RESEARCH PAPER

"Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning

Robert Pergl[1,2], Rob Hooft[3], Marek Suchánek[1], Vojtěch Knaisl[1] and Jan Slifka[1]

[1] Czech Technical University in Prague, Faculty of Information Technology, CZ

[2] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Life Sciences, CZ

[3] Dutch Techcentre for Life Sciences, NL

Corresponding author: Robert Pergl (perglr@fit.cvut.cz)

The Data Stewardship Wizard is a tool for data management planning that is focused on getting the most value out of data management planning for the project itself rather than on fulfilling obligations. It is based on FAIR Data Stewardship, in which each data-related decision in a project acts to optimize the Findability, Accessibility, Interoperability and/or Reusability of the data. The background to this philosophy is that the first reuser of the data is the researcher themselves. The tool encourages the consulting of expertise and experts, can help researchers avoid risks they did not know they would encounter by confronting them with practical experience from others, and can help them discover helpful technologies they did not know existed.

In this paper, we discuss the context and motivation for the tool, we explain its architecture and we present key functions, such as the knowledge model evolvability and migrations, assembling data management plans, metrics and evaluation of data management plans.

Keywords: Data Stewardship; Data Management Plan; FAIR

1 Introduction

We present a tool, the "Data Stewardship Wizard", that can bring together researchers, data stewards, and data experts pursuing better research through data management planning.

1.1 Changing Data Stewardship from Burden to Benefit

Especially in the last 10 years, manipulation of digital data resources has become very important for research projects in many research fields. It now forms such an obviously important fraction of research that it has become valuable to properly plan the allocation of people to as well as the budget for data management.

Data management plans (DMPs) are demanded from researchers by science funders and by research institutes, however the main motivation for them to ask for a plan at first appears to be different. Science funders request a DMP mainly because they demand that money spent on data collection will benefit other researchers; see e.g. National Science Foundation (2010): "Proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (DMP). This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results" and European Commission (2016): "The pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects". And research institutes request DMPs in order to be able to prove proper scientific conduct and reproducibility. As a consequence, currently the activity is seen as an obligation, a burden, by researchers and Data management planning is not as effective as desired (Smale et al. 2018).

One problem with the currently common approach to data management planning is that many researchers do not know the breadth of tools and expertise available to help with data management in their projects. And they have limited experience to warn them of the host of data-related risks that research projects can be exposed to. Experience that researchers have with digital data in a private setting does not properly prepare

for this: the common photo library at home is rather small in comparison to research data in many labs, it contains only 1 or 2 trivially related data types (JPG and RAW) instead of a multitude of complex related data, and is handled by a single person instead of a collaboration. Conversely, the experts providing data management services and expertise have difficulties being found by the researchers that need their expertise. Their expertise often solves problems many researchers do not know they will encounter, and therefore they can not look for a solution in the first place.

We try to solve those problems using our tool, the "Data Stewardship Wizard". We use the term "Data Stewardship" to indicate that the activity is not only taking place *during* the project, but extends to the long term *maintenance* of the resulting research data. We use the term "Wizard" to refer to the tool as an "expert system" providing context-dependent guidance to its users.

Our Data Stewardship Wizard targets to alleviate the negative view of data management planning by focusing primarily on the benefits of data management for the research project itself and the researcher, not on the obligations; for example by pointing out suitable tools that can help assemble and maintain the provenance metadata, or relevant data standards. The Data Stewardship Wizard clearly indicates the effect of each answer on the adherence to the principles describing that data should be Findable, Accessible, Interoperable and Reusable for machines and for humans (FAIR principles) (Wilkinson et al. 2016) in all its questions, thereby explicitly guiding researchers who are searching for good ways to make their results FAIRer. The Data Stewardship Wizard's questions cover the full breadth of expertise in order to show researchers all the different aspects of data management: IT, archival and data publication, sustainability and the entire FAIR data (Wilkinson et al. 2016) spectrum. And furthermore, the guidance available with the questions points to available experts and expertise exactly where the issue at hand is brought up in the questionnaire, so that researchers are encouraged to interact as early as possible with the experts that are relevant to their new project, not only when the project encounters a problem and it is too late to budget for a proper solution.

