RESEARCH ARTICLE

State of Research Data Management in Industry and Research Institutions in the Manufacturing Industry

An empirical analysis of the partners from the iDev4.0 Project

Dirk Ortloff [1]
Sabrina Anger [2]
Martin Schellenberger [3]

1. DO-IT-Service GmbH, Augsburg (before employed at camLine GmbH).
2. Hauptabteilung Halbleitertechnologie, Fraunhofer Institut für Integrierte Systeme und Bauelementetechnologie (IISB), Erlangen.
3. Intelligente Energiesysteme, Fraunhofer Institut für Integrierte Systeme und Bauelementetechnologie (IISB), Erlangen.

Date Received:
05.06.2023

Keywords:
RDM, State of affairs, Industry, Institutes, Inggrid

Data availability:
Data can be found here: NA

Software availability:
Software can be found here: NA

Abstract. The paper presents insights into the situation concerning research data management (RDM) in the high-tech manufacturing industry and respective research institutions. Besides standards and guidelines, data management and its degree of formalization play a decisive role in digital transformation in all organizations. The authors of this study benefited from the opportunity arising within the framework of the European collaborative project iDev4.0 to evaluate RDM in the industry as well as in research institutions of different sizes and orientations. The study focuses on RDM-related soft criteria (e.g., understanding, awareness, value assessment) but also the concrete implementation of RDM.
For this survey, the team conducted expert interviews and evaluated them using a qualitative analysis oriented to Mayring's approach. The results provide insight into the attitude of involved stakeholders towards RDM on the one hand and its practical implementation on the other. Identified commonalities, differences, and needs of the different parties are presented in this paper.

1 Introduction

Harmonization and standardization of data and handling of data is one fundamental aspect of "Industry 4.0" including the digitization of manufacturing. Standards and guidelines are enablers of digital transformation approaches in the industry. The availability of high-quality data throughout the whole data life cycle plays a decisive role in this respect. Therefore data management and its degree of formalization is another fundamental aspect of digital transformation. As such, it is also a central idea that legislation and funding agencies in their programs require this. The topic of formalized Research Data Management (RDM) surfaced during the EU project iDev40. iDev4.0 (for more details please refer to [1]) was one of the biggest recent European projects in the context of Industry 4.0 with the objective to develop and implement a digitalization strategy for the European electronic components and systems industry. The aim was to develop and

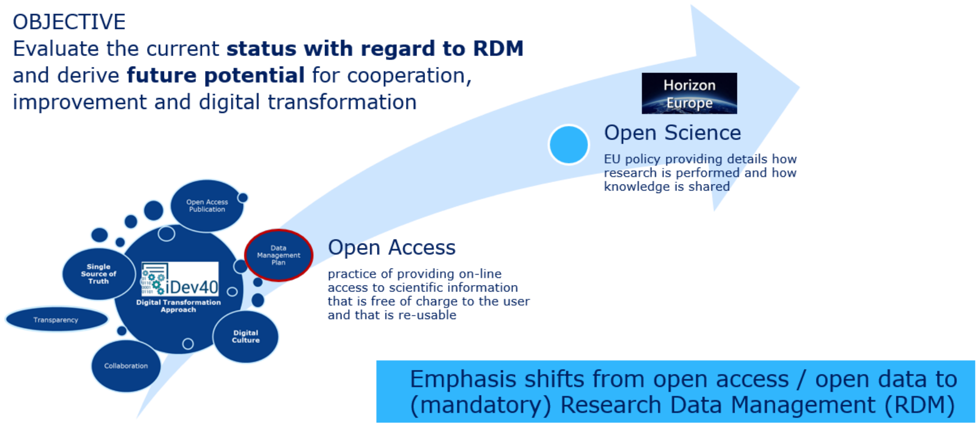


Figure 1: Objectives of the RDM interviews

implement solutions for data driven advanced analytics of largely heterogeneous databases and adopt artificial intelligence and deep learning algorithms in order to semi-automatically enrich contents and extract facts from unstructured contents. It was obvious that in order to achieve these goals, the development of the mandatory Data Management Plan must take into account the special challenges in the cooperation between industry and research institutions.

Thus, in the course of this project, a small team interested in that subject formed and investigated that topic further. This ”RDM-team” consisted of persons from Fraunhofer IISB and camLine GmbH.

