Philipp Conzett[1], Twan Goosen[2], Andrea Scharnhorst[3], Vyacheslav Tykhonov[4], Dieter Van Uytvanck[5], Jerry de Vries[6], Marion Wittenberg[7]

1 philipp.conzett@uit.no, UiT The Arctic University of Norway; ORCID: https://orcid.org/0000-0002-6754-7911
2 twan@clarin.eu, CLARIN ERIC
3, andrea.scharnhorst@dans.knaw.nl, DANS, Royal Netherlands Academy of Arts and Sciences
4 yacheslav.tykhonov@dans.knaw.nl, DANS, Royal Netherlands Academy of Arts and Sciences
5 dieter@clarin.eu, CLARIN ERIC
6 jerry.de.vries@dans.knaw.nl, DANS, Royal Netherlands Academy of Arts and Sciences
7 marion.wittenberg@dans.knaw.nl, DANS, Royal Netherlands Academy of Arts and Sciences

How to weave domain specific information sources into a large, FAIR data fabric for the Digital Humanities? The use of the Dataverse platform.

Good data curation and data management is a precondition for any replication of research. Research data and research tools are often intricately coupled. But increasingly, digital methodology in the humanities depends on the combination and re-use of data sources outside of their primary area of collection and curation. As those data sources need to be accessible (on-line) in a distributed manner, a holistic approach to curate them becomes more and more important. Scaling up of data in one domain - e.g. by linking different repositories together - is already a challenge. Cross-domain data operations in the field of digital humanities are confronted with the heterogeneity of humanities disciplines concerning their epistemological frameworks (what counts as a research question?) as well as the data or source materials they work with (text, image, objects, other media).

In this paper, we discuss how a data repository platform (Dataverse) which by default comes with a generic set of metadata can be adapted for the needs of a specific research community (CLARIN). While our concrete case concerns Dataverse and CMDI as the CLARIN metadata scheme, the problems we face are of a general nature. Many communities come with their own specific metadata schemes - for good reasons. But, we also want them to be found across domains. Archives, repositories, libraries are responsible to ensure the FAIRness of data (Findable-Accessible-Interoperable-Reusable [8]), of which Findability is one aspect. How to preserve the epistemological richness and specificity of data in the humanities, and still enable FAIR principles? As always the answer to this question has a technological and a social-organisational side, both of which we discuss further.

Nowadays the task of making data FAIR is shared among archives and research communities. But, they don't operate in isolation. Both are embedded in larger, complex structures of knowledge organisation and management, part of which can be called research infrastructure. Research infrastructures operate inside of domains and across domains. Currently, they aspire to enable data access and data curation in a way that digital humanities research can combine data from various sources, and of various kinds. Standardisation and certification are important means to ensure that data are interoperable and accessible in a reliable way. Digital humanities are continuously advancing their research driven ICT infrastructure, aspiring to create sophisticated, domain-specific services next to ensuring interoperability within and across domains. [1,2] Still, they also rely on partnership with generic building blocks of an infrastructure, such as libraries and (cross-domain) archives and repositories.

Part of the tasks executed in research infrastructures concerns processes of data management in general and data curation more specifically. One challenge for data operations lies in keeping a balance between principles of data fluidity on the one side (easy, seamless access, collective processes of analysis and enrichment) and reliable mechanisms for long-term preservation of selected parts of data on the other side. One could also see data fluidity as an express of 'user empowerment'. The researcher (research group, community or discipline) is the master of the cognitive process, and change, both of the content of data as well as their documentation, is an intrinsic element of research. Research at the same time needs to adhere to norms, values and standards agreed by a community and needs to challenge and violate them in the quest for new ideas. Regulations and standards are expressions for norms and values agreed upon at a certain point in time. They also form the basis to execute principles of FAIRness in a transparent, reproducible and durable manner.

