Research Data in the Fraunhofer Digital Project : Creating a FAIR Research Data Infrastructure and Culture

Oya Beyan , Fraunhofer FIT, Germany , beyan@fit.fraunhofer.de

Andrea Wuchner, Fraunhofer IRB, Germany, andrea.wuchner@irb.fraunhofer.de

Dirk Eisengräber-Pabst, Fraunhofer IRB, Germany, dirk.eisengraeber-pabst@irb.fraunhofer.de

Christoph Quix, Fraunhofer FIT, Germany, christoph.quix@fit.fraunhofer.de

Christian Zaschke, Fraunhofer IOSB, Germany, christian.zaschke@iosb.fraunhofer.de

Oliver Schumacher, Fraunhofer , Germany, oliver.schumacher@zv.fraunhofer.de

Abstract

The Fraunhofer Society, as the leading organization for applied research in Europe, conducts its research activities by 72 institutes and research units at locations throughout Germany. The Fraunhofer Digital project, a part of the Fraunhofer 2022 Agenda, introduces the vision, that research and administrative data will be linked, aggregated and analyzed in order to optimize the research and development processes and support management decisions. As a part of this vision, the Research Data in the Fraunhofer Data Space project develops methods to reuse the research data in a broader sense, and integrate research data silos in various institutes into Fraunhofer Data Space. In this work we present the concept, early outcomes, and a FAIRness assessment of the research data repository approach. In our assessment we used the FAIR metrics, and as a result developed a set of recommendations. These recommendations will be utilized to establish a digital research data infrastructure as well as a research data reuse culture in Fraunhofer. They can be also useful for other large scale institutions with heterogeneous research communities.

Keywords

Research Data Infrastructure; Data reuse; Fraunhofer Data Space; FAIR evaluation

I. Introduction

The Fraunhofer Society is the largest organization for applied research and development services in Europe with 72 institutes and research units spread throughout Germany, each focusing on different fields, predominantly in natural science or engineering studies. The majority of the more than 25 000 staff are qualified scientists and engineers who work with an annual research budget of 2.3 billion euros.

Since 2006, the Fraunhofer Society adopted the open access policy that implies granting free and long-term access to scientific findings and scientific literature. In recent years, an agenda for open data has been established, and starting in 2015, a research data management and an infrastructure project has been launched [1]. The Fraunhofer Institutes are largely independent in their research project planning and implementation. Therefore, the research data generated by different research communities are stored and maintained in heterogeneous repositories without any standard ways of access to them [2]. Although there are versatile networks, they currently do not support the discoverability, accessibility and reusability requirements of open science.

The need of having an overarching research data infrastructure is addressed by the Fraunhofer Digital project. As part of the Fraunhofer 2022 Agenda, research data and administrative will be linked, aggregated and analyzed in order to optimize the research and development processes and support management decisions in the institutes and in central levels.

This paper will present the initial result of the Fraunhofer Digital project concept phase for research data. The following section will introduce the concept and the projects which are a part of the research data infrastructure.

The next section will present an assessment of the designed system regarding to the FAIR principles of findability, accessibility, interoperability and reusability of research data. Finally, we will discuss a set of recommendations that developed as results of our assessment, which serve the establishment of a data sharing culture and infrastructure in large scale research institutions with decentralized heterogeneous research activities.

II. Research Data in the Fraunhofer Digital Project

The Fraunhofer Digital project is part of the Agenda 2022 which aims to increase the impact on business and society through excellence. The project aims to provide an efficient IT infrastructure to support emerging needs of research and innovation as well as digital business processes. It is a cross departmental project which involves six Fraunhofer institutes, and is composed of three main sub projects aims (i) to modernize the current automation of business processes, (ii) to make diverse Fraunhofer research data sources available in Fraunhofer Data Space and (iii) to develop a business intelligence framework to enable linking, aggregating, analysis and presentation of data.

