Managing data in cross-institutional projects

Zaza Nadja Lee Hansen[1], Filip Kruse[2], Jesper Boserup Thestrup[3]

Abstract

This paper provides guidelines for data management professionals and researchers on how FAIR data usage can help improve the planning, execution and overall success of a cross-institutional project. Cases from Danish cross-institutional projects are detailed to illustrate this point – as well as the lessons learnt with implementing FAIR data principles in such projects.

Key learnings from this paper are:

- Using FAIR data principles in cross-institutional projects can help manage the data used in the project in terms of knowledge sharing, access rights, use of templates, metadata and further sharing the data after the project has ended.
- To benefit the most from using FAIR data in a cross-institutional project it should be considered and planned for early in the project process.
- If FAIR is not considered early in the project process problems can arise such as a lot of time spent on converting formats, obtaining permissions and assigning metadata.
- It is necessary for researchers and research projects to have infrastructure and other services in place which support FAIR data usage.

Keywords

FAIR, data management, cross-institutional projects

1. Introduction

In order for research institutions to be competitive and to transfer knowledge into social and commercial gain there is an increasing need to share research data between multiple actors. However, when research data has to be shared across institutional boundaries challenges arise due to project members being used to different data management approaches.

FAIR data management is a requirement in many research projects, including Horizon 2020 applications (European Commission, 2018). The FAIR data principles (FORCE11, 2018) state that data has to be Findable, Accessible, Interoperable, and Re-usable; in other words that data should be as open as possible. However, some data cannot be made fully open – like healthcare data. However, it could be open to some degree. Hence, in the projects discussed in this paper, in particular the FAIR Across project, the participants were told to follow FAIR principles and make data as open as possible – but as closed as needed. In other words, given a good reason – for example that it would be a breach of the General Data Projection Regulation (GDPR), it is possible to make more data (semi-) open as openness is then seen on a scale and not as an either/or decision which for security reasons can make sure stakeholders go for the “safe” option and then say no to completely open data.

By implementing the FAIR principles data can more easily be shared in a cross-institutional project, thus making it more likely that project deliverables are done on time and give the expected outcome. This can be done by ensuring that all data related to the project follows some specific guidelines, standards and formats, agreed upon at the project start. Examples include where and how to store and share data, who is responsible for updating which data, which data format to use, including how and where to share metadata.

This paper will describe how the FAIR principles can be applied to research data with a specific focus on cross-institutional research projects. Furthermore, a series of adaptation cases for usage of FAIR data principles at Danish Universities will be presented.

2. The relevance of FAIR data in cross-institutional projects

2.1. The FAIR Across project

In 2018 a working group was created under DeiC – Danish e-Infrastructure Cooperation with representatives from Aalborg University, Copenhagen University, the Royal Library, Technical University of Denmark, Copenhagen Business School and the Danish National Archives. This working group investigated how FAIR data principles could be used to encourage researchers to share data across data types, disciplines and institutions. The working group interviewed researchers at several of the 8 Danish universities in order to create guidelines and material on the FAIR principles to help aid researchers share data when taking part in cross-institutional projects.

2.2. Key findings for relevance of FAIR data in cross-institutional projects

FAIR can be used to contribute towards the success of a cross-institutional project in three main ways:

1. Project planning of cross-institutional projects
2. Navigating project changes in cross-institutional projects
3. Sharing information in cross-institutional projects
4. Sharing data outside the project group

The primary focus areas within the FAIR data principles are making data Findable and Accessible so everyone in the project has access to the same data at the same time. The key findings from the project can be seen in table 1.

