

DATA STEWARDS AND DIGITAL PRESERVATION IN EVERYDAY RESEARCH PRACTICE

Esther Plomp

TU Delft
Netherlands
E.Plomp@tudelft.nl
<https://orcid.org/0000-0003-3625-1357>

Yan Wang

TU Delft
Netherlands
y.wang-16@tudelft.nl
<https://orcid.org/0000-0002-6317-7546>

J.S. Love

TU Delft
Netherlands
j.s.love@tudelft.nl

Abstract – Data Stewards at TU Delft promote digital preservation by incorporating actions in everyday research practice. Typical activities include requirement scoping, tool selection and policy drafting, all of which are tailored to a specific group, project or faculty. Here we discuss the situation of the Data Stewards within the university and examples of preservation work, including the creation of data repositories and a trial of Webrecorder.

Keywords – data stewardship, architecture, design, Webrecorder, policy

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?

I. THE TU DELFT DATA STEWARDS

Since the latter part of 2017, Delft University of Technology in the Netherlands has been running a pilot programme aimed at placing professional research data management (RDM) specialists into daily practice within university faculties [1]. These specialists hold the novel title of Data Steward, and they are charged with engaging research communities with research data issues and aiding them in defining what constitutes good data management within a specific discipline. Stewards are granted considerable latitude in identifying issues relevant to their communities as well as coming up with solutions on how to meet them and add value to daily research practices. As the entire domain of research data management falls within the Data Steward scope of interest, they partake of a broad range of data-related subjects: storage solutions, sharing technologies, legal & ethical concerns, etc. Digital preservation represents one key component of the Data Steward agenda with the stewards acting as mediators or facilitators, suggesting improvements to

work practices and gaining insight on implementation which can be translated into feedback for further development.

A credo of the Data Stewards is ‘incremental cultural change’, so large-scale shifts in work patterns are neither common nor necessarily viewed as desirable [2]. Instead, as members of a disciplinary research community, stewards join projects at various stages and act as analysts and consultants to make recommendations for improved durability of their research materials. The recommendations vary to suit person and project and may include actions such as suggesting software and tools, providing assistance with data sharing or helping draft documentation. Much of the work is simple advocacy, however, and often takes the form of raising awareness of why open formats are preferable or why and how to secure backup copies of data.

At this conference we aim to present a selection of cases to demonstrate how the TU Delft Data Stewards weave digital preservation into ongoing research practices. These are by no means an exhaustive list, and we are constantly innovating and iterating to improve operations. As jacks- and jills-of-all-trades, the Data Stewards are also by design not experts in digital preservation, and we would welcome discussions on how to better educate ourselves in the current tool sets available for digital preservation of research materials and how to translate this knowledge to our constituents.

II. RESEARCH DATA MANAGEMENT POLICIES AT TU DELFT

Since June 2018, there has been a central policy for research data management at TU Delft [3]. Part of this

policy is focused on making digital research data and outputs more accessible and sustainable. For example, the TU Delft library is expected to provide robust, high-quality infrastructure to facilitate data management and preservation. The TU Delft library hosts the DSA (Data Seal of Approval)-certified data repository, 4TU.ResearchData, that guarantees the availability of research data for at least fifteen years, in addition to advice and support for researchers in the preservation of their research outputs. The data repository has a dedicated team with the Data Stewards acting as satellites in their respective faculties.

In preparation for this central policy, a series of interviews and consultations were conducted with researchers from 2016 onward in order to gain a better understanding of prevailing perspectives on various issues relating to RDM [4]. It became clear from these consultations that practical guidance for diverse disciplinary practices was needed. In consultation with the Data Steward, members of each faculty at TU Delft are currently constructing faculty-specific RDM policies, in which the roles of the various faculty positions are outlined (principal investigators, PhD supervisors, PhDs, Department Heads and the Data Stewards). Researchers stand at the centre of the RDM policies and are consulted for their construction, an approach that is favoured at TU Delft [5]. In such a bottom-up approach, researchers are more inclined to invest their time and effort in long term preservation of their digital research materials [6]. Furthermore, the policy will reflect disciplinary practices and allow it to be a relevant document rather than being perceived as an administrative burden. The policies stimulate researchers to appropriately document and preserve their data by sharing their research outputs in a data repository (in accordance with the FAIR principles, [7]). For example, PhD students starting in 2019 will be required to deposit their underlying research, together with their PhD theses, before they are allowed to defend their thesis [3].

III. REACHING OUT TO RESEARCHERS

Being a large technical university, TU Delft hosts eight faculties covering a wide variety of science and engineering disciplines, as well as numerous interdisciplinary research groups. This composition poses challenges in providing preservation support at the university level, since one-size-fits-all solutions would never cover all needs per faculty or even per department. By being embedded in each faculty, Data Stewards have the chance to work closely with researchers, discover the challenges in their RDM

practices and subsequently understand the unique preservation requirements from each field via both top-down and bottom-up approaches. Top-down approaches include contacting researchers through the RDM policy engagement described above, faculty research councils, or graduate school and departmental executive secretaries. Bottom-up contact results from engagement with individual researchers on data management plans, providing researchers with information on RDM practices, and training sessions on tooling and long-term data archiving. Furthermore, researchers who advocate and implement good RDM practices are encouraged to become local Data Champions [8].

