project FAIRCORE4EOSC[44] it is likely that DANS will be migrating the registries and brokering of microservices for schematic and semantic transformation/enhancement to the EOSC[45]. DANS will continue to host some of the (micro)services, which ones are still a topic of debate. All future work of DANS will be shepherding transformations, enhancements, crosswalks, etc to a microservice/registry architecture.

All of our insights and workflows will be shared with the CLARIN and CLARIAH communities and we are looking for more collaborations on semantic mappings that could be used to get an appropriate ontology linkage not only on value level but also between fields available in CMDI Component Registry.

The *Common Framework* can help to support the enrichment of metadata, may aid the making of CLARIN datasets findable and accessible, and ultimately also supports Reusability. FAIR compliance automatic assessment tools, such as F-UJI[46], can be included in the *Common Framework* to evaluate the FAIRness of the metadata (Devaraju et al., 2020, 2021).

4 Conclusion

Our experimental work of building a *Common Framework* to expose CMDI metadata via a DANS discovery service relates to the migration of the DANS archive service to (a) newly to build DANS Data Station(s), which will serve as a basis for the discovery and OAI-PMH[47] harvesting services for the CLARIN researcher community and beyond.

[41] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data
[42] https://github.com/netwerk-digitaal-erfgoed
[43] https://graphql.org/
[44] https://dans.knaw.nl/en/news/consortium-led-by-dans-acquires-a-major-european-grant-to-make-eosc-more-fair/
[45] https://ec.europa.eu/
[46] https://www.f-uji.net/
[47] http://www.openarchives.org/pmh/

This paper describes the first two steps in our ETL-pipeline: 1) the extraction of the CMDI metadata from CMDI metadata files archived in DANS-EASY, and 2) the transformation of this metadata information. The third step of the pipeline, the loading of the transformed metadata into the new Data Station, is performed as a proof-of-concept in a separate Dataverse instance, one could say an envisioned Dataverse-based Data Station. The work is ongoing and the challenges we reflect upon, when addressed, are unavoidably 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 the event that the core set of CMDI metadata will be changed in the future. Its implementation in production services is still a challenge ahead.

To arrive at the proposed core set of CMDI metadata, we have analysed all linguistic and oral history datasets containing CMDI metadata stored in the DANS-EASY archive with the CMDI exploration tool. With the same tool we were able to transform each CMDI metadata file to the proposed core set.

To increase the FAIRness of the new metadata, 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 use case. A semi-automatic workflow, which uses a Skosmos API, was developed to query any Skosmos representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint was used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields.

To extend the semi-automatic workflow we started to explore the possibilities of a semantic gateway. We started a proof-of-concept with a semantic gateway lookup API. This API is able to return a list of standardised 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 complete the circle, we are currently 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 centres. Secondly, the original relationships between fields and concepts should be kept whereby the custom components should be added to a SKOS schema. If this is possible, then we should be able to reproduce the original CMDI metadata, which could be offered for download to any user without losing the authority and provenance of the original metadata.

The basis of our work lies in reusing and exploring new techniques, (micro)services and the basic ideas behind the *Common Framework* are not only to solve long standing problems, but also to build flexible solutions for different communities. This is the main reason for the setup of a microservice oriented pipeline. Being part of different communities has helped us to create a broad support base amongst these communities. In the meantime, multiple communities, organisations and projects are testing and exploring our experimental work and connecting it to their own infrastructures, providing us with feedback to improve the microservices leading to sustainable infrastructure.

This work has taught us that looking to the future and setting ourselves some big challenges not only leads to innovative ideas and solutions, but it also leads to further new challenges. These challenges are motivating us to build sustainable solutions with and for the communities by exploring new technologies. These new technologies furthermore allow us to circumvent existing technology-lock-ins and they also demonstrate how, via microservices and a distributed approach, new methods of aligning and enriching metadata can be created.

Implementing these solutions in a sustainable infrastructure for long-term preservation archives is yet another challenge which we did not discuss in this paper. An important aspect when it comes to the implementation is the ownership of (meta)data and recommended controlled vocabularies, provenance and authorisation. We have demonstrated how technologies such as machine-learning approaches can be used to clean, enrich and harmonise metadata. We have also indicated communities must be involved in these technological developments, and how to implement to meet their needs. However, it remains to be seen and investigated how these technologies will be used in daily work by data producers and consumers (Borgman et al., 2019) and how they change the work of data managers and archivists. Moreover, more work is needed to investigate what are possible consequences for the certification process, and in general, which monitoring and governance policies are required for an envisioned network of distributed service providers that is required to remain stable over time.

Acknowledgement

This work has been funded by the CLARIAH Plus project, https://www.clariah.nl/, more particular in Work Packages 2, 3 and partly 4.

