Originally published in: Skadiņa, Inguna/lEskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2018. 8-10 October 2018, Pisa, Italy. – Utrecht: CLARIN, 2018. Pp. 99-103.

CLARIN Data Management Activities in the PARTHENOS Context

Marnix van Berchum
Huygens ING, KNAW
Amsterdam, Netherlands
marnix.van.berchum@huygens.knaw.nl

Thorsten Trippel
University of Tübingen
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Abstract

Data Management is one of the core activities of all CLARIN centres providing data and services for the academia. In PARTHENOS, European initiatives and projects in the area of the humanities and social sciences assembled to compare policies and procedures. One of the areas of interest is data management. The data management landscape shows a lot of proliferation, for which an abstraction level is introduced to help centres, such as CLARIN centres, in the process of providing the best possible services to users with data management needs.

1 Introduction

Data management is the activity of creating, providing, maintaining and archiving research data over all stages of the research data life cycle (see Pennock, 2007). CLARIN centres working with data, operating repositories, and providing services to their users all work in the area of data management.

Each certified CLARIN-B centre provides some services (see Wittenburg et al 2013-2018, paragraphs 1a and 1d) and deals with the processes and technology required for data management. CLARIN does not operate independently. To avoid duplication of work, CLARIN makes use of open interfaces and compatibility layers to other systems. All certified CLARIN centres provide a level of trust documented by certification authorities (such as the Core Trust Seal, www.coretrustseal.org) and others to make the requirements and processes of data management transparent. Almas et al. (2016) describe the current diverse situation of data management and lay out possible ways for further development, including the establishment of policies. Independent of the documentation of current practices, all parties involved recognize the need to follow procedures and guidelines early on in the process, starting with a data management plan (DMP) ideally before data are created. The creation of DMPs is becoming more and more a requirement by research funders as well. Consequently CLARIN centres can provide assistance to their users in the process of DMP creation. To do this efficiently, they need to have an understanding on the requirements of funders for DMPs.

To synchronize and harmonize activities, CLARIN ERIC is part or the European PARTHENOS project (www.parthenos-project.eu). In this paper, we describe the current situation with regards to data management in CLARIN and PARTHENOS, the data policy implementation as required by funding organization and according to best scientific practice. We outline an abstraction level that is under discussion between partners in PARTHENOS and that can be applied by CLARIN centres.

2 Current situation with regards to data management in PARTHENOS and CLARIN

Amongst the PARTHENOS partners no common data management practices exist. The disciplines represented in the project – defined as 'the broad sector of Linguistic Studies, Humanities, Cultural Heritage, History, and Archaeology' – all have a different history with regard to data management. Some have archiving experience for objects and artefacts but less with digital born data (e.g. archeology), others are completely new to the field of data management.

Within PARTHENOS, CLARIN represents the disciplines dealing with all forms of language and text data. Although, as a well established research infrastructure, CLARIN has a long background in the handling (use, sustainably maintain, provide) of data, there is no overarching, fixed policy concerning the management of these data. This is contrary to the CLARIN Centre Assessment procedure which requires a finished, or rather 'at least initiated', Core Trust Seal application, implying several technical and organisational requirements for the CLARIN Centres (see CTS, 2017 in the references). These include requirements on backup solutions, persistent identification, minimal metadata, licence recommendation, and access restriction implementation. Individual CLARIN centres operating repositories may have additional requirements regarding data formats or have specific depositing agreements in place (see CCA, undated, in the references). Some CLARIN Centres even provide Data management plans as a service to scholars (see Trippel & Zinn (2015) or published handbooks and documentation (see for example Herold & Lemnitzer (2012), chapter 2) as reference for users.

Despite the organisation around a core technical infrastructure of tools and resources that imply ideas of data management, CLARIN has not yet recommended centralised data policies on its partners. From this background, CLARIN joined the activities by PARTHENOS, which has as one of the goals the design of Data Management Plan templates, which serve larger communities. For this purpose, PARTHENOS reviewed existing, discipline independent DMP templates provided by research funders.

