Rat für
Informations
Infrastrukturen

7

THE DATA QUALITY CHALLENGE

Recommendations for Sustainable Research in the Digital Turn

The Data Quality Challenge
Recommendations for Sustainable Research in the Digital Turn

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BMBF	(German) Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung)
CERN	Conseil européen pour la recherche nurcléaire
CTS	Core Trust Seal
DESY	German Electron Synchrotron (Deutsches Elektronen-Synchrotron)
DFG	German Research Foundation (Deutsche Forschungsgemeinschaft)
DHV	German Association of University Professors and Lecturers
DIN	German Institute for Standardization
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
GWK	Joint Science Conference (Gemeinsame Wissenschaftskonferenz)
HEI	Higher Education Institution
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
IUPAP	International Union of Pure and Applied Physics
MARC	Machine-Readable Cataloging
NFDI	National Research Data Infrastructure (Nationale Forschungsdateninfrastruktur)
PSI	Private Sector Information
RatSWD	German Data Forum (Rat für Sozial- und Wirtschaftsdaten)
RDM	Research Data Management
RfII	The German Council for Scientific Information Infrastructures (Rat für Informationsinfrastrukturen)
SFB	Collaborative Research Centre (Sonderforschungsbereich)
WR	German Council of Science and Humanities (Wissenschaftsrat)

EXECUTIVE SUMMARY

The qualitative dimension of scientific or scientifically generated data touches on the very autonomy of science and its operative subtleties: researchers themselves decide on the choice of their theoretical approaches and the methods and instruments they use in order to collect research data and achieve research results on this basis. The quality of the research data thus depends on standards which are inherent to research itself.

The digital turn in science, economy and society creates a new context here. Today, the quality problem arises in a new form due to the multitude of data generated and produced with digital technology as well as with their transfer and possibilities of use. In science, this affects all disciplines, subject areas and forms of research – although there are gradual differences depending on how "data-intensive" work has already been done in the past. However, new questions also arise at the interfaces between research and the previously separate infrastructure areas – the so-called "knowledge repositories", such as libraries, archives, collections, computer and data centres.

The topic of data quality is attracting even more attention today where the future of society is concerned: the simulations of climate and earth system research generated with large amounts of digital data, for example, are subject to public debate. Commercial companies are increasingly offering their own "hypothesis-free" data analyses, and public science is becoming dependent in some areas on infrastructure services and data from the private sector. Against this background, science and science policy have recognised data quality as a challenge that should no longer be underestimated. It has become the object of reflection on new framing and regulating measures.

In this position paper, the German Council for Scientific Information Infrastructures (RfII) examines current challenges for data quality in the scientific system and derives recommendations from them.

In the current situation (Chapter 1), the RfII draws attention to the fact that previous conceptualisations of data quality fit only conditionally to the specific needs of scientific knowledge production. For they often originate from a management (theory) context and must be translated with a sense of proportion into the logic of research processes. The RfII describes various concepts and instruments that are suitable for controlling data quality. What all existing data quality models have in common is that they do not provide sufficient information on how data and its documentation can be effectively linked with the requirements of digital research processes in different disciplines and fields.

In Chapter 2, the RfII traverses the stages of the scientific data life cycle and outlines the challenges facing research in the respective phases of data transformation in order to illustrate concrete problem situations. At the various "interfaces" – transitions between phases, from data collection to data sharing and publication – there are numerous difficulties to synchronise data protection and the associated increase in data quality between management and research practice. The RfII sees this as a multi-level task in which researchers and infrastructure staff must work together closely. Leading quality criteria, coupled with methodological standards, must come from the scientific communities and professional associations.

The challenge of data quality is amplified by other developments in the scientific system that affect research quality issues in general. In Chapter 3, the RfII refers to the so-called replication crisis and the much-discussed overload of the peer review system. The Council also sees the quantitative overstretching of publication requirements as a problem for an intensive commitment to data quality and – above all at the international level – a lack of science-compliant frameworks and regulations. Another problem that has been virulent and worsening for some time in the data-intensive research fields is the dependence on measuring instruments or hardware and software components of commercial manufacturers.

The RfII derives a bundle of recommendations from these assessments in Chapter 4, which call for a joint responsibility for quality or a responsibility to be exercised in dialogue. The addressees of these recommendations are:

- data producers, processors and various downstream users of research data,
- researchers, their specialist communities and infrastructure providers,
- Higher Education and and non-university research institutions as organisational designers, and
- the scientific organisations, the funding institutions and the specialised ministries at federal and *Länder* level, which also set the monetary and programmatic framework for the data quality efforts of the scientific community.

The RfII bases its recommendations on a process-oriented concept of data quality. This concept considers both transformations in the research processes and the current rapid changes in the technical possibilities for data processing. Openness and dynamics, but also a close connection to scientific methods and research forms must be essential guidelines of a data culture to be developed.

This includes understanding appropriate documentation of research data based on professional standards as a core scientific task and part of professional ethics. Securing and improving data quality is a fundamental value of good scientific practice. The RfII calls on the scientific communities to take this fundamental

value into account even more strongly than before in the methodological training of the disciplines and research fields. This also includes the task to integrate research processes and research infrastructures – including libraries, computer centres, etc. – more clearly than today. And working with research data deserves a higher degree of professional reputation.

In order to increase data quality, it is essential to consider the different interfaces to the data life cycle at every stage of the research process. This requires coherent data descriptions and declarations. Previously implicit knowledge must be explicitly documented and – as far as possible – made machine-readable. The RfII calls for the further development of appropriate technical assessments. Hardly transparent hardware and software properties or the disposing of infrastructure-relevant product lines by commercial providers complicate scientific efforts for high data quality. The RfII recommends considerable efforts on the part of scientific communities and learned societies to demand greater product transparency from suppliers.

The RfII sees the differentiation of scientific data products as an opportunity to give more recognition to the work with research data, to increase data quality and to make replication studies more attractive at the same time. Corresponding product forms range from the co-publication of research results and associated data sets to the creation of curated data collections, which can already contain applications for the further use of the data. The introduction of an independent scholarly review culture for research data would also be helpful – not as a niche product, but in a highly visible and trustworthy place, i.e. in respected journals.

Where this is not yet happening or is still in the process of being established, the RfII considers quality assurance for research and information infrastructures to be indispensable, for example through evaluations, based on scientific standards. In this way infrastructure facilities can be developed into centres of excellence over the long term (even in research areas that have hitherto been less data-intensive), that can encourage standard-setting in research. Different paths can be taken, including learning from partner institutions. Cooperation should also relate to the ongoing improvement and adaptation of the technical infrastructure, which must consistently meet the highest standards in order to keep research in Germany internationally competitive.

A basic prerequisite for data quality is the skills of those working in research and infrastructure. The RfII 2019 has published its own recommendations on this subject, DIGITAL COMPETENCIES – URGENTLY NEEDED!. The Council considers it just as important to break down the pillarisation in the training of scientists and of infrastructure personnel as it is to generally increase IT competence in all disciplines and to provide continuous training that traverses formal scientific qualification goals.

Funding policy today is only rudimentarily adjusted to the importance of data quality in the digital change. In principle, it must be possible to design the duration of funding projects more flexibly in order to give sufficient scope to data aspects early in the application phase. In this context, the qualitative yield of research (e.g. well-documented data sets) should be given preference over high quantity output in the evaluation of past research achievements. Until public funding bodies have expanded their programmes accordingly, the RfII sees great potential for foundations to act as catalysts with innovative funding formats for the further development of data quality. Furthermore, the RfII recommends establishing the production of innovative data products as an independent field for funding.

The RfII advises higher education and non-university research institutions to incorporate data quality as a core element in their research strategies. This should be linked to new collaborations that actively involve infrastructure areas such as computer centres, libraries and university as well as non-university collections. In particular, non-university research institutions with infrastructure tasks can play a leading role in the development of standards for data management in the future. Universities should also integrate research data as a topic more firmly in teaching activities and look for appropriate expertise in their professorial appointments. The RfII sees it as duty of the Federal Government and the *Länder* to actively support scientific institutions in the further development of data quality in Germany. In this regard, also the National Research Data Infra-structure (NFDI), a new initiative established by the Joint Science Conference (GWK), is an important player. In addition, the Federal Government and the *Länder* should continue their efforts to seek long-term options to secure the existence of successful but precariously financed research and information infrastructures at universities.

The European FAIR process has proven to be a successful way to raise awareness for minimum requirements for scientific data and their accessibility. The RfII advocates a substantive deepening of the FAIR process in order to further advance the integration and transfer of data in the European and international research area. The Council recommends linking FAIR more closely to discipline- and research-field-specific quality criteria in order to increase the quality and possible uses of “FAIRer” data. In addition to FAIR, a European campaign for good scientific data quality would be helpful. A marked increase in communication efforts is also necessary in science and society, in order to make the required explication of implicit knowledge a basic value of a global data culture.

INTRODUCTION: DATA QUALITY - AN UNDERESTIMATED ISSUE

Science[1] is based on a promise of quality: Research results are achieved on the basis of accepted methodological principles. The data used and generated in the research process meet high quality standards, which are set and controlled by the scientific communities themselves. In this context, the acquisition of new scientific knowledge is closely related to increasing the quality of data and data-based processes. This is also the basis for society's expectations of scientific performance.

Scientifically qualified knowledge in the global "digital turn"

Realising these elementary characteristics of the production of scientifically qualified knowledge under the conditions of a "Weltenwandel" (global transformation) driven by enforced digitalisation poses new challenges for research.[2] Data processing today is possible in all scientific disciplines and research fields with a high degree of automation in enormously complex and diverse forms as well as with technically almost unlimited networking options. Some research processes have changed drastically. Data volumes are growing. But there is also an increase in dependencies, verification problems and new forms of non-transparency, which make statements about data quality extremely demanding – especially beyond disciplinary boundaries. Research is in a special position in the digital transformation of the world: it is itself a massive driver of transformation – firstly in the fields of mathematics and information technology but also, through data-intensive technologies in the natural and engineering sciences. On the other hand, it is exposed to this transformation when it comes to adapting the understanding of methods and the handling of data in all academic disciplines.

New challenges for data quality

Only if the quality of data satisfies scientific demands even under these conditions, science can continue to deliver what society expects of it. Data sets and methods must also be transparent under digital conditions and be comprehensible to other scientists. The results obtained on this basis must be valid or even replicable under certain circumstances. Without substantial and well-documented data, research results and the innovations that stem from them will not be sustainable. This also affects the trust and support that a functioning science system needs in society: Both would erode in the medium term if justified doubts about the quality of scientific data arose. This makes the quality assurance of data and data processes a permanent challenge, both for the specific research scenarios in all disciplines and for the scientific system and its social role as a whole.

---

1 Speaking of "science" in the context of this position paper also includes the academic disciplines of arts and humanities.

2 The term "Weltenwandel " to describe the caesura that digitalisation means for science today is taken from RfII: Strohschneider (2018) – Neujahrsansprache.

But what exactly is data quality in a scientific context? In fact, for a long time science has only discussed what the term means under the general heading of "methods". A scientific quality of data can therefore not easily be defined with sufficient depth of focus. The RfII made its first attempt in 2016. Accordingly, the term data quality includes both general and typical characteristics of the data required from a methodological point of view as well as their additional suitability for further use, if necessary created by quality assurance measures.[3] More detailed quality standards and models for digital data originate from management theory, business informatics and considerations on industrial process optimisation in production cycles. Such models are only suitable for scientific research processes to a very limited extent, since science

Characteristics of data quality in the sciences

- proceeds methodically controlled, not fixed upon certain products, but open to results,
- wants to continuously increase quality on the way to the result (also for future, still unknown research questions),
- let knowledge and data circulate widely and, unlike a business, even shares them (ideally in altruistic fashion),
- archives data more sustainably than the economy (for the purposes of referencing results, documentation of research lines and time series formation in long-term studies), and
- re-uses data constantly in the sense of a cumulative progress of knowledge, or takes up "elder" data at unforeseeable intervals in order to answer new questions.

Old and new quality discourses in science

Quality criteria and standardised methods of data evaluation have always been discussed in science – even before the digital age was a common topic. There is also a public debate about general crisis phenomena in science that is leading up to the digitisation issue. These do not directly concern data quality, but questions of performance evaluation and good scientific practice. Among other things, it will be discussed whether and under what conditions scientific studies and research results must be comprehensible, repeatable or replicable in some areas.[4] In these discourses, too, trust in science, its capacity for self-organisation and suitable political frameworks for expanding its capacity to perform play an important role. The RfII sees a close connection between these developments and the previously underestimated issue of data quality.

3 See RfII (2016) – Enhancing Research Data Management, Glossary, p. 76. As a rule, the research methods that decide on the selection, collection, processing and transfer of the data are in turn dependent on theoretical assumptions or choices in individual disciplines and research fields or the different "schools" found in multi-paradigmatic fields of science. The embedding of method choice in theories also distinguishes scientific data work from data-based commercial research and "analysis".

4 For detail on this, see chapter 3.1.

Furthermore, the interest of states, governments and civil society in quality assured research data on a global scale has grown. On the one hand, research data is attracting general attention as an evidence base for non-scientific decision-making processes. On the other hand, they are regarded as a "raw material" for more rapid innovation cycles. All actors are thus becoming increasingly aware of how much the quality of data and data processes will coincide with the quality of science as a whole in the coming decades of the digital age. Digitality gives a new dimension to the topic of data quality, which urgently requires framework-setting action in science policy, but also an increase in attention and commitment in all scientific communities and scientific organisations.

Growing societal interest in data generated by science

There is a need for action – beyond the quality-neutral speech of a "use" of data – especially with regard to the quality of digital research methods. It is this quality that decides about

- the validity and connectivity of the later research results in the scientific communities (disciplinary and interdisciplinary);
- the success of the transfer into business and society – also in settings where commercial players now offer their own data sets and analytical procedures and are gaining in importance as providers of data, equipment and analytical tools for research;
- the society´s trust in the particular value and sustainability of scientific knowledge production – also in comparison to esoteric or more interest-driven forms of opinion-forming that are not based on data and that are not gained within the framework of intersubjectively verifiable standards and procedures.

Ensuring quality of digital research methods

The sheer volume of data that is potentially accessible today, as well as the technical possibilities to share and recombine it, opens up the opportunity for science to work on completely new research questions and fields. In this sense, increasing quantity also means greater comparability and expansion of the scope for the use of scientific methods. In order to be able to deal with these possibilities in a targeted manner, communication processes are necessary that not only cross disciplinary boundaries, but also overcome traditional institutional barriers. This means dividing lines between the research process including its actors in the narrower sense and the research-enabling institutions of the infrastructure (e.g. libraries, collections, archives, computer centres and data centres) and their personnel.[5] As a cross-sectional technology, digitality crosses the border between research and application. This can be observed, for example, in translational clinical research at the transition from laboratory

Opportunities from digitisation: new research questions and fields

[5] For this, also see RfII (2019) – Digital Competencies, p. 27 f., recommendation 4.5.

research to patient treatment or in engineering simulation research. A new type of requirements for data linkage demands new quality standards if they are not to be served "somehow", but are to be transparent, comparable and scientifically controllable. The need for research data that can be processed, understood and further processed at the highest level of quality across earlier dividing lines or that can be re-used or re-used downstream is therefore extremely high.

Raising quality consciousness as task for the future

With this position paper, the RfII raises the question of how the increased quality requirements posed on present and future data by newly developed digital methods can be met. Under the deliberately broad keyword "new data culture", he pointed out the important role of information infrastructures for the quality assurance of data as early as 2016.[6] With the recommendations presented here, the RfII emphasises that quality is essential for the future of science, especially in view of the exponential increase in data volumes in the digital age in all its disciplinary forms and connections. It is by no means just a matter of "continuing like this". Because: New cultures of open circulation and global sharing of digital data only have real scientific added value if there is a sharpened awareness of quality.

Science-wide quality discourse called for

A common discourse on quality as well as on binding quality assurance measures in science is required among all stakeholders. This needs to be based on a precise analysis of the changes that the digital turn entails for methodological research. The RfII is convinced that in the medium term such a discourse will also result in concrete research actions and will influence the practice of research institutions and science organisations as well as the establishing of science policy frameworks. In other words: the discourse will drive actions and strategic decisions for maintaining public confidence in scientists ability to develop quality standards which support the responsible transition toward a data-driven future.

6 See RfII (2016) – Enhancing Research Data Management, p. 46 f.

1 CURRENT SITUATION

1.1 DIGITAL TURN AND DATA QUALITY – WHAT HAS CHANGED?

The process of scientific knowledge production is designed to produce particularly qualified ("true") knowledge based on the best possible data and guided by evolving theories and methods. This knowledge is characterised by validity according to methodological standards and maximum verifiability. In the form of publications, it must prevail in academic debates inside and between scientific communities.

Digital research has experimental characteristics

The use of digital techniques in methods, organisation and communication of research does not fundamentally alter this normative model of knowledge production. However, changes in certain research forms can be observed in the factual and temporal dimension. Likewise, the digital transformation and the digitally driven enabling of data linkage stimulate new questions and allow "elder" problems to appear in a new light. In particular, the use of media and tools working on a digital basis as well as the collection and processing of data already digitally generated has a profound impact on many stages of the research process (cf. also 2.1). In many cases, the use of new digital tools in research has an experimental character. It produces results, for the validation of which standards and criteria have yet to be developed. Traditional mechanisms for ensuring quality aspects of research activities must now be reflected upon anew in all scientific disciplines.

So-called digital tools modify methods

Digital tools (especially computing, i.e. modelling and machine calculation, but also new data collection, presentation and analysis methods in science) have changed the methodological basis of entire research fields in recent years. Examples include particle accelerator technology, telescoping, digital remote sensing, tomography, geophysics, molecular biology, digital imaging and text analysis. Numerous fields in the natural, life and technical sciences have adapted to a world of almost completely digitised research objects. In other fields of research, digital sub-disciplines have been emerging for some time, such as computational physics, computational social sciences, digital humanities or bioinformatics, geoinformatics and archaeoinformatics.

Heterogeneity and interoperability as challenges

But here, too, the sheer amount of data generated does not only result in new options. There are also new challenges arising from the heterogeneity of the data and the rapid changes in the field of programming languages and software, for example with regard to the documentation and provenance of data as well as their quality for use in interdisciplinary contexts and with regard to the physical stability of data carriers over time. For the natural, life and technical sciences as well as for the social sciences, the adaption of their information infrastructures for the flexible provision of interfaces is largely uncharted

territory. In addition, there are the growing demands and requirements for contact with the "environment" of the scientific system: digitally available scientific data are increasingly becoming interesting for commercial, political and civil society applications (conversely, science also makes use of non-scientific data).

At the same time, the comprehensive digitisation of research processes is intensifying scientific quality issues, some of which were previously also virulent, but are now being raised to a new level by the exponential increase and availability of data. Table 1 below illustrates the challenges this has created:

Table 1: Growing global challenges of data quality assurance

■ Disproportionately large impact of small inattentions, errors and failures
■ Decisions on the usability of noisy mass data
■ Decontextualized uses of individual data sequences
■ Unclear or unrecognizable provenance of data (especially in the case of algorithm-generated outcomes and selections)
■ Non-transparent computational processes
■ Misdirecting algorithms (e.g. due to scaling problems)
■ Lack of training sets for the programming of machine learning (AI)
■ Complicated or impossible verification/validation of the practical value of oversized data batches
■ Division of labour along by now only weakly integrated process chains (so-called "pipelines") while working with scenarios or doing simulations
■ Growing dependence of knowledge work on proprietary software
■ Lack of archivability of the digital artefact
■ Presentation problems for the result dimension of complex computations and data condensation (e.g. through "visualization")
■ Data protection and other legal issues
■ Low-threshold manipulation possibilities
■ Hacking and cyber espionage
■ Targeted data sabotage
■ … and more.

Source: Own illustration.

Added scientific value through interdisciplinary data transfer

When it comes to the added value of digitisation in science, the transfer of data across disciplinary boundaries is particularly important. This is not a technical challenge in the narrow sense, but the internal order of data collections is decisive. For digital research data to be re-usable in interdisciplinary research processes, implicit knowledge must be explicated, because only explication determines the extent to which it can be used in (possible) other contexts. In other words, the conditions under which data was created, as well as its current state in a process of use and transformation, are described for further research. This comprehensive service, operationalised as "explication" in the following, is also the necessary basis for every form of machine data processing.

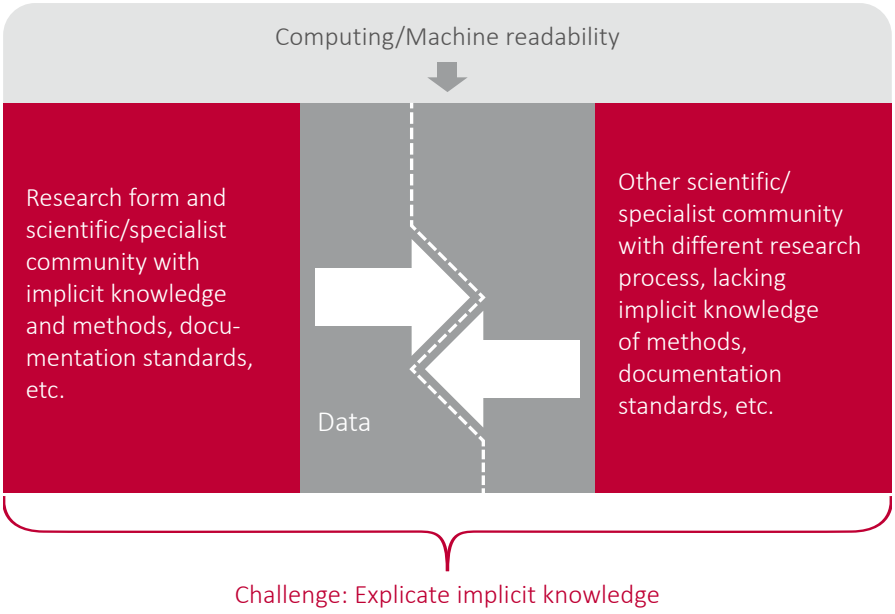

Figure 1: Digitisation increases the need for explicating implicit knowledge.
Source: Own illustration.

Progress in methodological knowledge involves the need to explicate implicit knowledge

The explication of the methodologically relevant knowledge which is contained in research data[7] is not a fundamentally new methodological imperative in the world of science. It has long been good scientific practice to store content data together with the information about its origin – metadata that summarises content (keywords) are part of this. The use of digital processes creates new needs for the explanation of data and their origins and uses. For example, the (unavoidable) "noise" in large streams of measured data and the (remaining) "blurriness" of digital data analysis methods must be identified. Likewise, the programs, transformation steps and parameters to which the data are bound must be explained – including information about software versions, hardware generations and, if applicable, laboratory or field conditions which were involved in generating data. The necessity of making data digitally linkable, which is imperative for today's research, also requires a strongly standardised explication with regard to the necessary "machine readability". In contrast to the analogue world, where humans can easily improvise in handling research data, here fixed machine-language structures are needed that rule out different handling situations as far as possible. The extent of the associated determinations and the routines tied to them can have an impact on research practice.