Table 1: Summary of benefits of FAIR data usage in cross-institutional projects

	How FAIR data usage can help	Expected benefits
Project planning of cross-institutional projects	• Define responsibilities for data management among the project partners, e.g. in a data management plan. • Agree on common standards / conventions for collecting, storing and documenting the data. • Ensure that all data from the project are shared on a common, secure platform.	• You save time introducing new members to the project. • You avoid data loss when someone leaves the project. • You enable easy reuse of data generated so far.
Navigating project changes in cross-institutional projects	• Document methods and decisions in a systematic manner, e.g. using common templates • Use common standards for collecting, storing and documenting data, e.g. on file names, file formats, table contents, etc. • Store all data in a structured way on a shared and secure system that is accessible for the project members.	• You can extend your analysis by finding new suitable datasets more quickly. • You can reuse existing datasets more easily for e.g. new analysis. • You minimize the risk of wasting time and money on duplicate work.
Sharing information in cross-institutional projects	• Make agreements on how research data are managed and shared between research partners. • Establish clear guidelines on how to document data in a consistent form, e.g. as templates for tables or notebooks. • Use a secure shared storage system with controlled access management for all data.	• You avoid misunderstandings regarding access to and use of the data. • You foster collaboration to make best use of the data, speed up the processes and reduce administration. • You improve the quality of the data and thus the scientific output from the project.
Sharing data outside the project group	• agree on a procedure for when data or findings are considered “complete”/”valid” and thus ready to be shared outside the project working group. • Establish clear guidelines on The criteria for data to be fully open, semi-open or closed, keeping in mind	• You make as much data open as possible – without fear of security issues/GDPR concerns • You foster further use of the data beyond the lifetime of the given project

[redacted]	that the goal is to make as much data as open as possible. • Include an IT security expert or GDPR expert or DPO as needed in creating the above guidelines. • Document the decisions for data openness in the project • Agree on what platforms/media data should be released on and a person responsible for this task

The findings were validated through interviews, workshops and a conference with data management professionals and researchers.

3. FAIR implementation cases

3.1 A brief outline of the project Data Management in Practice

The project *Data Management in Practice* (DMiP)[4] aimed at providing Danish researchers with an operational infrastructure for research data management. The general objective was to establish a Danish infrastructure setup with services covering all aspects of the lifecycle of research data: from application and initial planning, through discovering and selecting data and finally to the dissemination and sharing of results and data. Further, the setup should include facilities for training and education in research date management. Researchers' needs and demands should form the basis of the services and close connection to and cooperation with research projects already in progress was an integral part of the project setup - hence the "in Practice". Finally, the project should explore the role of research libraries regarding research data management.

The project organization can be described as a hybrid middle path between a purely case-based project with individual institutions each working on their own sub-projects, and a thematic project with institutions working within one or more broad themes. The structure chosen contained six themes: Data Management Planning; Data capture, storage and documentation; Data identification, citation and discovery; Select and deposit for long-term preservation; Training and marketing toolkits; and Sustainability. Each of the participating institutions worked on specific cases, such as ongoing research projects, well-defined data collections etc., all cases covering the entire data lifecycle. Each case should relate to the themes to be able to draw conclusions on both a more case-specific and a more general theme-specific level. The cases spanned the main academic fields of Humanities, Social Sciences, Science and Technology. For an overview of the project see Kruse & Thestrup (2018).

Below we shall explore selected cases and analyse which challenges FAIR data principles posed in the practical data management process, how these challenges were met and to what extent.

One important lesson from the process is that FAIR data principles have to be an integral part of the Data Management Plan and the project plan right from the outset.

3.2 Identifying case-specific examples of FAIR data principles

3.2.1. The LARM case

The aim of the LARM case was to enable researchers and students to use radio programs as research objects by establishing a digital archive with an associated research infrastructure with relevant tools (see Hansen et al. 2018, pp. 12-15)

The Royal Danish Library's LARM, the Sound Archive for Radio Media, offers researchers the LARM.fm (https://www.larm.fm/) platform as the workspace for research projects using radio data. This platform enables researchers to work with the material, add metadata, annotate and structure selections of items etc. An important aim of the project was to provide facilities to overview the new data, a solution for long-term preservation of data (location and format) and to ensure future reuse of data for other research projects.

Among the results of the LARM case, the following is of special interest: the development of the Royal Danish Library's DMP template (see Hansen, (2018b, 26-35[5] and 54-56[6]),, which is a central element in the Danish version of DMPOnline (https://dmponline.deic.dk/), preparations for a future data harvesting facility, and technical decisions on data formats etc.

The LARM.fm platform requires the researcher to confirm that all data she enters into the system such as new metadata, annotations etc. are shareable under a CCO license. Data entered is not issued with an author identification, which is a drawback, the data have an internal ID, while the metadata for the program includes a 'DOMS-ID', facilitating location in mediestream.dk[7]. Data sharing as an element of data identification, citation and discovery is not possible, however, for this data, as some of the media data is sensitive or protected by copyright. The annotations for the programs and the metadata for the annotations often also contain personal information or program descriptions from e.g. new agencies. This has led to the contingency of designing of a data repository with restricted access, which would still allow for reuse of data for research purposes, Library Controlled Access Repository (LCAR). Some of the challenges outlined above result from the data collections being large and inhomogeneous. In practical research, it will be possible for the researcher to select, download and share smaller data sets, after ensuring that the data is not protected or restricted in use, thus meeting several FAIR criteria. The Royal Danish Library's Library Open Access Repository (LOAR, https://loar.kb.dk/) will be well suited for this task.

