ing.grid preprints

RESEARCH ARTICLE

An Empirical Study of the State of Research Data Management in the Semiconductor Manufacturing Industry

An analysis of industry and research institutions in 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. Pi-Fab Mangement Group, Fraunhofer Institute of Integrated Systems and Device Technology IISB, Erlangen.
3. Intelligent Energy Systems, Fraunhofer Institute of Integrated Systems and Device Technology 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 [1]. 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 [2]) 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 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

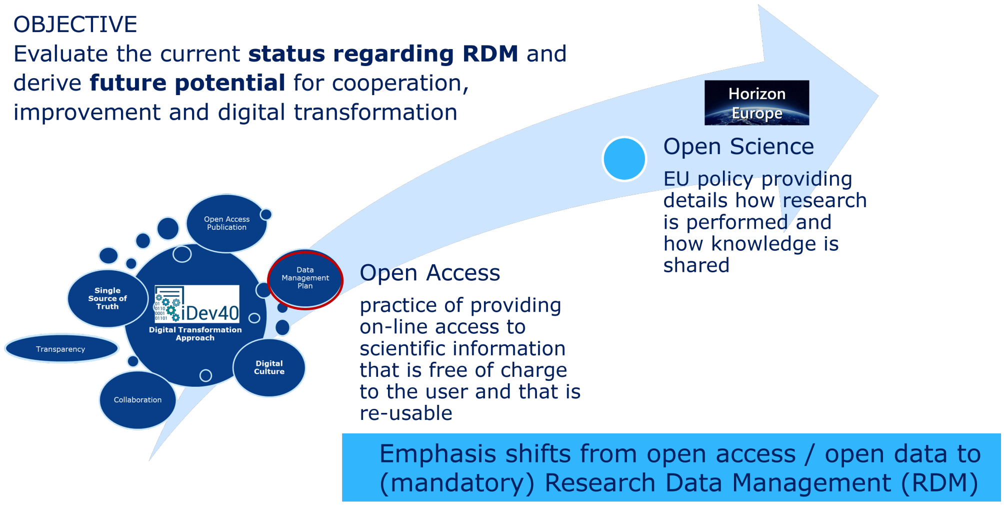


Figure 1: Objectives of the RDM interviews

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 collectively to volunteer to participate in these interviews. Additionally, the RDM-team contacted several partners of the project 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 [3].

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 [4]. Furthermore, Horizon Europe [5] 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 [6]. 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 ([7]; [8]).

For several industries, the research community and companies even have published the ideas, approaches, and benefits more than a decade ago (see, e.g., [9]; [10]; [11]), 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

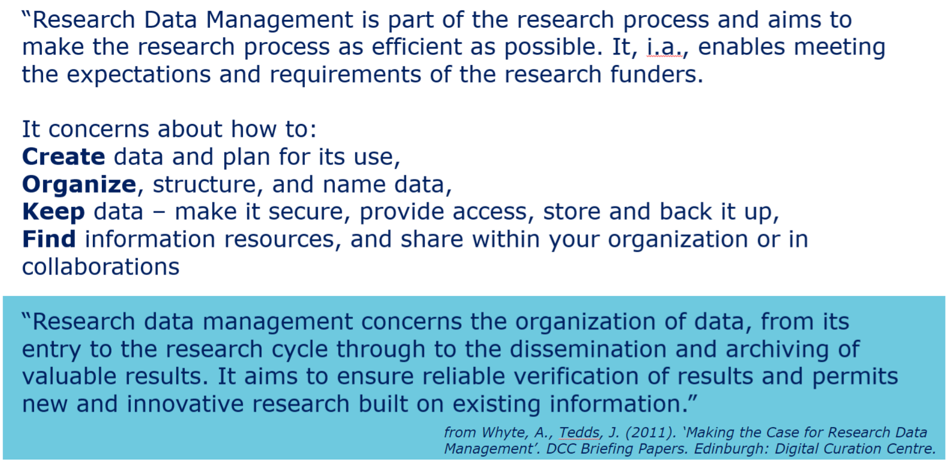


Figure 2: The intents and benefits of RDM

and infrastructure network in Germany [12]. 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 [13]. 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. 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, [14]), 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 2). 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 3.