The Research Data in the Fraunhofer Data Space is one of the sub projects that constitutes Fraunhofer Digital with a specific focus on research data. It targets the development of methods to reuse the research data in a broader sense, and integrate research data silos reside in Fraunhofer Institutes into Fraunhofer Data Space. The current Fraunhofer research data landscape is composed of a centralized system for publication data, a centralized patent management system and diverse systems which are owned and governed by research communities distributed to over 60 Fraunhofer institutes. As presented in Figure 1, the digital infrastructure will integrate the research data generated in various systems into databases, and will provide business intelligence to analyze them.

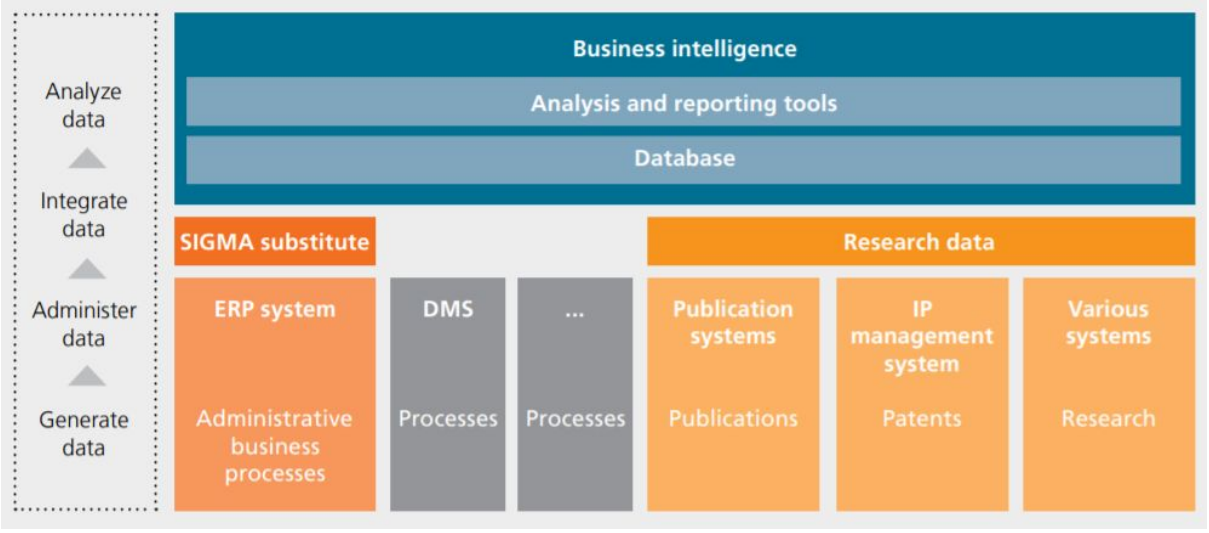

Fig1: The Research Data in the Fraunhofer Data Space as part of the Fraunhofer Digital project

The Fraunhofer Industrial Data Space will serve as a virtual secure data space to facilitate the exchange and linking of research data by providing standards and common governance models [3]. It will contain separately maintained research data sources, centralized and decentralized. Research data will be semantically described, and will be discoverable and accessible. The Fraunhofer reference architecture model [4] will be the base to describe research data assets, and it will be extended with required additional semantics. Connectors will be used to integrate data silos, whereas data apps will extend the functionalities of them with customized data integration, processing, linking and publishing. Brokers will enable the interaction between data consumers and

data providers in a secure environment. The project also has a perspective to share data between federated research data infrastructures by establishing consortia of data owners and infrastructure managers.

The project is divided into three phases, namely: analysis and concept; realization and implementation; operation and further development. The concept phase, which has started in 2018, has two main goals as follows: (i) to identify representative use cases on how research data is generated, stored and reused; and (ii) a prototypical implementation of an expandable core of Fraunhofer Data Space, covering a broad, representative selection of research data sources. This implementation, as a proof of concept, will verify the developed concepts and will help to evaluate conformance in existing and emerging research data management standards and recommendations, such as Horizon 2020, for data management plans or European Open Science Cloud.