Two aspects of manufacturing RDM were identified to be of interest:

1. RDM is playing a more and more important role in publicly funded projects regarding Open Access and Open Data. Both topics enforce the publication of the related research data in a well-documented and reusable manner.

2. The principles behind RDM for publication should be applied as a blueprint for organization-internal management of research data. The same is true for production data.

For substantiating the impression, the idea was born to investigate the current state of affairs within the iDev40 project partners. The RDM-team came up with the idea to analyze the awareness and practices through interviews guided by a prepared questionnaire. The iDev40 partners got asked to volunteer to participate in these interviews. Additionally, the RDM-team contacted several partners directly. Figure 1 highlights the main objectives of the RDM survey and its integration into the overall digital transformation pursued by iDev40. The results, therefore, contribute to the insights into the best practices in standardization and internal organization.

2 Activity classification concerning background and state-of-the-art and definition of terms

2.1 Background and state-of-the-art

In general, the vision and ideas about RDM are not new. There are already many efforts to promote the implementation of RDM in the context of open administrative data and in scientific institutions. Open Data laws have entered into force, e.g., the first Open Data law in Germany in 2017. It implements the demands from the G8 Action Plan for a legal Open Data regulation at the federal level. The paragraph instructs the authorities of the direct federal administration to publish the unprocessed, so-called "raw data" they have collected, with a few exceptions. This open administrative data ("Open Data") can be used by anyone free of charge and can be processed further in their administrative processes.

Therefore, the national metadata portal GovData (govdata.de) got established. A respective metadata standard was developed (DCAT-AP.de). The latter guarantees interoperability with the European Data Portal (data.europa.eu), which pursues the same purpose. Numerous programs, initiatives, and projects to promote open administrative data are ongoing in the D-A-CH region. Expansions to public sector information also regarding Open Data guidelines of Germany and the European Union (EU 2019/ 2024) are planned [2].

Also, the current push for broad RDM initiatives stems more from the legislation and the requirements of public funding organizations.

On the EU level, Open Data is pushed ahead even further by the approval of an Open Source Strategy by the European Commission [3]. Furthermore, Horizon Europe [4] mandates an Open Science policy (including mandatory Open Access publication and research data management (data management plan, metadata in line with FAIR principles) as the key novelty.

In the context of open administrative data, several contact points and guidelines exist to provide support for Open Data implementation. There are also many efforts to promote the implementation of RDM in scientific institutions, especially in those that create digital research data.

Germany's Federal Ministry of Education and Research, e.g., is currently funding 21 projects on research data management throughout Germany that look for solutions addressing identified challenges for RDM [5]. Many universities and research organizations already have pieces of training or guidelines concerning RDM available. They offer templates for implementation, e.g., of Data Management Plans, if existing templates of funding agencies (e.g., the H2020 templates for data management plans) cannot be used. Guidance for researchers concerning RDM implementation and underlying principles is also provided by openAIRE ([6]; [7]).

For several industries, the research community and companies even have published the ideas, approaches, and benefits more than a decade ago (see, e.g., [8]; [9]; [10]), too. Anyhow, the uptake in academia and industry was partly limited. Additionally, the breadth of today's scope was not yet fully addressed at that time.

To consolidate singular approaches and find a multidisciplinary solution the "Nationale Forschungsdateninfrastruktur (NFDI)" got proposed in 2016. NFDI got created as a nationwide competence and infrastructure network in Germany [11]. It intends to ensure the provision and indexing of research data for science. As one part of the German National Research Data Infrastructure (NFDI), besides other disciplines, the NFDI4Ing consortium aims to develop, disseminate, standardize, and provide methods and services to make engineering research data FAIR [12]. As one of the first consortia funded as part of the NFDI, NFDI4Ing has actively engaged engineers across all engineering research areas, including experienced infrastructure providers, since 2017. It now has more than 50 active members and participants and continues to grow. As technically appropriate, the RDM team established a dialog with this consortium beyond the project-related and confidential communication. However, the activities of the NFDI4Ing project appear to be mainly addressing the RDM topic from an academic point of view, at least for the moment. At least, that is the impression of the current RDM-team involvement in the NFDI4Ing, who recognized a certain disconnect between the process in academia and the perceived industry practices. There seem to be no practical solutions for industrial applications, combining internal data storage solutions with RDM solutions targeting the publication of research data, at least not yet.

Differentiating therefrom and complementary to other surveys (e.g., a survey performed by Springer Nature, for continuously published results see State of Open Data report, [13]), this iDev40 survey focused on RDM principles and approaches implemented in industry and industry-related R&D within the iDev40 consortium.