Explication is demanding

The fact that the spectrum of scientifically generated data is very heterogeneous shows just how difficult the required explication of the data and their genesis

7 The RfII uses a broad concept of research data here (as the Council has done so far). It includes both analogue and digital data and object collections, because the linking of analogue and digital data often plays an important role in research methods and scientific knowledge production in general.

is: from arbitrarily reproducible data sets, for example in genome analysis, to the one-dimensional, non-reproducible observational data generated in astrophysics. Also in the field of biodiversity research or archaeological excavations, sampling and documentation processes are often not repeatable. In such cases, digital data are unique specimens and a reference similar to everything "analogue" without substitution, provided that physical preservation as an object of collection is not possible or the physical representation of the object is lost due to a loss of the collection (e.g. war, natural disasters). Often, however, it is also the high investment costs that limit or make impossible the renewed collection of research data. The demand for an appropriate explication of research data goes hand in hand with the claim for optimal scientific valorisation, because the use of (existing) digital resources is not only time-efficient, it also ensures the comparability of research processes and results within the given framework.

Not only data quality assurance but also improvement

The issues raised here lead to the question of whether and how quality can be assured in digital research processes and how it can be improved in order to become the driver of a completely new dynamic in science and the innovation systems dependent on scientific data. This is accompanied by the question: How can data quality be "organised" for advanced digital work in science? Who assumes which responsibility and tasks, how can the experimental character of the use of digital tools be linked with existing quality discourses? And: Are the procedures of scientific quality control and the (self-)assurance of quality criteria still sufficient with regard to the dynamics of growth that enable accelerated digitisation? Which ones should be newly developed?

From the point of view of the RfII, it is inevitable to discuss the fact that innovations are necessary in the interplay between the quality of data and that of methods. Likewise, the task at hand would benefit from a binding terminology to describe data and recommendations for the development and assurance of data quality can be derived from it.

1.2 DATA QUALITY CONCEPTS – APPROACHES AND DESIGNS

Intrinsic motivation – scientific ethos

A central orientation point for data quality concepts in science is the intrinsic motivation of researchers: "Science as a profession". This professional ethos includes personal responsibility for maintaining intersubjectively verifiable quality of research, including the methods and procedures for obtaining research results. Today, scientific integrity is a key word in this regard. It requires compliance with the rules of "good scientific practice".[8] In addition, science

[8] To this end, DFG has recently presented new guidelines: DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis [Guidelines for ensuring good scientific practice].

relies on its own methodological culture. What scientific quality criteria are in the overall system is predominantly taught and learned within the framework of a long and thorough (specialist) socialisation as a scientist at academic institutions. In this context, also the habitus of the researcher is formed. Professional expertise and reputation are attributed individually to the researcher as a single person. This attribution is often based on internalised "tacit knowledge", which normally – once acquired and handled professionally – hardly requires any further explication.

Novel external stimuli

In the 20th century, driven by ambitious social and innovation policy goals and the international competition of states, economic systems and ideologies, intrinsic motivation is complemented by the external, structural control of science. Knowledge production is now being actively promoted, primarily through incentives and framed by large research programmes. In addition, institutional quality assurance mechanisms have found their way into publicly funded science. In this context, a knowledge process oriented towards the concept of truth and driven by curiosity, "methods", procedures of "examination" and by "scientific achievements" is not all that matters. The term "quality" appears as a kind of added value – including the associated quality measures and quality assurance procedures in research and teaching. This is initially done by analogy with the talk about manufacturing processes or the quality of products and processes.

A coherent interdisciplinary discourse on scientific "data quality" arises together with the increase in digital services since the 1990s. Since then, quality assurance has been a major issue. However, less is said about the topic of increasing the quality of (digital) data, which is particularly important for the scientific process, or about the role of quality in a more comprehensive transformation process of science and society as a whole.

In the discussion on the quality of research data, five concepts or guiding ideas for the (self)-control of scientific activity can be roughly distinguished, which are described in more detail below:

Five guiding ideas for quality control of research data

1. Setting standards and standardisations (e.g. ISO standards, technical standards), including the standardisation of quality management: the predominant control mode is primarily of a *juridicial* nature;
2. data validation and organisational or process-related operationalisation of data quality (e.g. through certificates or quality seals): the predominant idea of control here is primarily *organisational*, based on incentives and upgrading;
3. guidelines and policies as basic rules for the handling of research data (incl. associated data management plans): the internalisation of rules should rather be achieved through *contractual* or communication-oriented instruments;

4. ideal-typical and schematic descriptions of the processes to be optimised (e.g. data life cycles): Here, a primarily *procedural* ideal for the production of data quality is pursued;
5. definition and setting of pragmatic rules of thumb (above all the formula "fit for purpose") and definition of general principles (currently, for example, the FAIR data principles): In this model, *pragmatic* and primarily procedural control of quality developments is the main focus.

Origin of data quality concepts - primarily from business and administration

Comprehensive data quality and data quality management concepts originate to this day primarily from the economic sector and have generally been adapted in the area of public administration – and here also in publicly funded science. Parallel to this, there have been extremely successful efforts to develop quality standards and quality assurance procedures in research itself – especially in large-scale research (depending on huge research infrastructures and devices) and in the operation of large databases (cf. 1.2.3). However, cross-domain and original quality discourses explicitly tailored to the possibilities and challenges of digitisation for research data have so far only been produced to a limited extent. It is also an open question which political framework science needs without inhibiting or damaging the necessary autonomy of its self-regulation, i.e. the intrinsic motivation of researchers to further develop quality standards.

1.2.1 STANDARDISATION AND STANDARD SETTING

"De jure"- and "de facto" standards

In technology development, (minimum) quality measures are created classically through standardisation: "Good" is what corresponds to the standard and can therefore be used functionally across the board. The need for comprehensive standardisation that guarantees compatibility and quality originates from the world of machine components. Early, coordinated standardisation efforts are an achievement of industrial mechanical engineering. The standardisation of processes, which is carried out in a similar way, has also rapidly gained acceptance – namely agreed and unmistakable definitions together with detailed implementation rules. Standardisation in the field of information and communication technology (file formats and data carriers, data transmission, web technologies, interfaces) as well as in the field of documentation and indexing of scientific information (vocabularies, cataloguing, search services) are relevant for data quality in science. Technical "de jure" standards, such as the DIN standard, are applied in both areas, as are a large number of so-called "de facto" standards, which are disseminated and enforced via applications and acceptance.

Systems of order and standards in scientific practice

Standardisation through DIN and ISO

In digital or digitally supported research processes, science moves in almost all of its specialties within the standards of the German Institute for Standardization (DIN standards) and internationally, for example those of the three European Commissions for Standardization [9] or the International Organization for Standardization (ISO). These are predominantly standardisations that originate in regulatory requirements in industry and the service sector.

Scientific standardisation organisations

In the scientific documentation as well as in the subjects and disciplines, there are also scientific standardisation organisations, some of which are very strong, which set and maintain standards in a similar way – for example through controlled vocabularies (thesauri), reference models and norm files, taxonometries and nomenclatures as well as other classifications. These are also sometimes proposed for recognition as DIN or ISO standards, partly for pragmatic reasons in order to stabilise the process of further development, partly for the hoped-for higher impact and acceptance "in the system".

Metadata standards

Among the most prominent specifications for interoperable information systems in science are norms and standards for metadata in libraries. The development of metadata standards accelerated in the 1990s, but often did not follow an internationally obligatory guideline across domain boundaries. There was no limit to the type or quantity of resources that could be described by metadata, nor was there any limit to the number of cross-cutting metadata standards for each type of resource or subject domain. Today, for example, the Dublin Core Standard is widely accepted as the basis for the description of any type of document and the MARC file format for the exchange of bibliographic data between different institutions. In German-speaking countries, the Standardization Committee at the German National Library organises the use of uniform standards for indexing, formats and interfaces in libraries and decides on fundamental issues at the technical level.

The framework for digital research processes is fed from different sources

In the various scientific cultures – at different speeds – a gradually growing framework for digital research processes is emerging, which is partly based on established standardisation processes, partly driven by memorial and knowledge institutions, e.g. scientific archives, libraries and collections. The actors involved establish a basis for creating data in an already standardised way or migrating it into standardised systems in the future, or also offer translation rules between grown knowledge organisation systems. Since the 1990s, a variety has

9 European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), European Telecommunications Standards Institute (ETSI).

developed that corresponds to the breadth of methodological accesses, objects and forms of research.

Implementation still not coherent

In contrast, there is a lack of implementation in research practice in many areas, especially with regard to the application of standards in digital scientific documentation. In some cases there is a lack of feedback between the actors (for example between the standardisation committees of infrastructure providers, international expert committees and the scientific communities) or of sufficiently binding decision-making processes. Actors such as the globally active Research Data Alliance (RDA) are trying to remedy this deficit by advocating obligatory standards in the field of research data management.[10] In addition, relevant standards that make it possible to find and read information on the World Wide Web have an influence on scientific documentation.[11]

Standardisation of "data quality"

Orientation to "Total Data Quality Management"

Orientation points can be found in research and normative modelling on general quality management in management theory and business informatics. The work of Wang and Strong and the approaches of Total Data Quality Management, which also influenced the later elaboration of the FAIR principles (see 1.2.5), are still influential today. Here, data quality is defined demand-oriented, i.e. from the perspective of data use or "data consumption", and differentiated according to four characteristics:[12]

- *Intrinsic* data quality: Data have a quality of their own, for example, by being error-free, credible and objective.
- *Contextual* data quality: The quality of data results from its suitability for a context-specific purpose, but also, for example, from its relevance, timeliness and added value through linking.
- *Representational* data quality: Data quality is created when data is concise, consistent in its presentation formats, interpretable, and easy to understand.
- *Access-related* data quality: Data gain quality if they are accessible and editable and access to it is secure and will remain so in the future.

[10] The RDA was founded in 2013 "bottom up" from academia as a network of experts and is financially supported by numerous state and state-related actors. Its aim is to facilitate open exchange and re-use of data across technologies, disciplines and national borders.

[11] For example, the xml descriptive language (Extensible Markup Language) or the vocabularies on schema.org (https://schema.org/, last accessed: 30.08.2019). On the comprehensive standards and tools, see the website of the World Wide Web Consortium (W3C) – https://www.w3.org/standards/, (last accessed on: 30.08.2019).

[12] See Wang/Strong (1996) – What Data Quality Means to Data Consumers, p. 9 and p. 18 f.

A continuous quality definition, quality measurement and quality analysis should be carried out over the life cycle of data.[13]

Such requirements are operationalised even stronger in the ISO standard 8000 "Data Quality and Master Data Quality", published first in 2009, which originates from the area of eCommerce, or in the standard "Measurement of Data Quality (ISO/IEC 25024), which is part of a package of standards on software quality. The latter focuses primarily on the quality characteristics provenance, accuracy and completeness.

Limited dissemination in science

In science, this standardisation of data quality has found only limited resonance: The approaches derived from such standards can most likely be found in the quality assurance of cohort studies in medicine. Traditionally, formal assessments for data quality are also found in the engineering sciences. Also in the context of large research institutions such as CERN, in research with satellite data or in the environment of the large protein databases, corresponding assessments have been developed which have found resonance in the respective communities. Data-intensive businesses have developed approaches for explicit data governance in order to define responsibilities and formalised processes for handling data. Such concepts are to be found mainly where data form the business basis, as for example in the finance and credit industry. In science, formal procedures for data management have been developed within the framework of large longitudinal studies, which, among other things, have to fulfil extensive data protection requirements. Another variant of data governance can be found, for example, under the keyword "Good Clinical Data Management" in medical research.

Taken together, these approaches provide a direction in which later procedural sets of rules could move forward from science and with a direct scientific reference – such as the FAIR principles, for example.

1.2.2 VALIDATION AND CERTIFICATION

Certification's performance goal: establish trust in process quality

Certification procedures focus on the organisational or institutional implementation of quality standards. They lay down responsibilities. In general, these are conformity assessments which on the one hand create trust in the process quality of the generation, processing and storage of data. Also, certificates – e.g. in the form of quality seals – offer an orientation aid while accessing repositories and increase the willingness of researchers to transfer data. The

[13] Wang (1998) – Total Data Quality Management.

same applies to data from non-scientific institutions, such as statistical offices or social security institutions, which make their data available for scientific use. If they have a certification or alternatively an accreditation as a research data centre, this creates a clear advance of trust with regard to the validity of the data offered for scientific purposes.

In the field of digital information infrastructures, examples of successful, visible and generally accepted certifications include the international Core Trust Seal for "trustworthy repositories" (under the umbrella of the Research Data Alliance) or the accreditation procedure for research data centres at the German Council for Social and Economic Data (RatSWD).

Example: Core Trust Seal

The Core Trust Seal (CTS) is a peer-review-based self-evaluation process in which institutions evaluate their concepts and guidelines for data archiving according to a 16-point catalogue. In order to obtain the seal, an operator must, for example, document:

- which measures are taken to ensure the integrity and authenticity of the data as well as the long-term preservation of the interpretability of the data;
- which metadata standards are used (differentiated according to descriptive, structural and technical metadata); and
- the extent to which data is curated by the data archive.

This clearly relates to the quality of content data and metadata, even if data quality is not directly the subject of certification.

Example: accreditation by RatSWD

The accreditation of social and economic science data centres by the RatSWD focuses more on the accessibility of resources. It is also determined whether "data verification (for quality and consistency of the forwarded data)" is a task of the research data centre and which procedures are used. This is a milestone in the improvement of data quality in the economic and social sciences, insofar as "accessibility" and "(re-)usability" primarily refer to data which are collected outside of science and beyond scientific purposes, for example by statistical offices, social insurance institutions and public institutions for the regulation of the labour market. In the long term, accreditation by the RatSWD has led to comparable quality standards and thus to easier transfer of data and data sets collected in science and those collected outside the scientific domain. Some of the accredited research data centres have also been acquired certification as trustworthy repositories.

Table 2: Data repository certificates, sorted by frequency

	Name	Number of certified data repositories: [14]	Provider
1.	**Core Trust Seal (CTS)** Since 2017	62	Core Trust Seal Board, merger of WDS and DSA under the Research Data Alliance (international) umbrella
2.	**World Data System Certificate** Awarded until 2017, now Core Trust Seal (see no. 1)	55	ICSU World Data System (international)
3.	**German Data Forum Accreditation (RatSWD)**	32	German Data Forum (DE)
4.	**Data Seal of Approval (DSA)** Awarded until 2017, now Core Trust Seal (see no. 1)	31	DANS (NL) or Data Seal of Approval Board & General Assembly (international)
5.	**CLARIN Certificate**	27	CLARIN ERIC (EU), a research infrastructure in the ESFRI programme
6.	**DINI Certificate "Open Access repositories and publication services"**	6	DINI – German Initiative for Network Information (DE)
7.	**Nestor Seal for trustworthy digital archives** DIN 31644	1	Nestor Competency Network for Long-term Archival (DE)
8.	**Trustworthy Repositories Audit & Certification (TRAC)** ISO 16363	1	Consultative Committee for Space Data Systems (at origin), currently: ISO/TC 20/SC 13 Space Data and Information Transfer Systems (technical committee)

Source: Own illustration based on own analyses of data from re3data.org, as of 08.08.2019.

Worldwide only a few certified or accredited infrastructures

The number of research and information infrastructures that have been certified or accredited in a comparable way worldwide has so far been manageable: Of more than 2,300 registered data repositories in the database re3data.org, only a fraction is certified (see Table 2). This reflects the still low degree of institutionalisation and professionalisation, which was already diagnosed by the RfII in 2016.[15] The most project-based services simply lack the personnel to establish and describe the processes required for certification of quality assurance measures. In addition, CTS certification, for example, requires a university or university library to assume permanent institutional responsibility for a repository.

When it comes to the conformity of the data itself, there are a number of pragmatic approaches or tools for (automatically executable) compatibility tests. Its purpose is to enable users of services to check themselves quickly and

[14] Data repositories can have multiple seals, so each row does not sum to the total number of certified repositories (could be evaluated by the provider, inquiries welcome if required).

[15] See RfII (2016) – Enhancing Research Data Management, p. 26 ff., chapter 2.5.

easily, for example when uploading data. The automated validation of data and software is also used selectively in scientific quality assurance, for example in data archives or the peer review of publications.

1.2.3 RESEARCH DATA POLICIES AND DATA MANAGEMENT PLANS

Basic rules for handling research data

The actual research practice is often not reached by the setting of standards, definitions of standards or certificates. Here, other instruments of research data management (RDM), such as the so-called data management plans, are better suited as operationalised forms of guidelines and policies which serve as basic rules to handle research data in institutions or projects.

The beginnings of such commitments and guidelines date back to the 1990s. In addition to the voluntary commitments of internationally active research consortia (e.g. the Human Genome Project), similar requirements found their way into the guidelines for funding applications of research funding organisations or – starting from the Anglo-Saxon world – into universities. At the government level, too, there are now initiatives to regulate or improve the handling of research data, often in conjunction with e-Science, digitisation or Open Access strategies.

Open Access launches RDM policies

The increase in RDM-policies in recent years is due on the one hand to the fact that national research funding agencies took up the Open Access / Open Science paradigm relatively early.[16] The submission of a research data policy or data management plans are made mandatory step-by-step, for example in projects funded under the European Research Framework Programme Horizon 2020, and in some cases also in funding programmes of the DFG and the BMBF. As a result, the importance of specifications for RDM and the corresponding data management plans has increased significantly, at least in the area of data-intensive research institutes and university facilities that conduct a great deal of research with third-party funding. For the internal implementation of policies and plans within the organisation, extensive handouts, checklists and digital tools are provided and contact points for RDM consulting are being established.

Difficulties of operational implementation

The concrete implementation of the data management plans and other requirements for RDM – for example, making the data accessible or publishing the data – is usually the responsibility of the respective research project and

[16] On Open Science as a driver for the emergence of national and subnational regulations on research data and information infrastructures, see RfII (2017) – An International Comparison, p. 9. In Europe, for example, the Swiss National Science Foundation and the Norwegian Research Council have made requirements for research data management plans mandatory.

left often in the hands of the junior researchers or the third-party funded personnel, who must become familiar with RDM in a very short time. Apart from examples of subject-specific guidelines and voluntary commitments, the implementation of the rules laid down in RDM plans sometimes appears to be a (funding) policy-driven and externally imposed additional burden of research management that is not conducive to scientific work. In the intrinsic motivation of many researchers, RDM policies that contain concrete obligations for project implementation have thus far found little resonance.

Different depths of control

The RDM guidelines published to date show a great diversity in the depth of regulation with regard to the subjects covered. For example, not all regulations define the basic terms, clarify questions of data ownership or mention costs. More precise, however, are the statements on data management plans, on Open Access and on ethical issues. Increasingly, rules are being added to make primary data available to the public after an embargo period, which modifies the Open Access paradigm. Representatives of the Open Access idea criticise that research data guidelines only concern the disclosure of data and not the fair handling of data after publication. The fear of dishonest use by competitors (data parasitism) prevents many researchers from disclosing their data.[17]

1.2.4 OPTIMISING PROCESS CHAINS: THE DATA LIFE CYCLE

Plurality of concepts and focal points

Science theory and information science have developed circulation models as a reaction to the insight that research is based on research, in the sense that science permanently falls back on self-created knowledge and thus renews it. In order to illustrate the handling of research data – also from the point of view of quality – models of so-called data life cycles were developed from the mid-2000s onwards (for details of the quality challenges along the data life cycle, see chapter 2).

The data life cycle is a concept that describes the cyclical nature of the work with data of all kinds (including information) in its various stages in the process of scientific processing and use. The main steps of this cycle are data generation (e.g. measurements), data preparation, data evaluation/analysis, storage up to long-term archiving, as well as making data available through publication and subsequent use in further or new research contexts, which may also result from teaching.[18] Numerous variants of data life cycle models have emerged, which differ in terms of level of detail, subject specificity or operational objectives.

[17] E.g., Amann et al. (2019) – Toward Unrestricted Use of Data.

[18] On this, see RfII (2016) – Enhancing Research Data Management, Glossary, p. 76

Accordingly, the focus is set differently: on long-term archiving, the phase of data collection and evaluation up to post-use scenarios or also the clarification of responsibilities of professional actors and the division of labour between them ("data creator, data scientist, data manager and data librarian").[19]

Use value of cyclical models

Data life cycle models illustrate that data use and re-use generate new results in the form of research data. Data management along a life cycle must accordingly ensure that research results are reproducible at all stages. Furthermore, decisions have to be made at transition points between the phases of the life cycle (i.e. at the "interfaces") as to which data are to be stored, published independently as data sets or included in a publication, and how long they are to be kept available (see chapter 2). These decisions are currently e.g. made by research teams or individual researchers while managing their data, according to diverse criteria.

Operationalisation of the cycle in the context of research forms

An operationalisation by the professional associations according to at least rough relevance criteria in line with the respective form of research[20] would be helpful in this context, but is hardly available so far. For example, hermeneutic-interpreting forms of research still use non-digital media such as writings, images or natural objects as sources of knowledge production on a large scale today and will certainly continue to do so in the future. Their treatment in the data life cycle requires other efforts than the handling of measurement data from experimental research forms or survey data generated in the context of observational research forms. Also in the natural and engineering sciences, hybrid forms of digital measurement data and physical collections (tissue samples, drill cores etc.) play a role.

1.2.5 INTENDED USE AND GUIDING PRINCIPLES

The "fit for purpose" formula

A pragmatic short definition of data quality is the formula "fit for purpose", i.e. an orientation towards the purpose or intention of use. This central idea originally originates from quality assurance in industrial production processes and, at least in the Anglo-Saxon world, also constitutes a legally documented claim of the customer to the usability of a specific product that is purchased from a manufacturer.

Adaption of "fit for purpose" by science

The formula "fit for purpose" or "fit for use" is also used in many fields of science. In this way, data quality is defined comprehensively but openly as the

19 Swan/Brown (2008) – Skills, Role and Career Structure of Data Scientists, p. 1.

20 The German Council of Science and Humanities differentiates between a total of six research forms. In addition to those mentioned here are simulations, conceptual-theoretical and creative research forms. See WR (2012) – Empfehlungen zu Informationsinfrastrukturen, p. 35–38.

totality of properties and characteristics of data that make them suitable for a specific purpose. What is attractive about the blanket orientation towards the "purpose" (or the intention of the user) is that the context of (good) science or the question of methods, standards, etc. can be considered, but does not have to be specified in more detail. The short formula that data quality results from preparation for a particular purpose suggests, that the concept can be easily and operationalised in a suitable way. as Also, as a relational concept, it is in fact maximally flexible, since the "fitness for purpose" arises from the use and completely undefined user requirements, i.e. it can vary almost at will.

The FAIR principles and their operationalisation

The idea of a user-orientation with simultaneously open purposes also underlies programmes that demand pragmatic principles for the handling of research data in the sense of a voluntary commitment on the part of scientific institutions, but also of research projects. A prominent example of this are the FAIR Data Principles developed in 2014.

The four principles of FAIR (Findable, Accessible, Interoperable, Re-usable), which are summarised in the acronym, indicate pragmatic principles that must be fulfilled by sustainably reusable research data.[21] They aim in particular to improve machine readability.[22] The principles are clearly separated from the paradigm of "openness" in the current definition[23]: The FAIR principles can also be applied to data that is restricted for legal or other necessary reasons.[24]

The operationalisation of the FAIR principles by various infrastructure actors focuses to a large extent on the retrievability of research data and on the goal of enabling their use – i.e. on the "quality of services" (which provide the data). This alone is already seen as a clear benefit for science: FAIR data will be much easier to find across disciplinary and domain boundaries than in the past. In general, however, little is said about the scientific quality of these data (in terms of methods and research results based on the data). According to the FAIR principles, data should meet "community specific standards" in order to be "reusable" (criterion R.1.3, see table 3). Behind this requirement, however, there are complex conglomerations of open questions regarding the actual research practice in the respective communities, which demand further agreements and which cannot be answered by training alone (cf. also 1.2.1). How the criterion "reusable" could be specifically requested and implemented remains an open question.