This paper addresses both issues, the user empowerment and the long-term preservation, by looking into the use of a specific application, Dataverse, in the CLARIN infrastructure. CLARIN is a European Research Infrastructure for Language Resources and Technology, and positions itself as a resource domain rather than a research domain. It makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences. [3,4] Dataverse "is an open-source application for publishing, referencing, extracting and analyzing research data". [5] There are currently more than 60 Dataverse-based data repositories hosted by research organisations world-wide. One example is Dataverse.nl, which provides data repositories for 14 Dutch institutions. Another example is TROLLing, a domain-specific repository for linguistic data. Coordinated by Harvard, the Dataverse application is continuously further developed. In Europe, those developments take place for instance in European funded projects, such as SSHOC [6].

In the paper, we report on-going explorations to use Dataverse as intermediary service between research activities on the one side and long-term data preservation on the other side. More particularly, we discuss how Dataverse can be integrated into the CLARIN infrastructure. There are a couple of interesting features in Dataverse which make it a good candidate for a bridge between long-term archiving and collaborative data uses. Dataverse

draws on open source, community driven software development, which enables local explorations to become later integrated into centrally maintained versions (updates). It comes with a volatile way to extend metadata schemes beyond the obligatory fields, the so-called ‘citation block’. It is API-accessible and allows incorporating multilingual interfaces. Preview of data as well as in-file search are other interesting features. Dataverse can be used for long-term preservation itself, or coupled to a long-term archive (for instance, enabling the users to select data for long-term archiving by pushing a button).

The roundtable with experts from both CLARIN, Dataverse and a long-term archive, will discuss how the Dataverse Network platform can be used to expose language resources (following requirements of the CLARIN community) in an interoperable way, and at the same time enable long-term preservation for selected parts of those resources. CLARIN has deployed the Component MetaData Infrastructure (CMDI) as the standard metadata framework for the CLARIN community. While the framework enables a flexible use, core elements are crucial to allow for cross-repository discovery services. [7]
In the long-term archive DANS-EASY, the CLARIN collection encompasses 1002 datasets which all contain CMDI metadata, but are only indexed (in a searchable manner) with Dublin Core metadata. To make this collection interoperable with the CLARIN infrastructure is the goal of an on-going task. In the envisioned workflow, datasets from this collection will be migrated to a Dataverse instance, which is CLARIN compatible.

Bringing together developers, researchers, and service and infrastructure managers will enable us to present and discuss technical steps/problems of the above described workflow such as operations to map, transform, and enrich the metadata schemes, to automatically export information in files (of different formats) from the long-term archive into the Dataverse metadata structure, and automatic export into long-term archives for selected data. But, we also wish to address aspects of the social, communicative and legal processes which are necessarily part of the final implementation: who is responsible for the curation of the (meta)data, and the licenses? Who selects data sets for long-term archiving and so on.

We reflect at hand these very concrete questions, as well as the more generic aspects we introduced at the beginning of this paper. Namely, how to enable a productive division of labour between different stakeholders around tools and services provided for Digital Humanities research?

References

[1] J. Edmond (2015) Collaboration and Infrastructure. In: S. Schreibman, R. Siemens, J. Unworth (eds) New Companion to Digital Humanities.
[2] N. Harrower, M. Maryl, T. Biro, B. Immenhauser, ALLEA Working Group E-Humanities (2020) Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities, Digital Repository of Ireland. https://doi.org/10.7486/DRI.tq582c863
[3] https://www.clarin.eu/content/clarin-in-a-nutshell
[4] J. Odijk, A. van Hessen (eds) (2017) CLARIN in the Low Countries. ubiquity press London.

[5] M. Crosas (2011) The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine; Volume 17. Nr. 1-2, doi:10.1045/january2011-crosas
[6] https://www.sshopencloud.eu/
[7] D. Broeder , M. Windhouwer, D. Van Uytvanck, T. Goosen, T. Trippel CMDI: A Component Metadata Infrastructure. In the Proceedings of the Metadata 2012 Workshop on Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources. At LREC 2012, Istanbul, Turkey, (May 22, 2012).
[8] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. Scientific Data 3 (15 March 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

submission as round table
Theme 2: Replication, evaluation and quantitative analysis in the DH era