2.2 Definition of terms

The basis of this survey is the following understanding of the RDM team of research data management itself and the related terms:

The major benefit of formalized RDM is to ensure the usability of data during project execution and for a longer time afterward (see also Figure 3). Publication of research data allows verification and builds traceability and trust in the research results. In the context of public research, the terms ”Open Data”, ”Open Science”, and ”Open Access” come into play. These get explained in detail in Figure 2.

RDM is more than pure data management (see Figure 15). Rather, it is structured management of information (data with its context and meta-data) and knowledge. It even may include software tools and models if essential to reproduce the data analysis. The RDM-team’s understanding of formalized RDM and its benefits can be summarized as follows:

- Analyzability of data
- Exploitability for current problems and solutions/approaches
- Reuse of existing data for future problems
- Interpretation of existing data sets in light of new research questions
- Verification of results
- Derivation and documentation of lessons learned
- Transparency of scientific results and decisions, which builds trust

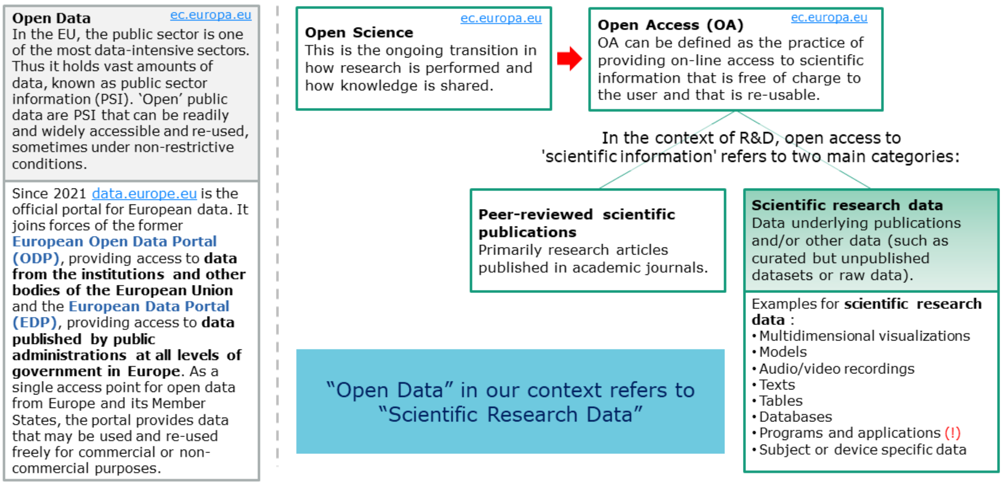


Figure 2: The context of Open Data, Open Science and Open Access

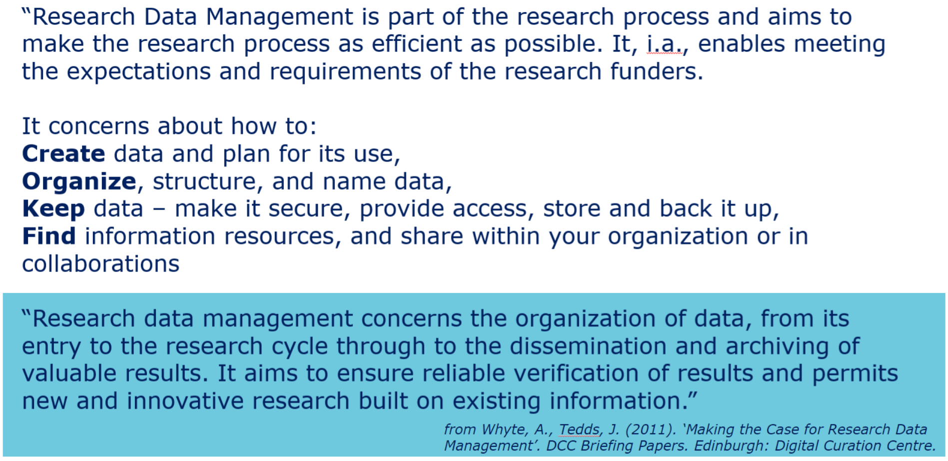
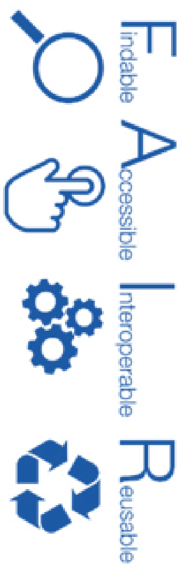


