Originally published in: De Smedt, Koenraad (Ed.): Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland.
– Linköping : Linköping University Electronic Press, 2015. Pp. 71-78.
(Linköping Electronic Conference Proceedings 123, NEALT Proceedings Series 28)

DMPTY – A Wizard for Generating Data Management Plans

Thorsten Trippel and **Claus Zinn**
Seminar für Sprachwissenschaft
Universität Tübingen, Germany
thorsten.trippel@uni-tuebingen.de
claus.zinn@uni-tuebingen.de

Abstract

To optimize the sharing and reuse of existing data, many funding organizations now require researchers to specify a management plan for research data. In such a plan, researchers are supposed to describe the entire life cycle of the research data they are going to produce, from data creation to formatting, interpretation, documentation, short-term storage, long-term archiving and data re-use. To support researchers with this task, we built DMPTY, a wizard that guides researchers through the essential aspects of managing data, elicits information from them, and finally, generates a document that can be further edited and linked to the original research proposal.

1 Introduction

All research depends on data. To address a research question, scientists may need to collect, interpret and analyse data. Often the first phase of scientific activity, data collection, is the most decisive, and also a time-consuming and human-resource-intensive task. It must be planned well enough so that a significant number of data points are available for subsequent inspection so that underlying research questions can be analysed thoroughly. When the analysis of data is yielding results of significant interest, the study is described in scientific parlance, and then submitted to a scientific conference or journal. Once reviewed and accepted for publication, the resulting article constitutes the formal act of sharing research results with the scientific community, and most articles in reputable publication outlets are archived for posteriority. While the results are now public, the underlying research data often remains private, and usually stays with the individual researcher or the research organization. This makes it hard for other researchers to find and to get access to the data, and limits the opportunity for them to reproduce the results, or to base secondary studies on the same data. In brief, the sharing and long-term archiving of research data has these four main benefits:

Reproducibility: One of the main principles of the scientific method is reproducibility: it shall be possible to replicate experimental results, in preference by redoing the analysis on the existing data rather than on newly collected data. This discourages fraud and tempering with research data.

Facilitation of secondary studies: With researchers having access to existing data sets, there is no need for a costly collection of new data, and therefore, it becomes easier for researchers to explore similar research questions, contrastive studies, or meta studies.

Attribution: It should be good scientific practise to give an explicit acknowledgement of ownership or authorship to the one who has collected the data. Scientific reputation shall not only be merited by findings, but also by having acquired underlying data.

Economy: Funding money and researchers' time shall not be wasted for collecting data sets if comparable data already exist. Open access to existing data also allows researchers to add to existing data sets, and hence might contribute towards a "Wikipedia effect", yielding increasingly rich resources.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-108605

To reap these benefits, research data should be accessible in public repositories, properly documented, with generous access rights, and possibly, in an easy-to-read, non-proprietary data format.

Funding agencies increasingly require grant applicants to complement their research plan with a plan for managing and sharing the research data that is going to be created during their research. In the United Kingdom, the Digital Curation Centre maintains a list of the major British funding bodies and their data policies [1]. In Germany, the national research foundation (DFG) expects data management plans, at least for larger collaborative research projects, and often funds personnel and necessary hard- and software to ensure that data is sustainably archived. The situation is similar in Switzerland where the Swiss National Science Foundation requires researchers that "the data collected with the aid of an SNSF grant must be made available also to other researchers for secondary research and integrated in recognised scientific data pools" [2]. Some universities see data management issues as an integral part to good scientific practise. In the *Guidelines for Research Integrity of the ETH Zurich*, Article 11 is about the collection, documentation and storage of primary data, primarily, with the purpose that all "research results derived from the primary data can be reproduced completely" [3]. The *Research Data Management Policy* of the University of Edinburgh states that "All new research proposals [...] must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication" [4].

Successful research proposals are centred around one or more research questions, the applicants apply sound research methodologies, and in empirically-driven research, this usually includes a convincing approach to gather and analyse research data to address the research questions. Individual research projects, by their very nature, take a short-term view: at the end of the project's lifetime, research questions have been addressed in the best possible way, and results properly written up and published. Usually, there is no long-term view, in particular, with respect to the research data gathered. What is the research data life cycle, how will data be stored (*e.g.*, using which format) and adequately documented? How can the data be cited and made accessible for the scientific community for the long term (archival)? With regard to accessibility, how is personal or confidential data be taken care of? Which licence should be chosen for the data to ensure that other researchers can access the data in the future? Does the data collection ensure that research results can be reproduced, or that follow-up studies can use the data with ease?