---

21 https://www.force11.org/group/fairgroup/fairprinciples (last accessed on: 30.08.2019).

22 See Wilkinson et al. (2016) – The FAIR Guiding Principles.

23 "Open data and content can be freely used, modified, and shared by anyone for any purpose", https://opendefinition.org/ (last accessed on: 30.08.2019).

24 Hodson et al. (2018) – Fair Data Action Plan. Interim Recommendations, p. 15 f.

Table 3: FAIR Data Principles 2016.

TO BE FINDABLE:	
F 1	(meta)data are assigned a globally unique and eternally persistent identifier.
F 2	data are described with rich metadata.
F 3	(meta)data are registered or indesed in a searchable resource.
F 4	metadata specify the data identifier.
TO BE ACCESSIBLE:	
A 1	(meta)data are retrievable by their identifier using a standardized communications protocol.
A 1.1	the protocol ist open, free and universally implementable.
A 1.2	the protocol allows for an authentication and authorization procedure, where necessary.
A 2	metadata are accessible, even when the data are no longer available.
TO BE INTEROPERABLE:	
I 1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I 2	(meta)data use vocabularies that follow FAIR principles.
I 3	(meta)data include qualified references to other (meta)data.
TO BE RE-USABLE:	
R 1	meta(data) have a plurality of accurate and relevant attributes.
R 1.1	(meta)data are released with a clear and accessible data usage license.
R 1.2	(meta)data are associated with their provenance.
R 1.3	(meta)data meet domain-relevant standards.

Source: Force 11.[25]

Dissemination of FAIR

The FAIR principles do not in themselves require (and do not constitute) scientific data quality. They do, however, offer a set of basic criteria that are suitable for guiding a process with the accepted goal of broad data usability and data use. This is also the conclusion reached by a commission of experts set up by the European Commission. It points out that the implementation of the FAIR principles may lead to further requirements.[26]

Due to their catchiness and perhaps also their limitation to usability requirements, the FAIR principles have been adapted very quickly by research policy makers. In the EOSC Implementation Roadmap of 2018, for example, the implementation of the FAIR principles plays a central role, and in Germany, too, compliance with the FAIR principles has already been set as a goal for the National Research Data Infrastructure (NFDI).

25 https://www.force11.org/group/fairgroup/fairprinciples (last accessed on: 30.08.2019).

26 "FAIR is not limited to its four constituent elements: it must also comprise appropriate openness, the accessibility of data, long-term stewardship, and other relevant features." Hodson et al. (2018) – Fair Data Action Plan. Interim Recommendations, p. 3.

1.3 BETWEEN TOP DOWN AND BOTTOM UP: THE QUEST FOR SCIENTIFICALLY APPROPRIATE DATA QUALITY

Science-related data quality requirements

In science – unlike in the data-intensive economy – the demands on data quality must be set in relation to the performance of scientific methods, to research already carried out and to the possibilities of future, currently unknown research. This means that data in science must meet highly complex requirements in both the time and the material dimension: In terms of their potential for explication, they ideally offer both the possibility of connecting to fundamentally new developments and the constant reference to previous research (as a benchmark for the progress of knowledge). Due to these extremely demanding requirements in the scientific system, agreeing on "standards" for the quality management of digital research data requires a complex interplay of bottom-up initiatives from research itself and top-down consultations – involving institutional stakeholders who organise the governance of science and thus its operational framework.

Challenges for definition and control

Policies for the quality of data that do justice to research thus cover a broad spectrum of political requirements, provisions of research funding, guidelines of individual institutions as well as standardisation in specific scientific communities or for certain field-specific subject areas.

Overall, it must be noted that data quality in science is not only difficult to define in conceptual terms, but is also very difficult to control both in a self-organised form and through external (political) frameworks. Nevertheless, quality requirements for digital data are set or demanded by science and for science in research policy today. The explicit standardisation of data quality (DIN/ISO) is of less importance. In the complex world of digital research, it does too little and, as a rather hierarchical regulatory regime, is hardly appropriate to the decentralised and dynamic structure of research processes. By contrast, pragmatic standards such as "fit for purpose" could turn out to be excessively flexible. The politically enforced unification (standardisation) of data properties through the establishment of principles such as FAIR, represents a middle way for achieving commitment throughout the communities and domains. Aspects of standardisation are combined with a pragmatic approach ("fit for purpose"), which focuses primarily on (technical) usability.

Thinking beyond FAIR data

However, FAIR's main focus in terms of quality is on the aspects of machine-readability and – closely related to this – the better finding and accessing of research data. So, the quality of data services and access procedures often takes priority. In addition, further reflection is needed to take into account the specificities of very different cultures of research and working practices in knowledge production to progress scientific quality in the ongoing debate (see chapter 4).

2 DATA QUALITY CHALLENGES IN PRACTICE

2.1 IDEAL AND REALITY: DATA QUALITY PROBLEMS IN RESEARCH

The fact that digital conditions give rise to a wealth of quality problems, some of which are new, is not only demonstrated by the efforts to arrive at standards for research data management which could be operationalised (cf. Chapter 1). It can also be traced much more concretely along the data life cycle. This raises the question of what the reality of data quality problems in science actually looks like "underneath" quality and quality management models.

Data transformation as dynamic process

The data life cycle describes the phases of data management that accompany the work steps in the research process – ideally: from data collection to scientific publication and archiving, which makes the data available for renewed use. For the re-use of digital data, the idea is generally established that the data should be independently understandable and processable, i.e. without the involvement of the experts who provide them.[27] In research, of course, it is essential to obtain a minimum of contextual knowledge about the data in order to understand their genesis and to be able to assess their potential (but also their limitations) for re-use. Moreover, almost every step in research involves processes in which data undergo transformations. Data thus have, so to speak, step by step different "states of aggregation".[28] The dynamics of data transformations are contrasted with terms such as data product or intermediate product, which suggest that there are stable and lockable states in the various work steps. Nevertheless, these are also possibly fragile or subject to subsequent corrections.

Under digital conditions, which – quite suggestively – are also described as "data continuum", ideally every step of the research process, including its "intermediate products", should be reflected and documented as transparently and explicitly as possible: "Data must be linked in a way that ensures the continuum can be traversed."[29] This is a basic idea that also meets the requirement of verifiability of scientific statements and the substantiation function of scientific data. The fact that the research design from which data originate is also subject to requirements, research questions require a coherent context and methods must be used professionally are further aspects that can have a

[27] See CCSDS (2012) – Reference Model OAIS, chapter 3.1.

[28] All procedures that process a record (that is, an entity) from a data source to create a new record are associated with a "transformation" of that data. In fact, digital research processes are less robust on this point than traditional ones involving objects with a more sustainable physical identity that survives different process steps (i.e. is not based at every step on new information processing).

[29] Field et al. (2013) – Common Challenges in Data Management, p. 6.

practical influence on data quality within the framework of a research process. Also in this respect, requirements can be located in the data life cycle at least in an ideal-typical way.

Leveraging a data life cycle model to illustrate multiple data quality challenges

In the following, the RfII uses the model of the data life cycle with critical (from a scientific point of view: self-critical) intent. For this purpose, a variant of the model is chosen which provides for a phase of data sharing and archiving at a very early stage, since Open Access and reusability are currently at the focus of science policy measures relating to research data management and publishing,[30] but also because important sustainability interests of science itself are affected by the publishing, sharing and long-term availability of data (cf. the following Figures 2 and 3). The cycle also takes into account constellations in which researchers access data from others in order to validate or supplement their own collection of data.

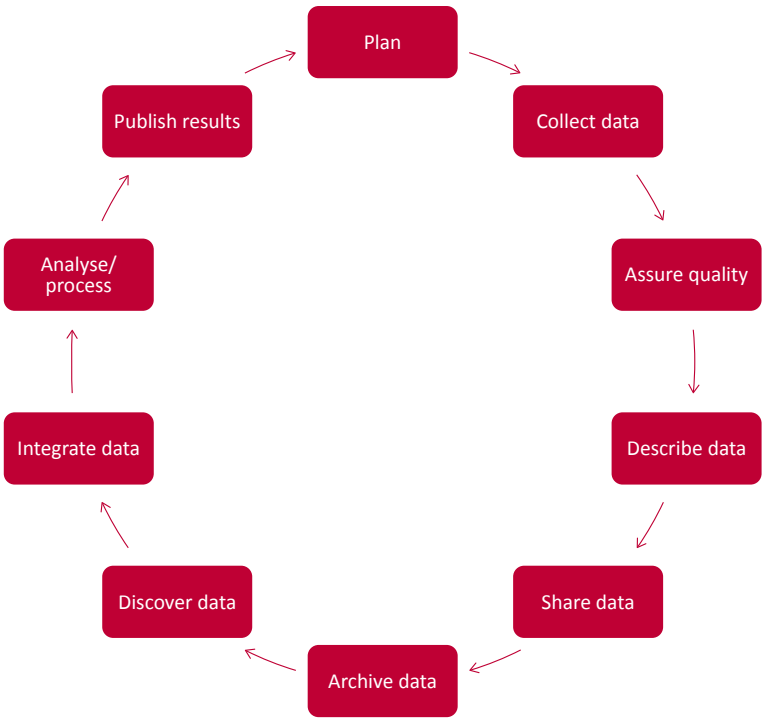

Figure 2: A Data Life Cycle.
Source: German Federation for Biological Data (2019).[31]

[30] See RfII (2019) – Statement on Open Data and Open Access.
[31] https://www.gfbio.org/training/materials/data-lifecycle/plan (last accessed on: 30.08.2019).

In this chapter, the data life cycle is examined in order to highlight problems which – under the conditions of digital change – in the day-to-day reality of research processes and research forms stand in the way of the implementation of ideal-typical quality objectives. In particular, the meshing of processing digital and non-digital data poses a challenge. For non-digital data will continue to exist in research processes – for example, physical objects, but also analogue recording methods and the habitualised intellectual techniques of the researchers themselves.

Quality ideals and challenges

In addition to the ideals of quality, the RfII proposes that challenges to the quality of data can also be identified through the life cycle. This will be attempted in the following – pragmatically and as realistically as possible. Due to the complexity of the research processes and the diversity of research forms, only a few highlights can be given to the problem areas mentioned.

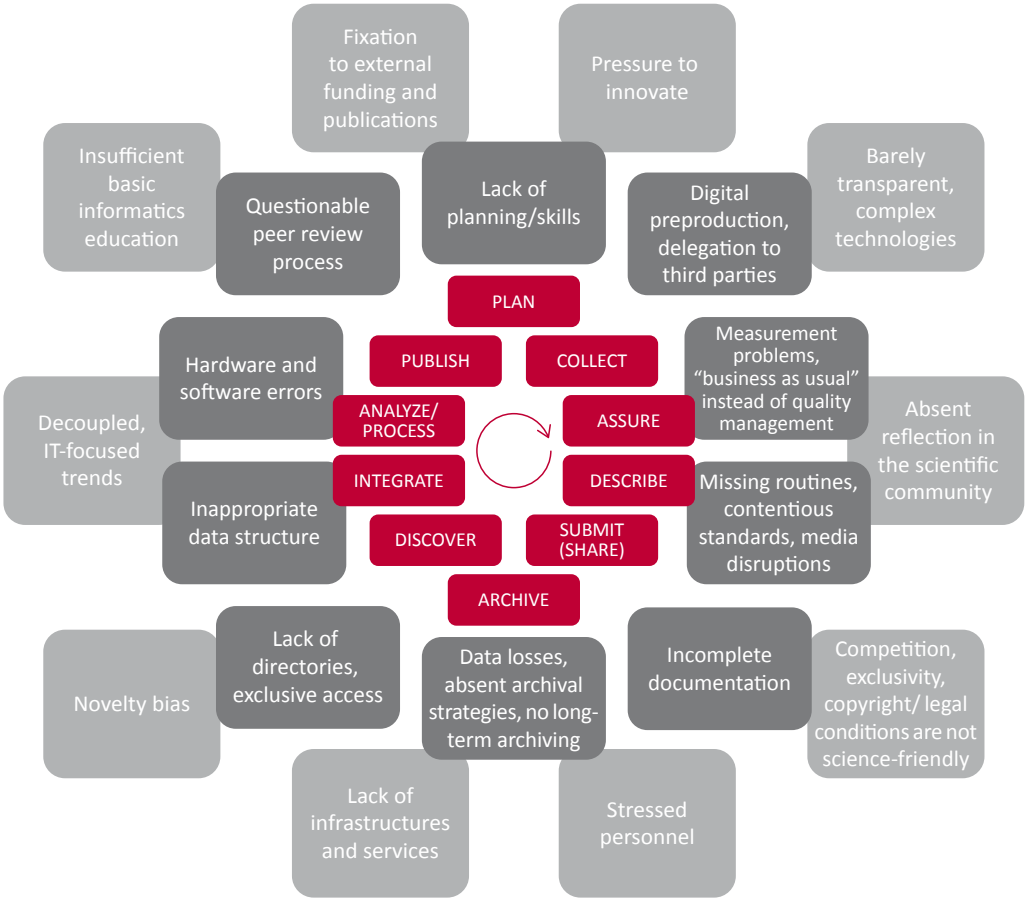

Figure 3: Challenges for Data Quality in the Data Life Cycle, a (Self) Critical View.
Key to figure 3
Inner ring: Problems and factors for data quality; middle ring: problems and aspects of data quality along the data life cycle; outer ring: obstructive framework conditions of science.
Source: Own illustration based on Figure 2.

2.1.1 COLLECTING DATA

Ideally, data are collected initially or, depending on research forms, are available from past collections or have otherwise been produced at some point. The methodological approaches and the research questions (e.g.: objectives of a study) differ, as do the pragmatics of data collection and the quality criteria taken into account. The basis can be empirical/observational practices, practices that "read" in the technical sense (i.e. transferring data) or practices that "read" in the hermeneutical sense. In some cases, completely new constellations have arisen as a result of digitalisation.

Digitality changes conditions for data collection

Automated data collection using digital sensor technology, the "empirical" reinterpretation of visual representations, graphics, texts, annotations, the production of simulation data, or the investigation of traces of use in digital systems (including machine-machine interactions) are examples of procedures that supplement established scientific procedures. For forms of research in which physical objects are traditionally captured by description, drawing or photography, automated methods of data collection, for example scanning or 3D mass digitisation, are available. Annotations can be assigned immediately and replace descriptions that have been added subsequently. Observation data have always been abstract representations, even in the context of empirical/observational research. Under digital conditions their quality depends on a number of device-related factors, which can also be described as forms of digital (pre-)production. Typical data quality problems are also already located here.

The following highlights illustrate some of the conditions for data collection that have been substantially changed by digitisation:

- *Non transparent (proprietary) instrument software:* Increasingly instruments are used whose results contain non transparent digital processing, at least where science does not have full access to instrument software. This is not only true in the classical research fields, where instrument-based measurements are traditionally used, but especially where observational data of third parties are used (e.g. tracking data). To the extent that the complex device parameters remain unknown company property, researchers are confronted with a "black box". Strictly speaking, only in a very mediated sense "scientific" results are obtained. Over changing software or device generations, data can only be scientifically valid and reusable if corresponding device knowledge or comparative studies are available. The "Guidelines for Ensuring Good Scientific Practice", which were updated by the DFG in 2019, give no indication of how research institutions and individual researchers should deal with this problem area.

- *Data of third parties with poor documentation:* Research also uses data collected by third parties. The data collection is not completely accessible for examination and the documentation of the relevant work steps may contain gaps that cannot be seriously filled, is missing or does not meet scientific standards without further research. Quality assurance tasks are similarly complex where research accesses data traces of a social network "freely" created in digital media by means of corresponding company services (e.g. programming interfaces). In some cases, commercial providers and platform operators are seeking cooperation with the scientific community for the purpose of data analysis and the improvement of algorithms, also because their own evaluations have proven to be inadequate or faulty.[32] Access that is granted, however, remains limited to a privileged circle of researchers – and these are often only allowed to work as contractually bound, so-called "embedded scientists", who, to varying degrees, are not fully informed by the companies about commercial filtering and selection mechanisms of data production or are restricted in a free analysis within the framework of disclosure policies.[33] Even in these cases, the data are not readily available for review, which diminishes their scientific quality.

- *Heterogeneous data models vs. "big data":* When collecting data, different data models (and thus, at an elementary level, already different technical, methodologically required or even research-related "logics") are confronted. Data models are developed by communities (for example, for the purpose of standardisation, but also for reasons of content), and they in turn influence the details of observation and data collection. Such a pre-configuration of data collection must be ingnored by broadly based, "overarching" Big Data approaches, in which structured and unstructured data are compiled and thus made explorable in the sense of a "collection". The quality of the data here depends on structuring processes of the actors, which are impossible to comprehend due to the sheer volume of data. In interdisciplinary research environments, complexity is increased by the fact that an answer to the question of (future) quality can only be found by combining differently collected data. The problem of (now: hidden) heterogeneity of origin also persists. This can lead not only to an unclear validity of statements, but also to artifacts in the analysis and requires appropriate attention.

32 For the negative example of Google Flu Trends, see the classic study by Lazer et al. (2014) – The Parable of Google Flu. In 2014, Twitter offered "data grants" to select research institutions that provided free data sets for research through a "certified data reseller partner." Facebook, together with academic partners, has recently developed an international cooperation model supported by various foundations, which, among other things, provides data for democracy research.("Social Science One", since April 2018, online under https://socialscience.one/our-facebook-partnership, last accessed on: 30.08.2019); see also King/Persily (2019) – A New Model.

33 See Pfaffenberger (2016) – Twitter als Basis wissenschaftlicher Studien, chapter 4.

- *Confusing multiplicity of ontologies/thesauri:* In observational and descriptive forms of research, controlled vocabularies are seen as a central means of quality assurance in the context of data acquisition (thesauri, standards data, ontologies). However, many of these vocabularies exist and research is now turning to sometimes very simple ("generic") standardisations. This is not conducive to the quality of the description of data when they are collected.

Prerequisite documentation of data in the various phases

In a digital working or research environment there are enormous ranges for the generation of data. Already in the stage of data collection, however, there are considerable documentation problems that stand in the way of a scientifically required comprehensibility of work steps. What is needed are documentation tools adapted to the procedures, which for example also integrate the records of electronic devices. Here, solutions are already available in individual communities to some extent; however, a comprehensive distribution is still to be achieved.

2.1.2 ASSURING QUALITY

Challenges for test procedures – is automation the solution?

An explicit quality assurance step in the data life cycle is a step in which data collection errors and "impurities" are eliminated. This is a genuine curation of the data. Quality problems can concern both the criteria and the instruments of such quality assurance measures.

- *Explication of criteria:* Selection, preparation and compression of the data can be based on assumed practices (depending on the context of data collection often also "manually") or on explicit criteria. Digitality forces a significantly higher degree of explication so that data can be processed by machine. Quality criteria can thus be applied more consistently on the one hand, but can also be more abstract (and thus unintentionally coarser) or riskier in application, for example where specifications are made for error tolerances or where "learning" (and thus changing) algorithms are used for quality assurance. Also, in scientific communities, there may at best be trends regarding suitable digital quality assurance steps and criteria – depending on the speed of digital change. These in turn must be chosen at risk. This is because they are only to a limited extent equivalent to the knowledge of (also) manual quality assurance, which has been established over decades, or they cannot be transferred in conjunction with the previously acquired knowledge.

- *Non transparent tools:* Automated testing methods can be used where data are no longer accessible due to the sheer volume of manual testing, or where this seems necessary for reasons of efficiency or is technically easy to do. Examples are logical checks for compliance with certain value intervals in tables, the so-called "missing data techniques" or automated validation tools (see 1.2.2). Where commercial service providers offer pre-programmed intermediate steps and science adopts such presettings untested, this can also have a negative impact on data quality.

2.1.3 ANNOTATING

Carefully documenting data is essential not only for data quality but also for traceability and thus the quality of research itself. At least in the early stages of data collection, the description cannot be completely separated from the purpose of the collection or the research method. Since in the digital world, data descriptions are again processed digitally, it has become common practice to refer to this additional information as "data about data", or "metadata". For metadata, research has formulated various, often discipline-specific and also transnational standards or metadata systems on an information science basis (cf. 1.2.1).

Metadata: Important added information – among others, for machine readability

The importance of metadata and of (possibly new) metadata standards for the quality of digital research data is rightly being emphasised.[34] In the digital world, additional information on machine processability must be added to data to a previously unknown extent. Automated processes do not actually work on data, but (solely) on metadata or "metadata about metadata", this is especially true for the field of Big Data. At the same time, it is obvious that a "complete" description of data is not possible and that a step-by-step improving documentation is necessary. Metadata therefore remain selective, changeable and are necessarily not complete. The practice is accordingly problematic.

- *Routines:* For the analogue world, science had established and progressively improved description routines for the most diverse types of data and media for centuries (tabulated measurements, protocols, transcriptions, directories for samples/archives, audio or slide libraries, library catalogues for texts, maps, images, etc.). Much of this is transferable to the digital world only to a limited extent, fits only partially, requires new digital tools or is simply not yet taught systematically.

[34] See among others DINI (2018) – Thesen zur Informations- und Kommunikationsinfrastruktur.

- *Unclear referencing data/metadata:* In database structures that have grown over a long period of time, it is not always immediately apparent what the referencing of data and metadata is (i.e. what exactly is being individualised). In a digitised image database, for example, do data refer to the physical object depicted or to representations (e.g. photographs, drawings, scans of the object)?[35] If the digital objects do not have a clear identification, problematic ambiguities arise in practice in the classification of the associated information. The copyrights of the described works may also be unclear. In a few cases, metadata can also be "works" in the legal sense, in any case if the texts are longer – which requires identification of the author(s).[36]

- *Provenance and transformation:*[37] Knowledge of the provenance, i.e. the origin or method of origin of the data, is essential information for scientific work. Data cannot be separated from information on software, codes or programming languages, often also on hardware. Corresponding descriptions increase the complexity of the annotation process. The acquisition of metadata for transformation steps of given data sets is even an issue throughout the entire data life cycle and not only during data acquisition and initial curation. The fact that the capture of metadata must be sufficiently static and dynamic in equal measure makes good documentation costly and requires a long-term organisational framework.

- *Manual and automated recording:* Where standards and good practice in writing are not agreed or explicitly formulated, it is up to the researcher, project or association itself to decide on recording methods, depth of coverage and scope of documentation. Metadata is often captured manually during the research process or is only determined afterwards (and then from other perspectives), for example in the preparation phase of a publication. This practice is considerably cumbersome and prone to errors. In the experimental sciences, research-supporting documentation software is increasingly being sought as an option, for example in the form of electronic laboratory books. In other research areas, the use of such tools is only just being discussed.[38]

---

35 To solve the problem, the issuing of a "persistent identifier" is currently recommended, such as a digital object identifier (DOI) or the Uniform Resource Name (URN) for Internet objects. However, how long the registration agencies currently operating in the market can maintain their service is an open question. The requirements for awarding a DOI are also under discussion.

36 On this, see Klimpel (2015) – Eigentum an Metadaten.

37 This refers to the translation, transformation or processing of analogue and digital data. Transformations are, for example, modelling, derived indicators or visualisations, etc. after measurement or survey.

38 See Peer/Green et al. (2014) – Committing to Data Quality Review, p. 275.

- *Versioning and dating:* Like data, metadata are typically subject to rapid changes in the digital world. Transformations of the data, error corrections or additional information for new purposes (e.g. archiving) make it necessary to update, date and version the metadata. Documenting the temporal shape of data, including the "datability" of changes in scientific processes, is a particular challenge of annotation.

