Supporting Sustainable Digital Data Workflows in the Art and Humanities

Vera Chiquet[1*]

1 University of Basel, Switzerland, Data and Service Center for the Humanities (DaSCH)

*Corresponding author: Vera.Chiquet@unibas.ch

Abstract

The Data and Service Center for the Humanities (DaSCH) operates as a platform for humanities research data and ensures access to this data and promotes the networking of data with other databases (linked open data), to add value for further research and the interested public. As a competence center for digital methods and long-term use of digital data, it supports the hermeneutically oriented humanities in the use of state-of-the-art digital research methods. The DaSCH focuses on qualitative data and associated digital objects (images, sound, video, etc.) in the cultural heritage field.

Long-term archiving or access is a major topic after the digital turn in the humanities, as many funding agencies such as the Swiss National Science Foundation and the European Commission are now requiring that a data management plan (DMP) be in place to receive research funding. This new imperative raises many questions in the scientific community. This paper points out the contributions of the DaSCH for digital humanities researchers and the advantages of interoperability.

keywords

sustainable workflows, best practices, research infrastructure, interoperability.

INTRODUCTION

The initiative launched at the Digital Humanities Lab of the University of Basel for the long-term availability of research data has culminated in a pilot phase for a digital infrastructure to support research and cultural heritage institutions concerning sustainable digitization. From 2013 onwards, this infrastructure was consolidated as a nationally funded Data and Service Center for the Humanities (DaSCH).

Since the start of the pilot project in 2009, the interdisciplinary exchange between experts in digital infrastructure and researchers from the humanities and cultural heritage has been a central achievement. A successful translation of subject-specific needs and technical possibilities is central to the successful and satisfactory implementation of digital research projects in arts and humanities. Only through interdisciplinary exchange the questions about research objects and their usage scenarios can properly be determined. Some processes have to be taken into account.

The observance of standards such as the "FAIR Principles" (Findable, Accessible, Interoperable, and Reusable), the "International Image Interoperability Framework" (IIIF), JSON-LD, as a self-explanatory data container, or persistent ARK identifiers (Archival Resource Keys) are important for many subsequent usage possibilities from the very beginning. Only if the whole workflow of digitization is kept in mind, long-term access to research data can be ensured, making them available for further research and thus facilitating the reuse of existing digitized material. If merely the internal, in-house processes are

followed, it often leads to isolated solutions - which in the long term is an obstacle to interesting opportunities for inter-organizational data exchange. To avoid the risk of an isolated, forgotten and neglected project you should follow the internationally recognized quality guidelines to allow your project to be reused and to play a vital role in networking with other digital resources.

I WORKFLOWS IN DIGITIZING RESEARCH PROJECTS

1.1 Data collection and documentation

First of all it seems to be a lot of extra effort to write a DMP, to think about the research data management and the FAIR principles. It is additional work that may not evolve directly from the key research question.

Much documentation is though simply good research practice. When starting with a new research project its documentation should be an issue right away, and not something that is written at the end of a project. Based on the experience that it is better to start sooner than later not only with the necessary documentation but also with the appropriate storage of data, we are developing a web-based application on which the research project can be introduced, managed and supplemented with data from the very beginning.

What data will be collected, how it will be collected, observed and generated and in which form its documentation will be recorded are key questions. It is important to deliver information to enable other researchers and re-users to understand your data. The documentation should be done consistently throughout the project. The data documentation gives contextual information about datasets. It specifies the aims and objectives of the original project and contains explanatory material including the data source, data collection methodology and process, dataset structure and technical information.

1.2 Ethics, legal and security issues

The implementation of the FAIR principles raises a lot of questions about rights permissions and accessibility in the digital field. The questions about what data will be collected, observed and generated will also outline if the data could be personal or sensitive, so the data should be protected. A further issue which should be considered from the beginning is the owners of the copyright and Intellectual Property Right (IPR) of all data. In many cases it is not possible to determine the legal situation at the beginning of a project – but the fact that this aspect is not being ignored until the end of a project is pivotal. Especially for research with pictures or artworks, this is a great asset. When the research outcome is ready for publication, you do not want to start negotiating with the rights holders or make a publication without images, if they are an essential part of your research data.

The DaSCH supports the open data usage license solutions but also offers limitation possibilities for cases where the legal situation has not yet been conclusively clarified or access restrictions are necessary because of specific reasons.

II FAIR-PRINCIPLES AND THE DASCH

The ‘FAIR Guiding Principles for scientific data management and stewardship’ were published in 2016 in Scientific Data. They intend to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. Because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data, the principles emphasize

machine-actionability (for example the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention).