We must not expect researchers to handle such questions, which are secondary to the research question, entirely on their own. Taking the long-term perspective requires a different set of skills, and it calls for a cooperation between researchers and a permanent research infrastructure. It is advisable that such cooperation is initiated at the early stage of a research project so that all aspects of the data life cycle are properly taken care of. Infrastructures such as CLARIN can assist researchers in managing their data.

The remainder of this paper is structured as follows. In Sect. 2 we describe the common elements of data management plans. Sect. 3 sets data management in the CLARIN context and defines the division of labour and shared responsibilities between data producer and data archive. In Sect. 4, we present the DMPTY wizard for data management planning. In Sect. 5, we report on a preliminary evaluation of DMPTY, and in Sect. 6, we discuss related work and conclude.

2 Common elements of data management plans

In Britain, the Digital Curation Centre (DCC) "provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data" (see `http://www.dcc.ac.uk/about-us`). The DCC, for instance, has a good summary page that overviews and links to data management plan requirements of a number of British funding agencies [5]. The DCC has also published a checklist for devising a data plan [6]. The checklist seems to be an amalgamation of the various plans, and with its broad base takes into account requirements from different scientific fields such as the Natural Sciences or the Humanities. The checklist is divided into eight different parts, and covers all of the essential aspects for managing research data:

1. **Administrative Data:** nature of research project, research questions, purpose of data collection, existing data policies of funder or research institution;

2. **Data Collection:** type, format and volume of data; impact on data sharing and long-term access; existing data for re-use; standards and methodologies, quality assurance; data versioning;
3. **Documentation and Metadata:** information needed for the data to be read and interpreted in the future; details on documenting data acquisition; use of metadata standards;
4. **Ethics and Legal Compliance:** consent for data preservation and sharing; protection of personal data; handling of sensitive data; data ownership; data license;
5. **Storage and Backup:** redundancy of storage and backup; responsibilities; use of third party facilities; access control; safe data transfer;
6. **Selection and Preservation:** criteria for data selection; time and effort for data preparation; foreseeable research uses for the data, preservation timeframe; repository location and costs;
7. **Data Sharing:** identification of potential users; timeframe for making data accessible; use of persistent identifiers, data sharing via repositories and other mechanisms;
8. **Responsibilities and Resources:** for DMP implementation, review, and revision at plan and item level, potentially shared across research partners; use of external expertise; costs.

The Directorate-General for Research & Innovation of the European Commission has also published data management guidelines for Horizon 2020 projects [7]. Following the guidelines, the DMP must describe how research data is handled during *and* after the project; how the data will be collected, processed or generated; and what methodology and standards will be followed. It must say whether and how data will be shared (preferably open access), and how data will be curated and preserved. While the use of a DMP is "required for projects participating in the Open Research Data Pilot", other projects are "invited to submit a DMP if it is relevant to their planned research" [7].

An interesting aspect of the EC guidelines is the emphasis on the dynamics of research data management: "the DMP is not a fixed document, but evolves during the lifespan of the project" [7, page 5]. As a consequence, projects that submit a DMP do so multiple times: they must provide a first version of the DMP within the first six months of the projects' start, and later on, periodically update their DMPs during the projects' lifetime. Moreover, in Horizon 2020, all costs related to research data management can be accounted for.

3 Data Management in the CLARIN Infrastructure

The CLARIN shared distributed infrastructure aims at making language resources, technology and expertise available to the Humanities and Social Sciences research communities. To streamline the inclusion of new data and tools, and to help researchers with managing their data, CLARIN-D now offers advice on data management plans and supports their execution. The CLARIN-D plan template mirrors the structure of the DCC checklist, but has a number of adaptations to best profit from the CLARIN infrastructure. As a first step, researchers are invited to select the CLARIN-D centre whose expertise matches best the type of resource being created during the project. This aims at ensuring that researchers get the best possible advice from a CLARIN-D centre of their choice.[1] Following the plan template, researchers are asked to contact their CLARIN centre of choice when starting to devise their research data plan.

With regards to the DCC plan, our plan adjusts to the CLARIN infrastructure as follows:

Data Collection: a policy on preferred non-proprietary data formats for all types of language-related resources (in line with the CLARIN-D User Guide, see [8]).