- *Quality assurance needs for metadata:* With the increasing importance of metadata in the research process, it is necessary that they also become subject of quality assurance processes. If data sets are available in which different levels of information are combined, the question of the effort and benefit of concretely proposed description systems arises logically and pragmatically. Currently, description paths and languages are also multiplying. Transdisciplinary registers for the unambiguous identification of so-called digital objects do exist, but their institutional sustainability is not yet clear. Transparent and subject-specific, scientifically developed, proprietary classification systems are also confronted with competition from automated procedures such as "search engines" optimised for commercial purposes and linked open data applications, which may rely on other links.

Time needed for annotation competes with other research tasks

All in all, the description effort is a decisive hurdle in practice. Metadata, which enable uninvolved third parties to make informed subsequent use of complex data sets, go far beyond what is documented at the time of data collection. For researchers, the curation effort is in competition with other tasks; in fact, it can slow down research processes considerably. This is compounded by the usual, sometimes short-term changes of institutional connections in early career phases, which are common in science. The acceptance and willingness to annotate research data also depends crucially on the question of what advantages are associated with this for the individual researcher (e.g. reputational gains, citations) and whether he or she receives professional support. However, appropriately trained support staff and suitable documentation aids are often lacking in everyday work.

2.1.4 SHARING DATA

Data that is generated in the course of the research process, or is available in enriched and modified form, is made available or left to scientists in many pragmatic ways (data sharing).

Data sharing: from "peer-to-peer" to "Open Science"

The traditional form of data sharing is peer-to-peer. However, the digital turn is accompanied by further expectations of data availability, up to and including

Open Science/open data, i.e. making the generated data available as a kind of scientific (if not universal) commons.[39]

Sharing data increases the comparability of studies

In principle, data sharing enables innovative research and comparisons with similar research in the entire scientific system. It is also an essential basis for the application of fundamental principles of good scientific practice – including the (temporary) validation and falsification of research results. Nevertheless, for understandable reasons, parts of the scientific community find it difficult to comply with the requirement to make "all" research data freely accessible. Not for every form of research and not for every intermediate stage and (interim) result of research is the openness of the data basis mandatory. It is useful and necessary where the data as such can have added value for further scientific progress – provided that they are not only accessible for further use, but also prepared for further research in terms of reusability and connectivity. If this is the case, then such data and data sets represent independent scientific products, which, just like the publication of the results, are to be evaluated as a genuine scientific achievement.

Data products are scientific achievements

Data products can be presented using dynamic systems (data corpora, data collections) or can be archived and stored in different forms ("archive packages"). They can be empirically collected data sets, which serve for new analyses or are aggregated with earlier data sets of a similar kind. A combination of data with associated analysis tools or applications is also possible. The product can also be a digital catalogue, which as a description, dating and interpretation of data represents an independent scientific achievement. In the context of data centres, product formats such as scientific/public use files[40] and data reports[41] are already known. In the context of scientific publishing, the accompanying published data typically have a function as argument or evidence for the presented results (cf. 2.1.9).

Digitality makes it much easier to share data. Whether and how the phase of sharing is actually practiced in science is of importance, because important quality assurance steps may take place here. Nevertheless, practical problems can be observed:

---

[39] The Knowledge Exchange initiative based on Whyte/Pryor specifies six data sharing modes that range between the poles "Private management" (sharing data with colleagues within a research group) to "public sharing" (making data available to any member of the public); See KE (2014) – Sowing the Seed, p. 22; Whyte/Pryor (2011) – Open Science in Practice, p. 207.

[40] For explanation of concepts, see Research Data Centres of Statistical Offices; https://www.forschungsdatenzentrum.de/en (last accessed on: 30.08.2019).

[41] An illustrative example are the standardised data reports of the Geoforschungszentrum Potsdam, which undergo an internal review before publication, see http://dataservices.gfz-potsdam.de/portal/about.html (last accessed on: 30.08.2019).

- *Access-related quality:* the various data products differ in the way they are accessible and processable, and whether access to them is and remains secure. Self-published data tend to be more vulnerable in terms of long-term availability. If the data products are published online, they may not be "open", i.e. available in licensed, machine-readable formats.[42] This makes further use more difficult (see 2.1.7). Furthermore, the majority of data obtained in the research process remains at the place of origin and is hardly accessible to third parties. This complicates re-use (see 2.1.7).

- *Contextual and presentation-related quality:* Data must be provided with information on the collection methods, processing and curation steps, tools and systems used and thus ultimately also statements on important aspects such as data integrity and authenticity.[43] In practice, this step is often complicated by the lack of agreed metadata standards and IT-supported tools (cf. also 2.1.1 to 2.1.3).

- *Data protection and rights of disposal:* The sharing of data in conformity with data protection law may require special processing steps that restrict the analyses of the data for certain purposes. In addition, many researchers are not clear how they can make their data accessible while at the same time safeguarding their interests. Many databases do not have a transparent marking of the rights or licences attached to the data. Furthermore, rights of disposal over data are often not sufficiently clarified or indicated.[44] This is particularly unsettling in areas where the data may have a commercial benefit. Objections by other project participants, questions of liability or undesirable forms of appropriation by third parties are a cause for concern.

- *Difficult quality assurance for "pure" data publications:* Data centres or research (data) infrastructures have established routines of internal quality assurance for products that are published on their own initiative. In publishing, so-called data journals[45] have become established in recent years, which publish articles about datasets ("dataset articles" or "data articles"). The regular publication of data takes place by parallel delivery of the data sets to data repositories or data centres. For reasons of capacity, quality assurance is often only carried out on a cursory basis (cf. in detail 2.1.9 and 3.1.2).

---

42 The definition of an "open work" includes, for example, an open license. An "open work" should also be downloadable via the Internet without charge and must be provided in a form readily processable by a computer; see https://opendefinition.org/od/2.1/en/ (last accessed on: 30.08.2019).

43 Individual data centres have developed exemplary documentations. See the "Product Types and Processing Levels" by European Space Agency ESA, https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/product-types-processing-levels (last accessed on: 30.08.2019).

44 Lauber-Rönsberg et al. (2018) – Rechtliche Rahmenbedingungen FDM.

45 Examples of data journals can be found under https://www.forschungsdaten.org/index.php/Data_Journals (last accessed on: 30.08.2019).

- *Own scientific reputation vs. collective curation:* Independent data publications are a form of digital edition, which can, however, be dynamic. One challenge is to version scientific achievements as "editions" and to assign them to individual researchers. In analogy to traditional publishing, concepts such as authorship and data citation (or "referencing") are discussed. On the other hand, it is argued that ideally data (products) are improved over time and through use.[46] Especially in the ongoing maintenance and enhancement of existing data sets there is considerable potential for data quality. However, this process potentially has many participants over a period of time in which there is not necessarily continuity in terms of personnel. The challenge of managing such a "continuous improvement process" is largely unsolved.

- *Unauthorised dissemination:* Cultures of sharing are under pressure, especially where peer-to-peer processes are exploited by third parties, through unfair competition or other forms of scientific misconduct, through appropriation by economic actors, through distorting media (e.g. net-public) reporting or through industrial espionage (see also Chapters 2.2 and 3.2).

Scientific culture changes slowly

All in all, data sharing offers a wide range of opportunities for scientific work – whether on a small scale (peer-to-peer) or in the form of well-designed, quality-assured data products with traceable processes. Nevertheless, the scientific culture is only slowly changing here,[47] and it is yet not protected by public policies towards the creation of suitable, exploitation-free spaces for scientists. Thus, a lack of appreciation for the creation of data products in comparison to classical publication and also an unclear, unprotected status of sharing come together. Moreover, the sharing of larger amounts of data or data in need of explanation can be quite costly for the data producers. Surveys on open data and data sharing among researchers also suggest that researchers prefer to arrange their data with a view to publishing the results only at the end of a project. In the best case, the provision/archiving of the underlying data is then carried out at this point in the data life cycle. Cultures of sharing are generally fragile. However, where the research process requires an early sharing of data – which is quasi part of research practice – this step is also lived out in early phases.[48]

[46] Parsons & Fox (2013) – Is Data Publication the Right Metaphor?, p. 39 f.

[47] Backed by various studies, among others KE (2014) – Sowing the Seed; Fecher et al. (2015) – Academic Data Sharing; Wouters/Haak (2017) – Open Data; Stuart et al. 2018 – Practical Challenges in Data Sharing.

[48] Examples can be found in research on large-scale instruments, such as the analysis of satellite data in astronomy and Earth observation, as well as in life sciences or text related research.

2.1.5 ARCHIVING

The long-term preservation of digital research data is an intensively discussed topic. If it is not practiced professionally, there is a risk of data loss, partly because digital storage media are short-lived compared with other media and the software versions on which data were created are outdated. If long-term archiving is practised, the effort and costs are likely to be considerable in view of rapidly increasing data volumes and a multitude of technical problems. It is also unclear what commercial partners can contribute to sustainable scientific development in the field of long-term archiving.

Long-term archiving as complex task

Archiving requires its own routines for recording data packets, converting them for secure storage and making them available again for use. An internationally accepted reference model (OAIS) already exists for this purpose, but it can only be implemented in a specialised institution with the appropriate personnel. There are considerable implementation problems on a broad scale.

- *Period of backup:* The interdisciplinary requirement for the storage of data in Germany is usually ten years.[49] The preservation of the primary data in question must be performed institutionally at the place of origin and includes both analogue and digital data as well as their combination. Neither the standards required for this nor the costs (in the digital sector) incurred, let alone a system of binding responsibilities, has evolved in the scientific system. Guidelines for research data management have so far assigned abstract formal responsibilities, but not concrete individual responsibilities. There is therefore likely to be a gap between the ideal and reality, even with regard to the minimum storage period. Especially for the historical sciences, a minimum requirement of this kind is no solution; solutions for genuine long-term archiving (storage for eternity) are indispensable, at least for a selection of digital artifacts.[50]

- *Data selection:* While rules for archiving and selection exist for the handling of documents from administrative processes, such transparent specifications are missing in many scientific fields. Which data must be archived and which data can or even must be deleted? Only with transparently formulated

[49] See DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis, p. 22, Leitlinie 17. This period is criticised for many fields as clearly too short, others consider it too long. Compared to the DFG's 2013 memorandum, the recently updated guidelines in the main text (Guideline 17) today speak of a "reasonable" period of storage. Only the explanation refers to a standard period of ten years for "raw data", but this can also be shortened in justified cases.

[50] Among other things, the RfII had already stimulated a expert discourse for differentiating between storage closely aligned with project duration and over significantly longer customised archiving periods. See RfII (2016) – Enhancing Research Data Management, p. 39 f., Recommendation 4.3.

rules can it be clearly stated how representative, meaningful and worth preserving an archived data stock is.

- *Rules and planning:* In many cases, the reality on the part of the data producers is that data is collected and processed without sufficient knowledge of the needs for long-term preservation, and the resources to proceed differently are also lacking. On the part of the data centres, the capacities and often necessary specialisations to curate the heterogeneous data stock of the diverse research processes are lacking. This is where demands for data governance, research data management plans and the development of an archiving strategy come into play. Often, however, the plans show how little of this is realistically feasible and can therefore be required.

- *Lack of infrastructure and services:* Although every research institution needs an archiving strategy and archiving facilities for both physical artefacts (samples etc.) and data in order to be able to store them for the minimum required period of ten years in accordance with the guidelines of good scientific practice, this task is often only partially solved due to a lack of infrastructure, depending on the location. The access of communities to suitable services also varies depending on the degree of self-organisation, the establishment of digital methods and international networking.

- *Curation and quality assurance:* A lack of personnel leads to insufficient preparation of data and metadata in the archiving process. In particular, harmonising the structure and semantics of the archived data sets can be a labour-intensive task in individual cases. However, both of these factors are essential in order to efficiently integrate data from different sources (for information on the requirements for data integration, see also 2.1.7). Many tasks require knowledge of the scientific domain and personal contact with the data producers. Often, basic metadata is not maintained until it is added to the archive and must be collected subsequently. If the data is used repeatedly, any changes or corrections to the data set must be documented and cross-references to the respective uses must be entered. Data archives and repositories in particular can make important contributions to the ongoing curation and maintenance of data corpora if they are appropriately staffed.

- *Technical infrastructure:* The quality of the basic infrastructure (hardware, operating systems, networks, security, standards, performance, data migration) also influences data quality. The necessary ongoing adjustments to the technical infrastructure can hardly be made, especially in smaller institutions. It can be similarly problematic if the necessary agility and willingness to innovate is not available in the personnel area. A further risk is the loss of data in the course of (re)storage, media and format changes (cf. 2.2).

From individual to collective responsibility

The long-term archiving of scientific data is not yet satisfactorily solved for the scientific system, both in organisational terms and with regard to available capacities (storage, costs). The National Research Data Infrastructure (NFDI) currently being established in Germany will only be able to provide partial solutions to the issue of long-term archiving. In view of considerable organisational and institutional shortcomings, shifting the archiving obligations to the shoulders of individual research, as is currently happening, is not expedient for ensuring data quality in the long term.

2.1.6 DISCOVERING DATA

Technical possibilities and disciplinary hurdles in finding data

IT-supported procedures enable information to be found in new and much more comprehensive ways than before. This opens up the possibility for science to use existing data collections for new research questions and analyses in addition to scientific literature. Thus, in the digital age, it is above all access to data that determines the depth, breadth and quality of scientific knowledge.

- *Searching and finding:* The potential that lies in the easy retrieval of data often remains unused, especially because existing collections are oriented along disciplinary boundaries and ultimately towards individual research communities. Efforts are being made to integrate these objects into existing catalogue systems and directories, but the scope of these efforts only covers part of the data landscape.[51] Despite broad agreement that good science is based on good data, scientists often fail to find the data. An additional complication in individual cases is that data in cooperation projects (or other contexts such as expeditions) are stored in archives that differ from one another in terms of subject or institution. This impairs their coherence and retrievability. This could be technically solved by interoperability of information systems and links between clearly identified data objects, as required by the FAIR principles, for example. In current practice, however, this is usually not the case, so that it is often impossible or at least difficult to retrieve data sets that belong together.

- *Access:* Often the associated scientific literature is the key to finding a data set. If the data is not published in a repository or otherwise, a personal request to the data producer is necessary (sharing data peer-to-peer, see 2.1.4). This is not always promising – contact persons may have changed, the preserving organisation cannot exercise its rights of disposal (or they

[51] The dataset search introduced in late 2018 by Google could help move this situation along. See Cousijn/Cruse/Fenner (2018) – Taking Discoverability.

are unclear), etc. Access to scientifically interesting data from the corporate sector is also often a matter of negotiation (see 2.1.1). In the case of sensitive data, for example personal data from medical or social science research, accessibility is further complicated by the legal framework for the protection of personality. Organisational models for access to such protected data exist, but they are neither consistently distributed throughout the scientific community nor homogeneous.[52]

2.1.7 INTEGRATING DATA

Linkability requires professional standards and data integrity

The combination of data from different contexts is especially attractive for scientific analyses. However, in individual cases this step involves considerable effort. Even if the data can be found and their use is permitted by law, both proprietary data formats and the often heterogeneous structure and semantics of the data make their aggregation and integration difficult. However, where data integration is practiced, important improvements in contextual quality, including the original data, may be achieved. Practical problems mainly concern questions of data structure and integrity.

- *Structure and contents of data:* The structure of the data (for example in a database) is often not the result of targeted planning, but has been pragmatically developed over time using existing methods (see 2.2). The contents of database fields, or more precisely their semantics, are particularly problematic. Exact definitions of terms as well as the relationship of terms to each other in the form of ontologies is essential for linking data sets. In some cases there are no technical specifications, in others – at the other extreme – there is a confusing multitude of applicable specifications and metadata standards, which makes it difficult even for specialists to build up widely available data sets. The research process and infrastructure management are not well integrated at this point.

- *Data integrity:* It is often technically not clear, for example, whether the integrity of archived data has been ensured during storage. In addition, there may be deficiencies in the experimental design which, for example, prevent statistical evaluation, as well as undocumented pre-processing which changed the data at an early stage (see 2.1.1). In the worst case, shortcomings from earlier phases of the data life cycle can hinder the synthesis of new knowledge.

[52] See the practice of research data centers in the social and economic sciences: RatSWD (2018) – Activities Report 2017.

- *Manual effort:* If data is not available in machine-readable form, integration requires a great deal of manual effort with a corresponding susceptibility to errors – for example, when data is extracted from the scientific literature and transferred to a database.

The linking of databases holds great potential for science. However, there is still a considerable discrepancy between the theoretical technical possibilities and the effort required in practice for such integration. In some cases, work is being undertaken in the form of qualification work (here the structuring of data is part of the learning process), in other cases entire projects are being designed around the integration of relevant data sets, for example in the agricultural or environmental sciences.

2.1.8 ANALYSING AND PROCESSING

Causes of data tampering

Problems in the analysis and processing of data are quite similar to those in the stages of data collection and quality assurance of data which were discussed earlier. Just as today's digital measuring devices generate algorithm-controlled transformation processes when collecting and measuring data, the algorithms in evaluation software and tools influence the results. The analysis of the behaviour of algorithms in complex application scenarios is itself still an open field of research. For a detailed examination of the possible causes of data tampering, a rough distinction can be made between hardware and software problems.

Hardware as a source of error

For many manufacturing processes, especially with regard to their replicability, differences in the computer hardware used already play a role. This affects the long-term storage of data, for example in the case of the readability of storage media. But also for processing procedures, hardware differences must be precisely documented and their effects made as controllable as possible. As examples can be mentioned:

- *Hardware errors:* The results of calculations may depend on specific implementation errors. This is partly due to bad software implementation, which only "randomly" shows the desired behaviour on a hardware, but also partly due to "correct" or known behaviour of the hardware.[53] If calculations depend on such factors, they cannot be reproduced on unaffected hardware. Especially for extensive calculations, errors in the memory chips used become more significant. ECC RAM devices that can detect and correct such errors

[53] An example is the "FDIV bug" of Intel's first generation Pentium processors, which in certain circumstances resulted in significantly lower floating-point accuracy.

within certain limits have become standard for servers and mainframes, but are less common on desktop computers and notebooks. Hardware errors can also be a security risk, so that, for example, the targeted manipulation of calculations by external attacks is possible.

- *Lack of emulation:* Programs developed for older hardware cannot easily be executed on newer hardware. Where the use of historical software is required, the programs have to be re-implemented, converted or made executable by emulators, which causes a lot of detail problems.[54] Where software for individual workstations is not purchased but licensed, the difficulties increase when the hardware is replaced.

- *Generation of artifacts:* Computer hardware differs, for example, in the methods used to implement floating-point arithmetic and to generate entropy/random numbers. Therefore, calculations on different hardware are not necessarily reproducible.

- *Physical environmental influences:* Especially in the case of converting analogue to digital signals (A/D conversion), physical environmental influences such as temperature, pressure, humidity, electromagnetic and radioactive radiation, but also vibrations and acceleration influence the processing. This can lead to increased noise, calculation errors and system crashes – for example due to "tipping bits" – which is why hardened hardware components have been developed for use in special environments.

- *Ageing:* In addition to the importance of subtle differences in environmental influences and the hardware components used, their ageing must also be taken into account. For example, even beyond ever-present artifacts and noise sources, sensors change their properties through aging and wear, which can lead to shifts in accuracy in different parts of the measured spectrum.

Professional computer centres take this into account. However, even with good hardware management, there remains a residual risk that must be determined and, if necessary, described.

[54] Worth mention here is possibly the impressive Visual6502 project that develops simulators for historical processors on the transistor level: http://www.visual6502.org/ (last accessed on: 30.08.2019).

Importance of software for data quality

Even more clearly than the hardware, the software used for data processing plays a decisive role in data quality:

- *Implementation errors:* Faulty implementations not only endanger the stability of software, but also the reliability of the data and analysis results obtained. The correctness of software cannot be guaranteed theoretically and can only be estimated with considerable effort, even with practical limits.[55] Implementation errors and lack of software maintenance can lead to loss of information and artifacts in data processing, which reduces the quality of the processed data.

- *Version differences:* Results of data processing and analysis can already differ across different versions of the same software. The problem is aggravated if identical procedures are provided by different software packages, as far as more complex processing procedures may differ in relevant implementation details. The differences relevant here are difficult to explain.

- *Blackboxing:* Research data processed using proprietary software and the results obtained from it can only be reproduced or replicated with complete documentation of the procedures used and possibly only by using the original software itself. In practice, the documentation of the procedures used may regularly prove to be insufficient. Moreover, results can only be reproduced with reasonable effort on the basis of the research data using the original software, if at all, which impairs the scientific verifiability.

- *Machine learning/learning algorithms:* The use of learning algorithms (artificial intelligence, machine learning) also limits the traceability and repeatability of calculations. For example, training processes can only be reproduced exactly when the original training data are available. As a rule, results cannot be explained satisfactorily because the models algorithmically derived from the training data are difficult or impossible to interpret by humans. Thus a detailed reconstruction of individual calculation steps is theoretically possible, but does not provide satisfactory explanations in practice. The explainability of results from such algorithms – meaning the explanatory and substantiating description of calculation results – is a dynamic field of

[55] The verification of software with regard to user requirements (validation) and formal-logical correctness (verification) have practically and theoretically narrow boundaries set for them. While verification is particularly sensitive to logical indecision problems, validation requires the most complete but practically difficult to achieve explication of user requirements, which moreover can shift continually throughout the entire software life cycle. See Liggesmeyer (2009) – Software-Qualität.

research. However, neither usable tools nor convincing theoretical concepts for explainability and comprehensibility are available so far.[56]

- *Security gaps:* Programs that are used to process and analyse research data are affected by security vulnerabilities just like other programs. In particular, this can disrupt the processing of data and impair its integrity and confidentiality. The confidentiality of research data plays an important role, for example, in compliance with legal regulations (data protection law, copyright/licensing conditions).

- *Application errors:* In addition to operating errors, unclear, poorly documented or unsuitable parameterisations (definition and selection of categories) should be mentioned here in particular. Even with correct operation and suitable parameterisation of the software, implicit knowledge can be expected to impair data quality, since implicit knowledge is usually not documented. For example, this can affect the deviation from a default setting, which is always used, but which is not documented. The quality of the processed data also depends on the fit of the algorithms used. In individual cases, this selection may follow dynamic trends and "fashions" in information technology rather than a requirements analysis based on the respective research question ("what is there and fits to some extent").

Documentation in the process stages can make sources of error transparent

Many of these problems cannot necessarily be solved by improved documentation, but at least they can be made transparent. The use of analysis tools that are open source and freely available to all scientists is another currently common approach in many scientific disciplines. However, this only applies to analyses that do not demand too much performance requirements. Furthermore, problems of reproducibility between different versions and hardware environments remain an open problem. The ongoing collaborative development of scientific community software, ideally including quality assurance procedures as they are established in some – mostly simulation and data intensive – disciplines, is a more far-reaching approach to create transparency and reproducibility.

[56] See for instance Lipton (2016) – The Mythos of Model Interpretability or Samek et al. (2017) – Explainable Artificial Intelligence.

2.1.9 SCIENTIFIC PUBLICATION

Publishing: not just results – also data

A prominent station in the ideal-typical data life cycle is the publication of the scientific results after completion of the analyses. Associated data are also made publicly available as evidence. This practice is promoted in the field of scientific journals by the specifications of the publishers. For example, the respective author guidelines may recommend additional publication of data or require explanations on data availability.[57] Another driver is the research data policies of the research funding agencies (cf. 1.2.3).