3 Data Policy Implementation

Current policy of most funders and research organizations is to rely on FAIR data (see FAIR, undated, in the references), i.e. data needs to be Findable, Accessible, Interoperable and Re-usable. CLARIN is dedicated to the FAIR principles documented by some infrastructure components: the VLO and FCS are good examples of making data in the CLARIN world findable, for details on CLARIN's efforts and mission with regards to FAIR, see de Jong, et al. (2018). With resolvable PIDs the data becomes accessible – even if access is restricted CLARIN provides access mechanisms with the Identity Provider and Shibboleth architecture. Interoperability is achieved by the utilization of standards such as the ISO TC 37 SC 4 endorsed standards or TEI. However, interoperability is dependent on the selection of the flavour of these standards and the support by software. The Language resource Switchboard (see Zinn, 2017) is a tool to provide opportunities to interoperability of data. Interoperability and Reusability are not only addressed by technology, but also by policies, legal restrictions, infrastructure components etc. Re-usability has the additional requirement that researchers are also allowed to reuse data and have access to the tools to do this. These policies and decisions need to be documented for data to become interoperable and reusable.

Reconstructing of FAIR relevant documentation with finished data sets is virtually impossible – especially when third parties were involved and hold rights for data and software. To avoid that, these aspects are documented before data is assembled or in the processes of data creation. As all of this is part of the data management plan, the data management plan is the key to successfully implement a FAIR data policy.

The situation with regards to the central nature of DMPs for a FAIR data policy is currently only partially reflected by the support for data management plans. In fact, there is obvious proliferation in communities and funding institutions with regards to their requirements for DMPs. The requirements range from no formal requirements besides providing DMPs (e.g. German Research Foundation) to detailed templates (e.g. Horizon 2020, see the Guidelines on FAIR Data Management in Horizon 2020 (2016) in the references). For the United Kingdom, Jones (2012) already provided a summary of eight (national) funding organizations and their requirements. Some academic institutions have their own requirement of good scholarly practice, for example at the University of Edinburgh (see Research Data Service in the references), providing additional requirements. To ease the situation the Data Curation Centre (DCC) in the UK provides a website with an interactive template for various funders, called DMPonline (see "DMPonline" (2010-2018) in the references). Though this is open source software which is extensible and many data management plans show significant overlap, DMPonline cannot cover all funders, disciplines and data centre related requirements. Despite this – conceptual –

shortcoming, the DMPs created with such a template have the benefit of documenting awareness of scholars in the data management process, which is a requirement for long-term digital preservation.

Within PARTHENOS a detailed report was created, documenting the proliferation, complexity and issues with regards to data management (see Hollander et al., 2017, Chapter 3). In the same document (ibid. Appendix III, Section 9.2) a unified template was created, which failed the usability test because of its complexity as the variety of descriptive levels including funders, data centres and disciplinary requirements are not sufficiently distinguished and the scholar is at loss with the template.

4 Abstraction for Data Management Plans: Data Management Protocols

For the creation and reuse of elements of the DMP templates, an abstraction layer could be helpful. Such an abstraction layer for DMPs takes common components of DMPs into account that depend on the funders' requirements and the requirements of data centres. This abstraction layer is termed DMP protocol, which could also be seen as a template for DMP templates used by funders and data centres alike.

The DMP protocol conceptually is a template for creating DMP templates and could be used by funders and data centres alike. The DMP protocol takes the DMP requirements of (1) a funder (2) a discipline (3) a data type (4) a data centre and (5) a DMP protocol to generate (6) a DMP template to be filled in for a concrete research project within a discipline and with a specific type of data, utilizing a specific data centre and tailored to the project reviewing needs of a funder. Though at present this DMP protocol is a conceptual model, it opens up the way to an implementation that helps to ease the DMP process by omitting unnecessary questions to a researcher and directing the attention to essential issues that are indeed project specific.