Documentation and Metadata: the selection of existing CMDI-based metadata schemes, or if necessary, the adaptation of existing ones to best describe the research data.

Ethics and Legal Compliance: encouraging the use of the CLARIN License Category Calculator[2].

[1] For identifying the most suitable CLARIN-D centre, researchers can consult the link `http://www.clarin-d.net/de/aufbereiten/clarin-zentrum-finden`.

[2] Available online at `https://www.clarin.eu/content/clarin-license-category-calculator`

Note that, while CLARIN encourages open access to research data, it is not imposed by the DMP.

Responsibilities and Resources: a budget estimate that accounts for all the personnel and financial resources, and which are shared between data producer and CLARIN-D archive.

Moreover, there is a ninth plan component that describes a time schedule that data producer and data archivist agree on: when is the research data (i) described with all metadata, (ii) ingested in a data repository, and (iii) made accessible to interested parties or the general public? Moreover, it defines how long the research data should be held, and when, if applicable, it should be deleted. The data management plan shall also be complemented with a *precontractual agreement* between the data producer and the CLARIN centre for archiving, and which captures the rights and obligations of each partner.[3]

4 The DMPTY Wizard for the Generation of CLARIN-supported DMPs

DMPTY is a browser-based wizard available at the German CLARIN-D portal and encodes the CLARIN-D plan template.[4] The wizard makes use of the Javascript framework *AngularJS*, see `angularjs.org`, where each of the nine plan steps is presented to the user as an HTML form, and where navigation between the nine forms is easily possible so that information can be provided in flexible order (see Fig. 1). Associated with the forms is an HTML document that represents the contents of the data management plan template. Whenever the user enters information into one of the web form elements, the underlying plan is instantiated appropriately. At any given time (but within a browser session), it is possible to generate the plan as a text document in one of the formats Word, rtf and LaTeX, using the *pandoc* tool, see `pandoc.org`.

Researchers can then edit the document to enter additional information to address, for instance, institution-specific ethical or legal questions, or to state a cooperation with third parties, or to respond to specific requirements of funding agencies that we have not anticipated. Researchers may also want to add cross-links to relevant parts of the corresponding research proposal, change the formatting *etc*.

At the time of writing, a beta version of DMPTY is publicly available; the wizard generates plans in German only, and it only lists CLARIN-D centres as cooperation partners. Upon a successful evaluation, DMPTY will also be capable of generating plans in English, and listing all CLARIN centres as archiving partner. So far, the scope of DMPTY is restricted to its application within the CLARIN world.

We are currently preparing an evaluation of DMPTY and seek interested researchers to participate in our evaluation study. We also intend do make DMPTY available on an HTTPS-enabled server to protect the privacy and integrity of all data exchanged between the web browser and the server.

5 First Evaluation Results and Discussion

DMPTY has been demonstrated in depth to interested researchers at a tutorial on data management plans at the November 2015 meeting of the German Research Data Alliance in Potsdam. During the 90 minutes session the following questions were brought forward from the audience (answers by the second author):

Question: Is the data provided by DMPTY users transmitted to the server, and how long is it kept there, and is the data safe?

Answer: For the generation of the plan document in Word, LaTeX, or RTF format, user data is transmitted to the server, and temporarily stored on the server's `tmp` directory. For the time being, an unsecured HTTP connection is used.

[3] The precontractual agreement precedes the *deposition agreement* that is signed just before research data is deposited at the CLARIN centre. The content of the two agreements may substantially differ. The precontractual agreement is seen similar to a letter of intent; and intentions may shift given the outcome of the research grant application, the reviewers' comments on the proposal, and the many unforeseen events that may happen during a research project. As a result, assumptions about the research data (*e.g.*, formats, methodologies, access policies) that were held at the time of the precontractual agreement may need to be revised at the depositing stage, and hence require a rewording in the deposition agreement.

[4] See `http://www.clarin-d.net/de/aufbereiten/datenmanagementplan-entwickeln`.

Figure 1: Screenshot of the DMPTY entry page. In the upper part, all nine parts of the plan template can be accessed; the lower part shows elements of the selected first part of the plan template.

The confidentiality issue is a serious one, as some researchers may fear to expose their grant application to an untrusted third party. Users asked if it were possible to develop and make available a stand-alone version of DMPTY that users can install on their local machine.