First step to solution: enhanced publications

- *Models:* A number of models have been developed in which data are published, for example, on enclosed data carriers (in the case of monographs) or as a supplement (for example in PDF format). Approaches to compile analysed materials in databases are also increasing. Around 2010, the term "enhanced publications"[58] was coined for a publication in which links can be established between the research publication and the underlying data publication via the so-called "Digital Object Identifiers" (DOI) in the relevant databases. The enhanced publication should at least help to improve the traceability of the presented results, but also to improve the subsequent use of the data by third parties. With regard to data quality, the additional contextual information on the data and the research question is helpful. Problems with data quality can be caused, for example, by the fact that the published files are public but can only be converted into a machine-readable form with manual effort. The accessibility is thus limited and the risk of transmission errors arises in the event of subsequent use. Current efforts are therefore moving in the direction of archiving data in repositories instead of supplement publications.[59]

- *Data origin (provenance):* In most cases, associated data are published in an aggregated or processed form, i.e. it has already undergone a series of transformations, or it is already selected data. Quality problems in enhanced publications can be caused by the fact that transformations that the data have undergone from the time of collection to the final published status are not documented. Possible sources of error are not traceable then. The

57 An example is the overview of different policy levels on the open data information page of SpringerNature, https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/12327096 (last accessed on: 30.08.2019). Journals of other publishers follow a similar line.

58 https://www.forschungsdaten.org/index.php?title=Forschungsdaten-Policies&oldid=3619 (last accessed on: 30.08.2019).

59 See Call of the Coalition for Publishing Data in the Earth and Space Sciences (already signed on to by a number of publishers), http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/ (last accessed on: 30.08.2019). Similar considerations can be found in other subjects.

software used is also relevant for assessing the quality during the further development of data sets, but in practice it is also rarely available.

- *Static vs. dynamic:* The current practice of enhanced publications means that the published data sets are largely static: Problems or errors identified later can be published as an erratum or noted in the metadata of a repository item. It is more likely, however, that such information remains undocumented, particularly because the persons involved leave the scientific community and, as a rule, no provision is made for the ongoing maintenance of data once it has been published.

- *Compliance:* The scope of the data provided and the accuracy of the documentation may be limited if researchers can only devote a limited amount of time to the publication of data due to other external factors, such as pressure to produce further articles or publications. An increase in data quality could be achieved by prioritising data preparation as part of scientific publishing – possibly at the price of a smaller number of articles or publications produced in one research unit. This is not yet effectively anchored in established reputation systems and is one of the common requirements for good research data management.

- *Quality assurance:* While dissertations, professional articles and book contributions undergo a standardised peer review or other forms of quality assurance, this is often not the case for the associated data sets: Many repositories allow files to be uploaded without further checking of their content or the question of whether they have undergone an internal review before publication. Also in the so-called data journals, a detailed technical review is usually omitted, it is simply too costly (see 2.1.4 and 3.1.2). Beyond the documentation function, the value of many of these data products may be low.

- *New barriers:* A data publication, which is primarily a supplement to the classic article or monograph, remains in the classic publication system and may "inherit" the findability and accessibility problems associated with the established subscription system.[60]

[60] Critiqued in Parsons & Fox (2013) – Is Data Publication the Right Metaphor?, p. 40. The publication of data as evidence for a scientific publication within the classical publication system does not contribute per se to more Open Science.

2.1.10 PLANNING

Challenging practice in research data management

Research data and publications are the basis for planning and applying for new research projects and data collections. As the data life cycle progresses, it becomes clear that data quality issues build on each other: Missing information from early phases, such as data collection, raises additional quality problems during archiving, analysis or publication. For this reason, there is an increasing demand to plan the documentation, management and archiving/publication of data already in the conception phase of a research project or to record them in a readable form (data management plans). The ideal – data management professionally planned and implemented – is confronted with practical problems which also have a negative effect on the assurance of data quality.

- *Acceptance:* Policies of the research funding agencies – and consequently also of the research institutions (see 1.2.3) – have not only increased the awareness of the topic of data management within the scientific community, it has also had a structural effect in the form of newly established RDM advisory offices or the introduction of supporting tools. However, there are also reports from RDM advisory centres that considerable efforts are necessary to win individual researchers for engagement in the creation of data management plans – there is obviously little demand for advisory services in this area. Formal requirements and the additional effort are mentioned as reducing acceptance.[61] The reason for the lack of acceptance of the planning task may be due to the often overemphasised rationale that data would have to be curated with a view to possible re-use by third parties. This is not considered a priority by all researchers.

- *Allocation of resources:* For a long time, research funding itself did not always consistently follow its own fundamental requirements in its approval policy. Funds for planned research data management and the associated curatorial expenses were not always considered worthy of support. In the basic implementation and realisation of the data management plans, one also encounters the fundamental problem that these activities are permanent tasks and not temporary activities. As an instrument, data management plans only have a quality-assuring effect if the associated tasks are organised and expenses are financed.

61 For example, Neuroth et al. (2012) – Langzeitarchivierung.

2.2 DATA INTEGRITY THROUGHOUT THE DATA LIFE CYCLE

Transitions and transfer points are important

Tracing the data life cycle has shown that typical factors in each phase can negatively affect the quality of data. Some of these are research data specific, others generic. Furthermore, it became clear that similar challenges occur in different phases, such as the documentation of transformations, changing media and formats or the creation of descriptions. In this respect, the transitions between two phases are particularly important, which can also be transfer points from one actor to another (for example, data producer/archive).

Data cycles must match research forms

Looking at the entire data life cycle, two aspects in particular are of importance: the partial perception or significance of the data life cycle for certain scientific and research communities and the forms of research practised there, and the important issue of data integration over the entire life cycle.

Objective: data integrity

In empirical/observational forms of research, the data life cycle is often shortened by leading directly from data collection to analysis and publication. The curation of data is not the main focus here, the importance of ongoing documentation is underestimated, so that information on the provenance of the data and the transformations made can be incomplete. In hermeneutic/descriptive research processes, data are often described in a minimalist bibliographic manner and are directly transferred to long-term preservation without data publication. Digital methods of analysis are only applied selectively. For memory institutions such as collections, archives and libraries, the focus is on the phases between "collecting" and "searching". One of the central challenges for the professional communities is therefore to analyse how exactly they deal with the data life cycle and where they are currently located.

This is all the more important as good science requires that the traceability, accuracy and consistency of data be maintained and guaranteed throughout its life cycle. The scientific object must remain unaltered, the substrate of the research work must be presentable and should also remain materially stable (as free from ageing as possible). "Data integrity" is a keyword in this context.[62] Data integrity is a decisive aspect for the design, implementation and use of any system that stores, processes or retrieves data. Questions of risk management, validation as well as legal issues, for example data protection, come into play here. Due to the multitude of actors (including non-scientific actors) and the extent of technological complexity, digital science is also confronted with challenges from the point of view of data integrity:

[62] Among other things, the German Council of Science and Humanities has made recommendations on the "integrity" of data processes, an organisation-wide and therefore particularly complex task for universities. See WR (2015) – Empfehlungen zu wissenschaftlicher Integrität.

- *Format changes:* In the course of processing data, these are regularly converted into other data formats, including both file formats in the narrower sense, which are processed by software components, and agreements on the use of data formats in the broader sense, for example, the coding of numbers and characters, the naming of fields in tables or formatting. A number of factors can affect data quality: The complexity or incommensurable paradigms in information modelling can affect the quality of encoded data when converting from one file format to another. Proprietary file formats may not be read and written completely or correctly by applications. Data loss due to compression procedures is also known. Image, audio and film data in particular are often stored in compressed form (e.g. as JPEG, MP3, MPEG) in order to reduce storage requirements. Due to multiple processing steps on compressed files (which are usually prone to data loss), the losses accumulate, so that such file formats are only suitable to a limited extent for long-term processing and re-use of data.

  Within individual processing steps, data formats can impair the quality of data. One example is the encoding of characters that cannot be processed correctly due to an incomplete implementation of the Unicode standard, for example, in the case of sorting. Another example is the use of unsuitable colour space models during digitisation.[63] Also described are the restrictions and properties resulting from floating-point arithmetic.[64]

- *Fragile data security:* Digitisation makes research data easily accessible. For example, as soon as data are shared, they can be comparatively easy violated and become object to “attacks”. The more openly science is accessible (which is a common goal in the digital age), the more it exposes itself and its data not only as a target for access, but also as a target for attack. Possible motives are compromising researcher’s and research institution’s reputation, the integrity of their research, as well as industrial espionage, piracy, robbery, but also other forms of criminal sabotage. Even military scenarios are conceivable. At present, the scientific system in Germany (beyond the usual protective measures) is not sufficiently prepared for this type of danger.

[63] For example, color space models in which high-resolution color values are designed for a coarser color space. A color space model with only 256 shades of gray is not sufficient for the processing and display of X-ray images.

[64] The widely used standard IEEE 754 for the presentation and processing of floating-point numbers has some unfavourable properties for certain calculations, which can lead to inaccurate results, in particular due to rounding errors. In multi-stage processing, the rounding errors can add up.

However, data integrity does not yet include all the necessary measures to protect against unauthorised modification. In its position paper PERFORMANCE FROM DIVERSITY, the RfII previously recommended that the responsible stakeholders should put much greater emphasis on technical-organisational measures for data security.[65]

While larger computing centres practice data integrity concepts as part of their mission, many smaller or self-organised research and information infrastructures are not capable of attaining this level of professionalisation.

2.3 INTEGRATING RESEARCH PROCESS AND DATA LIFE CYCLE

Contrary to what the illustrative model of the data life cycle with its steps and phases suggests, data management in the concrete research process proves to be a path with numerous hurdles. On the one hand, data quality arises in the research process and requires professional expertise. On the other hand, special expertise is required for numerous (information technology, technical and legal) interfaces.

Organisation at the interfaces

The complexity and controllability of the multitude of influencing factors and interfaces proves to be a challenge. There are various scenarios of division of labour in the process – between "people and machines", between "suppliers and customers", but also between researchers and science support personnel, who are partly active at the research site and partly in infrastructure facilities. Many of the tasks require specialisation. On the other hand, outsourcing of research data management tasks to specialised units and institutions also implies a certain dependency, or at least a partial relinquishment of control and responsibility on the part of the researchers for essential parts of their work. A further challenge is to make the very heterogeneous models and approaches for ensuring data quality usable for the respective concrete research task. Local and individual solutions are conceivable here. For science as a whole, however, consensus-building processes in the learned societies are ultimately needed to ensure that the solutions scale in a way that high quality is achieved in all research fields. Thirdly, adequate resources are needed. Ensuring and increasing data quality throughout the entire research process involves very comprehensive documentation tasks. Some of these can be facilitated with the help of software. However, well-trained specialists are indispensable along the process chain.

[65] RfII (2016) – Enhancing Research Data Management, p. 55 f., Recommendation 4.12.

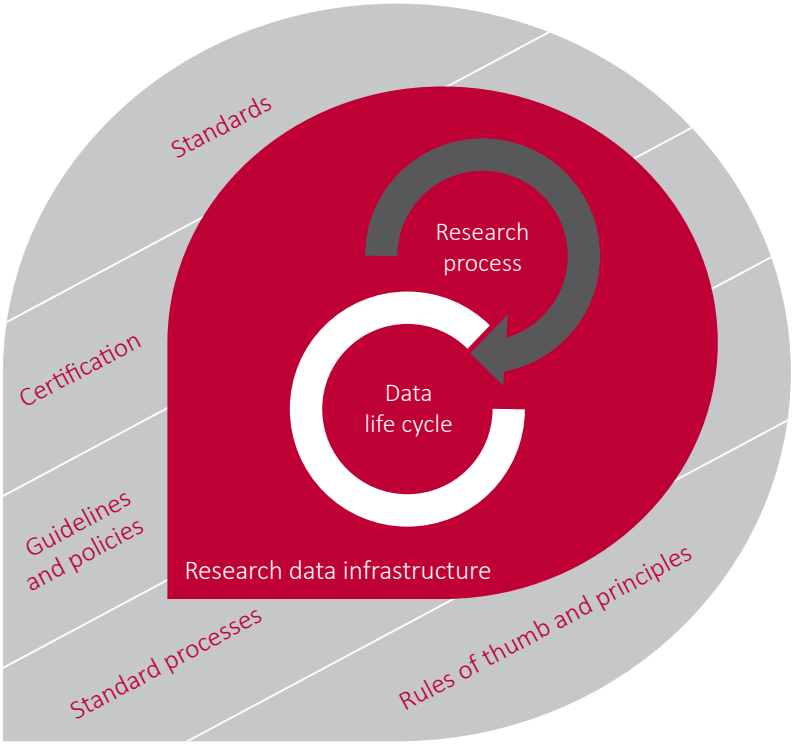

Figure 4: A multidimensional model of data quality.
Source: Own illustration.

Interlinking data life cycle and research process

The impression that the simple cyclical model gives the scientists the possibility to pass through the research data life cycle on a completely predetermined path without any problems in order to secure and improve data quality is an illusion. The data life cycle occurs alongside the process cycle of the respective concrete research tasks and the research personnel involved, i.e. the institutions; together, both form a multidimensional space with manifold cross-references between the two process cycles (cf. figure 4).

Research data management as a task at several levels of action

If the necessary regulatory level is taken into account, research data management with regard to data quality actually proves to be a complex multi-level task. Sustainable progress requires activities on the part of individual researchers, supported by an organisational and infrastructural environment, as well as on the level of (international) scientific associations, where the more fundamental discourses on methodological standards and scientific quality criteria take place. The RfII makes recommendations on this in chapter 4.

3 DATA QUALITY AND THE SCIENTIFIC SYSTEM

The description, enhancement and assurance of the quality of research data can follow heterogeneous theoretical models and approaches, as the analysis in chapter 1 showed. Just as diverse and multifaceted are the factors and processes that can increase, negatively influence or reduce data quality in the reality of digital research today. This was outlined in chapter 2 along the phases of the data life cycle.

Quality discourse and critical developments in the scientific system

A third view subsequently incorporates further developments in the scientific system (and its environment), which contribute to a critical examination of research performance and returns. The quality of the data produced in the context of scientific knowledge production is also affected. Several crisis-like tendencies are currently under discussion: There is a debate on the replication and reproducibility crisis with regard to published research results, of overloading the review system and of misplaced incentives which seem to be triggered by reputation measurement, introduced in many disciplines by means of publication indicators.

Drastic change in the publishing system

In the same way, the publication system, which is organised on a market-basis, is changing at a rapid pace: on the one hand, science is trying to publish data itself, while on the other hand, data companies are appropriating large market segments of the classic scientific publishing sector and data are disappearing behind payment barriers, which in part are not in the interests of science. Political demands, which postulate a fundamentally "open" publication obligation for data from publicly funded research, also directly affect the question of their quality.[66]

Digitisation joins existing crises as a systemic challenge

Some of the developments mentioned have already started before digitisation appeared on the scene as a "systemic" challenge for data quality in science. However, the digital options further aggravate the previous crisis developments. At the same time, digital change – at least at the technological level – also provides instruments that can help science to improve quality and emerge stronger than before from the publicly discussed "crises" phenomena. However, this requires not only innovations at the technological level (such as testing software or similar), but also the organisation of the entire scientific knowledge and knowledge valorisation process. There is no way around a systemic embedding of the data life cycle in order to answer the current "crises" and the question of what aggravates or "drives" them. Such a holistic embedding

[66] Plan S was intensively discussed in 2018. On the quality issue in the framework of implementing the PSI Directive (relating to public sector data) (2019) – Stellungnahme aktuelle Entwicklungen Open Data. On the PSI Directive in general, see chapter 3.2.1.

can be said to exist if the data life cycle as well as questions of harmonisation and standardisation of quality criteria are not considered in isolation, but in the context of both the research process and the interaction of scientific institutions (universities, non-university research and specific infrastructure facilities). The interdependencies induced by digitisation also include previously unknown non-scientific and commercial influences and access to science. They also include new, non-scientific forms of valorisation of and for research data, which science itself can only influence if it deals with this changed environment in a rapid and proactive manner.

Increased political attention to data quality

However, science must also be enabled to do so by the creation of appropriate science policy frameworks. Previous efforts to improve data quality in newly implemented information infrastructures – above all data collections created from concrete research, but also institutional repositories – have often been short-lived due to temporary project funding opportunities. Furthermore, the generation of high data quality was also taken for granted in science policy for a long time. Only recently has the topic of "sustainable research data management" attracted greater attention, which is currently reflected, for example, in the establishment of a National Research Data Infrastructure (NFDI) and similar efforts to shape the European Research Area (ERA), especially the European Open Science Cloud (EOSC).

Crises and drivers call for new arrangements in favour of data quality

In the following, crises and drivers that affect data quality in science are briefly discussed, although they are partly caused by the changing nature of digitalisation. Not only in the opinion of the RfII, their effects on scientific research have reached a magnitude that makes the creation of new arrangements in science and for science a priority task of the present – precisely because "quality" and thus also data quality is essential for the added value of scientific method-oriented work. This is a challenge for individual and collective action of researchers as well as for science funding and science policy.

3.1 CRISES AND DRIVERS IN THE SCIENTIFIC SYSTEM

The crises mentioned here include negative developments arising from science and the procedures for the production of scientific knowledge itself – which have become critically aggravated in recent years in view of the exponentially growing production of data and data availability.

3.1.1 LACK OF REPLICABILITY AND REPRODUCIBILITY OF RESEARCH RESULTS

In disciplines and fields of research in which replicability is methodologically possible and necessary, non-repeatable data analyses and experimental set-ups

are at least in a grey zone of scientific knowledge. "Replication crises" [67] are therefore defined as the phenomenon whereby scientific studies, experiments and simulations claim statistically significant correlations and causalities that cannot be confirmed in subsequent studies based on data generated under almost identical conditions. It is obvious that there is a problem of legitimacy here, especially for publicly funded and trusted science.

Research data play an important role for the replicability or reproducibility as a basis for statistics and other forms of scientific analysis. The quality of the data collected is by no means the only decisive factor. Also important is the way in which they are collected and handled, especially the quality of their (further) processing in the research process. Along the process chain of the individual research steps, transformation effects can easily occur, which, for example, under even slightly changed conditions in the infrastructure or the analyses tools (e.g. different hardware and software versions), can lead to results that are not reproducible.

Uncontrolled effects of data processing compromise replicability of studies

Thus, for example, it can be questioned whether numerous initial data are not already based on an experimental design in the course of the survey, which inevitably cannot be replicated in the statistical analysis and thus cannot lead to sustainable scientific results. Is it a case of having limited the sample size in the interest of making faster or cheaper progress? Do data come from invalid data sources even? Have unreliable service providers been used for data development and processing? Have the data been altered (possibly even without being noticed due to blackboxing effects) without this being made transparent and replicable for third parties? Or is working with "bad" data unproblematic if it is subsequently only "correctly" decoded by digital means? Are algorithms used whose behaviour can produce uncontrollable effects from a scientific point of view? Does a database software really allow the documentation of everything that needs to be documented – and if not, how to deal with compromises that have to be made? How to deal with translation problems between computer languages at "interfaces" and with negotiation processes in interface programming in general? And: Is there enough funded time in a research project for quality assurance and metadata creation?

Data intransparency can be removed by explication and binding to reference objects

In particular, questions as to why exactly in individual cases (possibly under the pressure of digitally accelerated, more confusing circumstances) not much is

[67] The RfII does not take sides in the debate about the adequate description of this quality problem – either as a replication, replicability or reproduction crisis with different demands on the conditions of repeatability of a study or experiment. For the Council, these terms express the same crisis tendency. To this extent, the terms are used here synonymous or are described as "repeatability".

happening indicate a new discourse on quality in science, which to a greater extent than before – and without standards having already been found – concerns questions of data quality and the process quality of their processing. The discourse on replication under digital conditions is one that is by no means easy to conclude or even arbitrarily to decide. Even within individual disciplines, the question can lead to frontline positions as to where exactly the limits of one's own claim to replication or reproducibility lie – in other words, where the methodologically "bad", the sloppy or the obviously error-prone organised (data) practice begins.[68] In particular, digitisation promotes a high degree of non-transparency: What do the many, enormously complex presettings look like, which have always been invested in digital processes in advance, but are never explained? And how is reproducibility to be ensured when not only a lack of explicability but also only "weak" forms of reproducibility are often the norm? The fact that, in the course of the digital turn, it is possible to access data material from other researchers and research groups stored in repositories in a relatively naive manner, in some cases without any further documentation of the contexts of creation and processing, has further fuelled the impression of a replication crisis even beyond experimental research (for example in the case of clinical data or in the area of making digital copies of cultural assets available). It is therefore all the more important today to link research results back to well-documented data in collections – at least where this is possible. This expressly includes analogue or physical collections and collection objects, which must be excellently documented by digital metadata in order to validate the research results based on their analysis. Where this happens, collections can be real drivers of good scientific practice.

3.1.2 PROBLEMS OF THE PEER REVIEW SYSTEM

In many scientific fields, peer review is established as the standard for the quality assessment not only of articles but also of the bibliometric "value" of a journal. The amount of peer review processes has increased enormously over the past decades – not counting the numerous evaluations and project reviews that researchers and scholars also have to deal with. The distribution of a growing number of submitted manuscripts among a relatively constant number of peer reviewers leads to an overload of the review system, which makes the timely publication of research results more difficult, but also threatens the quality level of peer work. If reviews become more demanding due to data sets that have also been submitted and are also to be reviewed, either the waiting period is extended further or the quality standard of the review that can be performed

[68] On this also see: DFG (2017) – Replizierbarkeit von Forschungsergebnissen.

under these conditions decreases. A "plausibility check" will in most cases be the only possible option for reviewers. The same applies to research results whose data basis can hardly be properly assessed due to its sheer quantity, for example in the field of complex simulations or high-resolution electro-microscopy.

High requirements for "data reviewing"

The requirements for the quality assessment of data are often much more complex than those for texts (see also 2.1.4 and 2.1.9). In case of doubt, reviewers must be able to fully understand the data extraction process, including the concrete digital processing steps. Researchers must provide very extensive documentation for this purpose. In a sense, the review crisis and the so-called replication crisis are critically intertwined.

Vague guidelines for review

The problem is discussed at various levels, both in terms of guidelines for peer review and alternative forms of evaluation. It is clear that technical and scientific peer review of data sets is costly and does not scale in terms of capacity given the rapid growth of published data sets. In addition, where data sets are already included in a peer review of scientific results, the review guidelines are often unclear. It is not uncommon for publishers and editorial boards to use review guidelines that are borrowed from the familiar field of results publications – originality of results and argumentation, contribution to technical/scientific progress, etc. – but which mislead and discourage both the researchers who submit data for review and their peers, who conduct the review. Guidelines specifically designed for data review, which rather contain criteria of accuracy of data collection and processing, the level of documentation and statement of metadata and – closely related to this – the requirements of interoperability of the data for further use (for further research as well as for the purpose of validation of the research already done with the data) are available, but are hardly used in scientific publishing.[69]

Data reviews as a stand-alone form of publication

A variant of the peer review is the emerging scholarly "data reviews" as an independent form of publication. Here, peers (subjectively) evaluate data collections or products from a scientific point of view.[70] In some cases, data repositories or other research infrastructures also take on partial tasks of quality control prior to publication, or take full responsibility for them if they make their own data sets publicly available. All of this only partially covers the landscape. Under the keyword "re-use is the peer-review of data", the idea is cultivated

[69] Carpenter (2017) – What Constitutes Peer Review of Data.