As a cross institutional project, Fraunhofer Digital Research Data project is built upon ongoing projects and works carried by participating Fraunhofer institutions. One of them is the extension of existing open access publication repositories, Publica and ePrints, which are the official institutional repositories of the Fraunhofer society to share research publications. The second project is the FORDATIS project, which has been started in 2016 as a part of the Fraunhofer Open Access Strategy 2020, with the goal to build a research data repository. The project is a part of the EU-Horizon 2020 project JERRI [1] and has been designed to create a meta-level describing research data and the data management plans. Other relevant work are data and metadata modelling as part of Fraunhofer Industrial Data Space project. The following subsections will give an overview on the existing projects and works, and will present their vision to support Fraunhofer Digital research data concept and infrastructure.

A. Publication Data Infrastructure

The Fraunhofer Publication Infrastructure began as a bibliographic database and was complemented 2003 by an Open-Access (OA) Repository. Today, the Publication Infrastructure (Publica and ePrints) serves an output of over 10,000 scientific papers, books and conference papers published every year. The growing OA Repository is fostered by a top-down OA-Strategy, an OA Publication fund and an OA-transformation task force. It is the data fundament for all services managed by the “Research Service & Open Science” Team (RSOS).
In the context of the project “Fraunhofer Digital” Publica and ePrints will be updated to a new technological platform. It will become the principle system for Fair Data Sharing of all Fraunhofer OA-Disseminates – including research data, scientific software, patents etc.. Disseminates like research data will not have to be held in the system itself but their metadata will. The wide diversity of resources will be linked to further entities of the research ecosystem like projects and researchers.
The new machine-friendly Fraunhofer Publica will be more research and resource centric as part of a globally networked research community. It will also provide automated processes and workflows to support researchers during the publication process. The Business Intelligence Ready Publica will not only act as first point of contact to the Fraunhofer Open Science Cloud, it also serves as a data source for Fraunhofer internal data handling.

B. The FORDATIS Research Data Repository

The FORDATIS Research Data Infrastructure project has been started by the Fraunhofer Society, with the goal to complement the existing publication infrastructure and related services with a Fraunhofer Research Data Repository named as “FORDATIS”. The FORDATIS supports the vision of making the results of publicly funded projects freely available to the public on a long-term basis, in order to make it possible to verify the research results and make the data more useful for further research. Funding organizations and science policy initiatives promote open research data such as Horizon 2020 EU Open research data pilot and provide guidelines for best practices to make data available, e.g., the Guidelines on FAIR Data Management in Horizon 2020. To fulfill these requirements the FORDATIS project has set the following goals for published research data: (i) construction of a research data repository to centrally record research data and provide them for reuse; (ii)

integration of the research data repository into national and international structures (e.g., OpenAIRE platform, European Open Science Cloud EOSC, Re3Data, Base); (iii) implementation of a DOI-Assignment in order to make the research data uniquely identifiable and citable, (iv) establishment and development of a permanent centralized service to ensure persistent access to data.

The first step of the implementation was the development of an Application Profile that contains all the metadata fields to describe research objects, namely both research data and software object types. The developed profile complies with EU and DOI requirements and with the DataCite-Standard [5]. In the next step, we realized that various objects, such as author and project, originally intended as attributes of the resource "research data", were also relevant for publications that are stored within the Fraunhofer Publica. Ideally, these data should be collected as separate objects with their own key, and linked to publications and research data. Therefore, the data models are revised according to these requirements (Figure 2).

At the current state of the project, the implementation choices are being discussed. The repositories DSpace 6.x and DSpace Cris are considered possible choices [6] [7]. Both systems are open source with a large developer community and are already being used as research data repositories in other institutions. Since the FORDATIS repository should get integrated in the existing publication infrastructure, therefore the final implementation choice will depend on the design choices of the Publica and ePrints systems.