Figure 3: The intents and benefits of RDM

- Prevents re-inventing the wheel

As mentioned before, the principles behind RDM may not only be used in the context of research projects but may serve as a blueprint for the everyday organization-internal management of data. Thus it also has a strong internal perspective. Formalized RDM ensures the usability of data during project execution and for a longer term afterward by sustainable data preparation and storage throughout the whole data life cycle. The most important underlying principles are summarized in the acronym FAIR, i.e., data treated under the FAIR principles must be findable, accessible, interoperable, and reusable, as defined by the Go FAIR Initiative [14]:



- Findable: The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata is essential for the automatic discovery of datasets and services, so this is an essential component of the “FAIRification” process.
- Accessible: Once the user finds the required data, one needs to know how they can be accessed, possibly including authentication and authorization.
- Interoperable: The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
- Reusable: The ultimate goal of the FAIR principles is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that results can be replicated and/or combined in different settings.

It is important to mention that the FAIR principles can be applied to treat data under Open Access and private data, i.e., FAIR is not equal to ”Open”: The ”A” in FAIR stands for ’Accessible under well-defined conditions’. There may be legitimate reasons to shield data and services generated with public funding from public accessing. These include personal privacy, national security, and competitiveness. The FAIR principles, although inspired by Open Science, explicitly and deliberately do not address moral and ethical issues about the openness of data. In the envisioned Internet of FAIR Data and Services, the degree to which any piece of data is available or even advertised as being available (via its metadata) is entirely at the discretion of the data owner.

3 Methodology of the interviews and the analysis

To be clear in advance, the evaluation of the situation concerning research data management within the iDev40 consortium was not conducted as a representative survey but as a spotlight investigation. For this qualitative survey, the RDM-team has chosen expert interviews as the survey methodology (see Figure 4). Due to the late start of this activity, the team conducted interviews with a smaller group of project partners only. The purpose was to tap into the relevant knowledge of the group of people dealing with data in their organizations. Throughout the survey, there was no need for these experts to provide (research) data beyond their answers or access to any data. As an initial step, the interviewers developed a questionnaire as a guideline for the expert interview to ensure the comparative structure and logic of the different expert interviews

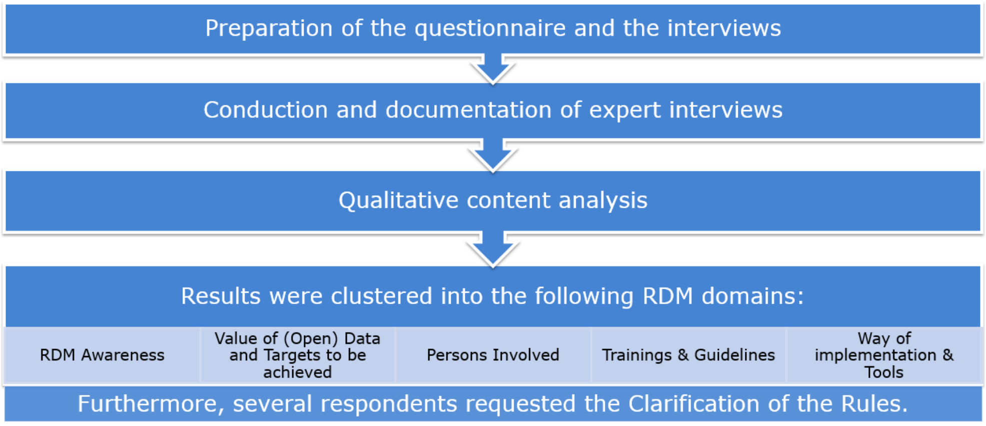


Figure 4: Methodology of the survey

on the one hand and comparable information content of the answers on the other.