[70] For example, the Digital Humanities have created its own review journal, dedicated to digital editions and resources, providing a critical forum for talking about them (including a list of criteria); https://www.i-d-e.de/publikationen/ride/ (last accessed on: 30.08.2019). Further examples can be found in economics and social sciences, but rather as sub-sections or headings in relevant journals such as the Journal of Contextual Economics – Schmollers Jahrbuch or the European Sociological Review.

that the intensity of re-use can serve as an indicator for high-quality data sets. However, this does not work in both directions: Little use of data sets is not an indicator of poor quality and, in case of doubt, quality problems that are not immediately apparent cost the re-user a lot of time or have a direct impact on the quality of the research based on them. The idea of "data reviews", which are published independently on a data set or a resource in the sense of scientific review, appears to be more trustworthy.

Collective curating of data sets

In line with the networking idea, a division of labour or collective curation of data sets is also conceivable, as organised by the Wikidata project, the Geo-Wiki platform for nature observation or – based on the commitment of academy members – with a view to curating historical German-language literature, the German Text Archive (DTA).[71] The inherent problem here seems to be the formation of a community that can and wants to reliably perform such a task over a long period of time. Overall, however, it can be stated that the limitations of the peer-review system in the area of data publications are currently still largely tolerated. Stricter conditions for authors, the consistent use of technical aids or a reform of the reviewing system on an international scale, which is indispensable for this purpose, have not yet been implemented.

3.1.3 UNDESIRABLE DEVELOPMENTS IN SCIENTIFIC PUBLISHING

Disincentives from quantitative research performance measurement

Behind the overloaded review system looms a quantitative overstretching of the publication output as a whole, driven by various factors. For a long time, misguided incentives in research funding and financing have led to a focus on quantity in science, including forms of informal performance requirements ("three publications per year"). Researchers in the qualification phase, in particular, prefer the rapid exploitation of a result in several stages to a more comprehensive, well-documented publication of results. On the other hand, a well-documented and machine-readable data collection as a supplement or even as an independent edition complementary to the publication of the results would require more or additional time and manpower. Another driver of this quantity-oriented development, which should not be underestimated, is the increasing international competition. In particular, catching up "science nations" are providing targeted incentives to increase the publication output for "their" scientific institutions and researchers, for example, in order to improve their position in international rankings and thus demonstrate their increasing performance (as reflected in these indicators).

[71] https://www.wikidata.org/wiki/Wikidata:Main_Page, https://www.geo-wiki.org/, http://www.deutschestextarchiv.de/ (all last accessed on: 30.08.2019).

Improving quality by reducing publication output?

While on the one hand, capacities have to be invested in reviewing, the increasing number of publications is accompanied by the justified concern that the quality of the publications themselves will decline, if only in terms of the time and effort that the authors have been able to invest. Researchers have therefore been discussing for some time now a significant reduction in the volume of publications in the sense of concentrating on a few but very high quality publications.[72] In terms of science policy, too, a scandal involving the falsification of cancer research in 1998 and plagiarism scandals for doctorates in 2011 have led to a critical reflection on the relationship between quantity and quality of scientific publications.[73] In 2015, the German Council of Science and Humanities once again impressively formulated a fundamental criticism of the current publication practice of "publish or perish" and its consequences.[74] The DFG has reacted to the excessively quantitative self-representation of researchers by drastically reducing the number of "own" publications to be listed in third-party funding proposals. The evaluation procedures and specifications of other research funding agencies are also moving towards requesting only a selection of the best (or directly relevant) publications as proof of past merit. However, attempts to make adjustments of this kind have not yet led to a fundamental overcoming of the sometimes one-sided, quantity-oriented publication behaviour, as the relevant international university and science rankings show. Moreover, reviewers are now informally obtaining the overall presentations that are no longer officially required in order to get an idea of the situation for their decision. A "cultural change" is not yet apparent here.

Data review may increase pressure in the system

If, in addition to the publication of results, a (possibly independent) publication of research data is required, the pressure on an already overstretched system increases. A competitive market for data publications following the already established pressure to show "publication activity" would have a direct impact on both researchers and reviewers if these publications have to be produced "on top" as another form of research output. By way of an all too pressing demand for the publication of data, a risk is emerging of aggravating the misguided development that has already occurred and not achieving the desired reproducibility and reusability of results.

---

72 In a survey conducted by the DHV, almost 82% of the participants agreed with a claim to halve the number of scientific publications. The claim was formulated by Helga Nowotny, former President of the European Research Council ERC. See The German Association of University Professors and Lecturers, Barometer, https://www.hochschulverband.de/frage-des-monats.html?&tx_jkpoll_pi1%5Bno_cache%5D=1&tx_jkpoll_pi1%5Buid%5D=118#_ (last accessed on: 30.08.2019).

73 Kleiner (2010) – Qualität statt Quantität.

74 "The German Council of Science and Humanities reiterates that untraceable quantities of publications counteract the very meaning of the publication obligation, which was originally used to communicate and verify new research contributions by the scientific community. All stakeholders are called upon to promote the long-term necessary change towards a more qualitative research evaluation and thus a reduction in the volume of publications". WR (2015) – Empfehlungen zu wissenschaftlicher Integrität, p. 32.

3.2 CRITICAL EFFECTS OF UNSUFFICIENT FRAMEWORK-SETTING FOR SCIENCE

Consequences of internationalisation for data quality

In view of the diverse international co-operations and the resulting global interdependencies in politics, law, business and public life, the framework conditions of science and its organisations are also changing.[75] Complex international relations raise new questions that go far beyond the fundamental issue of digitalisation and concern the ideal of free access to education, knowledge and cultural assets as well as the international networking of research. Science operates globally, but it must nevertheless take national boundary conditions into account. In a changing political landscape, for example, are mirror servers advisable for data backup in another country or on another continent? What role do global, commercial providers of data space (or cloud services) play? How should national scientific institutions and organisations position themselves in relation to the demand for "openness" of data, if accessible data is transferred to commercial use in other countries, possibly appropriated and/or enriched there and then no longer openly available?[76] How can research data that circulates freely be effectively protected against manipulation and sabotage and, conversely, how can researchers and scholars be protected against the inadvertent reception of digitally generated false information? The "open" paradigm can itself be in crisis here, because it presupposes two things:

a. a methodologically reliable attitude on the part of the users of research data and thus a globally identical and possibly idealised norm of "scientificity",

b. comparable (and internationally connectable) national efforts to build infrastructure for truly quality-tested "open" data.

Asymmetry in access to research data

The currently predominant form of output-oriented scientific competition, on the other hand, can lead to asymmetries in access to high-quality research agendas if individuals first gain a competitive advantage by using the disclosure of other people's data for their own benefit, but either do not use the model of "openness" at all or only superficially.

Call for data sovereignty

Furthermore, the economic use of "open" research data under the condition of international competition can lead to distortions of competition: Individual nations use the research data made available, but keep their own under lock and key. This problem is discussed in connection with the establishment of

75 See WR (2018) – Internationalisierung von Hochschulen.

76 It is necessary to check here whether this can be effectively countered with appropriate licences or whether this has other "side effects"– see Creative Commons Licence CC-BY-NC, https://creativecommons.org/licenses/by-nc/3.0/de/ (last accessed on: 30.08.2019).

the EOSC. For Germany, "data sovereignty" was demanded in this context; the RfII took up this keyword primarily in the sense of a sovereignty of the German scientific system with regard to its own research data.[77] In its statement on current developments in the area of Open Data and Open Access, the RfII calls for political strategies for a system of governance which offers scientists and researchers, but also economic actors, security of action with regard to the sharing and use of research data.

3.2.1 VAGUE REGULATORY FRAMEWORKS FOR "OPENNESS"

Search for legally compliant standards

Numerous legal regulations are currently being adapted to the new conditions of a digitised world. For Germany and Europe, the research and economic policy side is striving for "openness" of publicly funded data sets. As a result, the scientific community is also faced with new questions of knowledge and proof of data provenance, authorship, security and protection of data (as well as demands for "ownership" or even property rights to data). For researchers, the search for standards that also meet legal requirements begins with the question of how they can and must document their scientific achievements in the collection and curation of data sets, also with a view to long-term preservation (cf. 2.1.5). When are rights violated in the use of other data and how can one protect one's own data from misuse or harmful use? The release of data also has complex legal implications. Since digital data can be used in multiple ways, the image of "publication" may be misleading.[78] "Use" can be many things. Who may then have access to the data and how?[79] The general uncertainty of a legally compliant handling of data is also recognised as an obstacle to good scientific practice. A real crisis phenomenon is that researchers, due to a lack of tangible legal contours, often pass on their data at will or even let it get lost and, especially when using commercial (free or inexpensive) data services, do not even bother with the "small print". In this way, research data flow into the sphere (and eventually ownership) of large data corporations and into other zones of uncontrolled re-use.

Research data as a regulatory subject

The influence of external regulations on research activities is also made tangible in the relevant data policies of the European Union. According to the PSI Directive adopted in June 2019[80], research data are for the first time also included in the public sector data, which in principle should be non-discriminatory and re-usable under transparent conditions. The aim is to make publicly available

77 RfII (2016) – Enhancing Research Data Management, p. 30.
78 Parsons & Fox (2013) – Is Data Publication the Right Metaphor?
79 See Lauber-Rönsberg et al. (2018) – Rechtliche Rahmenbedingungen FDM.
80 EU (2019) – Neufassung PSI-Richtlinie 2019/1024/EU.

research data from government-funded research usable for further application, especially for commercial purposes. In addition, however, the Directive also requires the Member States to submit their own strategy for Open Access to research data. In a statement, the RfII has welcomed the associated intention to achieve harmonisation in the area of research data repositories which are now commonplace in Europe.[81] The RfII recommends that instead of quantitative growth, the qualitative goal of high-quality data stocks and data services should be pursued. Undifferentiated "publication obligations" – for example, also for all preliminary intermediate products on the way to a result – would not correspond to the specific performance requirements and the responsibility for quality of science. In this sense, "openness" cannot be an end in itself.

3.2.2 DEPENDENCE ON COMMERCIAL PRODUCTS AND SERVICES

The increasing dependence of the scientific production process on digital services, infrastructure products and tools, which usually come from commercial providers, also drives the discourse on data quality. There is no public debate (yet) of a crisis caused by black boxing. However, the digital turn gives new and changed weight to the fact that research must of course often work closely with the manufacturers of special equipment.

Cycles of hardware and software obsolescence are accelerating

The cycle of obsolescence of research equipment, which has always been well known in the natural and engineering sciences, is clearly surpassed by the much faster-paced cycle of software obsolescence. Similarly, software is only partially created or optimised for scientific use. Moreover, evolutionary processes in the software sector often take into account complex interdependencies of entire software worlds; corporate strategies are correspondingly volatile and hardly transparent for science as a customer. This has far-reaching implications, for example, for the difficulties described above in the replication or reproducibility of studies and experiments (cf. 3.1.1). The use of different software versions on the same devices often leads to differing results. Therefore, many institutions continue to operate old systems with some effort (for hardware and software problems see also 2.1.8).

Negotiation of conditions of use as a collective task of science

The documentation of research data sets always includes information on the device and software versions used - for all processing stages in the data life cycle. This requirement can hardly be met in practice if there is a dependency on the documentation of the com-mercial providers. Access to commercial software documentation for scientific purposes is rare. Ideally, the software

[81] RfII (2019) – Statement on Open Data and Open Access.

used in research should be designed and published according to scientific criteria – which is especially the case if the product in question was developed in close cooperation with researchers or by scientists themselves. The problem could also be alleviated by negotiating better conditions of use for the products of commercial partners. However, both are collective tasks; individual researchers often have neither the time nor the resources for this in their everyday research.[82] Sometimes there is also a lack of awareness that minimising dependencies (and, where this is not possible, documenting them) makes an important contribution to good scientific practice.

Data security as a problem field

Another problem in this context concerns data security. If external providers discontinue certain operating systems and applications or their maintenance, this also affects computers and internet-capable (measuring) devices in the research field. If, for whatever reason, it is not possible to update to a newer Windows version, for example, many research institutions are faced with the problem that they have to disconnect internet-capable devices from online access or transfer them to specially protected networks. Such elaborate (emergency) solutions present considerable challenges, especially for smaller institutions.

Risk of data loss

The described bottlenecks in IT equipment and IT management also concern long-term archiving. Without a forward-looking strategy, smaller research institutions or university chairs in particular often risk complete data loss when data is stored on obsolete storage media. Resources and personnel have so far only been available in large-scale research and infrastructure facilities. The problems of digital long-term archiving can certainly be addressed within the framework of the establishment of the NFDI. An in-house solution is hardly in sight, not least for technical reasons. There is also an urgent need for action in view of the acquisition of the necessary personnel.

3.2.3 PRECARIOUS FINANCIAL PROSPECTS FOR SERVICES

Tight financial scope for long-term security

A framework condition for the production, processing and, above all, long-term preservation of high-quality research data has so far been its public financing, which has hardly been in the public eye. In Germany in particular, there is little scope for further developing and financing research and information infrastructures that have developed out of university research as ongoing projects within the universities themselves.[83] In university libraries, too, more recent

[82] There are many case studies of such dependencies, not least when it comes to accessing data from commercial (or regulatory) providers. For science to gain access often requires protracted negotiations.

[83] RfII (2016) – Enhancing Research Data Management, p. 32 ff., Recommendation 4.1.

approaches to infrastructure development are highly dependent on fixed-term project funding. More opportunities for long-term development and expansion have so far been offered by the large science organisations and their infrastructure facilities, which have been pushing international standards with greater staying power and whose secure planning horizons offer good conditions for long-term quality assurance.

Difficulties in maintaining research and information infrastructures at HEIs

Wherever exemplary information infrastructures with excellent data quality have been established at Higher Education Institutions (HEIs, especially universities) – from epidemiological studies and cancer registers to large social science survey studies and linguistic databases – the only way to maintain the structure after many years of variable project funding was often to move on to a non-university institutionalisation. At the numerous breaking points in this process (marked by the respective project sponsors), important personnel and thus also scientific data and infrastructure competence has repeatedly been lost. These erratic processes have often been addressed in science as not helpful when it comes to establishing and maintaining permanent research-related services around research data – also and especially in the area of "small disciplines", which cannot be transferred to non-university institutes without difficulty. Smaller data collections, in particular, which were created in research, are often completely lost when the holder of the responsible professorship leaves or retires. In the wake of the European ESFRI process, numerous improvements have also been made in the humanities and social sciences in recent years. However, a coherent policy to maintain excellent research and information infrastructures at universities is not yet in place. The establishment of the NFDI is a first step towards at least showing project-financed and organised services the way in which they could approach or even integrate into already institutionally secured contexts in the medium term. Universities have recently expressed to be very open to follow this path and to actively contribute to the design of the NFDI.

3.3 LATENT PROBLEMS IN SCIENTIFIC PRACTICE

It is essential for scientific practice that it not only ensures the quality of basic processes such as logging, documentation, editing and representation/presentation in a subject- and task-specific manner, but that it also reflects this practice.[84] This is essentially about the traceability of analyses, but also about the adaptation of existing data collections to changing research questions and methods. There are similar problems in the subsequent digitisation (retro-conversion) of already standardised data such as texts or images, but also

84 Daston/Galison (2007) – Objektivität.

in catalogues in libraries and archives. The necessity of editing or curating such data – terms that come from the contexts of philological editing or research in museums – are once again coming to the fore. At the same time, however, other problems typical for digitisation are also emerging:

- heterogeneity and short life of tools used,
- despite abstract reflection, the de facto non-implementation of criteria related to the transition from analogue to digital processes,
- absent documentation culture for "digitaly-driven" (ancillary) decisions,
- unclear method reference to newly introduced tools (some of which were not developed for science, but for other purposes and were merely inherited),[85]
- an unavoidable loss due to digital storage (for example, material-bound traces of use as information carriers are omitted).

Cooperation between technical research and scientific libraries necessary

Likewise, the effort required to keep digitised collections usable over a longer period of time through professional long-term archiving has also proven to be high. Libraries and archives in particular need not only resources, but also urgently require informed support from the scientific community in order to maintain that task area permanently.[86] Sustainable maintenance would, for example, require that relevant object-related data be systematically compiled during the research process and not subsequently requested, so that this task area is not experienced as a crisis-ridden challenge. In this context, academic libraries in particular do not see themselves merely as suppliers of specialist information. Rather, they increasingly define their role as a driver in the digital transformation of science, who plays an active role in securing and enhancing data quality, especially in the area of metadata creation and maintenance. This includes the desire for a recognisable interest in cooperation on the part of researchers as well.[87]

In addition to generic challenges, which all disciplines and the forms of research practiced in them face to almost the same extent, there are also specific challenges for data quality depending on the form of research. The constellations from which these challenges have developed, the most urgent points in the data life cycle and the instruments or institutional precautions with which solutions are sought vary greatly.

85 An example is the use of search engines or recommender software in the catalog area or text and image representations that are optimised for portable devices.

86 To this end, RfII has recommended that infrastructure areas and research should also be closely linked in terms of personnel. See RfII (2019) – Digital Competencies, p. 27 f. especially Recommendation 4.5.

87 On this see DBV (2018) – Positionspapier Wissenschaftliche Bibliotheken 2025.

3.3.1 HERMENEUTIC-INTERPRETIVE FORMS OF RESEARCH

Increased interest in "classical" data sets

In the humanities, "philology" has a history dating back to pre-modern times and cultivates, among other things, editions of texts and images in "corpus works" and catalogues or dictionaries. In the 19th century, large edition projects were created. These projects, which were carried out as long-term undertakings, came under criticism when digitisation made fast and unlimited access to information possible. Edition projects with traditions dating back to the 19th century no longer seemed to meet current expectations and the use of resources for these projects was questioned. The struggle for a meaningful connection between the traditional handling of the material and the use of new digital technologies is, however, not limited to editions of texts and pictures, but also concerns technical knowledge (for example, about building materials, instrument making or classical ways of producing paint) or cartography (including historical aerial photographs). Quality assurance under the conditions of digitisation therefore requires reflection on basic processes and standards as well as questions of systematic (re)development of "analogue" knowledge. The interest in digital data collections is therefore often accompanied by a partly increased interest in (to be linked) classical data collections. For scientific and medical collections, not least in the context of the history of science and in the museum sector, a reflection on methods is characteristic, which sees a special quality feature in the combination of "analogue" and "digital".[88]

First digitisation wave was not yet conceived sustainably

Initially, however, digitisation was not accompanied by a "new" discourse on quality. In the 1990s, digital and automated methods of documentation and presentation were rather seen as an opportunity for more quantity. They simply seemed to offer ways of systematically accessing far greater quantities of information than in traditional processes. Cost-performance ratios for making information accessible seemed to be much better than in the analogue world. The ease of data transfer was also impressive. Increased speeds and data volumes thus seemed to be decisive effects of digitisation. The "digitisation wave" of the first years therefore led to a latent quality crisis in many disciplines – and especially in the humanities and cultural studies: Much was invested in making data readily accessible – but often in formats that impair its long-term usability. The initial euphoria has thus given way to the realisation that, for digital data, too, the technical and scientific quality is ultimately decisive. This quality, in turn, must be actively ensured through conceptual approaches and editorial intervention.

88 For example, see activities of the specialist group on documentation of the The German Museums Association: https://www.museumsbund.de/fachgruppen-und-arbeitskreise/fachgruppe-dokumentation/arbeitsgebiete/ or of the clearing house for scientific university collections: https://wissenschaftliche-sammlungen.de/de/ (both last accessed on: 30.08.2019).

3.3.2 OBSERVATIONAL AND EXPERIMENTAL RESEARCH FORMS PLUS SIMULATIONS

Early focus on the development of databases and repositories

In some sub-disciplines of the natural sciences, databases or repositories have been established for the delivery of research data, for example in the earth and environmental sciences, genome research or medicine. One example is the World Data System with its predecessors, the "World Data Centres", founded in the 1950s.[89] In the life sciences, an agreement has been in force since the mid-1990s to publish new gene sequences in one of the three worldwide data repositories within 24 hours if possible. Since the mid-1990s, the Cochrane Centres for Evidence-Based Medicine, with members from more than 130 countries, have also been involved in producing and making available health information generated from scientific studies that is free of commercial funding (for example, from the pharmaceutical industry).[90]

Pressure to publish vs. effort to prepare data for re-use

However, the scientific publication culture in the natural sciences is still predominantly characterised by the rapid introduction of results in article form into the research discourse. As a rule, this practice does not allow the underlying data to be presented in detail; only a selection of already processed data shapes the discourse and serves as a reference (see 2.1.9). In this way, the data on which the analysis process is based on are hardly available for other research questions, and even where data are published, there are considerable hurdles for further studies to handle these data. Data – for example in chemistry – are available in a standardised form with reference to the international standardisation committees, but they are not machine-readable and have to be extracted from publications manually, i.e. with a high level of personnel effort (see below). If they are published in databases, they are often not provided with context information (metadata) to the extent desirable today – this information is then, again at great expense, taken from the corresponding publications or requested from the implementing institution. It is not always clear, however, who archives the underlying data and may grant access. In this way, the research discussion may – despite intensive processing of data – move on a level that abstracts from the original data to a (too) high degree.

Controlling possibilities for accessing data as a first priority

In the social and economic sciences, the establishment of central databases and repositories for statistically relevant data on social development came much later. The initial problem here was that data collected by public institutions such as statistical offices and social insurance agencies had to be made available

89 See ICSU- World Data System, http://www.icsu-wds.org/organization (last accessed on: 30.08.2019).

90 See website of the Cochrane Foundation, https://www.cochrane.de/de/cochrane (last accessed on: 30.08.2019).

to research in the first place. The same principle – driven by the possibilities of digital remote access – was also made possible for survey data from large panel studies, which had previously only been made available to other researchers "on site" by the people or institutions collecting the data, sometimes with high requirements for data use. Here, too, the initial motive of opening up or accessing the data and data sets was preceded by an intensive quality check. Wherever empirical data are made available today in research data centres – especially in non-university research institutions – the focus of access to active quality assurance has already changed considerably in the course of preparation for the research interests of third parties. This change also includes the increasing influence of elected representatives from the scientific communities on the design of the survey samples (e.g. questionnaire design, special topics for subpopulations within the framework of panels with constant survey worlds) – and thus, in general, on the question of which section of reality the primary data should represent.

Standards needed for machine-readable data

In parts of the natural sciences, the international professional associations have long played an important role. They define standards and nomenclatures for the use of specialist terminology and measured values in tables and formulae in committees set up specifically for this purpose, which are characterised by a high level of acceptance in research practice. However, associations such as the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Pure and Applied Physics (IUPAP) are now also reacting to the fact that standardised forms of describing data are not yet an adequate basis for re-usable data. This requires adequate standards for machine-readable formula and data sets which, in case of doubt, allow for quick traceability and thus replicability of a research experiment. In addition to the already established reference databases – in which only a selection of verified data is kept for validation purposes – the establishment of data repositories is being discussed in order to make research results openly available.[91]

Innovative forms of research documentation are in demand

In experimental research, it is difficult to substitute the classical quality assurance (and guarantor) function of the laboratory book. The same applies, for example, to excavation books in the field of classical studies or research diaries on participatory observation in ethnological and social anthropological field research. Reproducibility of results is, of course, not possible or only possible to a limited extent without a reconstruction of the experiment or the data collection itself. It is not only the experimental sciences that lack tools for suitable documentation, some of which are adapted to digital methods. "Electronic laboratory books" are available (as commercial software products), but they

91 Koepler et al. (Aug 2018) – Thesenpapier NFDI4Chem.

compete with each other and are not always adapted to the specific needs of the discipline. Laboratory books also cannot adequately document work steps, some of which are fully automated. This also applies analogously to the above-mentioned forms of documentation in the cultural and social sciences.