3.2.2. The Netlab case

The aim of the Netlab case was to use the Royal Danish Library's webarchive, Netarkivet.dk, to map the historical development of the Danish web (see Hansen et al. 2018, pp. 16-20). To achieve this aim it was necessary to develop an infrastructure with provisions for corpus creation, tools, workspace, storage etc.

The Royal Danish Library's webarchive, Netarkivet.dk, poses much the same problems for the researchers as LARM. As material and data harvested from the web contain personal information or copyrighted material, the archive is only accessible to researchers who have requested and been granted special permission to use the collection for specific research purposes. Analogous to access to the archive, sharing of data generated through research in the archive is prohibited if it contains personal information or material otherwise restricted. Deposit of such data in LCAR will still be an option.

Regarding data identification and discovery the LOAR and LCAR repositories mentioned above supplies data with DataCite DOIs thus making data searchable through DataCite's services. By the use of OAI-PMH metadata is available for harvesting or searching through the Danish Data Archive https://www.sa.dk/en/

Both LARM and the Danish netarchive ask the question of ownership of data. The Royal Danish Library owns data in the netarchive and collective or individual rights owners own data in LARM. Research data created from these sources can be the product of individual or group efforts. In order to clarify legal issues of ownership and copyright to research data a legal framework was drawn up, the Model Agreement on Data Management in Cooperative Research Projects[8].

Based on the two cases outlined above we may conclude firstly that to make data FAIR it is necessary to take these principles into account at the earliest possible stage of the research process. All the solutions were developed ex post as answers to challenges occurring in the course of ongoing research projects. Secondly, that while access to research data for obvious reasons can be restricted - without necessarily preventing its reuse - care should be taken to make metadata findable and accessible for external users and non-researchers. These external users should for example be able to search metadata and should be given information on how to obtain access to data if they find data, which is not public accessible.

3.2.3. The Kierkegaard case

In the Kierkegaard case (Hansen et. al. 2018a, pp. 21-25) the project participants were involved in a process keeping data available for future researchers. Søren Kierkegaard's collected writings were republished from 1997 to 2009 in 55 volumes. As part of this publication an electronic version was made available on www.sks.dk. The online version is based on a format called Kierkegaard Normalformat (KN1). This format was developed as part of the publication process. The goal of this DMiP case to preserve the data in a format, which would allow researchers to annotate and collaborate with the text and ensure that the data can be reused in 50 years (Hansen et. al. 2018, p. 23).

This goal was achieved by reformatting the data into TEI (Text Encoding Initiative[9]). TEI is an international standard, which allows long-term preservation and gives researchers the possibility of collaboration. About 80 % could be reformatted automatically. The remaining 20% had to be manipulated by a computer scientist in order to ensure that all data could be saved in the TEI format (Hansen et al. 2018, p. 23).

In order to share the data the project group tested a Dataverse[10] server. Dataverse is a software, which makes it possible to upload and share data. The data is now available if a researcher requests access. But the data is stored on a platform called ERDA[11] which offers no direct access to the data for the general public. Since the data is not shared the datasets are not issued with a DOI and metadata has not been shared (Hansen et al. 2018a, p. 24).

The case underlines that FAIR as a principle need to be taken into consideration early in a given process in order to ensure that data can be reused. It must be planned to store the data in an open format, in an early phase of the research project. The case shows that it can be necessary to allocate considerable resources to make data FAIR, especially in a case like this where the data format has been designed years ago. The case also shows that the data must be stored within an infrastructure that can add adequate and open metadata to make data findable.

3.2.4. The DTU Wind Energy and DTU Space cases

Two of the Science case of the DMiP project concerned data from DTU Wind Energy (http://www.vindenergi.dtu.dk/english) and DTU Space (http://www.space.dtu.dk/english) . Regarding the Wind Energy project, the project group should "document, catalogue and archive datasets to make them available wherever possible". The staff worked at the DTU Bibliometrics and Data Management office (Hansen et al, 2018, p. 38).