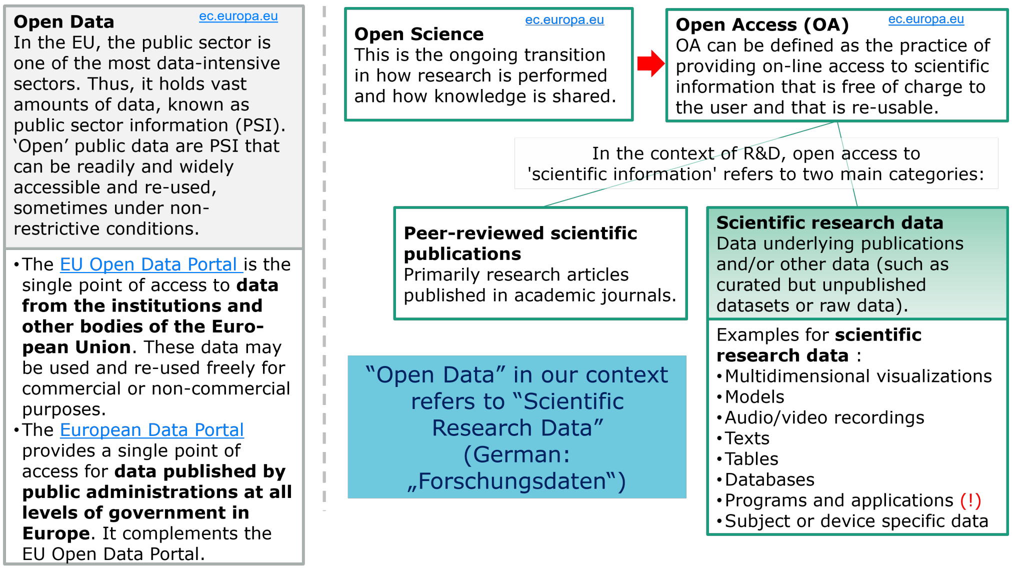
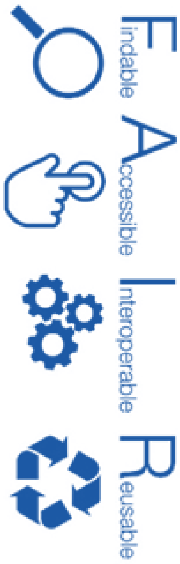


Figure 3: The context of Open Data, Open Science and Open Access

RDM is more than pure data management (see Figure 3). 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
- 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 [15]:



- 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 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 study considering the state of open data [4], but questions and content have been modified with reference to industry-related aspects and have been enriched and supplemented by the RDM team’s 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:

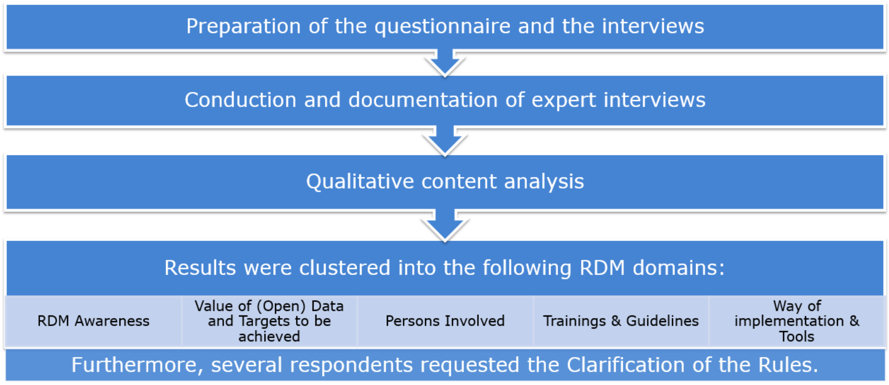


Figure 4: Methodology of the survey

- 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
  KAI is an industrial research center having proven experience in the coordination of interdisciplinary research projects. Its core competences are in the area of power electronics reliability test concept development, advanced semiconductor materials research, statistical lifetime modeling and multi-physics FEM simulation. https://www.k-ai.at/
- Elmos Semiconductor SE
  Elmos develops, manufactures and sells semiconductors primarily for use in cars, with this components being relevant for communication, measurements, regulation, and control of safety, comfort, drive, and network functions. https://www.elmos.com/
- Infineon Technologies AG (Austria, Dresden)
- Infineon Technologies AG (Munich)
  Infineon is a leading global supplier of semiconductor solutions for power systems and IoT. Infineon solutions come into application in the fields of automotive, industrial power