1.2 How the Data Stewardship Wizard Changes Data Management Planning

Research projects have very different data management demands. Many data management planning tools, (e.g. DMPonline and DMPtool (Sallans and Donnelly 2012), together writing *roaDMaP* (Simms et al. 2018), which are the most well known and most developed tools around) nevertheless are constructed to ask the same questions (together constituting a 'template') of each user. The variability comes from the descriptive answers.

A study was performed in November 2017 in Australia (Smale et al. 2018), where a random sample of 834 completed DMPs from the university's DMP database for evaluation across several criteria. DMPs were assessed for detail and quality of information provided about physical and digital data storage. The results showed that "few DMPs provided specific useful information about the research data nominally being described".

In the Data Stewardship Wizard, we have taken another approach: most questions are closed questions with a limited set of possible answers. And based on the answer that is selected by the user, follow-up questions will be added to the questionnaire. Also, some answers may be obtained from linked services, such as FAIRsharing.org (Sansone et al. 2019) containing a curated database of standards, policies and databases, mainly (but not limited to) life science (**Figure 1**).

The questioning in the Data Stewardship Wizard is modelled after the conversation a researcher could have with a data management expert: the questions asked by the expert would depend on previous answers and be relevant to the project; likewise only questions relevant to the project will be in the questionnaire generated by the Data Stewardship Wizard.

This approach has several advantages over a flat questionnaire as currently used in other tools:

- Filtering the questions this way makes it possible to add additional expertise in additional questions without unnecessarily burdening projects for which this expertise is irrelevant, for example on dealing with privacy-sensitive information. A broad coverage of expertise ensures a good coverage of topics relevant to different projects; the Data Stewardship Wizard does not need to limit its questions to topics that are relevant to the majority of research projects.
- We can avoid complex questions (e.g. starting with "describe how you will" and accompanied by guidance containing a list of aspects that need to be addressed, for examples see the annotated H2020 template (European Commission 2018)) that require the data steward to write a piece of text. Instead, all of the aspects of the complex question can be addressed in a context-dependent way in separate closed questions. Being asked to write composite answers researchers often adapt DMPs from other projects rather than starting their answers from scratch.

Figure 1: Example of the questionnaire illustrating the closed-questions approach and autocomplete from *FAIRsharing.org*.

– DMP reports can be generated from the provided answers into various templates provided by specific funders and institutions, thus easing their preparation and also ensuring consistent language quality.
– The information entered into the Data Stewardship wizard is mostly structured, which means that specific information can be used algorithmically, e.g. to sum up storage requirements specified in a large number of DMPs. The plans are said to be *machine actionable*. (Simms et al. 2017) (see Section 3).

The breadth of the coverage in the Data Stewardship Wizard questionnaires also is a help to experienced project data stewards: they can use the Data Stewardship Wizard as a *checklist for their projects*, not unlike expert pilots or surgeons using checklists in order to make sure not to forget any part of their routine. Data Stewards working this way can concentrate their expertise on the specific challenges of the project at hand.

In the rest of this paper we present what our Data Stewardship Wizard offers to researchers and data stewards, and how this is technically achieved.

2 Methods and Results
The Data Stewardship Wizard consists of an open source web questionnaire tool, an expert system embodied in a so-called *Knowledge Model*, and a system to maintain knowledge models as depicted in **Figure 2**.

2.1 Web Questionnaire Tool
At the Czech Technical University in Prague, we developed a web tool to present hierarchical data management questionnaires, storing intermediate results in a database. The core of this web tool is a dynamic form engine developed by Pergl (2018).