By reaching out to researchers, some insights into the varieties of research data they possess and challenges that they face are revealed, and it becomes clear that data diversity is often found within the same discipline. Figures 1 and 2 are word clouds representing primary research data types according to researchers in the fields of architecture and housing study respectively, both situated within the Faculty of Architecture and the Built Environment. They highlight the variation in research data even at the local/departmental level. Such overviews constitute essential input for building preservation infrastructures.



Figure 1 Research data types in architecture



Figure 2 Research data types in housing study

Another practical challenge faced by researchers is the lack of capacity in preparing datasets for publishing and archiving. Wide application of technology has led to increasing amounts of collaborative, interdisciplinary

research. For instance, scanning technology developed in an aerospace engineering group could contribute to heritage studies, and satellite imagery processed by geoscience researchers can be used in research on disaster management. However, these research opportunities gives rise to new challenges concerning data structuring and preservation. Sometimes, perhaps even often, new projects start with insufficient technological knowledge and skills from both researchers and research support to perform the desired tasks. Given the array of technologies and research instruments available, some researchers are struggling with structuring data sources, cleaning datasets or preparing documentation. As part of ongoing efforts to address these issues, the Data Stewards organize training workshops for the stewards themselves and for researchers on basic software skills, such as Software Carpentry [12]. At the same time, an increasing number of projects also demand a more customized preservation environment to ensure the reuse and reproducibility of the research. The case on the digital humanities at TU Delft will explain more about this preservation challenge.

IV. CASES

A. *Webrecorder*

Since early December 2018, the Data Stewards have been conducting a small pilot aimed at preserving research websites using Webrecorder (<http://webrecorder.io/>). The primary goal of this pilot is to perform preservation actions on a set of content-rich digital-born research outputs (websites) which currently have a high risk for loss. Pilots of this variety are initiated ad hoc by instigation of the Data Steward.

For this pilot we compiled a quick survey of existing websites produced by research groups involving members of the Faculty of Industrial Design Engineering (IDE) at TU Delft. The list was not intended to be exhaustive but rather to identify representative work and can be expanded as desired. Inclusion criteria were broad, and consideration for the pilot required only that a website be somehow affiliated with or produced by a member of IDE and that it be hosted outside of the TU Delft network (pages within the TU Delft domain currently have an institutional preservation solution through Archiefweb: <http://tudelft.archiefweb.eu/>). The cursory inventory yielded almost twenty sites of varying content and complexity. Of these, four were selected for a test with Webrecorder. Webrecorder was selected for this purpose following a recommendation obtained at the 'Memory Makers' DPC event in Amsterdam in November 2018. WARC files for portions of all four sites

were produced using Webrecorder. These provided a short-term preservation solution as well as tangible, usable products with which content producers could engage. These sites currently being tested are:

- CRISP, a design sector framework programme: <http://selemca.camera-vu.nl/index.html>
- Design & Emotion (D&E), a design society: <http://www.designandemotion.org/>
- DINED, an ongoing anthropometry production: <https://dined.io.tudelft.nl/en>
- 'Involving the CROWD ...', a site for a conference held in 2016: <https://museumsandcrowds.wordpress.com/>

Of these four sites, it was determined that Webrecorder was a suitable tool for preserving two, namely the CRISP and Design & Emotion sites, both of which are no longer currently being maintained but were deemed to possess continuing value. The 2016 conference site was judged to be temporary and not worth retention, while the DINED site continues to be developed and proved to be too complex in structure for Webrecorder to capture even superficially.

The most valuable insights from discussions with contributors to all four sites included the reasoning behind why these sites should be preserved (or not) and in what format. Within the design industry, CRISP was viewed as a flagship project, and its outputs continue to be consulted by practitioners in the field. Similarly the D&E Society provides historical context to the current Delft Institute of Positive Design. Webrecorder also sufficed to meet the demand that these sites remain integral and easily accessible online rather than as a complex set of HTML pages and associated media. At present the WARC files are accessible through the IDE Data Steward's personal Webrecorder account, and plans are being made to have a more centralized location as well as a backup copy of these sites in faculty archives.

B. *Digital Humanities at TU Delft and GIS Platform*

Although the primary research emphases at TU Delft lie in the sciences, a number of disparate research groups around the university also conduct research in the humanities. These include the history of architecture and urban planning, the Museum Futures Lab, preservation technology for imagery and heritage and a number of others. Unlike many other universities where the humanities have their own, separate research centers, these groups are embedded within major science and engineering disciplines. The demand for creating a digital humanities community across multiple

engineering disciplines at TU Delft was recognized and picked up by the Data Stewards in November 2018. Since then, several community-driven events have been organized. This initiative is led by researchers from the heritage groups in the Faculty of Architecture and the Built Environment, and Data Stewards from a few faculties followed it closely with the purpose to steer discussions and capture requirements on digital preservation.