and energy, sensors systems and connected secure systems. At the Dresden site, Infineon operates chip production on 200mm silicon wafers as well as the world's first high-volume fab for power semiconductors on 300 mm silicon wafers. At the Villach site in Austria, main activities are in wide bandgap semiconductor technologies for small and energy efficient devices. At Infineon's largest research and development site in Munich, activities focus in particular on technological integration, chip card ICs, power electronics for automotive and industrial applications, and the development of production processes. https://www.infineon.com/cms/de/

- AVL List GmbH
  AVL List is one of the world's leading mobility technology companies for development, simulation, testing and trial in the automotive sector and other industries. https://www.avl.com/en-de
- JEMA Energy SA
  JEMA Energy located in Spain is a benchmark company in the energy sector offering design and manufacturing of energy conversion systems. JEMA is active in energy storage, electromobility, hydrogen, solar energy and nuclear fusion. https://www.jemaenergy.com/en/jema-en/
- Fraunhofer Institute for Integrated Systems and Device Technology IISB
  As part of the Fraunhofer-Gesellschaft, the Fraunhofer Institute for Integrated Systems and Device Technology IISB conducts applied research and development in the field of electronic systems for application in, e.g., electric mobility, aerospace, Industry 4.0, power grids or energy technology. With its two business areas, semiconductors and power electronics, the institute uniquely covers the entire value chain - from basic material to whole power electronic systems. https://www.iisb.fraunhofer.de/

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 which is why it is not expedient to present the complete questionnaire here. 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.

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

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 [1]. 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 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 several 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.

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

Most of the interviewed partners are aware of the **Open Access requirements and the FAIR principles**. Most of the partners are strongly aware of the **benefits provided by extended data usage and the respective demands** for structured data management in the organization. At this point, the awareness even increases at all organization hierarchy levels. This rising awareness paves the way towards more and more structured RDM. However, awareness and the respective judgment of the benefits on the different hierarchy levels differ from organization to organization. In most cases, employees in research and development are aware of the value of data and the necessity for data management and analysis. In principle, most of the partners are familiar with the FAIR principles and intuitively follow them in their own data management. However, related formal business processes are available in a minority of cases. Almost all organizations participating in the survey have internal data experts. On the other hand, data management experts are mostly only present in larger organizations.

It is important to note that the **value of structured RDM** is well-established for internal knowledge management. Concerning **external sharing and complying with Open Access requirements**, there is a broad spectrum of adoption. Academic and applied research-focused organizations quite openly adopt Open Access paradigms, while the industry has high doubts regarding confidentiality. The latter might correlate with the industrial environment and the tradition of the interviewed partners. Anyway, it is a crucial point to consider and address.

A heterogeneous picture evolved when the **awareness of the benefits generated by holistic data management / RDM** was evaluated. Some organizations fully acknowledge the benefits of RDM and are leveraging them, despite the extra efforts for, e.g., meta-data enrichment, data consolidation, and integration. These organizations consider the extra efforts as a central vehicle to reduce manual labor, improve knowledge management and quality, and for being able to reuse data. Additionally, they recon this data as a foundation for AI-enabled business processes, autonomous manufacturing, etc. On the other hand, some entities consider this type of holistic RDM activities and efforts mostly overhead.