2.2 The Knowledge Model
The expert content of the Data Stewardship Wizard comes originally from a mind map (Buzan and Buzan 2006) collected by Hooft (2019). It captures years of experience with various projects and organisations in the life science domain, obtained through accidental encounters as well as interviews. The mind map consists of nested questions that researchers planning a project can ask themselves, possible answers, guidance based on expert experience, and links to external resources. Development of the mind map still continues. In its current version, the mind map contains over 600 nodes in 5 levels of depth. It also includes cross-links connecting nodes in different parts where a pure tree structure can not capture the relationships between subjects. All in all, it contains extensive organised experience, but not in a form that is structured and accessible to data stewards and researchers. The desire to exploit the collected value in the mind map resulted in the idea to build a software tool which, when combined with the Web Questionnaire Tool, has become the Data Stewardship Wizard.
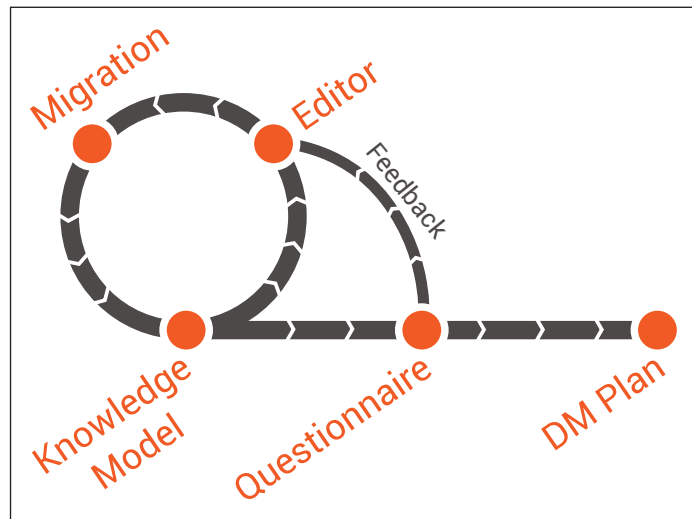
Figure 2: Workflow in the Data Stewardship Wizard (Suchánek et al. 2019): The standard *Knowledge Model* is adapted by a data steward for their own research field and/or institute using the *Editor*. The *Migration* tool is used in this process to integrate changes by other data stewards. From a customized Knowledge Model, a researcher will create a *Questionnaire* and answer the relevant questions for their project. Each question contains the option for the researcher to provide *Feedback* about the questionnaire to the data steward who created it. At any step in the answering process, the researcher can use a template in the system to create an actual *Data Management Plan.*

For the development of the Data Stewardship Wizard, the contents of the mind map have been translated into what we call the standard Knowledge Model. The standard Knowledge Model currently contains a few hundred questions. Updates to the Data Stewardship mind map are represented in the Data Stewardship Wizard as new versions of the standard Knowledge Model. The standard Knowledge Model is structured into six chapters that follow the research data life cycle (UK Data Service 2012). Not all chapters have the same depth of coverage: we are still collecting experience from users and experts, and also (e.g. through workshops) encourage others to contribute by adding their own expertise in the form of new questions.

In parallel to the development of the Data Stewardship Wizard, Prof. Barend Mons published a book (Mons 2018) based on the questions in the data stewardship mind map, elaborating experience and insight in the form of "What's up?", "Do", "Don't" sections for each question. Under permission of the publisher, these sections have been linked to the equivalent questions in the standard Knowledge Model and can be displayed in our instance of the Data Stewardship Wizard.

2.3 A System to Maintain Knowledge Models
The Data Stewardship Wizard contains a system to maintain knowledge models and to adapt them to individual institutes or infrastructures.

Using the Data Stewardship Wizard, data stewards can customize the standard knowledge model we provide and build their own knowledge model (**Figure 4**) with special questions or guidance for their own discipline or for their own institute. Customizations of a knowledge model can add questions, but also remove or modify existing questions. Customizations can be shared with other data stewards through a registry of knowledge models, so that users can build upon each other's work.

Structure of a Knowledge Model The Knowledge Model consists of a list of chapters at the top level. Chapters group together questions connected to a similar topic. Each chapter contains a tree-like structure of questions and answers.

We support several types of questions:

- *Options* – the question has a list of answers. Each answer can have follow-up questions assigned, which allows building a recursive tree of any depth.
- *A list of items* – researchers can fill-in several items. Each item can have a subtree of follow-up questions.