Within this new community, some group members have elected to generate specifications for a GIS platform. This platform is intended to supplement TU Delft's 4TU.Researchdata repository by allowing an expanded range of search mechanisms and improved visualizations of GIS data. In a requirement gathering meeting in early March 2019, several existing open source geoplatform solutions (such as QGIS <https://www.qgis.org/en/site/about/index.html>, GeoBlacklight <https://geoblacklight.org/>, GeoNode <http://geonode.org/gallery/>), and current research projects involving spatial data were showcased and discussed. One insight derived from this activity is that maps are perceived as both research data and research outcomes by project members. Therefore both functions require consideration when designing preservation solutions. For an open repository of geographic data and related research outcomes, textual descriptions or simple metadata attached to the datasets were deemed insufficient for making datasets reusable. Instead, it was decided that potential users need a visualization of the datasets in order to understand their contents before downloading and re-using them. During the discussions of the platform, we were also able to have an open discussion about the merits of preparing datasets in open formats and that this was a more fundamental need beyond the software tools employed by any individual or group. With this understanding in place, or at least introduced, the envisioned geographic platform will be more accessible to any interested parties through the use of interoperable formats. Following a series of fruitful discussions, a list of features for the platform was compiled, which, together with the insights obtained, could serve as baseline requirements and initial input for other service providers at the university to proceed with selection and implementation.

V. DISCUSSION AND FUTURE WORK

The cases presented here represent only two examples of the types of digital preservation activities which are needed for research work at TU Delft. Data Stewards are in a fortunate position to gather and

interpret further requirements as part of daily research, and they stand in a prime position to discover practical challenges to digital preservation within the university research environment as well as negotiate between researcher needs and the design of digital preservation solutions. Ideally this will all be done incrementally and iteratively in order to discover and reinforce good practice wherever possible. As part of this process, the Data Stewards would like to seek out new avenues by which they might gain relevant knowledge concerning good practices recommended by preservation experts.

REFERENCES

- [1] Teperek, M. "Vision for Data Stewardship at TU Delft", 2018. Retrieved from <http://doi.org/10.5281/zenodo.1251989>
- [2] Teperek, M., & Dunning, A. "The main obstacles to better research data management and sharing are cultural. But change is in our hands", November 14 2018. Retrieved from <https://openworking.tudl.tudelft.nl/2018/11/14/the-main-obstacles-to-better-research-data-management-and-sharing-are-cultural-but-change-is-in-our-hands/>
- [3] Dunning, A. "TU Delft Research Data Framework Policy", 2018. Retrieved from <https://doi.org/10.5281/zenodo.2573160>
- [4] Teperek, M., Cruz, M. J., Verbakel, E., Böhmer, J., & Dunning, A. "Data Stewardship addressing disciplinary data management needs." *International Journal of Digital Curation*, vol. 13 no. 1, pp. 141–149, 2018. <https://doi.org/10.2218/ijdc.v13i1.604>
- [5] "TU Delft Strategic Framework 2018-2024", 2018. Retrieved from [https://d1rkab7tlqy5f1.cloudfront.net/TUDelft/Over TU Delft/Strategie/Towards%20a%20new%20strategy/TU%20Delft%20Strategic%20Framework%202018-2024%20%28EN%29.pdf](https://d1rkab7tlqy5f1.cloudfront.net/TUDelft/Over%20TU%20Delft/Strategie/Towards%20a%20new%20strategy/TU%20Delft%20Strategic%20Framework%202018-2024%20%28EN%29.pdf)
- [6] Teperek, M., Higman, R., & Kingsley, D. "Is Democracy the Right System? Collaborative Approaches to Building an Engaged RDM Community". *International Journal of Digital Curation*, vol. 12 no. 2, pp. 86–95, 2017. <https://doi.org/10.2218/ijdc.v12i2.561>
- [7] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* vol. 3, art. no. 160018, 2016. <https://doi.org/10.1038/sdata.2016.18>
- [8] Plomp, E. "Data Champion kick off meeting", 14 January 2019. Retrieved from <https://openworking.tudl.tudelft.nl/2019/01/14/data-champion-kick-off-meeting/>
- [9] M. King, B. Zhu, and S. Tang, "Optimal path planning," *Mobile Robots*, vol. 8, no. 2, pp. 520-531, March 2001.
- [10] H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.
- [11] M. King and B. Zhu, "Gaming strategies," in *Path Planning to the West*, vol. II, S. Tang and M. King, Eds. Xian: Jiaoda Press, 1998, pp. 158-176.
- [12] Kurapati, S., & Teperek, M. "4TU.Centre for Research Data partners with The Carpentries: Impressions from the first workshop at TU Delft". 14 March 2019. Retrieved from <https://openworking.tudl.tudelft.nl/2018/12/16/4tu-centre-for-research-data-partners-with-the-carpentries-impressions-from-the-first-workshop-at-tu-delft/>