3.1 Preparation of the questionnaire and the interviews

The interview guideline was developed based on a literature study concerning the current state-of-the-art in RDM. Furthermore, the RDM-team added its own experience about the subject matter. A collaborative brainstorming activity collected the questions from within the team. These about 60 questions got combined into a questionnaire grouped into the following categories:

- Basic RDM understanding and awareness concerning the topic
- Implementation of the FAIR principles / the principles of RDM
- Assessment of the value and benefits
- The role of RDM in the organization
- User groups of RDM-like approaches
- Implementation of RDM solution
- Other topics

The outline of questions was presented to the consortium when introducing the RDM survey activity to the iDev40 partners at the M36 general assembly meeting. It provided the basis for choosing suitable experts to be interviewed. Experts in data management fill different positions in the hierarchy of the organizations taking part in the survey, as became clear by preliminary talks. The choice of the experts to be interviewed was left to the organizations involved. In total, seven interviews were conducted within WP5 with the following voluntarily participating iDev40 partners:

- KAI Kompetenzzentrum Automobil- und Industrieelektronik GmbH
- Elmos Semiconductor SE
- Infineon Technologies AG (Austria, Dresden)
- Infineon Technologies AG (Munich)

Figure 5: Overview of important RDM-related aspects identified by the survey

- AVL List GmbH
- JEMA Energy SA
- Fraunhofer Institute for Integrated Systems and Device Technology IISB

3.2 Interview method

Due to the Corona pandemic, no personal and more exhaustive interviews were possible. Therefore, the RDM-team conducted the interviews with the other project partners as online meetings. Each meeting had one moderator. After an initial introduction to the RDM topic and its context, an open question & answer session commenced, leaving room for feedback and remarks of the experts. Moreover, all interviewers, typically two to three, asked questions in a loose order. The purpose was to try to achieve comparable information content in all interviews. Next to the prepared questions open questions resulting from the conversation were asked. The interviewers thus used the questionnaire as a rough outline but did not necessarily follow it question by question. All interviewers took notes about the answers given by the interviewees. After the meetings, the team members consolidated their notes into final answers to the questions of the guiding questionnaire. After an internal review, the respective contact person from the RDM-team asked the interviewees to review, amend and approve the filled questionnaire. The filled questionnaire was sent to the interviewees to release them for analysis. All filled, consolidated, and revised questionnaires became part of the analysis.

3.3 Analysis method

The Interviews got analyzed in two qualitative ways. Firstly, a word cloud got generated visualizing frequently used words in a larger font. As a result, one gets a rough visual overview of the major topics in connection with research data management (see Figure 5 in "Survey Results").

Secondly, a structured analysis aims to gain insight into the commonalities and differences of data management in the interviewed organizations by complete indexing of the interview contents. Therefore, the RDM-team applied the principles of qualitative content analyses introduced by

P. Mayring [15]. For these, one can fall back to the structured Q&A documentation mentioned above. It is based on the same questions in all cases and therefore allows the comparison of different answers. However, similar issues were touched on in several questions. Therefore, it was necessary to resort to using categories for interview analysis, as Mayring proposes. Some of the interview documentation introduced newly emerging content. The categories for analysis got formed inductively instead of just sticking to the ones used for grouping the questions. By forming related subcategories, these categories got enriched. Afterward, text modules from the answers were assigned to them. This assignment is often based on so-called code words or modules. Text coding and respective paraphrasing are the basis for the comparative analysis of the interviews because this approach enables analysis in a more generalized form. The answers in the individual subcategories were presented next to each other in tabular form to compare and interpret them. Based on that arrangement, the derivation and formulation of the overall statement per topic and category got performed.

Due to the open interviews, it was impossible to prevent multiple assignments of the answers to several categories in all cases. However, the interviewers rate this as uncritical in this case. It is targeted neither to present a complete view on research data management in industry nor to perform any quantitative analysis. In this case of coded assessment of the situation of RDM at iDev40 partners, some blurring of the categories is therefore tolerable. Nevertheless, maintaing a sufficient degree of the systematic and rule-governed procedure could be achieved. The performed analysis ensured transparency, comprehensibility, and validity of the survey results.

4 Survey Results

Figure 5 provides a rough, visual evaluation of the interviews by a wordcloud generated from the transcribed content of all filled questionnaires. It visualizes the frequency of words mentioned during the interviews. In a nutshell, the interviews concerning Research Data Management (RDM) delivered a broad spectrum of results in several domains of RDM. While the value and awareness of internal scientific data management are well established, the adoption is partly lacking due to various reasons and concerns. For industrial partners, there are concerns regarding IP protection and data security internally as well as externally. The historic silos inside the organizations still can be found and protection against members of other silos still exists. For external sharing, there are even more IP-related concerns. For most RDM domains, including training & guidelines, establishing infrastructure, etc., there is room for improving the current practices. Starting points for improvement were also given in the interviews: establishing proper templates, guidelines, and training for data collection, analysis, and sharing. Beyond that, improving the practices requires a cultural shift in many of the interviewed organizations. In many areas, tools that are more suitable would improve the research data infrastructure. Although most respondents acknowledge the necessity for sharing data, information, and knowledge, infrastructure limitations prevent proper adoption and execution. The following subsections describe the summarized results of the study in more detail. In doing so, it addresses the RDM domains mentioned in Figure 4.