The data was gathered over a period of at least 20 years and contains information, which can be used by other researchers, governments and industry. The group realized that the datasets consist of very different types of data such as measurements, experiments and models. The data was stored on different types of media and without sufficient metadata (Hansen et al., 2018, p. 37).

The project group worked with several topics. Evaluation and usage of an existing Data Management Plan (DMP) template was used to define criteria for repositories and longtime preservation. Then the group compiled and shared two datasets via Zenodo[12] (Hansen et al. 2018, pp. 40-42).

Based on the case the DMiP report concludes that a DMP describing the data and how to store and share is easy to create via the tool DMPOnline, but requires a good template for any given DMP. Zenodo uses the DataCite metadata schema and offers the possibility of exporting the metadata via OAI-PMH. This allows for harvesting and transfer of metadata to other data catalogues. The use of a wide array of data types and formats stresses the importance of choosing a metadata standard suitable for data discovery now and in the future.. The case showed that FAIR principles have to be part of the research project as early as possible in order to ensure that data can be shared together with relevant metadata.

The data from the DTU Space case contained information on the magnetic field of Earth collected by 17 ground stations. The stations are located in the South Atlantic, Greenland and Denmark. The data is transferred from the stations to servers on DTU Space. The goals in this DMiP case was to "increase visibility of the group's research", to give access to data and make the data citable (Hansen et al. 2018a, 43-44).

As part of the process, the researchers produced several versions of a DMP. The final DMP was used to design the work of the DMiP project case. An infrastructure to share the data was described, but as of 1. April 2019, it is not yet in operation, but the existing infrastructure is supplemented with necessary guidelines etc. The researchers have begun uploading their data to the DTU Data Repository (http://data.dtu.dk) adding DOIs, thus improving data discovery. The case gave the project members improved insight in how to plan a workflow and share data, which can be used when other research projects are planned. Regarding FAIR principles, the case demonstrated that issues such as infrastructure and metadata must be taken into consideration in order to be able to share data.

3.2.5. The Kepler Case

Since 2008 the KEPLER mission (https://www.nasa.gov/mission_pages/kepler/main/index.html) collected data on stars and extrasolar planets orbiting the stars. This has generated over 100 TB of data, which must be shared and preserved for future research. The DMiP project explored how to ensure longtime preservation and sufficient metadata. The case showed that researchers have to consider issues like formats, metadata and infrastructure as a whole, and thus also the FAIR principles, very early in the research project. To share and longtime preserve 110 TB of data requires knowhow on formats and infrastructure in order to ensure the desired outcomes For example the suggested solution to ensure longtime preservation was to store data in a format called BAGIT[13], where all information regarding a given star would be stored as individual datasets with a unique DOI per dataset.

3.3 Services established

As a result of the DMIP project several services were established in order to provide researchers with access to data infrastructures: the Danish version of DMPOnline and the open access repository LOAR (Library Open Access Repository). LOAR offers 5 years preservation of up to 10 GB of research data free of charge for researchers from Danish universities. Preservation for longer periods or larger data sets is possible for a fee. Researchers are expected to share the data using Creative Commons licenses. Thus LOAR facilitates reuse of data. A restricted access repository, LCAR (Library Restricted Access Repository) is developed but as of April 15. 2019 awaits decision for launching. Both of the latter services use the free open source software DSpace (https://duraspace.org/dspace/) for the building of repositories and are similar in structure and terms of service, except for terms of access. Further, in order to create a common legal framework a model agreement[14] on handling data was designed (Hansen et al. 2018a, p. 8). As can be seen from the cases mentioned above, FAIR has to be integrated in the planning of any given research project in order to ensure access to data with a minimum of work. This requirement relates closely to data formats, metadata, and legal constraints for access, but also access to infrastructure and general legal frameworks.

4. Discussion

Implementation of the FAIR data principles can help a cross-disciplinary project to deliver better results and solutions with more efficient use of project resources. FAIR data principles support and facilitate knowledge discovery and sharing as roles and responsibilities are clearly indicated and data is delivered according to standards and issued with the necessary metadata markings. However, we saw that it is vital that the project participants agree to use FAIR principles from the start of the project and that the infrastructure is in place as it is very time-consuming and costly to add this later.