This is why the DaSCH provides tools and long-term access to research data in accordance with the FAIR principles and international standards for interoperability (IIIF, JSON-LD, ARK…). As a national research infrastructure it offers services to individual scholars as well as to projects of research consortia from all areas of the Humanities dealing with qualitative research data. It supports scholars in generating new and re-using existing digital research data for cutting-edge research, to promote the reuse and cooperation between different research projects.

The "DaSCH" got selected in 2019 for FAIRsFAIR funding to obtain the CoreTrustSeal certification, which promotes as a community based, non-governmental, and non-profit organization sustainable and trustworthy data infrastructures. FAIRsFAIR - Fostering Fair Data Practices in Europe aims to supply practical solutions for the use of FAIR data principles throughout the research data life cycle. It emphasizes on fostering FAIR data culture and the uptake of good practices in making data FAIR. By providing a platform for using and implementing the FAIR principles in the day-to-day work of European research data providers and repositories, FAIRsFAIR will play a key role in the development of global standards.

Research-IT groups at universities or platforms like FAIRsFAIR can be helpful to find the repository that fits the research project. Different projects need different archiving solutions. This is especially the case in Arts and Humanities, as every project is individual and sometimes very specific. An important step is to decide which is the suitable digital archiving solution. Currently, there are different repositories in Switzerland that support research projects. Even though the DaSCH is a national infrastructure, it can not be the perfect repository for every project in the Humanities. For quantitative or Social Science research projects for example, FORS may be the right alternative. Once the cooperation between researchers and repository is established, the technical expertise and support should be covered by the repository, so that the researchers should not have to worry too much about the storage and preservation of their data.

2.1 Data storage and preservation

Data storage on laptops or hard drives, for example, is not safe enough. Storage through research infrastructures is safer. The implementation of the data access rights and permissions and the use of format standards should then be handled by the research infrastructure that stores your data. The research infrastructure should enable long-term preservation for the datasets beyond the lifetime of the project.

Findable: The findability is defined by using Persistent Identifiers (PID) at best down to single objects (like DOI, ARK). This should enable to trace the complete change history for objects and data fields. The data should be described with rich metadata. The metadata should also be assigned to a globally unique and eternally persistent identifier, to be indexed in a searchable online resource.

Accessible: The (meta)data should be retrievable by their identifier using a standardized communication protocol. The protocol should be open, free, and universally implementable. It should also allow for an authentication and authorization procedure, where necessary. The metadata have to stay accessible, even when the data are no longer available. For the long-term archival strategy, the data should be stored in multiple redundant copies on

different media. Even if the data can not be made public, the metadata should be accessible online.

2.2 Data sharing and reuse

Data has to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output. Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. The research infrastructures that will take over the role as a repository, should conform to the FAIR Data Principles.

Interoperable: To generate interoperable digital resources they should be based on a broadly applicable language for knowledge representation, which includes qualified references to other (meta)data. By using open-source software it is possible for the users to study, change, and distribute the software, thus it enables open collaboration. By using RESTful APIs the repository provides interoperability for search, metadata, and data.

Reusable: The (meta)data should be released with a clear and accessible data usage license (CC0, CC BY…). For cases where necessary (legal reasons, embargo period) rights and permission systems can be integrated. It should follow the domain-relevant community standards and provide rich documentation and transparency of the provenance of the data.

Concerning the aspects of data storage and preservation, as well as data sharing and reuse, which are very technical, researchers can hand over responsibility to the research infrastructures. By cooperating with digital repositories these issues should be covered by these institutions. The research data infrastructures should know best which standards are sustainable, which formats are maintained as well as take over technology watch and migration strategies. It is precisely in the course of digitalization that cooperations like this have become inevitable, which can be a great opportunity for connecting humanities scholars and software scientists for a broader, socially relevant research.

IV REFERENCES

http1 https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8793

http2 Force11: The FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples

http3 https://www.go-fair.org/fair-principles/

http4 Swiss National Science Foundation: Explanation of the FAIR data principles: http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation-_SNSF_logo.pdf

http5 European Commission: Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants-_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

http6 Checklist: How FAIR is your data? https://zenodo.org/record/1065991

http7 https://www.fairsfair.eu/

http8 https://www.coretrustseal.org/

http9 https://iiif.io/

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

ANNEX REPOSITORIES

https://www.re3data.org/

https://dasch.swiss/

https://forscenter.ch/