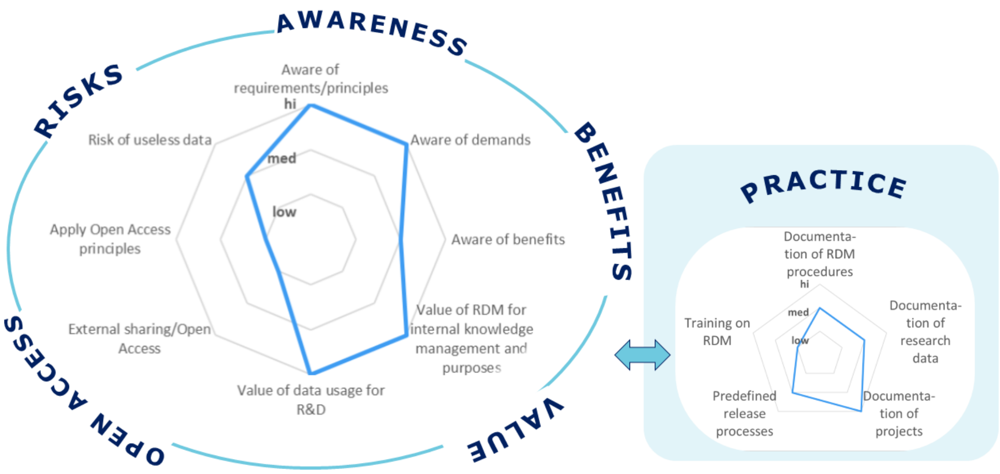


Figure 6: Summary about awareness, value and openness in relation to the current practices

4.1 Results concerning RDM Awareness, Value of (Open) Data & Targets to be achieved

A summary of the findings in this area is presented on the left side of Figure 6.

High, and even increasing, awareness about:

- Open Access requirements
- The FAIR principles
- Benefits of extended data usage and structured data management

Value of internally collecting, analyzing, and (re)using data is well established.

The value of formal lessons learned is known, but some partners still perform knowledge management on an informal basis.

There is still improvement potential with the integration of data silos exists.

There is a mixed picture concerning awareness of the benefits and value of holistic/structured RDM.

- Some partners are even happy to spend extra effort for, e.g., meta-data enrichment.
- Some partners recon this data as a foundation for AI-enabled business processes and autonomous manufacturing, data as a business (model) enabler.
- Others consider it an overhead.

Openness concerning externally sharing scientific data

- Well adopted for academic and applied research: Applying Open Access principles is only valued by academic entities.
- The industry is very protective (partly even completely forbidden) concerning data sharing due to concerns about IP-protection.

Risks of the uselessness of data are only partly recognized; Awareness of issues like Dark Data, Data Quality (VDI 3714-2) only limited.

4.2 Results concerning Persons Involved in RDM in the organizations

- Breath of stakeholder involvement: partners are well aware of the necessity, and broad contribution is established, partly even via specialized roles and teams.
- Consents: When set up consistently and executed diligently, the benefits are evenly spread through the departments. In general, there is a trend toward more open and broader involvement; opening up for external stakeholders slowly as well.
- Change towards a sharing culture is in progress but still protracted, the process needs to take the concerns of all stakeholders and the transfer of associated business models into account.

4.3 Results concerning Trainings and Guidelines

Guidance for documentation and practice (as summarized on the right side of Figure 6):

- Documentation of projects, lessons learned, etc. well established, however, formalization differs from organization to organization.
- Documentation of RDM procedures, tools, etc. varies broadly from details of how to retrieve data to almost no guidelines.
- Documentation of research data (i.e., data itself and data management procedures) is reasonably widespread but with a broad spectrum of structurization (from semantic web modeling until informal collection of files on a server)

Following predefined release procedures is found in most organizations; still, some follow more ad-hoc approaches

Training:

- Formal RDM training can hardly be found; mostly training on the job (demand- and application-oriented) or training on how to publish according to Open Science.
- Formalized training and learning processes are a major concern for the future of most organizations.