– *An open question* – a question where researchers fill in the answer, we support ordinary types such as string, number or date.

Most questions in the standard Knowledge Model are of the type *Options.* Such a question defines a list of possible answers. Each of these answers can be an anchor for followup questions. A similar technique is used for questions that require a list of items as answer: here the question itself is the anchor for followup questions, which will be repeated by the Data Stewardship Wizard for every item in the list.

Questions also specify guidance information

– a *Text* that explains the question context.
– an *Expert* representing contact information about a person who can help with answering that question. It is meant to be used especially within knowledge models for specific institutions.
– external *References* are resources that can help researchers with answering the question. References to pages of the book "Data Stewardship for Open Science" (Mons 2018) (**Figure 3**) are special instances of this.
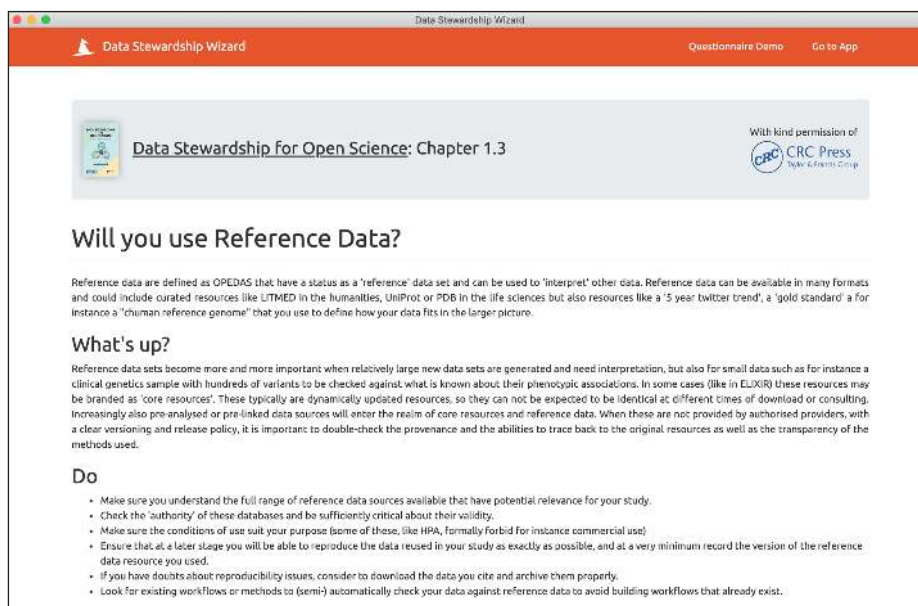
Figure 3: Example of a reference to the book "Data Stewardship for Open Science" (Mons 2018).
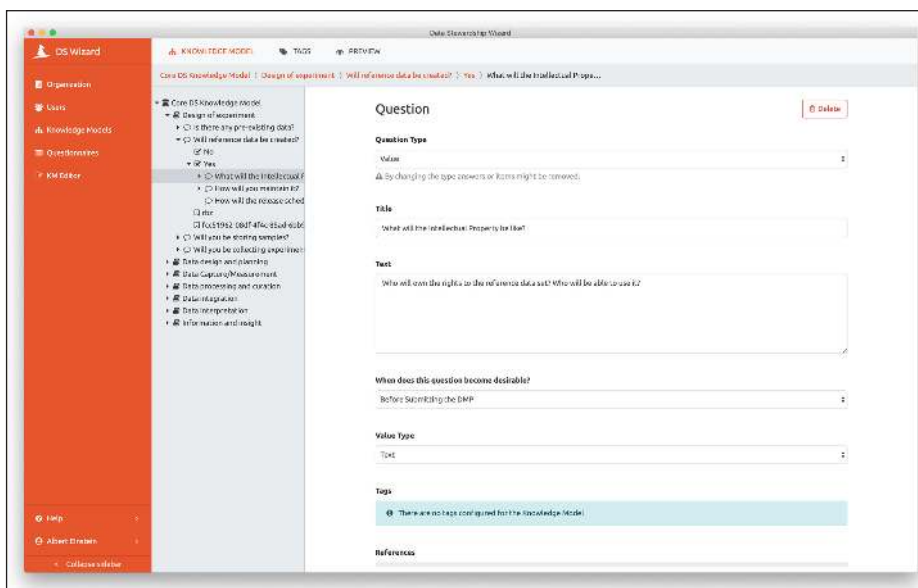

Figure 4: The Knowledge Model Editor.