The problem we repeatedly encountered during the DMiP project in relation to making data from the cases in the project FAIR was that data can be sensitive, contains private or personal information or is protected by copyright. Such ethical and legal issues can to some extent be addressed from the start. But not if the use of this type of data occurs during the project. Temporary solutions could be necessary to consider: anonymization, access to limited parts of the data, access to metadata only, etc. This could be helpful to researchers interested in the project.

The use of open access repositories such as The Royal Danish Library's LOAR could also encourage the researcher as data owner to incorporate FAIR principles also in the process of choosing repository. Again, this stresses the importance of issue of infrastructure.

5. Conclusion

This paper has provided guidelines for data management professionals and researchers on how FAIR data principles can help improve the planning, execution and overall success of a cross-disciplinary project and have shown cases of how FAIR data principles can be used and the benefits and challenges with doing so.

In conclusion, the implementation of FAIR data principles can be a useful tool to help ensure the success of a cross-disciplinary project and is a great tool to aid in knowledge sharing as well among the project participants as with external parties. However, it is important to decide upon using FAIR principles before or at least as early on in the project as possible to lessen cost and minimize difficulties in their use and also to ensure that infrastructure and services are in fact in place to support this type of knowledge sharing between the institutions participating in the project.

Acknowledgements

Thanks to Anne Sofie Fink Kjeldgaard, section head at the Danish National Achieves, for input, feedback and general support.

Reference List

FORCE11. *FAIR data principles*, https://www.force11.org/group/fairgroup/fairprinciples. Accessed 31/08/2018.

Hansen, K. K. et. al. (2018a). *Data Management in Practice Results and Evaluation*. Copenhagen: DEFF. DOI: 10.7146/aul.243.174. http://ebooks.au.dk/index.php/aul/catalog/book/243

Hansen, K. K. et. al. (2018b). *Data Management in Practice Supplementary Files.* Copenhagen: DEFF. doi: 10.7146/aul.244.175. http://ebooks.au.dk/index.php/aul/catalog/book/244

EUROPEAN COMMISSION (2016). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020*. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf . Accessed 04/15/2019

Kruse, F. and J. B. Thestrup (2018). *Data Management in Practice – Knowing and Walking the Path. ERCIM NEWS*, no. 114: 40-41. https://ercim-news.ercim.eu/en114/r-i/data-management-in-practice-knowing-and-walking-the-path. Accessed 04/15/2019

[1] Project manager at the Danish National Archives

[2] Senior Advisor at The Royal Danish Library

[3] Communications Officer at The Royal Danish Library

[4] The project partners: RUC - Roskilde University, KB - The Royal Library (merged in 2017 with the State and University Library as The *Royal Danish Library)*, DDA - Danish Data Archive, DTIC – DTU Library, Technical Information Center of Denmark, SB - State and University Library, now The Royal Danish Library, AUB - Aalborg University Library, SUB - University Library of Southern Denmark. DEFF, Denmark's Electronic Research Library and the participating institutions funded the project evenly. The project period was March 2015 - June 2017, final report January 2018: Hansen et al.: Data Management in Practice, Results and Evaluation, available at: http://ebooks.au.dk/index.php/aul/catalog/book/243 Accessed 04/15/2019.

[5] In Danish

[6] In Danish

[7] Mediestream.dk provides online access to Danish cultural heritage, newspaper, radio and television programs etc. DOMS (Digital Object Management System) is an in-house custom-built system for preservation of cultural heritage metadata.

[8] http://www.au.dk/samarbejde/erhvervssamarbejde/samarbejde-med-forskere/modelaftale-for-samarbejde-om-forskningsdata/ The Model Agreement has been approved by Aarhus University.

[9] http://www.tei-c.org/index.xml, Accessed 31/08/2018.

[10] See more on the software here: https://dataverse.org/. Accessed 31/08/2018.

[11] Copenhagen University's Electronic Research Data Archive (ERDA)

[12] http://doi.org/10.5281/zenodo.160136 and http://doi.org/10.5281/zenodo.161966. Both accessed 04/15/2019.

[13] BAGIT is a file packaging format designed for storing and transferring digital content. http://www.digitalpreservation.gov/series/challenge/data-transfer-tools.html. Accessed 04/10/2019

[14] https://www.deic.dk/da/news/2017-08-15/modelaftale. Only in Danish. Accessed 04/15/2019.