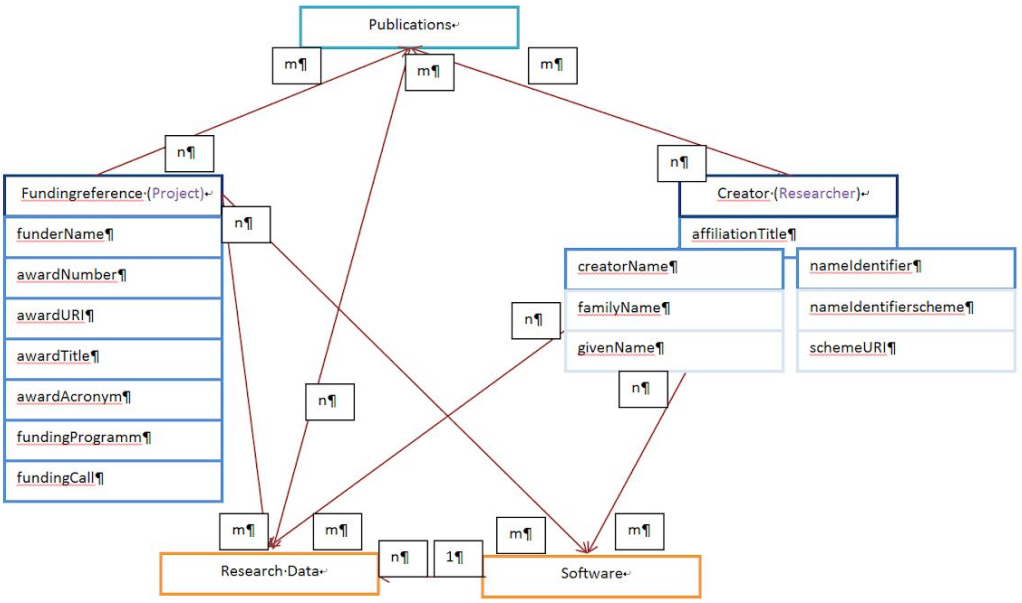

Figure 2: FORDATIS/Publica-Object-Model

C. Research Data Provenance and Research Objects Metadata

One of the challenges of integrating the research data into Data Space is identifying and understanding the original sources of research data. As almost in every research organization, this is also in Fraunhofer a very heterogeneous structure, including file systems, database systems, or proprietary data stores of specific applications (e.g., control software of a measurement device in a lab). To get an overview of the available research data sources and the research objects contained in them, we are currently using questionnaires to collect the basic information. The collected information will be formalized in a metamodel which is based on the Information Model of the Industrial Data Space (IDS) [8].

However, current work covers only the descriptive and administrative metadata of research data. To enable interoperability of research data, we also need to capture the structural metadata. Structural metadata describes the structure of data containers and also contains the schema of a relational database or XML document, the tree structure of a JSON document, or just a list of columns of a CSV file. This information can then be used by data integration systems to exchange data between different organizational units. As it is already available in the data containers, this metadata can be extracted automatically from the sources [9].

In the next phases of the project, we plan to make the metadata available in a comprehensive metadata management system, which allows the curation, semantic enrichment and the annotation of the extracted metadata.

D. Long Term Archiving

In addition to the process of persisting and describing data using metadata, a long-term archiving concept for the Research Data has to be elaborated. It needs to considered which data needs to be archived in which way. Also the process of accessing archived data has to be described.

To make data searchable and retrievable by different communities working in various domains, domain specific metadata has to be provided. Such specific metadata enables the provision of the relevant data to the requester in an efficient way. To allow this, it at first needs to be determined which is the right set of metadata for a specific community. After that, we need to extract these information out of the existing research data and its corresponding metadata and make them available for the search and retrieval functionality.

When working with sensible data also access restrictions have to be taken into account. Some data might be of interest for different persons working in the same or in different institutions. So it would be very useful if those persons could easily access this data. On the other hand, if this data is sensible in some kind of way (e.g., personal data or security-related data) there need to be protective measures to restrict access to this data. Therefore, the concept should also allow easy access for authorized persons and effectively protect the data as needed.

In the long term, metadata models are subject of change. This changes also affect the long-term archiving. If such changes emerge, conversion strategies have to applied. One solution could be to transform the queries performed on the data storage which would affect mainly the intermediate layer between the requester and the storage. Another way could be to change the data representation on the storage side which would imply changes on the data storage level. We have to determine the best approach for the Fraunhofer environment.

III. FAIRness Assessment of the Fraunhofer Research Data Repository

One of the targets of the concept phase is to evaluate the Fraunhofer research data management approach regarding the best practices, recommendations and guidelines. FAIR principles provide guidelines to make digital research objects findable, accessible, interoperable and reusable both for machines and humans. The goal of research data in Fraunhofer Data Space is to enable business intelligence to consume research data without any additional manual processes. Therefore, to evaluate the compliance to the FAIR principles in early stages of the process was important to provide feedback to the conceptual approach.