3.3.3 DATA CURATION FOR USE BEYOND DISCIPLINARY AND DOMAIN-SPECIFIC BOUNDARIES

Quality assurance is also necessary for non-scientific target groups

Public access to research data or the "openness" of scientific databases (cf. 3.1.3) is increasingly associated with the expectation of quality assurance, in order to make data available to interested scientists and scholars from other disciplines as well as external specialists, for example science journalists. It is not yet clear to what extent a scientifically desirable quality assessment can be carried out for these extended target groups in the breadth required for this purpose. No one in the natural sciences or in engineering disputes the necessity of editing or curating data. However, the ways to open up data for a user group that extends beyond the domain are difficult and resource-intensive – even if the result can be worthwhile, as the example of satellite observation of polar ice shows: While the relevant observation data were only accessible to a small circle of experts in the 1980s, climate research and later biodiversity research also began to work with these data in the 1990s – and then in an aggregated or modelled form for these purposes, such as time series. Today, special data preparations are also available for journalists and the interested public.[92]

Comparability and connectivity of data publications

Progress in the field of subject-related data curation requires time and resources to negotiate corresponding principles in the scientific communities and their learned societies and to operationalise and implement them in universities and research institutions. In many places, there is simply a lack of supporting experts for the documentation of digital research processes.[93] In addition, there is the task of developing forms of data publication that can also be used meaningfully by researchers outside their own discipline. Digital processes also call into question existing scientific standards (partly through commercial offers). Thus, under the keyword "Semantic Web", competing proposals for keywords appear as a replacement for the "cascading catalogue system". Nevertheless, controlled vocabularies or thesauri (cf. 1.2.1), which were developed in the analogue world, are also (further) used for advanced text-mining and pattern recognition processes. This is necessary to guarantee comparability and

92 Baker et al. (2015) – Scientific Knowledge Mobilization.

93 The RfII made its recommendation in 2019 on career and educational perspectives, see RfII (2019) – Digital Competencies.

connectivity. Comparability and connectivity are required on several levels and make the problem of generating data quality both urgent and essential. For example, the harmonious integration of “old” and “new” data represents a major challenge for the development of research data. As described above, the obsolescence of storage media, among other things, causes major problems for the continuity of long-term studies. The transfer and linking of data not only across subject boundaries, but also across the narrower sphere of the scientific system, also poses enormous quality demands. For example, research into the major widespread diseases requires the linking of health data from the medical field with data from socio-economic panel studies and with geo-referenced data in order to obtain information on disease and disease progression patterns for different populations. Last but not least, the comparability and connectivity of data from different research sources are essential for the training sets of machine learning processes, i.e. for artificial intelligence (AI) learning. Qualitatively “poor” or non-connectable data sets cannot lead to valid learning outcomes.

4 RECOMMENDATIONS FOR DEVELOPING DATA QUALITY IN SCIENCE

Definitions and uses of the term "data quality" are characterised by a high degree of heterogeneity and diversity. On the one hand, the word has a technical meaning that is accessible to small-scale standardisation; on the other hand, however, it equally affects the ethos of all scientific activity. The development of research data infrastructures requires agreements of the data to be stored, transmitted and processed – both with regard to the methods used for their generation as well as their types and forms. This concerns for one thing requirements regarding the retrieval, usability and proper attribution of the data as well as requirements regarding the structure or format and the quality of the data itself in a broad sense. Yet it is precisely these specifications, which show the difficulty to achieve a convincing (acceptable), useful (functional), sufficiently broad (comprehensive) and at the same time also in detail meaningful concept of data quality. The required concept should be able to reflect the rapid technological and scientific change and should enable general consensus at least in the short to medium term.

4.1 TOWARD A DYNAMIC, PROCESS-ORIENTED DATA QUALITY CONCEPT

Multidimensional understanding of data quality required

The recommendations of the RfII assume a multidimensional understanding of data quality, which for one part includes the data life cycle and – as a second process cycle – the respective research process as well as the institutions and individual researchers involved (see Figure 3). The Council therefore refrains from a simple (normative) definition of the term data quality. Such an approach would run the risk that solutions for the creation and handling of data quality would (a) appear to be under-complex, i.e. disappointing, (b) inappropriate for the concrete application and thus not acceptable, or (c) could be postponed into the future. From the perspective of science policy, this results in an interminable process. It must also be taken into account that the radical change driven by digitisation also changes the research process itself and thus the immediate context in which data quality can be meaningfully discussed. Research itself is both the driver and the subject of the current transformation. The RfII proposes instead to make the development of a concept of data quality, which can only be defined provisionally, the subject of a continuous methodological discourse in all scientific communities.

All stakeholders in the scientific system are addressed

The respective recommendations are addressed to all stakeholders in the scientific system who shape the research process and its framework conditions through guidelines and consensus-building: In addition to the individual researchers themselves, these include the scientific communities and their learned societies, the scientific organisations, the research funding agencies, but also the bodies of scholarly communication, especially the research journals with their reviewers and editing committees. A forward-looking science policy will provide these addressees with the material means to act, both within the framework of a reliable institutional basic funding and through specific performance incentives (conform with the scientific reputation system). Coping with the additional efforts required in the course of digitisation in many disciplines will be essential in order to use and secure the flood of generated and available data at a high level, beyond individual groups as well as disciplinary and domain boundaries.

4.1.1 RECOGNISING DOCUMENTATION AS A CORE ELEMENT

Disclosure of procedural steps and analytical tools is necessary in research

Digitisation requires not only to "use" data or to apply tools "on" data, but also to regard the quality assurance of the data (in the respective technical manner required) as part of the research process. The core element of a sufficiently dynamic understanding of data quality is the precise documentation and disclosure of the measures, tools, the research software used and the procedural steps for generating, processing and making the data available.

4.1.2 DEVELOPING A SCIENTIFIC DATA CULTURE BASED ON PROCESSES

Aligning methods with the entire data life cycle

The added value (and unique selling point) of a scientific data culture is that its methodological efforts are in principle focused on the entire data life cycle. The RfII is therefore convinced that scientific information infrastructures must demand and ensure data quality so that the entire scientific system can indeed advance at the highest level of research-driven progress. Lastly, working out discipline- and domain-specific as well as generic procedures and benchmarks to ensure scientific data quality will be a permanent task for the upcoming National Research Data Infrastructure (NFDI) in Germany and the European Open Science Cloud (EOSC). Successful efforts of several scientific communities in the establishment of globally used open source programs can serve as an example.

4.2 INTEGRATION INTO THE SCIENTIFIC UNDERSTANDING OF METHODOLOGY

Researchers are the relevant stakeholders who, in their own interest, must initiate good scientific practice within the methodological canon of their respective disciplines and research forms, directly linked to the quality of their research data. The associated individual and collective responsibility is an essential component of the scientific "professional ethics".[94]

4.2.1 EMPHASISING DATA QUALITY AS A CORE VALUE OF SCIENTIFIC BEST PRACTICE

The RfII recommends that the topic of data quality, as an indispensable basic value of good scientific practice, be anchored even more sustainably than before in the understanding of scientific methodology.

Scientific data must be characterised by transparent quality assurance

- Accepted methodological standards and quality-assured processing in the ongoing knowledge production process give scientific data – in the midst of the "flood of data" that characterises the digital age in general – their special validity and thus give science one of its essential unique selling points. In this sense, the RfII generally regards commitment to creating or improving data quality as a contribution to strengthening society's trust in science and as a genuine scientific achievement (cf. 4.7.1).

- The minimum requirements for all scientific communities and disciplines, but also for the technical infrastructure services working in close proximity to science, include basic knowledge of data protection regulations. Compliance with the relevant legal requirements is an essential component of a first-class data collecting culture. This will also include a culture of explication (see also 4.8.3), which should initially be reflected in a culture of explaining remaining uncertainties or potential sources of error in a data set. Observance of such general minimum standards would significantly enhance the data collection phase, which is often (and wrongly) seen as a mere "technique".

4.2.2 FUNDING REPLICATION STUDIES

The RfII recommends that the scientific communities and their learned societies develop discipline- and research-field-specific criteria for data quality, which

Replication studies are an incentive for improving data quality

94 See DFG (2019) – Leitlinien zur Sicherung guter wissenschaftlicher Praxis p. 9 f., Leitlinie 2.

integrate them more closely into the respective understanding of research methods, where this is not yet the case. Where the research process allows this or where the disciplinary culture requires it, this will include the targeted promotion of replication studies that would be suitable for ensuring the validity of data, data sets and "data products" (see recommendation 4.4). Replication studies could thus offer a direct incentive to increase data quality and the associated detailed documentation of research data. Also, their upgrading would promote the internalisation and updating of the basic rules of good scientific practice at all stages of a scientific career. Last but not least, the funding granted for such studies should also be seen as an encouragement and award researchers, who conduct them at a high methodological level, with appropriate reputation.

4.2.3 ENSURING SCIENTIFIC DATA QUALITY CONTINUALLY DURING THE RESEARCH PROCESS

Understanding quality assurance as a permanent challenge

Securing and improving data quality is a permanent challenge throughout the entire research process and must be reflected accordingly by the researchers. Information sciences and infrastructure facilities can set technical principles or standards as guidelines for quality. Ultimately, however, the scientific quality of data is characterised by a plurality and also by a dynamic of criteria. These criteria can only be defined and described in detail by the scientific communities themselves and must be integrated into research practice. Harmonisation and standard-setting in research data management require specifications on the part of research or subject-specific answers to questions of quality. However, interdisciplinary connectivity and potential transfer possibilities must always be kept in mind: Broad (scientific) re-usability is not a final criterion, but an important aspect of data quality.

Tasks for NFDI and EOSC

The RfII sees precisely this aspect of quality assurance of databases, also with interdisciplinary use in mind, as one of the most important tasks of the future NFDI- and also EOSC-consortia: Their task is to find a balance between the setting of general guidelines across disciplines and domains and the subject-specific quality discourses. As a matter of principle, harmonisation and standard-setting – unless they concern purely technical issues – should always come from the scientific communities or, within the framework of the NFDI, be linked to their feedback.

4.2.4 TAKING RESPONSIBILITY FOR OPERATIONALISING QUALITY CRITERIA

The scientific communities have the responsibility

The RfII encourages the scientific communities to take active responsibility for the definition and description of the plurality and dynamics of quality criteria, especially in the context of the necessary negotiation processes within and

between the NFDI consortia. The Council recommends that the close connection between digital research processes and digital tools and services (i.e. the "infrastructure dimension") be reflected more intensively in the scientific methodological discourses. This applies in particular to disciplines that work with qualitative research methods or in hermeneutically interpreting and observing forms of research – especially large parts of the humanities and social sciences. Data based, for example, on field observations or open surveys and interviews pose different challenges for the methodological quality and documentation of their developmental contexts than those based on source interpretations. There are also major differences between the natural sciences: Data sets obtained via imaging procedures using sensors, detectors or scanners pose different challenges to quality assurance than data recorded about biochemical reactions in laboratory settings or those that arise in the context of translational medicine, for example in the interaction of epidemiological studies and individualised therapies. In many cases, the focus of data quality concepts will be the linking of "analogue" data and digital artifacts – for example in the engineering sciences, but also in other empirical research. NFDI Consortia should specifically take responsibility for promoting mutual understanding and working towards common guidelines for good research data management, despite the differences between the various disciplines. However, certainly there can be no "one size fits all" solutions across disciplines throughout the entire data life cycle.

Developing models for RDM plans

In order to put the quality related guidelines for data management into practice in individual disciplines or subject groups, the RfII recommends that models for research data management plans (RDM plans) be further developed. These should be discussed between scientific disciplines, universities and non-university research institutions, especially those charged with infrastructure tasks. The respective models, which would also differentiate between different forms of research, could be adapted locally to the requirements and tasks of individual project contexts. The RfII does not regard RDM plans as an additional bureaucratic burden for the research process. In fact, they can relieve the research process, provided they are adequately adapted to the disciplinary needs and forms of research: They document the effort researchers invest in quality and are an important basis for the adequate sizing of third-party funding and for compliance with the principles of good scientific practice.

4.2.5 RESEARCH INFRASTRUCTURES AS ENABLING STRUCTURES

Data quality depends on infrastructure quality

The quality of research data is also determined by state-of-the-art research infrastructures used to analyse, store and share (namely digital) data. Similarly, the quality of research data is a crucial yardstick for the development of good infrastructures. This interplay between data quality and infrastructure quality must be recognised much more explicitly in many academic disciplines than has

been the case to date. It is part of the basic understanding of the respective subject and should be taught early during the course of studies. Knowledge of data-generating, -processing and -storing infrastructures – from important institutional facilities, process steps (data life cycle) to hardware and software knowledge – should be included in the understanding of methodology in the disciplines (see also 4.6 on human resources development). A good understanding of the phases of the data life cycle as well as the interdependency between research processes and the related infrastructure would help researchers give due consideration to data documentation already during the research process, which is indispensable for the subsequent long-term archiving and maintenance of a data collection.

4.2.6 HIGHER RECOGNITION FOR WORK WITH RESEARCH DATA

Recognizing data quality assurance as an indispensable performance task in the scientific system

The RfII considers it indispensable that skills and capacities in the scientific community is actively developed and valued, corresponding to a growing awareness of the challenges to good scientific practice associated with digitisation. The RfII therefore welcomes the national and international discourse on the recognition of quality assurance as part of digital "methods", which is being promoted by numerous scientific communities. Corresponding approaches, such as those to be developed within the framework of the NFDI and various EOSC initiatives, must lead to the sustainable legitimisation of quality assurance and the recognition of related academic activities. The latter are not only "good practice", but indispensable efforts in the research process. They deserve recognition and reputation because they are an essential basis for the validity and robustness of research results. The production of data quality is a "positive goal" of research in all disciplines. Achieving this goal should be cultivated as a reputation-effective task throughout the entire data life cycle.

Boosting quality assurance as a topic in scientific training

Wherever the "traditionally" high quality criteria of publicly funded research are challenged by a high volume of and tolerance for "unchecked" data that may belong to a wide range of socially relevant issues, science can and should take this as an opportunity to renegotiate the framework for its own conditions of success. This also means that quality assurance of the data must be taken seriously as a natural part of scientific training in all disciplines and that appropriate courses of study and training must be expanded further. The RfII recommends to integrate additional courses into the curricula of the disciplinary and sub-disciplinary communities, instead of establishing new transdisciplinary chairs without contact to the respective disciplinary research base. Ultimately, only close integration into (sub-)disciplinary curricula will guarantee the desired enhancement of the subject- or field-specific understanding of methods.

4.3 ACCEPTING QUALITY ASSURANCE IN THE COURSE OF THE DATA LIFE CYCLE AS A GENUINE SCIENTIFIC TASK

Documentation tasks are part of good scientific practice

In the data life cycle, specific problems arise at all phases and at all interfaces, which have a negative impact on data quality and can subsequently be passed on in the cycle. At the end of the chain, research results may be jeopardised by inaccuracies, errors, bias or a lack of sustainability. Similarly, important conditions such as validity, reproducibility, authenticity etc. are subject to good data processing. The idea of good scientific practice in this respect remains related to the framework of an equally "well" lived culture of responsibility in the scientific communities. It also requires to state explicitly how it can be lived at and between each phase in the data life cycle (which would be adapted, depending on the scientific (sub-)discipline). Both the scientific communities and individual researchers should acknowledge documentation in every phase of the data life cycle as an essential contribution to good scientific practice, and reflect on the consequences of data-related decisions in the research process with a view to subsequent archiving, accessibility, maintenance and valorisation for other scientific issues (also beyond their own research field).

4.3.1 CLARIFYING DATA DESCRIPTION AND DECLARATION REQUIREMENTS

Ensuring professional assessment of data quality

The analytical parts of this position paper (Chapters 1 and 2) have shown that the claim to scientific validity results in fundamental requirements for data description and declaration: The generation of data must be described and declared in such a way that an assessment of its quality is possible at least within the respective technical or methodological domain. These requirements apply regardless of whether the data are archived in accordance with good scientific practice or made available in any form as a product for subsequent use by third parties.

The RfII recommends that the following requirements be specified:

- Researchers must make a decision – with professional support if necessary – as to where to focus on the quantity and where on the quality of the stored data. Even highly "noisy" data sets are scientifically valuable. The quality assurance steps that have been carried out in each case and any remaining imponderables must be clearly documented.

- The provenance of the data must be documented and – where the research process permits it – be traceable. The guiding principle is that of a "continuum of provenance" along the data life cycle, including information on the software and codes used and important transformation steps such as anonymisation.

- For personal data, data protection guarantees and appropriate labelling are essential: This also includes consent management that is comprehensible to third parties, for example, whether data may be passed on to third parties for validation purposes or for further research.

- Rights of disposal must be documented: Who is entitled to hand over the data in case of need or to decide on a publication or even authorise corrections? Likewise, the conditions for scientific use or, if applicable, economic valorisation must be documented.

Sufficient data description with metadata is mandatory

The following applies to digital research in particular: data does not speak for itself. Its quality can only be assessed if its context of origin (proveniance) is sufficiently described by metadata. Only through differentiated quality concepts (beyond problematic simplifications) can science contribute to countering populist "fake" messages.

4.3.2 IMPROVING COMMUNICATION AT THE INTERFACES

Actively designing data quality - improving interface management

The RfII recommends improving communication at the interfaces between stakeholders in the different phases of the data life cycle. This should be done at all institutional levels of the scientific system and actively supported through forward-looking research management. In addition to individual researchers, the RfII sees the responsibility for this primarily with the scientific communities and science organisations. HEIs and non-university research organisations should use the opportunities for action opened up by target agreements and the "system of pacts" [95] to collaborate on data management strategies across the data life cycle. They should also pool resources in a targeted manner. All stakeholders are also called upon to create transparency in good time with regard to the efforts and costs of data work at the respective stages of the data cycle. "Data quality" is a topic that needs to be actively shaped to a high degree.

In the following, the RfII refers to individual steps that should be taken into account when implementing improved interface management throughout the data life cycle:

[95] The "system of pacts" includes three treaties (pacts) between the Federal Governement and the *Länder* as the main funding bodies for the publicly financed scientific system in Germany: the treaty on HEIs ("Zukunftsvertrag Studium und Lehre", the treaty on non-university research institutions ("Pakt für Forschung und Innovation") and the treaty on quality improvement in higher education teaching ("Innovation in der Hochschullehre"). Common to all three treaties is the guaranteed increase in public subsidies for HEIs and non-university-institutes until 2030, which is, however, tied to certain performance targets.

- When collecting and generating data, data producers should anticipate later steps of the data life cycle according to the possibilities of their scientific culture and be professionally supported by contact persons in their institutional environments. Within the framework of the NFDI and the EOSC, for example, research actors should exchange views on the development of sustainable data governance that is adequate for the respective subject or subject group, wherever there is a need and no corresponding concepts or implementations are yet available. Corresponding strategies should enable a joint assumption of responsibility (who is approachable/responsible for the respective phases in the data life cycle, which services are available, etc.).

- Wherever possible, the RfII recommends agreements regarding accepted formats, vocabularies and ontologies or the use of uniform templates. The nomenclature and standardisation committees of international scientific associations carry out the corresponding coordination work and issue binding recommendations that are internationally accepted and applied. In disciplines and subject groups in which research is organised in a more idiosyncratic, multiparadigmatic or simply less cooperative way, the relevant scientific communities and their learned societies should at least examine how a consensus on overarching standards of data description could be achieved in order to ensure not only retrievability but also a broader scientific connectivity of data. The FAIR principles offer good guidance in this context.

- Capacities should be created in the scientific system to provide advice on the legal aspects of data handling, i.e. on specific questions of author's rights, copyrights and other valorisation provisions as well as data protection, public service law, "intellectual property" and good scientific practice.[96] These advisory capacities should in particular be geared towards the generally international character of data law issues. They do not necessarily have to be set up locally, but can also be installed nationwide or in networks. The NFDI consortia would be important actors at the national level to provide appropriate impulses.

4.3.3 DEVELOPING TECHNICAL ASSESSMENTS AND USING THEM SYSTEMATICALLY

Using IT-supported data quality testing more consistently

The RfII takes the view that the potential of IT-supported procedures for integrity/consistency and quality checks of data is not sufficiently used. They could

[96] See RfII (2016) – Enhancing Research Data Management, p. 55 ff., Recommendation 4.11 and 4.12.

facilitate the required documentation steps in the research process considerably and thus ensure a noticeable progress in data management.

Within the framework of research funding at national and European level and in coordination with corresponding initiatives of the future NFDI, targeted projects for the further development and testing of services and procedures should be supported. In the future, proven procedures and methods of data verification should be made available to the scientific community in the sense of best practice examples, beyond individual institutions and specialised domains.

Procedures to be considered in the context of future research funding would be, for example (without any claim to completeness):

- Automated quality checks for digital data ("validators") and plausibility checks,
- procedures for securing the data provenance documentation ("provenance tracking"),
- procedures for documenting data transformations,
- troubleshooting procedures,
- methods of an evaluation of data carriers, also "historically" reaching far back in time, with regard to the type of their usage methods in order to make derived data sets comparable, and
- procedures for detecting data tampering and data sabotage.

4.3.4 CONTROL AND TRANSPARENCY OF SOFTWARE PRODUCTS

Blackboxing effects compromise data quality

A major problem in the use of proprietary hardware and software in the scientific research process is the ignorance of researchers about the "inner workings" of the machines they use for experiments and analyses, and their influence on data generation, processing and analysis. In the worst case, a research result would be a simple artifact of the mechanical "inner working" of a device or the processing algorithm of the software used. Research results obtained in ignorance of the "within-put" of the instruments used are not replicable and counteract the ideal of good scientific practice (see 2.1.8). The phenomenon known as blackboxing is not easy to rectify, as science often lacks the resources to push own instrument developments, which it could then control. At least in the field of software, own open source developments can help, in which the source code is made transparent. Nevertheless, blackboxing can be expected to remain a critical phenomenon in science permanently.

Scientific communities must jointly articulate requirements for manufacturers

The commercial manufacturers of research equipment and laboratory tools refer to their property rights and try to protect themselves against technology piracy. Nevertheless, there are options for action, at least where scientific and

clinical institutions are the sole purchasers of industrial products. In addition to joint testing of the functions and consequences of different products (benchmarking), scientific communities, universities and science organisations should join forces and – similar to the debate on publication oligopolies – identify and articulate common needs.

The RfII recommends in this context:

- a broad disclosure of non-transparency of external devices and products as well as the results of disciplinary benchmarking processes, with which the researchers concerned often already help themselves;
- selecting the "better" product for a domain or research field in terms of transparency as jointly as possible;
- the development of scientific in-house solutions – akin to building blocks for scientific infrastructure in the sense of "commons" – where the size of the intra-science sales market justifies this also from an economic point of view; and,
- creating a clearing house in the German scientific system for negotiating the purchase, maintenance and usage conditions with manufacturers based on a binding agreement (on the need for joint discussion groups or consortia for the negotiation of infrastructure see recommendation 4.5.4).

4.4 DESIGNING AND DIFFERENTIATING DATA PRODUCTS

Data products increase scientific value creation

The RfII sees great potential for scientific value creation in well-documented and curated "data products". Various forms of data products have already emerged (see also 2.1.4). Firstly, such data products have the status of independent knowledge products, without which scientific breakthroughs and the transfer of research findings into the innovation system would be unthinkable. Secondly, data products also function as pieces of evidence in the scientific system. Databases and data centres, for example, can thus also be viewed in the tradition of scientific collections: They fulfil an important function of validation and reinsurance of research results and thus ensure stability over time and sustainability of scientifically proven knowledge in general.