2.4 Assembling of Data Management Plans

A key function of any tool for data management planning is to assemble all answers into a data management plan (DMP). rSmale et al. (2018) reports that existing DMPs, which usually consist of pieces of text provided directly from the researchers, are difficult to read and contain often incomplete sentences. This is not the approach taken for the Data Stewardship Wizard, for which the vast majority of questions are closed questions and very little text is provided by the researcher. Instead, the Data Stewardship Wizard generates textual DMPs from a questionnaire using an encoded **DMP template** in HTML, PDF, MS Word and LaTeX formats. The standard template represents all the questions and answers in the questionnaire. It can be customised using the Jinja2 (Ronacher 2019) template language.

We are currently working on more complex **assemblers**. These assemblers may for example prepare a concise DMP required by a funder containing just key topics. An example of such a template is the one based on Science Europe recommendations (Science Europe 2018). We are working on an assembler that will generate a Science Europe DMP from the questionnaire. Other templates, like the one for Horizon 2020, may follow.

This functionality lifts the weight of essay-style DMP writing from the shoulders of the researchers – based on user's answers, the assembler is able to formulate proper English sentences, thus resulting in high-quality DMPs not only from the perspective of completeness, but also readability. Furthermore, if at any point the need arises to switch templates, the Data Stewardship Wizard will be able to generate a new DMP in seconds, without requiring the researcher to answer a new set of questions.

2.5 Automated DMP Evaluation through Metrics

After extensive discussions with Dutch funder ZonMw on the process of DMP evaluations, we have gone one step further: in order to be able to judge how well a DMP satisfies the demands of a funder, each closed question in the standard knowledge model is enriched with metrics indicating how much each answer contributes to *Findable, Accessible, Interoperable, Reusable* data, how much it guarantees to deliver data that is as *Open* as possible, and also how much the answer otherwise expresses compliance with current *Good Data Management Practice*. Our approach differs slightly from that taken by Wilkinson et al. who synthesized specific questions to gauge the existing FAIRness of data (Wilkinson et al. 2018). Our approach does not require the researcher to answer specific questions, and it tries to predict how FAIR the data will become if the DMP is executed. Note also that the researcher answering the questions in the questionnaire will be shown an indication of which answers will lead to the best metrics: the questionnaire is meant to be a *guide* that helps the researcher achieve the FAIRest data, not an *examination*.

Based on their answers to the questionnaire, the researcher and funder can get a summary report of the metrics, as shown in **Figure 5**.

Figure 5: A summary report from the Data Stewardship Wizard showing six quantitative metrics, predicting the FAIRness and Openness of the resulting data based on a partially filled data management questionnaire.

We soon hope to be working with funders to test the use of these reports to automatically evaluate DMPs in project applications (Wittenburg et al. 2019).

3 Discussion and Conclusions

We presented the motivation, philosophy, and features of the Data Stewardship Wizard in the context of the current and emerging trends towards FAIR and Open research data. Apart from being a comprehensive tool for preparing DMPs, our Data Stewardship Wizard sets itself apart from other solutions by:

1. being based on FAIR Data Stewardship from day one,
2. providing predictive automated metrics for FAIR and Open Science practice,
3. being well prepared for machine-actionability of DMPs,
4. being able to fulfil different DMP templates based on a single questionnaire,
5. and providing scalability and evolvability of the Knowledge Model.

The first three points directly address the FAIR movement and became a basis for discussions about leveraging the Data Stewardship Wizard in a FAIR tools ecosystem. Such a case is one of our projects called the "FAIR Funders Pilot" which is described in Wittenburg et al. (2019). This is an initiative towards automated, machine-actionable DMPs included in projects proposals enabling and effective and efficient workflow for both funders and researchers.