One attendee suggested to define a sophisticated Microsoft Word or LaTeX template that users could fill out using their favoured word processing software, without fearing Browser-related session time-outs or data breaches (data loss or theft).

Another attendee suggested more advanced but desktop-based software, where data management plans are devised by starting in terms of the data that is being used or created during a research project. Such software should be less text-driven; rather, a visual programming language should be defined where users create data management plans by manipulating plan elements graphically, following the visual programming paradigm.

Question: How about the completeness and adequateness of the plan template across different disciplines (in light of our previous discussion of different requirements for DMPs across European

funding agencies and research institutions)?

Answer: DMPTY's plan management template largely draws upon the DCC checklist, which we described in Sect. 2, and which seems to be an amalgamation of the various data management plan templates provided by different UK funding agencies. As such, it takes into account requirements from different scientific fields. Our adaptation of the template for the CLARIN infrastructure is seen both as advantage and drawback, depending on whether potential users are part of the CLARIN infrastructure, or not.

We have seen that British funding organisations offer detailed checklists for writing data management plans. The situation in Germany is rather less developed; the German Research Foundation and the Federal Ministry of Education and Research offer little guidance in this respect. Here, the use of DMPTY gives researchers a hands-on approach to become aware of all the different aspects of data management in a research project, but completeness and adequateness issues must be addressed between scientists, their research institutions and their funding organisations, and by taking into account the nature of the research proposals in question.

(Rhetorical) Question: What is the purpose of data management plans, other than securing funding?

Answer: While it is hard to plan a three to five-year research project, it is also hard to carry it out as planned. The same holds for data management plans. For the research aspect, many funding agencies require researchers to submit regular progress reports (in Germany, often mid-term and final progress reports). Should the reporting be extended to include the data management aspects of research projects?

Clearly, a data management plan should be more than just a "letter of intent" that is not acted upon once funding has been secured. The plan should be regularly consulted, its steps executed, and the execution monitored. In DMPTY, the plan includes steps where interactions with third parties (the CLARIN centres) are necessary at given points in time. This puts some pressure on researchers with regards to plan execution and monitoring. Time will tell whether funding agencies will ask grant holders to complement their research reports with the addressing of data management issues.

In this respect, DMPTY (and other data management plan wizards) might develop into more sophisticated planning and plan execution systems, supporting researchers in devising a plan, but also executing it, and when necessary, in adapting the plan given changing requirements. DMPTY may hence develop into some sort of project management software that helps managing the entire life cycle of a project's research data.

It must be noted that none of the twenty-plus participants of the tutorial session had any experience with writing data management plans. Some reported having experimented with a number of DMP tools (such as DMPTY), in part, to prepare for the tutorial. Overall, participants seemed to appreciate the importance of addressing data management issues in a systematic manner. Also, most participants seemed to wish receiving more guidance from their funding organsations, but this may be due to the tutorial's German audience, and thus specific to the German research sector.

6 Related Work and Conclusion

With data management plans becoming increasingly necessary to attract funding, there are now a number of tools available that help researchers to address all relevant data management issues. The DCC provides a web-based wizard (DMPOnline) to help researchers devising data management plans, see [9]. Once researchers have selected a funding organisation for their research proposal, and optionally, their home research organisation, a corresponding data management plan template is created, which the wizard then follows step by step. DMPOnline users can thus devise a DMP that adheres to both funder and institutional requirements. The DMPOnline software is written in Ruby on Rails (see `rubyonrails.org`), is open source, and is available on Github for download, see [10]. The software's design facilitates the adaption of DMPOnline to cater for institution-specific customization (*e.g.*, customized logo, guidance,

and boilerplate text); in fact, over a dozen of UK institutions have already customized DMPOnline to better fit their needs.

The second DINI/nestor workshop was devoted to data management plans [11]. The workshop's programme featured, among others, a talk from the German Research Foundation on policy issues, several presentations on the nature and benefits of data management plans, and also talks on examples of successful data management, *e.g.*, for biological data and earth system science data. During the workshop, there were also a number of tools presented: the DMP Webtool from the University of Bielefeld, see [13], and the TUB-DMP tool of the Technical University of Berlin, see [14]. Both tools offer a plan template that largely draws from the WissGrid checklist [12]; also, both tools support the Horizon 2020 checklist [7]. Both tools are in house developments that aim at connecting data management planning with the existing research infrastructures of the respective institutions. The DMP Webtool is based upon the content management system Drupal (see `drupal.org`); the TUB-DMP tool of the TU Berlin is developed in php (see `php.net`).