4.4 Results concerning the Way of Implementation and the Tools

Data storage and management:

- Mostly traditional storage mechanisms (file servers, Sharepoint servers, relational databases, etc.) prevail.
- Only a few organizations adopted data lakes etc.
- Some organizations are working with semantic networks and exhaustive meta-data
- Version control is established in most cases.

- Although the value is widely accepted, structure RDM and data curation is not a priority.

The tool landscape is mostly a zoo of partly homegrown, MS Office, etc. tools; seldom adoption of a tool from major suppliers, and if, then by large companies.

Data security and IP protection are a paramount priority to all organizations.

Externalizing data in terms of Open Access is not a priority and is seen skeptically because of IP protection concerns, however, improvement activities are ongoing.

5 Conclusions

Beyond the insight into the current state of affairs in the interviewed organizations, one major finding from the interviews is that several organizations requested that funding agencies and legislation should define the requirements more clearly and strictly. They ask for guidelines on how collaboration, publication, and IP protection should work together. They ask for more crisp definitions and execution guidelines concerning Open Data, Open Source, Open Access, and public sharing of scientific data.

6 Acknowledgements

This project has received funding from the ECSEL Joint Undertaking under grant agreement No 783163. The JU receives support from the European Union’s Horizon 2020 research and innovation programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania.

7 Roles and contributions

Dirk Ortloff: Formal Analysis, Writing – review & editing

Sabrina Anger: Conceptualization, Writing – original draft

Martin Schellenberger: Formal Analysis, Writing – review & editing

References

[1] E. E. project iDev40. “Integrated development 4.0.” (2023), [Online]. Available: http://www.idev40.eu.

[2] B. des Inneren und für Heimat. “Open data.” (2021), [Online]. Available: https://www.bmi.bund.de/DE/themen/moderne-verwaltung/open-government/open-data/open-data-node.html.

[3] E. Commision. “Open data guidelines of the european union (eu 2020 - 2023).” (2020), [Online]. Available: https://commission.europa.eu/news/european-commission-adopts-its-new-open-source-software-strategy-2020-2023-2020-10-21_en.

[4] E. Commision. “Horizon europe.” (2020), [Online]. Available: https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en.

[5] B. für Bildung und Forschung. “Forschungsdatenmanagement.” (2023), [Online]. Available: https://www.bildung-forschung.digital/digitalezukunft/de/wissen/forschungsdaten/forschungsdatenmanagement-fuer-freie-nutzbarkeit/forschungsdatenmanagement.

[6] OpenAIRE. “How to create a data management plan.” (2023), [Online]. Available: https://www.openaire.eu/how-to-create-a-data-management-plan.

[7] Forschungsdaten.info. “Forschung und daten managen.” (2023), [Online]. Available: https://forschungsdaten.info.

[8] J. Barkai. “Accelerating science-led innovation for competitive advantage.” (2012), [Online]. Available: https://www.tradepub.com/free-offer/idc-manufacturing-insights-accelerating-science-led-innovation-for-competitive-advantage/w_aaaa2562.

[9] K. Hahn, T. Schmidt, M. Mielke, R. Bruck, and D. Ortloff, “Micro and nano product engineering using data management for silicon-based fabrication process development,” in *Nanotechnology, 2009. IEEE-NANO 2009. 9th IEEE Conference on*, Jul. 2009, pp. 337–340.

[10] D. Ortloff, T. Schmift, K. Hahn, T. Bienek, G. Janczyk, and R. Brück, *MEMS Product Engineering*. Springer-Verlag Wien, May 2014, ISBN: 978-3-7091-0706-5.

[11] N. F. e.V. “Nationale forschungsdaten infrastruktur.” (2023), [Online]. Available: https://www.nfdi.de.

[12] NFDI4ing. “Nfdi4ing – the national research data infrastructure for engineering sciences.” (2023), [Online]. Available: https://nfdi4ing.de.

[13] D. Science, M. Hahnel, L. D. McIntosh, *et al.*, “The state of open data 2020,” Dec. 2020. DOI: 10.6084/m9.figshare.13227875.v2. [Online]. Available: https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2020/13227875.

[14] GoFAIR. “Fair principles.” (2023), [Online]. Available: https://www.go-fair.org/fair-principles/e.

[15] P. Mayring, *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Klagenfurt, 2014, p. 143. [Online]. Available: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173.