Smale et al. (2018) observed several problems of DMP practice, mostly related to a low quality of DMPs both in content and style. We showed how the Data Stewardship Wizard can help alleviate these. Also, this study called for a possibility to "enable future reviews/updates", which is effectively addressed by our Knowledge Model migrations. Integration with other business systems similarly to the "FAIR Funders Pilot" is also mentioned in the study, as well as "exportable as PDFs for funders, publishers and institutional reporting purposes".

3.1 Availability

Although the Data Stewardship Wizard is an academic product, the project and various technical aspects are handled in a state-of-the art way by programmers having extensive experience from practice. We take pride in using good Open Source software development practice, which includes that the code is available in a public source code repository, contains complete documentation, and is subjected to continuous integration. On-site installation is possible by compiling from the source code, but also through Docker containers (Data Stewardship Wizard 2019). We solicit contributions both to the code and to the knowledge model from the world community of data stewards. A list of options for hosting of a Data Stewardship Wizard for your own institute can be found on the Data Stewardship Wizard web site, https://ds-wizard.org/.

Acknowledgements

Competing Interests

References

Buzan, T and **Buzan, B.** 2006. *The Mind Map Book*. BBC Active.
Data Stewardship Wizard. 2019. DSW Docker Repositories. URL: https://cloud.docker.com/u/datastewardshipwizard/.
European Commission. 2016. Open access & Data management – H2020 Online Manual. URL: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm.

European Commission. 2018. Template Horizon 2020 Data Management Plan (DMP): Annotated version for the use of participants under Societal Challenge 1. URL: http://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotateden.pdf.

Hooft, R. 2019. Data Stewardship Mindmap. DOI: https://doi.org/10.5281/zenodo.2614819

Mons, B. 2018. *Data Stewardship for Open Science: Implementing FAIR Principles.* CRC Press. DOI: https://doi.org/10.1201/9781315380711

National Science Foundation. 2010. ENG Guidance on Data Management Plans. URL: https://www.nsf.gov/eng/general/dmp.jsp.

Pergl, R. 2018. Form Engine. URL: https://github.com/ds-wizard/ds-form-engine.

Ronacher, A. 2019. Jinja2 The Python Template Engine. URL: http://jinja.pocoo.org/.

Sansone, S-A, et al. 2019. FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4): 358–367. DOI: https://doi.org/10.1038/s41587-019-0080-8

Sallans, A and **Donnelly, M.** 2012. DMP Online and DMPTool: Different Strategies Towards a Shared Goal. *IJDC*, 7(2): 123–129. DOI: https://doi.org/10.2218/ijdc.v7i2.235

Science Europe. 2018. Science Europe Guidance Document. URL: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf.

Simms, S, Jones, S, Mietchen, D and **Miksa, T.** 2017. Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes*, 3: e13086. DOI: https://doi.org/10.3897/rio.3.e13086

Simms, S, et al. 2018. DMPRoadmap wiki. URL: https://github.com/DMPRoadmap/roadmap/wiki.

Smale, N, Unsworth, KJ, Denyer, G and **Barr, DP.** 2018. The history, advocacy and efficacy of data management plans. *bioRxiv*, 443499. DOI: https://doi.org/10.1101/443499

Suchánek, M, et al. 2019. Data Stewardship Wizard – Diagrams. URL: https://github.com/ds-wizard/dsw-diagrams.

UK Data Service. 2012. Research data lifecycle. URL: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx.

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3. DOI: https://doi.org/10.1038/sdata.2016.18

Wilkinson, MD, et al. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5. URL: https://www.nature.com/articles/sdata2018118. DOI: https://doi.org/10.1038/sdata.2018.118

Wittenburg, P, et al. 2019. The FAIR Funder pilot programme to make it easy for funders to require and for grantees to produce FAIR Data. *arXiv:1902.11162*. URL: http://arxiv.org/abs/1902.11162.