The **value of intensively collecting, analyzing, and (re)using data for internal purposes** is well-established in almost all interviewed organizations. They all try to apply learnings derived from the collected data to improve new products or revisions thereof. Knowledge sharing, also between different departments, is the current state of practice in most organizations. For some, it is part of a structured RDM, and lessons learned process, for a minority still information exchange on an informal basis. Depending on the interviewed organization, the awareness of the power of integrating various data pools to find new correlations and knowledge varied. Here is certainly potential for improvement in some entities. Most respondents have concrete ideas of the value the data itself and data management have for them internally. Data is the basis for respective development activities carried out by the interviewed research organizations. This is why the availability of high qualitative data is decisive for these organizations. However, data is no less valuable for the companies investigated in the expert interviews. Here data is stated as the basis for development activities and learnings in manufacturing. Data availability is the prerequisite to enable retrospection in production. The value added is driven by understanding and interpretation of rich datasets. Based on data and respective data analyses, organizations can identify manufacturing issues and support the application of machine learning. The application of AI might reveal additional, formerly hidden information. Using the data exhaustively can facilitate failure understanding and root cause analysis. This can prevent making the same mistake twice and contribute to solving problems. Thus, the proper use of data, in general, will improve products and reduce the overall development effort.

When investigating the **openness to apply external Open Access principles**, the RDM team found a slightly different perspective. As indicated in the previous paragraph about Awareness, the more industry-centric an organization is, the more difficulties exist in externally sharing research data. In extreme cases, it is more or less forbidden to share any data, even with collaborating partners. Here guidelines foresee only sharing data sheets and such. In cases where no other option exists, sharing of old data or anonymized data was the maximum allowed. In contrast to the prevalent value add of sharing data within an organization, the value of shared data is not that obvious in the case of open (research) data. Open Access data is of value for research organizations because, on the one hand, it enables collaboration and probably also new business models. On the other hand, it improves the attractiveness of the research organization in hiring and training students and by upgrading its publications. For enterprises, there is a low value in openly sharing data. This is due to the risk of IP loss in favor of a direct competitor. Additionally, there is the risk of violating confidentiality and compliance. Sharing data with partners must build on confidence and a contractual basis for all interviewed partners. In extreme cases, it is more or less forbidden to share any data, even with collaborating partners. Here guidelines foresee only sharing data sheets and such. In cases where no other option exists, sharing of old data or anonymized data was the maximum allowed.

Per the description above, the **value of intensive data usage for improved R&D** is a well-established practice. Data management is the key to using data beneficially. If implemented

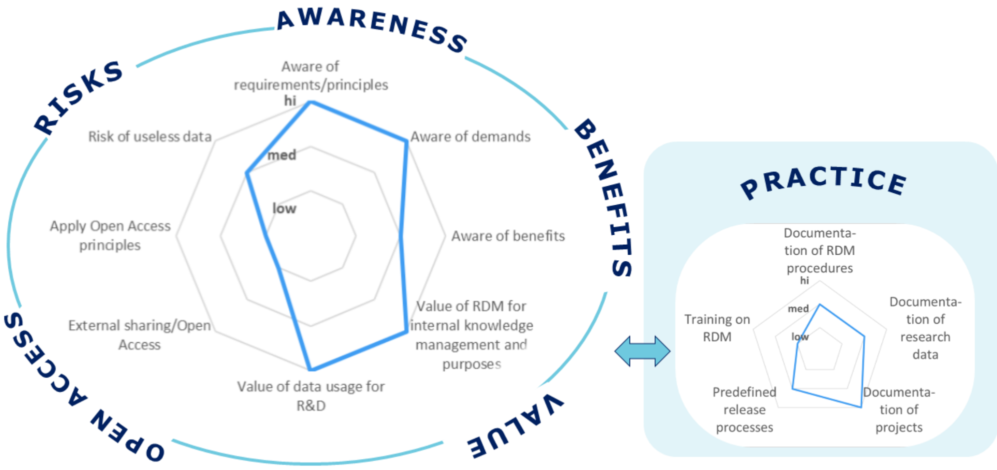


Figure 6: Summary about awareness, value and openness in relation to the current practices

according to the ideas of the respondents, it enables the availability, findability, understandability, and reusability of data for various purposes and across the whole organization. Structurization, harmonization, and related documentation of context information reduce the manual efforts for data access and use. Proper data management practices link data, merge information, and safeguard knowledge. The complete and long-term traceability even throughout the whole value chain can be assured. Building on respective standards, sharing data is simplified thus, fostering collaboration even in cases of independent task processing. A suitable data management solution provides all levels of information, i.e., data, context, and documentation, in a user-friendly way. Proper data management practices avoid non-productive steps. They do not generate additional effort for the provision, retrieval, and use of data. Supporting information for everyday work is just available on demand.