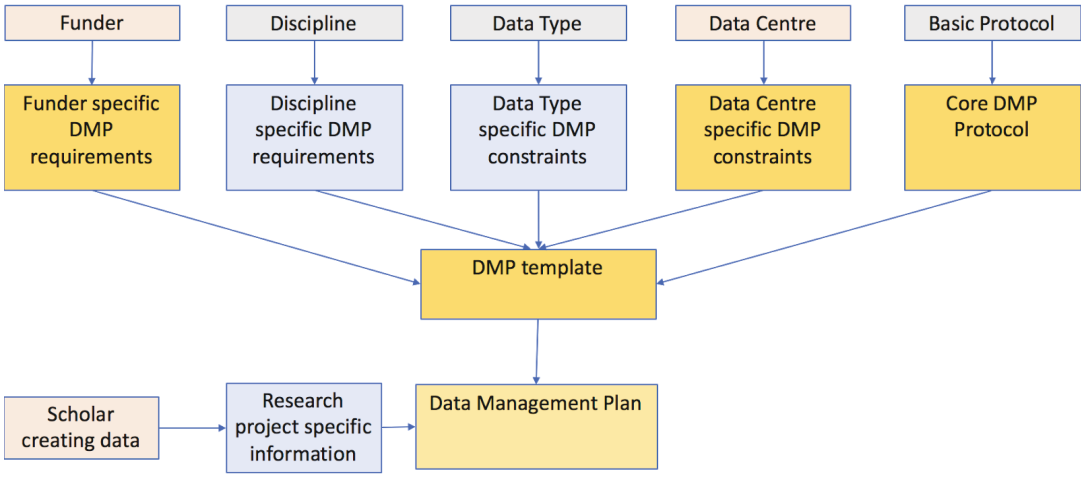

Figure 1 Schematic structure of a data management protocol for creating DMP templates and DMPs

The funder specific requirements, data centre specific constraints and core protocol in this model are defined by policies of institutions, while the discipline specific requirements and data types are based on practices and have to be extrapolated by experts on the field and in data modelling, hence they are intrinsic to the discipline and data type. Similarly, the research project information directly relates to the project, which is defined by the scholar, hence this information is intrinsic to the project. It seems obvious that there is a difference in the level of formalization possible on the input side. Formalising the policies of data centres and funders together with a core protocol are institutional decisions and can be achieved, if the institutions decide to do so. The discipline and data type specifics are harder to formalize as they rely on implicit best practices in a discipline, interpretation and formalization.

5 Future Work: CLARIN implementation of the DMP Protocol using CMDI

The abstraction model for data management plans shows the five different components going into a template plus the project specific information. To implement these, it would be possible to define components or partial documents for each and integrate it by means of for example standard XML technologies (XML include). As the documents are not necessarily prefilled, another option would be to utilize ISO/DIS 24622-2 based CMDI to define (CMDI-)components for each, such as a funder specific component, a discipline specific component, a data type specific component, etc. The result would be a CMDI-Profile dependent on all the components, which results in an XSchema for a DMP for each of these. A sample profile would be the developmental DMP-Protocol profile (CLARIN component registry ID p_1527668176067, see https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1527668176067/xsd). Using standard CLARIN technology such as the COMEDI CMDI editor (Lyse et al., 2015), a DMP could then be written as a CMDI document, to be transformed into a DMP in a layouted format, for example based on HTML.

This implementation might look awkward as it is reusing a technology that is used for data descriptions elsewhere. However, the details specified in the DMP should also be included in the metadata description of the research data created within a project, hence the CMD-modelled information of a DMP will be reused and updated in the metadata for the research data. Other parts of the DMP are not part of the metadata created in the project, such as details on DMP budgets, repository assessment information, etc. The CMDI implementation of DMPs looks promising and should be further explored. As the competence and technology is available within CLARIN, the CLARIN researchers within PARTHENOS will work on that.

6 Conclusion

There is a gap between disciplinary requirements and wishes concerning data management, funder requirements, and ideas by interdisciplinary initiatives such as PARTHENOS. The latter project would be the right platform to compare and assess the different data management practices across the humanities, but it is still complicated to reach common, generic data management policies or services. As a disciplinary infrastructure CLARIN should define a default policy for DMP for its centres and require that in the Centre Assessment Process. This policy could feed into the work of PARTHENOS.