We assessed the design principles and implementation choices regarding the FAIR principles, and identified challenges and recommendations. Although FAIR principles are well known, the metrics to measure FAIRness are still in progress. Recently, a definition of the FAIRness measurement published by the FAIR Metrics group. The group developed a framework and a core set of FAIRness indicators [10]. In our work we used 14 exemplary universal metrics, developed by this group, covering each of the FAIR sub-principles. A summary

matrix of the evaluation outcome is presented in Figure 3. Fully satisfied metrics are colored green, partially satisfied ones yellow, and failed ones red.

Figure 3. The FAIR Metrics fulfillment level

Metrics addressing the findability aspects are fully satisfied by the Fraunhofer Research Data infrastructure. The FORDATIS systems implements DOIs as persistence identifiers and machine readability of the data is ensured by application profiles in Dublin Core format. The application profile provides dedicated fields to include the DOI of the data it describes. DSpace Lucene solutions will ensure indexing the metadata in searchable resources. One of the shortcoming of the current implementation is that it supports only high level metadata of research objects.

Accessibility principles are only partially satisfied. Our approach implements open protocols and provides access authorization and authentication. However, there is no policy defined to describe processes and mechanisms to access restricted content yet. Therefore, the access authorization metric is partially satisfied. Another failure point is related to metadata longevity solutions. The current implementation does not have a metadata longevity approach. This shortcoming will be address in collaboration with the long term archiving group.

Interoperability metrics aim to evaluate both metadata and data. The FORDATIS systems follows best practices regarding to metadata. Data resources are described with formal and accessible knowledge representation language and vocabulary. However, no mechanism exists to monitor how individual data sets satisfies these metrics. Regarding the linking data sets, the current application profile supports qualified references.

Reusability evaluation measures the existence of an usage license, availability of detail provenance and certification of the resources. Since there is no certification body exist yet, last metric is unachievable for any data infrastructure. The creative common licenses are used by the FORDATIS system. The system keeps provenance information via Dublin Core vocabulary such as creator, publisher, description and dates. However, there is no contextual provenance related to why and how data is produced.

IV. Discussion and Recommendations

As the outcome of the assessment of the Fraunhofer Digital Research Data concept and implementation choices with FAIR principles, a set of recommendations have been developed. In many cases, these recommendations did not directly arise from a non-satisfied metric, but addressed the challenges of widespread adaptation of research data reuse culture. The Fraunhofer society, with over 60 institutions, embrace many scientific

communities with different needs and requirements. The following recommendations were developed to assist a research data infrastructure, which strives to support a variety of use cases arise from these communities.

Rec 01 - Persistence Policy and Related Services : The FORDATIS system has a persistence identifier system based on DOIs. However, it is important to understand the advantages and limitations of different identifier schemas. Researchers may have diverse needs and they may need guidance to identify the right PID schemas; and in some cases they may need support to establish different PID schemas for their projects. Fraunhofer can provide guidelines, consultancy support, and PID system as a service.

Rec 02 - Research Data Policies: Research data policy documents describes policies, terms of use and guidelines for the reuse of data. Fraunhofer research communities have their own approach and restrictions regarding data reuse policies. However, identifying these requirements and harmonizing heterogeneous policies at the institutional level is a challenging task. An institutional data policy framework can ensure the coherence across community level data policies. A policy registry system and related services can help researchers to specify their own data reuse policies and link with the institutional strategies. A common structure of policies and guidelines to customize the specifications and semantic annotation capabilities can help researchers to develop and publish their data plans.

Rec 03 - Rich and Machine-readable Metadata: The current practice of the machine-readable metadata is limited with high level descriptions of research data sources. Most of the context specific information captured in description fields is in a natural language. Since metadata curation is time-consuming, the only way to achieve desired rich metadata is the extensive use of automated tools. Fraunhofer can support researchers by investing in tools to capture and formalize metadata data as a part of their daily activities. These tools, coupling with the project management and data stewardship workflow, can enable Fraunhofer researchers to generate machine readable semantic metadata, and can optimize the business processes of creation and publishing of the metadata. The new generation machine executable data management plans and their supporting tools might be considered as a solution.

Rec 04 - Metadata Longevity Plan: As a part of preservation policy, a long term archiving strategy and metadata longevity plan should be developed. These plans should include items such as policy and procedures for digital archiving, backup schedules, and preservation of fair objects.

Rec 05 - Research Data Objects: The scope of the research data objects should be defined. Data, metadata, algorithms, materials, or hardware can be considered as research data. The knowledge representation language and supported data types should be identified.

Rec 06 - Licensing: A rich option of data licensing should be provided for researchers, including creating their own licenses. A guidance to select the right licenses should be provided.

Rec 07 - Contextual Provenance: Contextual provenance is important for the reuse of the data. It defines the context the data is produced in and the relevance of the data for the reuse purpose. This metadata can be used to evaluate the quality or fitness to the use from the perspective of data consumer. However, it is challenging to define the scope of the context and methods to collect metadata. Different research communities may have different needs and perspectives regarding the context. A community-driven approach can be adopted and contextual provenance can be prototyped in selected use case.

Rec 08 - Data quality: High quality research data are a cornerstone of scientific knowledge. FAIR principles provides only limited guideline regarding the quality of data itself. New standards for the measurement of the data quality should be studied.

V. Conclusion

Fraunhofer Digital Research Data project develops a research data infrastructure to integrate and to link heterogeneous research data repositories to the Fraunhofer Data Space. The framework supports business intelligence as well as reuse of data. The project is in the conceptual phase and is composed of multiple projects executed by different institutions from the Fraunhofer society.

This paper presents the Fraunhofer approach to research data infrastructure, and reports the outcome of a self-assessment of current design and implementation choices with FAIR metrics. As a result of self-assessment, a set of recommendations were developed, and feedback was given to the conceptual design.

The developed recommendations might help institutions who embrace heterogeneous research communities with different requirements of research data management and sharing. The aim is twofold: to establish a research data reuse culture and to support digital research data infrastructure.

References

[1] Küsters, U., & Klages, T. (2018). Fostering Open Science at Fraunhofer.
[2] Küsters, U., & Erben-Russ, M. (2012). Forschungsinformationssysteme bei Fraunhofer. FORSCHUNGSINFORMATION IN DEUTSCHLAND: ANFORDERUNGEN, STAND UND NUTZEN EXISTIERENDER FORSCHUNGSINFORMATIONSSYSTEME, 37.
[3] Otto, B., Auer, S., Cirullies, J., Jürjens, J., Menz, N., Schon, J., & Wenzel, S. (2016). Industrial data space: digital souvereignity over data. Fraunhofer White Paper.
[4] Otto, B., et al. "Reference architecture model for the Industrial Data Space." Fraunhofer-Gesellschaft, Munich (2017).
[5] Ammann, N., Nielsen, L. H., Peters, C. S., & de Smaele, T. M. (2011). DataCite metadata schema for the publication and citation of research data. DataCite. http://web. archive. org/web/20160825024134/http://schema. datacite. org/meta/kernel-3/doc/DataCite-MetadataKernel_v3, 1.
[6] DSpace 6.x Documentation: https://wiki.duraspace.org/display/DSDOC6x/DSpace+6.x+Documentation [Accessed July 10. 2018]
[7] Palmer, D. T., Bollini, A., Mornati, S., & Mennielli, M. (2014). DSpace-CRIS@ HKU: Achieving visibility with a CERIF compliant open source system. Procedia Computer Science, 33, 118-123.
[8] Chakrabarti, A., Quix, C., Geisler, S., Pullmann, J., Khromov, A., & Jarke, M. Goal-Oriented Modelling of Relations and Dependencies in Data Marketplaces.
[9] Quix, C., Hai, R., & Vatov, I. (2016). GEMMS: A Generic and Extensible Metadata Management System for Data Lakes. In CAiSE Forum (pp. 129-136).
[10] Wilkinson, M. D., Sansone, S. A., Schultes, E., Doorn, P., da Silva Santos, L. O. B., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. Scientific data, 5.