However, a **risk of the uselessness of data** may remain despite correct data annotation according to the responses. Here Dark Data and data quality aspects (for more details, please refer to, e.g., VDI 3714-2) are significant sources of uncertainty and risk. Nevertheless, especially large enterprises are convinced that not using data will be at least a competitive disadvantage in the future. Depending on the interviewed organization, the awareness of the power of integrating various data pools to find new correlations and knowledge varied. Here, there is certainly some potential for improvement in some entities.

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

4.2 Results concerning Persons Involved in RDM in the organizations

The interviewers found a coherent picture when investigating the breadth of **stakeholders participating in the data usage, management, and analytics efforts**. In all organizations, various departments and stakeholders contribute to RDM-related efforts. In some entities, specialized roles like data curators, semantic web teams, etc., become involved in the efforts. Accordingly, the interviewees considered the benefits widely and evenly spread through the departments and roles of the organization. However, only if the RDM efforts are set up consistently across the entity and executed diligently.

On the flip side, the breadth of people and roles involved in data collection, management, and usage sparks concerns about data security. Tight data access control and security are critical tasks for most organizations, internally and even more so externally.

Different stakeholders initiate the **transition towards Open Science or Open Access**. The triggering stakeholder depends on the organization interviewed. For some enterprises, the quality departments or the customers require more research data management. For more research-oriented organizations, it is the funding authority demanding data management plans and Open Access.

One major roadblock in establishing beneficial RDM practices and tools is the **change toward a sharing culture**. All participating organizations mentioned this issue. Transforming the mindset of people towards “1+1=3” in terms of sharing knowledge is a stony road and the most critical success factor in transitioning towards Open Science.

One central aspect of this transition is to involve all stakeholders very early in the process. Furthermore, it is essential not to fall into the trap of digitalizing existing business processes and procedures one to one. Instead, assessing and reworking them carefully while digitizing them is paramount.

4.3 Results concerning Trainings and Guidelines

The interviews revealed that the extent of formalized RDM is different in the questioned organizations as it is briefly summarized on the right side of Figure 6 Firstly, there are various focuses on the relevance, nature, and comprehensiveness of documentation. The interviews discovered a broad spectrum concerning training, documentation, and guidelines, similar to previous categories. **For the documentation of RDM requirements, tools, and procedures**, some organizations have exhaustive documentation. Some organizations even document how to perform information retrieval, execute lessons learned workshops, document the research data, etc. At the other end of the spectrum, some academic and industry participants are rather informal about RDM documentation.

Concerning the **documentation of research data** itself, the way of documentation spans the whole range of rudimentary textual documentation of files stored on a server via more or less thoroughly described and documented research data and code stored on file servers, SharePoint, or in databases towards an advanced data management solution based on fixed principles. The latter is, e.g., guided by a defined processor it is a semantic representation based on master data management guidelines and knowledge graphs on the meta-level according to the interviews. Although partly under construction, there already are or will be comprehensive guidelines for data management, e.g. for semantic web modeling. In other cases, data description is at least in parts based on templates, whereas formal guidance is completely lacking in the remaining ones.

Documentation for development activities and projects is well established in all cases. Project information gets linked with context information, where and how to find data associated. Lessons learned are well documented, especially in design/redesign processes, where they are decisive. For overarching aspects, they are partly in a more individual form. However, a formal lesson learned business process is established only in one case. Formal support for operational decisions is more often available. However, the interviewed partners are aware of the benefit provided by experience gained in the past. That is why efforts have there already started toward the interpretation of textual information. The text is read and interpreted by machines from lessons learned documents.

Predefined release processes are established for those partners who provide Open Access to their data or publish on Open Access platforms. It is interesting to note that the volume and type of documentation practices do not always correlate with the amount of structure given by templates and guidelines. Some organizations do quite exhaustive documentation while only working with a limited set of templates and guidelines.