With a considerable number of different initiatives in the area of data management plans, researchers may be confronted with an increasing number of requirements for a "proper" management of research data. There is a danger that requirements defined by the researcher's home institution may conflict with those of the funding organisation, and that both require the use of respective DMP wizards. When research data is deposited at a third institution, say a CLARIN centre, a potential third wizard (DMPTY) enters the scene. Here, however, we anticipate that requirements for data management plans will stabilize across research institutions, funding organisations, other parties involved, and also across disciplines. Also, a single DMP tool will need to be able to process multiple templates and guidelines from different research institutions and funding organisations. The DMP Online Tool of the DCC already follows this direction, and we anticipate a market consolidation where few, feature-rich tools will survive. DMPTY's design makes it easy to adapt a given plan template, or to include secondary ones. Also, DMPTY's design anticipates a follow-up editing phase, where researchers use their favoured text processing software to adapt the resulting DMP for their needs.

Conclusion. On a grand scale, managing research data poses significant challenges for research infrastructures, see [15]. On the individual level, researchers face additional work. With data management plans becoming an accepted part of good scientific practice, researchers must take into account all questions concerning their research data at an early stage of their research projects. Clearly, specifying and executing data management plans consumes resources, but the investment will pay off. DMPTY lowers the burden for researchers to develop their own plan, it guides them through all relevant aspects of such plans, and helps streamlining the cooperation with the CLARIN infrastructure. With CLARIN involved, researchers get support for the management of their data during the data's entire life cycle; touching base at regular intervals with CLARIN guarantees that the plans are up to their needs and properly executed. It also ensures that the appropriate resources (personnel, equipment) are accounted for. As a result, the number of high-quality accessible research data is bound to increase, which makes it easier for researchers to reap the benefits of sustainable data archiving and data re-use.

Acknowledgements. In November 2015, the authors co-organised a tutorial on data management plans, which was held during the German RDA meeting in Potsdam. We would like to thank all participants for their valuable feedback on DMPTY, and their comments on research data management in general. Also, we would like to thank the anonymous referees for their comments, which helped improve this paper considerably.

References

[1] Digital Curation Centre (DCC). Funders' data plan requirements. See http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[2] Guidelines at the Swiss National Science Foundation. See http://www.snf.ch/sitecollectiondocuments/allg_reglement_e.pdf, Article 44(b).

[3] Guidelines for Research Integrity at the ETH Zurich. See https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf.

[4] Research Data Management Policy at the University of Edinburgh. See http://www.ed.ac.uk/schools-departments/information-services/research-support/data-management/data-management-planning.

[5] Funder Requirements at the Digital Curation Centre (DCC). See http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[6] Checklist for a Data Management Plan. v.4.0, Edinburgh: Digital Curation Centre, 2013. Available online: http://www.dcc.ac.uk/resources/data-management-plans.

[7] Guidelines on Data Management in Horizon 2020 (Version 2.1). European Commission, Directorate-General for Research & Innovation. See http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, February 2016.

[8] CLARIN-D User Guide, v.1.01. See http://www.clarin-d.net/de/sprachressourcen-und-dienste/benutzerhandbuch

[9] Digital Curation Centre. DMPOnline, a tool to devise data management plans. See https://dmponline.dcc.ac.uk.

[10] DMPonline_v4. Software maintained at https://github.com/DigitalCurationCentre/DMPonline_v4.

[11] Second DINI/nestor Workshop, Berlin, 2015. See http://www.forschungsdaten.org/index.php/DINI-nestor-WS2.

[12] J. Ludwig and H. Enke (Eds.) Leitfaden zum Forschungsdaten-Management. Verlag Werner Hülsbusch, Glückstadt, 2013. See http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf.

[13] The DMP Webtool, University of Bielefeld, see https://data.uni-bielefeld.de/de/data-management-plan.

[14] The TUB-DMP Tool, Technical University of Berlin, see https://www.szf.tu-berlin.de/menue/dienste_tools/datenmanagementplan_tub_dmp

[15] B. Almas *et al.*, Data Management Trends, Principles and Components - What Needs to be Done Next? V6.1. EUDAT, 2015. See http://hdl.handle.net/11304/f638f422-f619-11e4-ac7e-860aa0063d1f.

All links were accessed on March 01, 2016.