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*. Zenodo. http://doi.org/10.5281/zenodo.4651421

Borgman, C., Scharnhorst, A., Golshan, M. S. 2019. Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70(8):888-904. DOI: 10.1002/asi.24172; preprint version: https://arxiv.org/abs/1802.02689

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.* 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., White, A. 2020. *FAIRsFAIR Data Object Assessment Metrics.* Working paper. Zenodo. 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., Diepenbroek, M. 2021. From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. *Data Science Journal*, 20(1):4. https://doi.org/10.5334/dsj-2021-004

Durco, M., Lorenzini, M., Sugimoto, G., 2018. Something will be connected - Semantic mapping from CMDI to Parthenos Entities. *Selected papers from the CLARIN Annual Conference 2017, Linköping Electronic Conference Proceedings,* 147(3): 25-35

Durco, M., & Windhouwer, M. 2014. The CMD Cloud. In *Proceedings of LREC 2014: Ninth International Conference on Language Resources and Evaluation.* http://www.lrec-conf.org/proceedings/lrec2014/pdf/156_Paper.pdf

European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., Ojstersek, M., Choirat, C., van de Sanden, M., Coppens, F. 2021. *EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture.* Publications Office. https://data.europa.eu/doi/10.2777/620649

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Durco, M., Schonefeld, O. 2015. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. *Selected Papers from the CLARIN Conference 2014, Linköping Electronic Conference Proceedings,* 116(004):36-53. http://www.ep.liu.se/ecp/116/004/ecp15116004.pdf

Guéret, C., Chambers, T., Reijnhoudt, L., van der Most, F., Scharnhorst, A. 2013. *Genericity versus expressivity - an exercise in semantic interoperable research information systems for Web Science [Digital Libraries].* Working paper. http://arxiv.org/abs/1304.5743

Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., van Uytvanck, D. 2014. *CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres.* Working paper. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLARINs_virtual_language_observatory_under_scrutiny_2014.pdf

Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., Bonino, L. 2020, *D2.5 FAIR Semantics Recommendations Second Iteration.* Working paper. Zenodo. https://doi.org/10.5281/zenodo.4314321

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

McIlwaine, I. C. 2010. Universal Decimal Classification (UDC), in M. J. Bates, M. N. Maack (eds.), *Encyclopedia of Library and Information Sciences, Third Edition (3rd ed.).* CRC Press.1(1):5432-5439. https://doi.org/10.1081/E-ELIS3-120043532

Myers, J., Tykhonov, V. 2021. *Proposal on the ontologies and external controlled vocabularies support in Dataverse.* Working paper. Zenodo. https://doi.org/10.5281/zenodo.5845540

Odijk, J. E. J. M. 2016. Linguistic research using CLARIN. *Lingua*, 178:1-4

Shlomo, N. 2019. Overview of Data Linkage Methods for Policy Design and Evaluation, in Crato, N., Paruolo, P. (Eds.) *Data-Driven Policy Impact Evaluation*. Springer, Cham. https://doi.org/10.1007/978-3-319-78461-8_4

Scharnhorst, A., Smiraglia, R. P. 2021. The Need for Knowledge Organization. Introduction to *Linking Knowledge: Linked Open Data for Knowledge Organization (Chapter 1)*. In R. P. Smiraglia; A. Scharnhorst (Eds.), *Linking Knowledge*. Ergon –Nomos, Baden-Baden, pp. 1-23. https://doi.org/10.5771/9783956506611-1

Smiraglia, R. P., Scharnhorst, A., Akdag Salah, A.,Gao, C. 2013. UDC in Action, in A. Slavic, A. Akdag Salah, & S. Davies (Eds.), *Classification and visualization: interfaces to knowledge,* pp. 259–270. Ergon Verlag, Würzburg. Preprint available at http://arxiv.org/abs/1306.3783

Smiraglia, R.P., Scharnhorst, A. 2021. *Linking Knowledge. Linked Open Data for Knowledge Organization and Visualization.* Ergon-Nomos, Baden-Baden. https://doi.org/10.5771/9783956506611

Svenonius, E. 2000. *The Intellectual Foundation of Information Organization*. The MIT Press, Cambridge, USA.

Tennis, J. T. 2018. Intellectual history, history of ideas, and subject ontogeny. In *Challenges and Opportunities for Knowledge Organization in the Digital Age*. *Proceedings of the Fifteenth International ISKO Conference.* pp. 308 – 313**.** Ergon, Baden-Baden. https://doi.org/10.5771/9783956504211-308

Tykhonov, V. 2021. *Controlled vocabularies and ontologies in Dataverse data repository*. Presentation at the *Dataverse Community Meeting 2021*. https://doi.org/10.5281/zenodo.5838161

Tykhonov, V., de Vries, J., Scharnhorst, A., Admiraal, F., Indarto, E., Priddy, M. 2021. *Flexible metadata schemes for research data repositories.* Presentation at the *CLARIN Annual conference 2021*. Zenodo. https://doi.org/10.5281/zenodo.5838156

Tykhonov, V., Scharnhorst., A. 2021. *Flexibility in Metadata Schemes and Standardisation: the Case of CMDI and DANS Research Data Repositories*. Presentation in the series *ISKO Knowledge Organisation Research Observatory,* November 24, 2021. Zenodo. https://doi.org/10.5281/zenodo.5838109

Tykhonov, V. 2021. *CLARIN CMDI use case and flexible metadata schemes*. Presentation at the *CLARIAH interest group Linked Open Data*, 4 November 2021. https://doi.org/10.5281/zenodo.5838132

Wals, H. 2021. *Focus on FAIR: DANS 2021-2025.* The Hague. Viewed 26 April 2022. https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/DANS-2021-2025/UK_DANS20212025.pdf

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

Wang, Y., Kadiyala, H., Rubin, J. 2021. Promises and challenges of microservices: an exploratory study. *Empirical Software Engineering* 26:63. https://doi.org/10.1007/s10664-020-09910-y

Wilkinson, M. D., Dumontier, M., Aalbersberg, I., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Nature, 3,* 160018. https://doi.org/10.1038/sdata.2016.18

Windhouwer, M., Broeder, D., & van Uytvanck, D. 2012. A CMD core model for CLARIN web services, in *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, pp. 41-48. http://hdl.handle.net/11858/00-001M-0000-000F-A418-0