Secondly, **training on RDM is organized differently**. At the moment, strictly speaking, training activities do not focus on research data management and effective learning processes in most cases. In individual cases, master data and semantic web courses exist. Furthermore, online training for scientific publishing in Open Access, including publication of related data, is established. Nevertheless, most partners indicated that pieces of training get individually provided as needed and under other aspects. There are demand- and application-oriented internal classes for software and tools available in all cases. Additional training needs are met by personal advice and more or less frequent exchange meetings and workshops. One partner stated that it would be desirable to use existing templates and solutions more consistently and deepen the exchange of best practices even within the organization. All partners agree that empowering employees to work with data effortlessly has to occur soon. Training and learning processes need to be formalized, structured, and supported by guidelines or even machines. Data management and documentation standards are to be created within the organizations, at least. Additionally, these should be applicable for data exchange. Furthermore, optimal collaboration and exchange of experiences between employees is vital and needs to be promoted.

In addition to their efforts, most of the partners moreover wish for external supportive expertise and guidelines in the realm of RDM.

4.4 Results concerning the Way of Implementation and the Tools

For the **implementation of data storage and management**, the traditional storage mechanisms still prevail. Most interviewed organizations rely upon predefined structures on file servers, SharePoint servers, and relational databases. Therefore, predominantly heterogeneous, siloed storage structures are in use while only a minority of organizations adopted more state-of-the-art approaches like data lakes. Despite that, most of the interviewed organizations adopted some level of version control for the data sets to achieve a certain level of traceability and audibility. Semantic enrichment of the data to foster reuse is not yet widely adopted. Concerning standardizing and regulating storage structures, the resulting picture is similar to the storage approach. On one side of the spectrum, organizations use guidelines for organizing and documenting all R&D data in semantic networks, strict templates & structures. On the other side, the interviews uncovered informal ways of relying on the best guess and attitude of the researchers, partly driven by customers. One of the findings of the interviews is that only a minority of organizations have dedicated roles like data curators. Systematically addressing the RDM principles seems not to be a priority, despite the broadly shared acceptance of the paramount value of data and its management described above. But, the FAIR principles seem to be acknowledged and mostly followed.

When investigating the **tool landscape used for internal storage, retrieval, and analysis of data**, the landscape is very diverse as well. More advanced, and typically larger, organizations use integrated storage & analytics solutions of major software providers. Some are already using Ontologies, Artificial Intelligence, and Machine Learning solutions in production. Small and medium organizations often rely on MS Office tools and partly informal knowledge exchange in the coffee corner. Anyhow, the importance of data analysis, following different levels of sophistication, seems to be broadly accepted and established across all interviewed iDev40 partners.

The paramount concern for most organizations is **data security and access control**. R&D data is the core of intellectual property. Therefore, fine-grained control of who has access to which data and to what extent is a central concern. As a result, most organizations manage this via a centralized group and have very high barriers granting access, especially for external people/organizations. For collaborating with external partners, typically specific contracts get negotiated to regulate access, sharing, and IP on a per-project basis.

Externalizing data in terms of Open Access is limited due to often unsolved IP protection issues. Some more academic-oriented institutes have established structures to publish parts of the data on open platforms like Zenodo or have established their open platform following the FAIR principles. If data gets published on such platforms, mostly anonymized or older data is published. That is again due to confidentially concerns. The good news is that most interviewed organizations have actively run programs and projects to extend the possibilities for more open and intensified collaboration. These development projects range from building structured templates for consistently applying the FAIR principles to Natural Language Processing analysis of thesis, 8D-reports, or FMEA studies. These improvement projects aim to facilitate data, information, and knowledge sharing and reuse by integrating data silos. The overall targets are to improve Open Access and FAIR practices and semantic enrichment of data.

4.5 Additional findings from the interviews

As mentioned above, the interviewed iDev40 consortium members are fully aware of the requirements regarding data management imposed on them from outside. May they be imposed by either their customers or by funding agencies. However, it also became quite prevalent in the interviews that there are still some open questions to be clarified and some requirements to be fulfilled concerning the provision of data in general and open research data in particular. It was clearly stated that it is not enough to address respective issues internally and that it is especially not possible to solve the open questions within the organization only. One major finding from the interviews is that several organizations requested that funding agencies and legislation should define clearer and more strict requirements. 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.

5 Conclusions

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 is still existent. 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, more suitable tools 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.

One additional major finding from the interviews is that several organizations requested that funding agencies and legislation should define clearer and more strict requirements because of open data policies.

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] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.

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

[11] 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.

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

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

[14] 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.

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