The abstraction of the DMP protocol (paragraph 4) should be taken into account when CLARIN defines the default policy. The CLARIN template should cover on a general level the disciplinary and data type specific aspects of the DMP. Each centre can fill in their data centre specific needs. Funder requirements would have to be left out at present, as there is too much of a variety there. However, a template for CLARIN could build on top of the HORIZON 2020 template as a starting point.

Reference

Almas, B.; Bicarregui, J.; Blatecky, A.; Hill, S.; Lannom, L.; Pennington, R.; Stotzka, R.; Treloar, A.; Wilkinson, R.; Wittenburg, P.; Yunqiang, Z. (2016): Data Management Trends, Principles and Components - What Needs to be Done Next? Research Data Alliance (RDA). Available at http://hdl.handle.net/11304/7721971a-23f3-4eed-8df8-739ff0f2bc6e.

CCA (undated). Assessment procedure. Available at https://www.clarin.eu/content/assessment-procedure.

CTS (2017). Core Trustworthy Data Repositories Extended Guidance: Core Trustworthy Data Repositories Requirements for 2017–2019, Extended Guidance. Version 1.0: October 2017. Available at https://www.coretrustseal.org/wp-content/uploads/2017/01/20171026-CTS-Extended-Guidance-v1.0.pdf.

Peter Doorn (ed., 2018) 'Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management': D/2018/13.324/1, available at: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

DMPonline. (2010-2018). Digital Curation Centre (DCC). Available at https://dmponline.dcc.ac.uk/.

FAIR (undated): THE FAIR DATA PRINCIPLES. FORCE11. Available at https://www.force11.org/group/fairgroup/fairprinciples.

Guidelines on FAIR Data Management in Horizon 2020 (2016). Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Herold, A.; Lemnitzer, L. (2012): CLARIN-D User Guide. BBAW, Berlin. Available at http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf.

Hollander, H. et al. (2017) Report on Guidelines for Common Policies Implementation. Project Deliverable D 3.1. Available at http://www.parthenos-project.eu/Download/Deliverables/D3.1_Guidelines_for_Common_Policies_Implementation.pdf.

ISO/DIS 24622-2 (2018) Language resource management -- Component metadata infrastructure (CMDI) -- Part 2: The component metadata specification language. Draft International Standard.

De Jong, F.; Maegaard, B.; de Smedt, K. Fišer; van Uytvanck, D. (2018): CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In N. Calzolari, et al. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan. Available at http://www.lrec-conf.org/proceedings/lrec2018/pdf/575.pdf .

Jones, S. (2012): Summary of UK research funders' expectations for the content of data management and sharing plans. Digital Curation Centre (DCC). Available at http://www.dcc.ac.uk/sites/default/files/documents/resource/policy/FundersDataPlanReqs_v4%204.pdf.

Lyse, G. I.; Meurer, P.; De Smedt, K. (2015): COMEDI: A component metadata editor. In: Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014 (8), Soesterberg, The Netherlands, 82-98. Available at https://www.clarin.eu/sites/default/files/cac2014_submission_13_0.pdf.

Pennock, M. (2007): Digital Curation: A life-cycle approach to managing and preserving usable digital information". In: Library & Archives Journal, (1). Available at http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf.

Research Data Service. Website of the University of Edinburgh's research data services for local data management services., The University of Edinburgh. Available at https://www.ed.ac.uk/information-services/research-support/research-data-service.

Trippel, T.; Zinn, C. (2015): DMPTY - A Wizard For Generating Data Management Plans. In: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland, (123), 71-78. Available at http://www.ep.liu.se/ecp/123/006/ecp15123006.pdf.

Wittenburg, P. et al. (2013-2018) CLARIN B Centre Checklist, Version 6, Last Update: 2018-02-07, Status Approved by the Centre Committee, https://office.clarin.eu/v/CE-2013-0095-B-centre-checklist-v6.pdf.

Zinn, C. (2017) The CLARIN Language Resource Switchboard. CLARIN Annual Conference 2016. https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf.