

# A Conceptualization of Research Data in the Context of Industry-Academia Collaborations

Whitepaper on behalf of the Section Industry Engagement  
of the National Research Data Infrastructure (NFDI)

written by

**Florian Stahl**, University of Mannheim

**Kai Hoff**, Stifterverband

**Andreas Hamann**, University of Mannheim

in February 2025

**Additional Contributors:** Christian Busse (DKFZ Heidelberg), Barbara Ebert (German Federation for Biological Data), Juliane Fluck (ZB MED), Andreas Förster (DECHEMA), Georg Rehm (DFKI), Bernard Seeger (University of Marburg), Sylvia Thun (Berlin Institute of Health)

## Contact Persons:

Prof. Dr. Florian Stahl [Spokesperson of Section Industry Engagement]  
Email: [florian.stahl@uni-mannheim.de](mailto:florian.stahl@uni-mannheim.de)

Prof. Dr. Chris Eberl [Co-Spokesperson of Section Industry Engagement]  
Email: [chris.eberl@iwm.fraunhofer.de](mailto:chris.eberl@iwm.fraunhofer.de)

# Table of Contents

Executive Summary .....	III
1 Introduction .....	1
2 Potential Scope of Research Data .....	2
2.1 Existing Perspectives on the Term “Research Data” .....	2
2.2 Implications for Data in Academia and Companies.....	3
2.2.1 Primary Data .....	3
2.2.2 Secondary Data.....	4
3 Factors Influencing and Limiting the Suitability of Data as Research Data in Different Research Disciplines.....	5
3.1 Legal Restrictions.....	5
3.2 Disciplinary Relevance of Data .....	5
3.3 Data Quality .....	7
4 Fostering Industry-Academia Collaborations for Research Data Sharing.....	8
4.1 Supporting Services by the NFDI and its Section Industry Engagement .....	8
4.2 Clearly Defined Data Governance Processes.....	9
4.3 Data Exchange Through Data Spaces .....	10
4.4 Using Generative AI for Data Sharing .....	11
Appendix.....	12
References .....	14

# Executive Summary

Vast knowledge generated in industry and academia is based on the analysis of data. Therefore, access to data is fundamental to creating new insights and knowledge. While many public organizations, private companies, and academic institutions possess relevant data, the uncertainty about what data researchers in industry and academia require is a major obstacle in sharing this data to allow for even more insights. As various conceptualizations of research data exist, implications for whether and when data can be considered as “research data” in the context of industry-academia collaborations are fuzzy and unclear. Therefore, this whitepaper discusses existing academic, industry, and corporate perspectives on what the concept of research data in the context of industry-academia collaboration encompasses<sup>1</sup>.

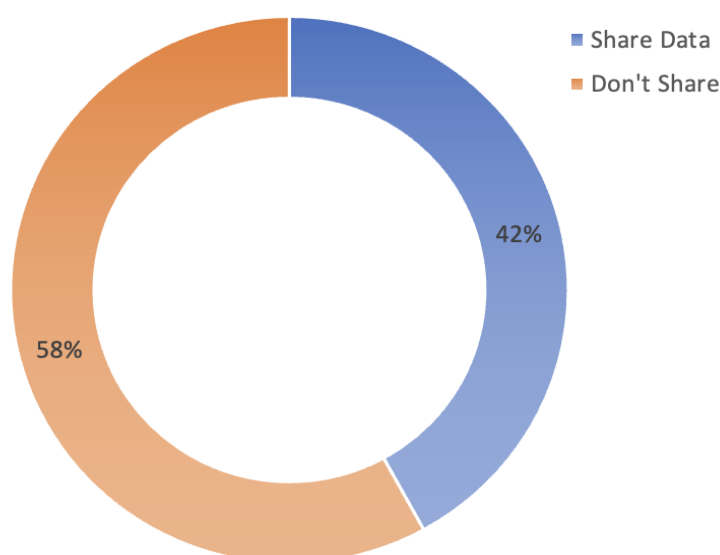
We show that, theoretically, any type of data can become research data. However, in practice, several factors can affect the usability, relevance, and value of data for research. These factors include legal obstacles, disciplinary differences in research relevance of data, and insufficient data quality that can reduce the usability and relevance of data for research, ultimately hindering the possibility of turning data from industry-academia collaborations into research data. We show that several services provided by the NFDI and its Section Industry Engagement, data governance best practices, recent generative AI developments, and currently developed data spaces exist to foster the sharing of relevant research data between industry and academia.

---

<sup>1</sup> We do not aim to create a new definition of the term “research data” here. Instead, this paper merely summarizes how research data has been defined in the past and, based on these different perspectives, infers what research data in the context of industry-academia can all encompass.

# 1 Introduction

Access to research data is fundamental for academic and corporate (R&D) research. The nature of research data goes beyond mere information; it embodies the empirical basis on which knowledge is built, theories are formulated, and insights are derived. Nowadays, digital technologies facilitate the rapid generation and dissemination of data in various fields, especially of unstructured data like text, images, video, and audio. These data types are increasingly relevant for researchers to generate new research insights and knowledge. Academic institutions, companies, or organizations that usually own the data should increasingly collaborate to remain internationally relevant. Currently, the potential of using company data for such research purposes is far from exhausted, as less than half (42%) of companies release data for research purposes (Büchel and Engels 2023; see Figure 1).



**Figure 1:** Share of Companies Engaging in Data Sharing in 2022 (Büchel and Engels 2023)

Given the vast amount of data academic institutions, companies, and public organizations possess, a reason for the lack of data sharing may be the uncertainty of which of this data is relevant for academic and corporate research. This concern is illustrated by the fact that various definitions of “research data” currently exist (e.g., DFG 2015, 2023; EU Directive 2019/1024). A consistent understanding of research data can help overcome initial barriers to data exchange and sharing, both for scientists and industry stakeholders.

This whitepaper summarizes predominant conceptualizations of the term “research data”, with a particular focus on the context of industry-academia collaborations. We aim to foster a clear understanding of different terminologies of research data and infer in how far data can turn into research data. Thus, we will first provide an overview of currently existing different perspectives on the scope of the term “research data”. Based on this understanding, we will discuss what research data in the context of industry-academia collaborations could all encompass and ultimately describe factors that determine whether data can become research data.

## 2 Potential Scope of Research Data

### 2.1 Existing Perspectives on the Term “Research Data”

To discuss the concept of research data in industry-academia collaborations, we first outline what “research data” generally encompasses. Different perspectives and scopes of the term exist.

For instance, the German Research Foundation (DFG 2023) defines research data as:

*“All data created, collected, and processed during the research process.”*

In addition, the German Data Usage Act specifies in §3 that:

*“Research data are records in digital form, other than scientific publications, that are collected or generated in the course of scientific research activities and used as evidence in the research process or that are generally considered necessary in the research community for the validation of research findings and results.”*

Thus, these two definitions emphasize the use of any data during the research process, and they are used as proof to justify research results.

The German Council for Scientific Information Infrastructures (2017) further elaborates on these conceptualizations by explicitly pointing out that research data is not just the data created during a scientific inquiry but also any other type of data used during research:

*“Data not obtained through direct scientific activity but that is used by science for the purpose of research to form the methodological foundation of the specific research process is also research data. This is the case, for example, when official statistics or other data from public authorities or products from non-scientific service providers are processed scientifically. That research data also includes the research tools used as well as the traces of scientific activity continuously generated.”*

Other conceptualizations of research data rather rely on an enumeration of potential data types for research data. For instance, the EU Directive 2019/1024 (European Union 2019) specifies:

*“Research data includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. It also includes meta data, specifications and other digital objects.”*

Relatedly, DFG (2015) does not only label “measurement data, laboratory values, audiovisual information, texts, survey data, objects from collections or samples” but also “methodological test procedures like questionnaires, software, and simulations” as research data.

Finally, DFG (2023) differentiates between primary and secondary research data. Primary data are raw data directly generated through an experiment or simulation or through program codes for the specific

research endeavor. In contrast, secondary data are created through the continued processing of primary data. Moreover, metadata are data that allow the description, management, and classification of data and are, thus, also closely related to the concept of research data.

Overall, the usage of data during the research process appears to be crucial for data to be called research data. As various examples of research data exist, it appears that basically any data used during a research project can be labeled as research data.

## 2.2 Implications for Data in Academia and Companies

Based on the latest research data definition of the German Research Foundation (DFG 2023), which considers all collected, generated, and used data during the research process as research data, we infer that basically any corporate data used during a research project can be considered research data. There are several ways in which data in academia and industries can become part of a research project – either through primary creation directly for the research project or by sharing secondary, already existing data that is potentially pre-processed and then reused for the specified research analysis. Leaning on the differentiation of primary and secondary research data by DFG (2023), in the following, we further specify how primary and secondary research data can be generated in the context of industry-academia collaborations.

### 2.2.1 Primary Data

The generation of new research data involves the collection of new data explicitly for the given research project. Several possibilities to create new research data exist, for instance:

- 1) **Observational methods:** Researchers systematically observe and record natural phenomena, behaviors, or events in a collaboration with an academic institution or a company. This could be an observation of internal machine- or human-related processes, for instance.
- 2) **Experiments:** Experimental designs manipulate independent variables to assess their effects on dependent variables under controlled conditions. Researchers can use randomized controlled trials, quasi-experimental designs, and factorial designs to evaluate treatment effects in laboratory or field settings and establish causal relationships. In companies, for instance, A/B tests could be run on a website or to improve a corporate process.
- 3) **Surveys and questionnaires:** Surveys and questionnaires are used to collect self-reported data using structured instruments, interviews, or surveys administered to research participants. Survey methods collect quantitative or qualitative data on specific populations' attitudes, beliefs, behaviors, or perceptions, facilitating exploratory research, hypothesis testing, and theory development. Such surveys could be filled out by a company's employees or customers.
- 4) **Simulations:** Researchers can use mathematical models and computer simulations to model complex corporate phenomena, processes, or systems in virtual environments. Techniques such as agent-based modeling or finite element analysis allow the exploration of behavior, predicting system dynamics, and simulating hypothetical scenarios based on corporate information.

Therefore, this definition of primary data overlaps with existing definitions of research data, where new data is created during the research process. The process of obtaining primary data can be carried out

exclusively and solely by academic institutions, by companies, or in cooperation between academic institutions and private organizations or companies.

## 2.2.2 Secondary Data

Data does not necessarily have to be newly created for a research project to become research data but can also be reused from existing sources for such purposes. The data may either be publicly available or proprietarily owned, requiring data sharing agreements.

Publicly available data is primarily aggregate data about corporate performance (e.g., overall firm performance and overall economic data). It can, e.g., be obtained from official statistics, research data centers, research institutes, and databases of public institutions (see Appendix A1 for an overview). Such public data sources are an integral part of academic and corporate research as they provide a wealth of data that enables research projects and investigation, empirical analysis, innovations and theoretical advances in various disciplines.

The latter approach circumvents the need for a direct one-to-one collaboration between industry and academia, as the data is already publicly available for research purposes. However, academic institutions, companies, and public organizations generate a wide range of data in their activities that may not be public (see Figure 2 for an overview of non-public corporate data). Even if this data is not publicly available, it may be relevant for research in multiple fields, and its use in research can also provide reciprocal added value for companies.



Figure 2: Type of Data Collected and Stored in Companies (own illustration)

For instance, material properties data (e.g., tensile strength, hardness, and chemical compound data) may be used for research purposes, aiding in the development of new materials or improving existing ones. Also, financial data (e.g., insights from balance sheets and income statements) provide opportunities for analyzing corporate financial performance and modeling market behavior. Similarly, customer data (e.g., purchase histories and demographic profiles) is essential for studying consumer behavior and evaluating marketing strategies. Employee performance evaluation can similarly be relevant, particularly for academic and corporate research in business and economics. Depending on the industry, however, also other research areas can benefit from, e.g., operational data (e.g., production metrics, logistics information or sustainability and impact management). Moreover, legal data,

including labor contracts and incident reports, helps assess regulatory compliance and analyze associated business risks. Environmental data also plays a critical role in assessing market trends.

Conversely, academic institutions have secondary data on, e.g., research processes or standardized material properties which is highly relevant for various companies in their internal research. However, this data is often difficult for companies to access, due to restrictions on data sharing, lack of standardized access mechanisms, or the proprietary nature of institutional research data.

## 3 Factors Influencing and Limiting the Suitability of Data as Research Data in Different Research Disciplines

As any data can potentially become relevant to academic and corporate research, either now or in the future, all types of data might be considered and be classified as research data. However, several factors limit the actual suitability of data to become relevant to research. In this section, we want to limit and restrict research data to data from different perspectives.

### 3.1 Legal Restrictions

Sensitive or proprietary information in datasets may raise ethical, legal, or privacy issues that limit their accessibility and use in scientific research. For example, proprietary algorithms in technology or customer transaction data in financial institutions may be restricted due to privacy concerns or intellectual property rights. Such data is therefore more difficult to be shared and worked with in research, as strict anonymization procedures must be maintained.

While making research data public has been fostered for many years (e.g., European Union 2019), we argue that consequently not all data could become public after a research project. For instance, making pricing data of a company publicly available would allow competitors to adjust its prices accordingly and potentially collude to achieve higher profits. This would, however, be illegal.

Moreover, the same data types may not be equally shareable across academia and industries as different regulations exist. For instance, the financial and pharmaceutical industries are heavily regulated, which affects their propensity to share data. Consequently, data on innovations might, for instance, be sharable in the telecommunications industry but not in the pharmaceutical one.

### 3.2 Disciplinary Relevance of Data

Not all types of corporate data are relevant for all types of research disciplines (see Table 1). Thus, in individual industry-academia collaborations, some data may be more relevant to researchers than others, depending on the individual collaboration partners and available data.



Area	Type of Data	Collaboration Approach	Research Cooperation Goals
<b>Natural Sciences</b> (e.g., physics, chemistry, biology)	<ul style="list-style-type: none"> <li>• Climate and weather data</li> <li>• Astronomical data</li> <li>• Environmental data</li> <li>• Genetic sequencing data</li> </ul>	<ul style="list-style-type: none"> <li>• Primary data (e.g., experiment, observation)</li> <li>• Secondary data from open databases &amp; proprietary research collaborations (e.g., in a bilateral collaboration)</li> </ul>	Testing hypotheses and theories through empirical observation and experimentation
<b>Social Sciences</b> (e.g., business & economics, sociology, psychology, political science)	<ul style="list-style-type: none"> <li>• Survey data</li> <li>• Social media data</li> <li>• Behavioral data</li> <li>• Economic indicators and financial data</li> </ul>	<ul style="list-style-type: none"> <li>• Primary data (e.g., A/B test, survey of employees &amp; customers)</li> <li>• Secondary data from open databases (e.g., for social media data) and proprietary research collaborations</li> </ul>	Exploring human behavior, social structures, and cultural dynamics
<b>Computer Science and Information Technology</b> (e.g., cybersecurity, AI, computer vision)	<ul style="list-style-type: none"> <li>• Image and video datasets</li> <li>• Network traffic logs</li> <li>• Social media data</li> <li>• Web usage logs</li> </ul>	<ul style="list-style-type: none"> <li>• Secondary data from open databases and proprietary research collaborations</li> </ul>	Advancing technology and user experience through the development of algorithms, systems, and methodologies
<b>Urban Planning and Geography</b> (e.g., mobility, socio-economic inequality, climate change)	<ul style="list-style-type: none"> <li>• GPS tracking data</li> <li>• Mobile phone data</li> <li>• Public transportation records</li> <li>• Satellite imagery</li> </ul>	<ul style="list-style-type: none"> <li>• Secondary data from open mobility databases and proprietary research collaborations</li> </ul>	Understand the dynamics of human environment interactions within urban areas, and developing sustainable solutions for enhancing the resilience of cities and regions
<b>Humanities and Linguistics</b> (e.g., literature, history, linguistics, anthropology)	<ul style="list-style-type: none"> <li>• Textual data</li> <li>• Linguistic corpora</li> <li>• Archeological data</li> <li>• Cultural heritage data</li> <li>• Linguistic survey data</li> </ul>	<ul style="list-style-type: none"> <li>• Secondary data from open linguistic corpora and proprietary data from collaborations with specialized institutions</li> </ul>	Analyzing human expressions, cultural artifacts, and historical narratives
<b>Medical and Health Sciences</b> (e.g., medicine, public health, pharmacology, epidemiology)	<ul style="list-style-type: none"> <li>• Clinical trial data</li> <li>• Epidemiological data</li> <li>• Medical imaging data</li> <li>• Genetic and genomic data</li> <li>• Electronic health records</li> <li>• Medical device data</li> </ul>	<ul style="list-style-type: none"> <li>• Primary data (e.g., clinical trials)</li> <li>• Secondary data from aggregate health statics from open databases and proprietary, anonymized patient data from hospitals or health institutions</li> </ul>	Investigating disease mechanisms, treatment efficacy, and preventive measures Drug development/repurposing, biomarker research
<b>Engineering &amp; Material Science</b> (e.g., mechanical & civil engineering, material science)	<ul style="list-style-type: none"> <li>• Engineering design data</li> <li>• Material properties data</li> <li>• Sensor data</li> <li>• Manufacturing process data</li> </ul>	<ul style="list-style-type: none"> <li>• Primary data (e.g., process &amp; machine observations within a company)</li> <li>• Secondary data (e.g., automatically collected proprietary machine data)</li> </ul>	Designing and innovating engineering systems, materials, and manufacturing processes

**Table 1:** Examples of Disciplinary Perspectives on Research-Relevant Corporate Data (own illustration)

### 3.3 Data Quality

Even if the data is legally allowed to be shared and relevant for the research project at hand, the data itself may not meet research standards and, thus, ultimately not qualify as research data. According to Wang and Strong (1996), data quality refers to the fitness of use by data consumers, that is, how successfully the data serves the goal of the researcher and meets his or her expectations. The level of data quality is thus use case- and consumer-dependent (Tayi and Ballou 1998). Four data quality dimensions have been identified by Wang and Strong (1996):

#### 1. Intrinsic Data Quality

This type describes how far the stored data values represent the real-world ones. It consists of four dimensions – the believability, accuracy, objectivity, and reputation of the data. Therefore, corporate data has not only to be accurate and objectively correct but also credible and reputable from a researcher's point of view. Otherwise, it is less suitable for becoming research data.

#### 2. Contextual Data Quality

The more applicable the data is to the given circumstances of the research user or the individual research collaboration, the higher the contextual data quality is. This concept evaluates data suitability across five dimensions: value-added, relevancy, timeliness, completeness, and data volume, recognizing how data quality varies with users and tasks. For instance, in environmental research, ecological studies require precise measurements of parameters like temperature and humidity, while broader economic analyses may need less detailed data. Hence, corporate data must meet the criteria of value, relevance, timeliness, completeness, and sufficiency for the task.

#### 3. Representational Data Quality

Refers to how clear and readable the respective data is. In detail, it includes the dimensions of interpretability, ease of understanding, representational consistency, and concise representation of the data. Therefore, if the given corporate data is well represented and good to understand, it becomes more relevant for research purposes.

#### 4. Accessibility Data Quality

Addresses to what extent the data is receivable for the researcher. Note that this dimension also intersects with the accessibility dimension of the FAIR concept. The two dimensions of accessibility and access security are included in this data quality category. Therefore, highly qualitative data needs to be inherently valuable, suitable for the given task, well represented, and available to the data consumer (Wang and Strong 1996). E.g., data formats can differ by data type (e.g., text, numerical data, multimedia data), research area, research sub-discipline, and research instrument (see Appendix A2 for an overview). This multitude of standard formats can hinder the seamless integration and exchange of information between companies and researchers. If the data is not accessible to researchers due to the file format or access restrictions, the data does not become relevant for research purposes.

## 4 Fostering Industry-Academia Collaborations for Research Data Sharing

Various conceptualizations of research data exist. Implications for whether and when data is research data are thus fuzzy and unclear. Based on existing definitions of research data, we show that in theory all types of data can become research data. However, in practice, legal obstacles, disciplinary differences in research relevance of data, and data quality aspects can limit the relevance of data, ultimately hindering the possibility of turning all data into research data.

Data sharing between industry and academia allows researchers to create insights from (research) data. Therefore, fostering industry-academia collaborations for research data sharing is desirable to create more value of existing data. The following sub-sections illustrate services and practices that can help industrial and academic institutions to collaborate and exchange relevant (research) data.

### 4.1 Supporting Services by the NFDI and its Section Industry Engagement

Beyond existing laws, encouraging the further sharing of such relevant research data from industry to researchers requires fostering a culture of collaboration and mutual benefit. One approach is to establish partnerships between industry and academia, where both parties contribute data and expertise towards common research goals. Legal certainty, funding, and accessibility criteria must be met to make data in these collaborations usable. Ultimately, emphasizing the societal and scientific benefits of data sharing, e.g., advancing knowledge, driving innovation, and addressing pressing challenges, can further incentivize the industry to participate in sharing relevant research data.

Given the new legislations in the EU and Germany, we argue that now is a good time to advance the integration of data exchange between industry and academia. Research data are becoming increasingly important in the development of innovations. The National Research Data Infrastructure provides various services that can effectively support an industry-academia cooperation:

#### 1. Data Accessibility and Interoperability Services

The NFDI establishes standardized protocols and formats for data storage, management, and exchange. This ensures that data generated by academia can be easily accessed and utilized by industry partners, fostering collaboration on research projects and facilitating the translation of academic findings into practical applications.

#### 2. Comprehensive Metadata Catalogs

The NFDI develops metadata catalogs that provide comprehensive information about datasets, including their origin, structure, and usage permissions. These catalogs help researchers from academia and industry discover relevant datasets for their projects, promoting collaboration and reducing duplication of efforts.

### **3. Data Security and Integrity**

The NFDI implements robust security measures to protect sensitive data while ensuring its integrity and authenticity. This is essential for building trust between academia and industry partners, encouraging the sharing of proprietary data for collaborative research endeavors.

### **4. Collaborative Research Platforms**

The NFDI develops collaborative research platforms that facilitate communication and project management between academia and industry partners (see, e.g., [researchDATAmarketplace.com](https://researchDATAmarketplace.com)). These platforms streamline the exchange of ideas, resources, and expertise, accelerating the pace of innovation and driving collaborative research initiatives forward. They are embedded in other national and international initiatives.

### **5. Mediating Services**

Mediators can help to bring both business and academic partners together. The Section Industry Engagement does this between National Research Data Infrastructure (NFDI) consortia and industry partners.

The services provided by the NFDI already enable effective collaboration in research projects between industry and academia, promote the exchange of expertise, and assist in converting scientific discoveries into practical applications that benefit society. To further expand and enhance these interfaces, commitment from industry, academia, and policymakers is necessary to support and sustain these collaborative efforts, fostering continuous innovation and mutual growth.

## **4.2 Clearly Defined Data Governance Processes**

In addition to the technical and operational services provided by the NFDI, established data governance frameworks play a crucial role in creating shared standards and principles for managing data effectively. In the context of industry-academia collaborations, a data governance framework provides a structured approach to managing and protecting shared data assets, ensuring their proper handling throughout their lifecycle.

A data governance framework encompasses policies, processes, roles, responsibilities, and metrics. Policies form the foundation by establishing guidelines and standards for data management, with a particular focus on quality, privacy, and security to meet the diverse regulatory and ethical requirements of both industry and academia. Processes operationalize these policies through steps such as data classification, lineage tracking, and incident response, maintaining the integrity and usability of shared data. Roles and responsibilities are clearly defined, ensuring accountability across all stakeholders. For example, data stewards in academia may oversee the quality and integrity of data, while data custodians in industry manage the technical environment for secure sharing. Lastly, metrics play a pivotal role in evaluating the framework's effectiveness, with indicators such as data accuracy and compliance rates offering insights into performance and areas for improvement.

Established frameworks and standards offer guidance for managing data governance in industry-academia collaborations. The DAMA-DMBOK (Data Management Body of Knowledge), for instance, offers comprehensive best practices for managing data as a strategic asset. Its structured approach covers key areas such as data architecture, quality management, and security, making it particularly relevant for ensuring that shared data between academia and industry remains organized, high-quality, and secure. By providing a foundation for managing data lifecycles, this framework supports the seamless integration of data resources from multiple stakeholders.

## 4.3 Data Exchange Through Data Spaces

Beyond the services offered by the NFDI and effective data governance frameworks, data spaces are an emerging concept that plays a pivotal role in the exchange of data between different stakeholders, including academic institutions and industry partners (see, e.g., International Data Spaces Association 2024a; International Data Spaces Association 2024b). Data spaces are secure environments where industrial and academic partners can share, access, and collaborate on data while maintaining control over their data assets. That is, in a data space, data providers can make their data available under specific conditions, and data users can access and utilize this data for various purposes, including research, innovation, and development, leading to several beneficial outcomes.

Thus, incorporating data spaces into the existing framework for data sharing between academia and industry can significantly enhance the accessibility, security, and utility of research data.

Some data spaces have already demonstrated their potential in fostering effective collaborations between industry and academia. Notable examples include the International Data Spaces and the Catena-X Automotive Network, both of which illustrate how structured data sharing can drive innovation and mutual benefits.

### 1. International Data Spaces (IDS)

The International Data Spaces (IDS) initiative, which is led by the International Data Spaces Association, provides a roadmap for secure and sovereign data exchange across organizational boundaries. IDS promotes data sharing based on trust, interoperability, and compliance with legal frameworks such as GDPR. It operates on a decentralized architecture, ensuring that data providers retain control over their assets while facilitating seamless collaboration.

The IDS has successfully enabled collaborations in domains such as manufacturing and logistics. For instance, the Smart Connected Supplier Network connects high-tech manufacturing companies and their IT suppliers, utilizing IDS principles to enable secure and seamless data exchange within the supply chain. Academic researchers leverage this shared data to study and optimize production processes, fostering innovation and creating more efficient workflows in supply chain management. Another example is the Fraunhofer Supply Chain Manager, developed in collaboration with Volkswagen AG, ThyssenKrupp, and Fraunhofer ISST, which uses IDS standards to enhance transparency in automotive supply chains. Through a secure data-sharing platform, academia gains access to real-time supply chain data, enabling researchers to develop advanced models for improving operational efficiency and addressing bottlenecks.

## 2. Catena-X Automotive Network

The Catena-X Automotive Network represents a pioneering data space within the automotive sector, designed to enable secure and efficient data sharing across the entire automotive value chain. Developed as part of the European GAIA-X initiative, Catena-X integrates manufacturers, suppliers, and academic partners into a collaborative ecosystem. The network facilitates seamless data exchange by ensuring interoperability, data sovereignty, and adherence to stringent security standards. Through shared access to lifecycle data, Catena-X addresses critical industry challenges, such as sustainability, supply chain resilience, and digital transformation.

The Catena-X data space can facilitate collaborations between industry and academia, particularly in the areas of sustainability and autonomous driving. In the realm of sustainability, Catena-X has developed solutions aimed at achieving end-to-end transparency and aggregation of CO<sub>2</sub> values for produced vehicles, including components installed at all tier levels. This initiative enables academic researchers to access comprehensive data on carbon emissions across the automotive supply services in Industry 4.0, facilitating secure data exchange between industry partners and academic institutions. Such collaboration accelerates research and development in autonomous driving technologies by providing researchers with access to critical data necessary for advancing vehicle automation systems and fostering the development of strategies to reduce environmental impact. Regarding autonomous driving, Catena-X has established a collaborative and open data network for the German and European automotive industry. This network represents the first large-scale deployment of Gaia-X compliant (see Fraunhofer 2024a; Fraunhofer 2024b).

## 4.4 Using Generative AI for Data Sharing

Finally, generative AI technologies are expected to play a central role in the future of secure and efficient data sharing between industry and academia. By addressing critical challenges such as data privacy, regulatory compliance, and interoperability, these technologies may transform the way collaborative research is conducted.

On the one hand, generative models, such as Generative Adversarial Networks (GANs) and advanced language models, enable the creation of synthetic data sets that reproduce the statistical properties of real data while protecting sensitive information. This capability allows proprietary or confidential data to be shared securely without compromising privacy.

On the other hand, generative AI automates key processes such as data anonymization, standardization, and transformation, ensuring that data from multiple sources can be seamlessly integrated for collaborative research initiatives.

Therefore, as generative AI technologies continue to evolve, their integration into industry-academic collaborations will increase trust, ensure compliance, and reduce barriers to data sharing, accelerating innovation and fostering impactful research partnerships.

# Appendix

## Appendix A1: Exemplary Sources of Publicly Accessible Corporate Research Data (own illustration)

Data Provider	Description	Structure	Exemplary Providers in Germany
<b>Official Statistics: Federal and State Statistical Offices</b>	Data collected, analyzed, and published by government agencies for informing policy and research	Breakdown of statistics by subject area, e.g.,: <ul style="list-style-type: none"> <li>• Healthcare</li> <li>• Buildings and housing</li> <li>• Environment</li> <li>• Manufacturing</li> <li>• Financial and other services</li> </ul>	<ul style="list-style-type: none"> <li>• Federal Statistical Office</li> <li>• Federal Employment Agency</li> </ul>
<b>Research Data Centers</b>	Specialized facilities established by government agencies, research institutions, or universities to provide access to and facilitate the analysis of sensitive or restricted datasets for research purposes	Data available by: <ul style="list-style-type: none"> <li>• Economic sectors (manufacturing, energy, environment, agriculture, services, other economic sectors)</li> <li>• Products</li> <li>• Reporting years</li> <li>• Basic statistics</li> <li>• Description</li> <li>• Data access</li> </ul>	<ul style="list-style-type: none"> <li>• German Foundation Center</li> <li>• Federal Bank</li> <li>• Research Data Centre of the Federal Statistical Office</li> </ul>
<b>Research Institutes</b>	Organizations dedicated to conducting scientific research, often focusing on specific fields or interdisciplinary topics, and contributing to knowledge advancement through experiments, studies, and publications	<ul style="list-style-type: none"> <li>• Typically organized into departments or research units focusing on specific or interdisciplinary fields</li> <li>• Governed by a leadership team, including a director or head researcher, with support staff such as administrators and technicians</li> </ul>	<ul style="list-style-type: none"> <li>• ifo – Leibniz Institute Institute for Economic Research</li> <li>• RWI – Leibniz Institute for Economic Research</li> <li>• IWH – International Science Forum Heidelberg</li> </ul>
<b>Databases of Public Institutions</b>	Centralized collections of structured information maintained by government agencies or public institutions, offering access to various records, datasets, and documents pertaining to their areas of operation and jurisdiction	<ul style="list-style-type: none"> <li>• Organized into categories or sections based on the type of information stored</li> <li>• Managed by database administrators or information specialists responsible for data organization, maintenance, and access control</li> </ul>	<ul style="list-style-type: none"> <li>• Commercial register</li> <li>• Federal Financial Supervisory Authority (BaFin)</li> <li>• Bundesbank</li> </ul>

## Appendix A2: Exemplary Different Data Formats Across Scientific Disciplines and Data Types (own illustration)

Discipline	Text Formats	Numeric Formats	Multimedia Formats	Research Field-Specific Formats	Instrument-Specific Formats
Natural Sciences	FITS (astronomy)	NetCDF (environmental sciences) HDF5 (high energy physics)	AVI (astronomy) TIFF (microscopy)	CIF (chemistry) PDB file format for 3D protein structures	SPE (spectroscopy)
Social Sciences	CSV (survey data) SPSS (statistical analysis)	SAS (statistical analysis) Stata (econometrics)	WAV (audio recordings) MP4 (video recordings)	DDI (social science metadata) ANVIL (qualitative data analysis)	E-Prime (experimental psychology) Qualtrics (online surveys)
Humanities and Linguistics	TEL (text encoding) XML (linguistic corpora)	CSV (lexical data) JSON (text analysis)	MPEG (audiovisual recordings) PNG (image archives)	EA (linguistic annotations) RDF (semantic web data)	ELAN (Linguistic Annotations) Praat (phonetics analysis)
Life, Medical, and Health Sciences	FASTA (genomics) HL7 (healthcare data exchange) FHIR (electronic health records)	ECG (electrocardiography)	MRI (medical imaging) CT (computer tomography)	LOINC (laboratory observations) SNOMED CT (clinical terminology)	EHR (electronic health record systems) DICOM (medical imaging)
Engineering and Material Sciences	XML (engineering specifications) TXT (engineering documentation)	MATLAB (numerical simulations) Excel (experimental data)	CAD (computer aided design files) GIF (structural animations)	STEP (3D CAD models) G-code (CNC machine instructions)	SCADA (supervisory control and data acquisition) RTDS (real time digital simulations)



# References

- Büchel, Jan and Barbara Engels (2023), "Data Sharing in Deutschland," *IW-Trends*, 50 (2), 19-37.
- DFG (2015), "Guidelines on the Handling of Research Data," (accessed February 02, 2025), <https://www.dfg.de/resource/blob/172098/4ababf7a149da4247d018931587d76d6/guidelines-research-data-data.pdf>.
- (2023), "Handlungsempfehlungen zum Umgang mit Forschungsdaten," (accessed February 02, 2025), <https://www.dfg.de/resource/blob/173258/64ce3b7ee138953d13d80530d891bfa3/fk-materialwissenschaft-werkstofftechnik-2023-data.pdf>.
- European Union (2019), "DIRECTIVE (EU) 2019/1024 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," (accessed February 02, 2025), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L1024>.
- Fraunhofer (2024a), "Catena-X Automotive Network", (accessed February 02, 2025), <https://www.isst.fraunhofer.de/de/abteilungen/industrial-manufacturing/projekte/CatenaX.html>.
- (2024b), "Fraunhofer supports in the BMWI-funded "Catena-X Automotive Network" (accessed February 02, 2025), [https://www.iml.fraunhofer.de/en/fields\\_of\\_activity/enterprise-logistics/supply\\_chain\\_engineering/research-projects/catena-x-automotive-network.html](https://www.iml.fraunhofer.de/en/fields_of_activity/enterprise-logistics/supply_chain_engineering/research-projects/catena-x-automotive-network.html).
- German Council for Scientific Information Infrastructures (2021), "Research Data, Research Data Management (V2)," (accessed February 02, 2025), <https://rfii.de/en/topics/>.
- German Data Usage Act (2021), "Section 3 Definitions," (accessed February 02, 2025), [https://www.gesetze-im-internet.de/dng/\\_\\_\\_3.html](https://www.gesetze-im-internet.de/dng/___3.html).
- International Data Spaces Association (2024a), "International Data Spaces – The future of the data economy is here," (accessed February 02, 2025), <https://internationaldataspaces.org>.
- (2024b), "Use case are IDS in action," (accessed February 02, 2025), <https://internationaldataspaces.org/make/use-cases-overview/>.
- Tayi, Giri K. and Donald P. Ballou (1998), "Examining data quality," *Communications of the ACM*, 41 (2), 54-7.
- Wang, Richard Y. and Diane M. Strong (1996), "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, 12 (4), 5-33.