4.4.1 DIFFERENTIATING AND DISTRIBUTING DATA PRODUCTS

According scientific recognition for creating data products

The RfII recommends that the researchers and their scientific communities as well as the science organisations, together with the information infrastructures, provide sustainable support for the production of data products. These should be valued as independent scientific achievements. In order to meet the

above-mentioned requirements in the scientific system, the purpose and target group for these products must be clearly definable; ideally, they should also be based on explicit, scientifically accepted standards. How NFDI and EOSC can promote and make available suitable forms of data products for different scientific and disciplinary purposes should be decided in the medium term within the framework of the NFDI consortia. Even in the short term, the RfII considers it the duty of research funders to provide incentives for the development of data products (see 4.7.1). Future data products can be oriented towards the following emerging formats:

a. Delivery of a data set to an existing scientific collection
The delivery of a data set to an – in the best case certified – archive or a data collection with continuous curation represents the lowest threshold variant of a data product. As a rule, the submitting researcher will have to prepare the data set in such a way that it is compatible with the guidelines or rules for submission to the collection. The RfII recommends that the scientific communities of those disciplines and subject groups in which potentially re-usable data are generated in the research process should support this practice as a minimum requirement for responsible data protection.

b. Co-publication of results and the related data set ("enhanced publication")
This combination of publication of results and a suitably linked data set is fundamentally useful and worthy of support in terms of quality assurance of research. However, the practice of supplement publication in PDF format, which is still widely practised, does not meet the requirements of accessibility and interoperability. Data sets for "enhanced publications" should – with the same care and context-related quality – be prepared in a machine-readable format so that they can be used more easily for any replication studies or subsequent research questions (see also Section 2.1.9). It should also be noted, however, that the problems of publication and reviewing in the area of "enhanced publications" are increasing (see 3.1.2 and 3.1.3). In the view of the RfII, there is an urgent need to improve the reviewing practice for data in particular. IT-supported procedures for checking the technical data quality are a possible part of the solution. Internal quality assurance procedures, which data sets undergo prior to publication, can be a useful supplement to external peer review, as long as they are regulated transparently and are comprehensible to third parties.

c. Digital editions
The creation of "digital data editions" goes far beyond enhanced publications in terms of the effort and quality of the documentation. The aim of an "edition" is to set up data beyond its documentation function or the possibility of using it repeatedly for similar purposes in such a way that it can be used over a longer period of time and for as many research questions as possible. This includes, among other things, a time-stable setup (possibly long-term archivability),

cross-domain annotation with metadata, interactive linking with external archive material or the analysis and visualisation of text phenomena using digital tools and services. Such a data product can be created not only for language and image data, but also for measurement values, quantitative survey data, etc., in order to show the context of the data sets and to enable assessment. The RfII recommends that this type of product be promoted as a recognised scientific service for data documentation. Standards for suitable formats should be developed in the scientific communities, and questions of technical preservation and long-term availability should also be clarified.

d. Data reports

Data Reports are data products that present and describe data material used or usable for research from a provider perspective (here: scientists). They are common instruments used by large-scale research facilities, infrastructure-providing non-university research institutes or large research data centres to supply the interested expert public with continuous and quality-assured data reporting from long-range research – from panel studies to data from ongoing accelerator experiments and astrophysical observational data from large radio telescope facilities. The RfII recommends that such reports should also be provided within the framework of other long-term research projects (e.g. SFBs or Clusters of Excellence) in the sense of continuous data monitoring, and that their use should be examined depending on the specifics of the subject and field.

e. Building curated data collections

The RfII considers the development and maintenance of a curated data collection to be one of the most comprehensive data products. A curated collection is characterised by dynamic maintenance and processing of research data, which is closely oriented towards current research questions and often collectively organised. The development of such data sets requires knowledge of and compliance with standards already at the time of data collection. New product formats and standards can also develop from curation, which in turn represent independent scientific achievements. At the same time, curators must be open to new standards-setting developments that come from science itself. The RfII welcomes the fact that, within the framework of the systematic collection of research data, its preparation for further research purposes or for the public is often of great importance, for example in the form of scientific or public use files or the presentation of data in combination with software or easy-to-use software applications.

4.4.2 DIFFERENTIATED RELEASE AND DISCLOSURE OF DATA

Considering different disciplines and forms of research

The disclosure of research data in the form of data products is a demanding and complex task. Especially in the humanities and social sciences (but also in

parts of the life sciences and clinical research), it must be decided on a case-by-case basis whether and to what extent the obligation to publish and disclose collected data raises the threshold for successful field access or the willingness of individuals and groups to be interviewed. This must be verified by means of suitable empirical studies. Especially in the field of hypothesis-free research on large amounts of digital data, both personal reference and other critical forms of data use (e.g. exposing social groups) are very easily possible. The already initiated expert discourse on consent management and, if necessary, on the "ethical" limits of data analysis is also important and necessary with regard to the use of social and language data by basic research in information technology. Here, too, it must be taken into account that the tracking of data traces left on the Internet by individuals for research purposes can lead to the violation of personal rights and further limit field access in the medium term. The recommendation to create data products should therefore not per se be equated with a publication. In individual cases, suitable and appropriate access regulations should be found.

4.4.3 PROMOTING A CULTURE OF REVIEW FOR RESEARCH DATA

Getting data reviews out of the niche - launching in leading journals

To the extent that data products are getting established as equivalent formats to the publication of results, the RfII also considers the promotion of an appropriate review culture to increase the awareness of these resources in the scientific communities and to stimulate interaction with research users. In suitable fields of research, it may be useful to have independent boards, platforms and processes for the review of research data.[97] In most cases, however, more space should be allocated to the review of research data in the leading journals of the scientific communities, or separate sections created, in order to strongly encourage the integration of data quality into the general understanding of methods.

4.4.4 FAIR PARTNERSHIPS WITH SERVICE PROVIDERS

Designing the market for data products along scientific lines

The preparation of data requires time and – depending on the format of the data product – in-format expertise of varying depth. This cannot always be fully provided by researchers, project participants or even scientific institutions. It is therefore foreseeable that a market for the creation of data products will develop, similar to that for research software. Corresponding trends can be observed at established publishing houses as well as in the area of science-related spin-offs. All parties involved are called upon to ensure a high degree of fairness

97 Section 3.1.2 provides a few examples.

in the interest of open access to research data and results. Scientific institutions and their sponsors must ensure through appropriate contracts that commercial partners and service providers keep the data entrusted to them accessible and – in relation to third parties – protect the researchers' data records from possible unauthorised access. In this context, the RfII points out the importance of the sovereignty of research over "its" data as an essential basis for the functioning of the scientific system and the expectations of society associated with it.

4.5 RESEARCH AND INFORMATION INFRASTRUCTURES AS GUARANTORS FOR QUALITY ASSURANCE

Deepening dialogue between infrastructure experts and research stakeholders

The RfII recommends that the data archives and repositories regard themselves as competence centres for the handling of scientific data and thus as institutional forms of quality assurance and quality promotion in the scientific system and, if necessary, develop strategically in this direction. Where this is not yet the case, they should be integrated systematically into the research process by scientists and academics to provide support. Under no circumstances should they be treated as simple "storages" or "data reservoirs" that are merely filled in order to meet, for example, the requirements of the institute's own specifications or those of research funding bodies. On the contrary: for a comprehensive data culture in the scientific system, a permanent dialogue between representatives of the information infrastructure and research actors is absolutely essential.

4.5.1 RELIABLE INSTITUTIONAL LINKS AS A NECESSARY FRAMEWORK

Merging services with the research process

The RfII is committed to ensuring that in the digital turn data-related services increasingly merge with the research process itself and that this happens via mutual exchange between infrastructures and scientific communities. This has long been the case in research institutes with large-scale facilities such as CERN or DESY, and is one of the essential conditions for global research success in decoding the elementary particles of matter. In the social sciences, self-organised and certified research data centres at public data resource providers (such as the statistical offices) and Leibniz Institutes with infrastructure tasks have also succeeded in establishing sustainable information infrastructures that are relatively easily accessible to the scientific communities. Even beyond the data life cycle, the learned societies have a say in the process and how the collection and processing of research data is to be documented in a comparatively transparent manner. In many other fields, research data management projects are rather precarious in their status. This is particularly true for smaller project contexts at HEIs and university libraries, which regularly face the question of their existence after the end of the usual limited project funding. This is not a sustainable state of affairs, since smaller, local information infrastructure

projects in particular are usually directly involved in research processes or have emerged from them. The RfII has clearly addressed the problem of precarious projects in 2016.[98] The Council still sees a great need for action, which the NFDI cannot solve either, because it was not set up to ensure the continuity of existing projects.

Nevertheless, future NFDI consortia can take responsibility for identifying what should be done. The RfII recommends using the NFDI structure to promote solutions for the long-term continuity of services through networking and communication. By linking up with larger organisational units of the NFDI, smaller local projects could participate in instruments of quality assurance and best practice learning. Appropriate certification procedures can help to establish common standards in the area.

4.5.2 INSTITUTIONAL QUALITY ASSURANCE

Organising performance and portfolio evaluations

The RfII recommends that those institutions that support research and information infrastructure and are not yet included in formal performance and portfolio development schemes a) also undergo such procedures and systems at reasonable intervals as part of their continuous improvement processes or b) take these as orientations for further action (with justifiable administrative effort). Examples of such procedures are evaluation procedures in the Leibniz Association or the certification procedures described in Chapter 1. The institutional and personnel structures should be aligned in such a way that it is possible to react reliably over time to requirements of technological and methodological change. “Evaluation” means here assistance to achieve improvements. The acquisition of quality seals should also be systematically promoted and demanded. They make quality efforts visible in the sense of a joint of all the organisation´s members – researchers and infrastructure specialists alike.

4.5.3 STANDARDS AND CRITERIA

Provider organisations are responsible for enforcing and implementing standards

The RfII regards the providers of research and information infrastructures as central stakeholders for the implementation and (co-)enforcement of quality criteria and standards. Accordingly, they should provide basic specifications for the data they collect and thus promote the necessary standardisation that is technically or scientifically necessary. In this context, they check the completeness

98 This is the subject of the first recommendation in RfII (2016) – Enhancing Research Data Management, p. 32 ff.

of the documentation and promote a culture of explication, document the changes made to data (records) transparently and carefully and ensure technological connectivity beyond the boundaries of domains and institutions. They are also in a good position to develop into competence centres for the creation of quality-tested data products and publications. The numerous research publications for which authors have drawn on quality-tested data sets from such competence centres clearly demonstrate their potential as enabling institutions.

4.5.4 TECHNICAL INFRASTRUCTURE

Securing the material basis of the research infrastructure is necessary

The quality of research data is related to the quality of the technical infrastructures on which data are processed and stored (cf. Chapter 2.1.5). In addition to concrete hardware and software errors that are difficult to detect, changes of versions or simply the discontinuation of certain series or product lines as well as aging or environment-related damage to components can lead to a massive impairment of data quality. Replicability may also be called into question. In order to rectify this situation, all research institutions and also smaller research units at universities must be enabled to process and save data using state-of-the-art components and be assisted by professional personnel. The RfII takes the view that public research funding must enable HEIs and non-university research institutes to counteract the obsolescence of infrastructure components and storage media. This also includes providing sufficient financial resources for the required personnel in the long term and the necessary structural quality of research buildings in which the information infrastructures are hosted. For their part, HEIs and non-university research institutes are called upon to give appropriate priority to the maintenance of the technical data infrastructure and the personnel required for this. Smaller research units can cooperate strategically for data storage with appropriately equipped and professionalised institutions at local and regional level, such as computer centres, large university libraries or non-university research institutes. This is how economies of scale can be achieved.[99]

Initiating working groups between providers and users

In addition, it seems necessary to increase cooperation and contractual agreements with the commercial manufacturers of technical research environments. Dealing with possible version changes or the discontinuation of product lines, including support, must be regulated early on and in the interest of science. In this context, the RfII recommends the establishment of groups or consortia in which further infrastructure-induced quality aspects for data – such as the

[99] For further recommendations concerning the long-term archiving of research data as well as the technical infrastructure and the implications for tasks of the future NFDI, see RfII (2016) – Enhancing Research Data Management, p. 39-42.

"black boxing"-problem – can be continuously negotiated between research, infrastructure providers and commercial providers (see also recommendation 4.3.4 on the establishment of a clearing house). Here, too, fairness must be ensured with regard to the corresponding contract design, which ensures the sovereignty of the scientific community over its own data (see recommendation 4.4.4).

4.6 DIGITAL SKILLS AS REQUIREMENTS FOR GOOD DATA MANAGEMENT

Anchoring data quality in study programmes, training and continuing education

The RfII considers the providers of research and information infrastructures to have a duty to play an active role in the communication and further development of data- and method-related competences and also to continuously train their own staff. This includes anchoring the important role of data quality for the entire research process – from the collection of data to its use in scientific publications and transfer processes – as an essential content in training and study programmes as well as in further education and training based on these programmes. In this way, study, training and further education programmes will contribute significantly to good scientific practice and help to promote good scientific work and to detect and avoid misconduct at an early stage. In the area of human resources development, an investment in data competence always also contributes to improving the quality of science and confidence in scientific knowledge as a whole.

In its recommendations DIGITAL COMPETENCIES – URGENTLY NEEDED! the RfII has provided several suggestions for human resources development in science, which also aim to improve the quality of research data throughout the data life cycle.[100]

4.6.1 BREAKING UP PILLARISATION BY TASK-RELATED TRAINING

Interlinking training centres – developing task-based offers

The RfII sees an obstacle in tendencies towards institutional pillarisation of the data production and distribution chain – more precisely: in the hitherto impermeable separation of the spheres of responsibility into a technical-administrative part, a science supporting part and a purely scientific part, with different working cultures of the staff and different regulations concerning personnel's work autonomy, remuneration and terms of employment. The RfII is convinced that only a better integration of the units involved in data production and provision

[100] See RfII (2019) – Digital Competencies, chapter 4.

in HEIs, non-university research institutes, academic libraries and computer centres – and thus: an increased permeability of personnel categories – as well as massive efforts in the training and further education of staff beyond the present boundaries of personnel categories can contribute to maintaining and improving data quality in research. With this in mind, the RfII has recommended that training materials be developed in a task-oriented manner and implemented within the framework of determined "qualification alliances" of the German science organisations. Already during training and studies, the transfer of perspectives should be made possible through internships and work shadowing and later through temporary job rotation across organisational boundaries. The RfII sees a need, particularly on the part of researchers, to emphasise the importance of infrastructures and infrastructural work for good scientific practice and high-quality research data and to create an awareness of the importance of these tasks.

4.6.2 EXPANDING INFORMATION SCIENCE EXPERTISE

Intensifing the teaching of IT-based competencies across the board

A major challenge for maintaining high data quality is the competent handling of the hardware environments and software components which are used to collect, process, analyse and store research data for subsequent use (catchphrase "Laboratory 4.0" or similar terms). In addition to scientific or disciplinary knowledge, technical and technological IT-based skills are also highly required, which are not always taught in disciplinary study and training. Qualified personnel with basic knowledge of information science is needed, as well as the establishment of new divisions of labour. Here, personnel with the appropriate IT know-how can interact with the research personnel in specialised scientific disciplines who use digital processes or operate equipment in a targeted and project-related manner. For this purpose, silo formations between infrastructure facilities and research units must also be broken down.

4.6.3 ENSURING DATA MANAGEMENT IN RESEARCH

Ensuring data protection in research despite high staff turnover

From the point of view of data quality, one of the important aspects of staff development is the internationalisation of research processes. This goes along with a high fluctuation of internationally recruited personnel, especially in HEIs and non-university research institutions. The RfII recommends that the securing of the data with which international scientific personnel – but of course also junior and senior researchers from the national environment – have worked at the respective institution during their (usually limited) stay be made a task for the institutes´ or facilities´ leading management. Uncontrolled entrainment of primary data and intermediate products or "data cemeteries", which are unusable for further use and valorisation after the researchers have left, should

be avoided. In order to ensure this, all research institutions should have rules for good scientific data management and, at the beginning, assign concrete responsibility for data in a staff meeting, even for short-term employments. When the employment relationship is terminated, it should be checked whether all those involved have fulfilled their data responsibility.

4.6.4 COMBINING DATA QUALITY AND COMMUNICATION COMPETENCE

Strengthening external communication

Trust in the quality of research data, both within and outside the scientific community, does not come about automatically. In today's age of a digital "flood of data" and pseudo-evidence created in part by media manipulation, science cannot hope that its nimbus will generate social trust on its own. The media scandalisation of cases of scientific misconduct contributes to the fact that even individual cases compromise science as a whole. Trust in the quality of the data with which research works, which it generates and on which it bases its findings, must therefore be actively won. The RfII recommends – in addition to the many points already mentioned – that the ability to communicate externally on aspects of data quality be professionalised in both the scientific and the science-supporting areas. People involved in the data life cycle must not only know what they are doing, but also be able to explain it. Scientific infrastructure services should therefore be considered as an integral part of the public relations work of scientific organisations. The RfII sees the ability to acquire the necessary media competence as an important component of human resources development, which must be cultivated at an early stage in all training institutes and "on the job".

4.7 FUNDING POLICY AND ORGANISATIONAL REQUIREMENTS FOR QUALITY DEVELOPMENT

In this position paper, the RfII deliberately places the quality of research data at the centre of scientific and science policy interest. This is based on the assumption that even in the age of the digital turn, good research and excellent research performance in all disciplines and subject groups can only succeed on the basis of a reliable and quality-assured database – from the initial collection of empirical research data and text editions to the re-use of research data for further subsequent research or for quality assessment of research already carried out (replication). Against this background, the RfII also makes recommendations for funding policy, for HEI and non-university research institutes and for the science policy of the *Länder* and the Federal Government.

4.7.1 SUPPORTING THE QUALITY OF RESEARCH DATA IN TERMS OF FUNDING POLICY

Developing funding policy further on - scheduling resources for data management

In order to anchor data quality more firmly in project funding, foundations will be addressed in particular, next to the DFG and the BMBF which are the two most important national research funding bodies in Germany. They could make a major contribution to the further development of research data quality in Germany. Foundations should – as they have often done in the field of teaching – experiment with new funding formats that are suitable to gather empirical knowledge on which further public research funding can build. Furthermore, the RfII considers data quality to be a cross-cutting issue that touches on all disciplines and research topics. Appropriate requirements should be made in all forms of public research funding in the application process, and their compliance should be monitored during and after the funding period.

a. Prizes/awards for contributions to the further development of data quality
In today's research process, there is a lack of explicit incentives – across all disciplines and subject groups and especially at HEIs – to deal with the topic of data quality and its further development as a scientific core task. At best, it is a by-product of the work of professorships and chairs with a methodological denomination. And even there, it is not the main criterion that would decide on the filling of such a position. Foundations could provide appropriate incentives to make the work with data standards, data management and appropriate IT solutions for improving data quality more visible and reward it through prizes and awards.

When it comes to data quality, research data management has the dubious reputation of being a technocratic, not very creative affair, which a promising young scientist, for example, who is striving for a career, prefers not to touch. Too much management, too little reputation is the short formula. The range of tasks that is elementary for ensuring good scientific practice throughout the entire research process is therefore often "delegated" to support staff. Good research data management is not a trivial matter and should be carried out jointly by technical-administrative and scientific personnel – whereby the research interest must always guide the action. Here, too, it is important to make best practices and outstanding achievements more visible through public awards in order to increase the attractiveness of the task for researchers as well. The RfII sees this as an explicit field of activity for business-related foundations – not least because of the proximity to certified quality processes and standardisation procedures in industry.

b. Funding innovative data products

Data publications and other forms of data products are not yet widely used in research and publishing. Here, too, it is important to provide incentives within the framework of research funding in order to establish data processing at a high quality level as a legitimate and necessary product of the research process and at the same level as the publication of results. The RfII particularly encourages foundations to make advance efforts with funding programmes for the development of data products that use new representation technologies but also consider the long-term preservation and usability of these products.

c. Extending project durations for data documentation

In the medium term, however, data products in their various forms should also become the standard for reporting on a successfully completed research project in publicly funded research. Since this cannot be achieved within the usual project duration of three years in the context of individual funding, the funding agencies would have to create opportunities to apply for additional funds for editing or curating or to extend the overall project funding period.

d. Giving quality preference over quantity

Procedures for evaluating research funding applications should explicitly provide for the importance of qualitative parameters also for data work. In addition, sensitivity should be sharpened to the fact that the quality of publications also depends on a well-documented data basis. A smaller number of well-documented publications backed up with data should be given greater value in the current account than a large number of publications in which the data base is conventionally or superficially documented.

4.7.2 RECOMMENDATIONS FOR HIGHER EDUCATION INSTITUTIONS (HEIS)

Incorporating data quality in research strategies, appointments and study regulations

In science policy terms, the universities are regarded as the “heart chambers” of the science system. Accordingly, an increasingly important field for research, such as the further development of data quality, should also be proactively addressed and pursued at universities and other HEIs. The RfII argues that the issue of data quality should be addressed with high priority, particularly in the HEI context, since this will also directly affect the quality of teaching and thus the educational function of universities and other HEIs for society in the medium term. The RfII welcomes the universities’ joint commitment to the NFDI and their desire to actively shape the national network.

a. Making data quality a subject of research strategy

Many HEIs and non-university research institutes have made it their institutional development strategy to enable excellent research performance in profile-forming areas. The discussion about data management plans has also long since reached the HEIs; guidelines for research data management have been adopted in many places and are on the way to implementation. Nevertheless, the topic of "data quality" has not yet become an integral part of the research strategy (and possibly also of the research priorities that are promoted internally in a special way) of the universities and other HEIs.

The RfII recommends that data quality from an institutional point of view (e.g. by means of targeted development of local data competence centres within the framework of research priorities or by networking across locations) should ensure that methods and data culture are promoted. This is often done where infrastructure-supporting institutions are already networked into the universities through joint appointments and staff involvement in teaching, and is unanimously welcomed by the RfII.

b. Considering data expertise in appointment proceedings

When appointing new scientists to professorships and chairs, the RfII recommends that services rendered in the establishment, maintenance and research-related networking of information infrastructures should be explicitly considered as important contributions to research. Adequate evaluation of achievements in the development of data quality and information infrastructure should always be considered: Researchers who are active in this field make their institution attractive for other researchers. Professors who succeed them, and especially young researchers, will then be able to embed their research lines in an already well-developed environment.

c. Incorporating data quality in study regulations

In order to increase the information and media competence of students as a whole and, in particular, to sustainably anchor the awareness of data quality in the methodological understanding of the scientific communities, it is necessary to integrate the topic into degree programmes as early as possible.[101] How data is created, how it is processed and which horizons of valorisation and use must be considered, which legal, political and ethical framework conditions play a role in this context, must be part of the basic knowledge of every student. The RfII recommends that, in addition to a subject-related methodological component, the social implications of dealing with data be taught in the respective courses of study.

[101] See RfII (2019) – Digital Competencies, p. 22 f., recommendation 4.2.

4.7.3 RECOMMENDATIONS FOR NON-UNIVERSITY AND DEPARTMENTAL RESEARCH INSTITUTES

Further increasing permeability - promoting close cooperation with HEIs

The non-university research institutes and departmental research play an important role in Germany in the further development of data quality. Many of them are infrastructure-supporting institutions themselves and have massively expanded research with their own infrastructures and their opening up to the scientific communities in recent years. The RfII recommends that they continue to resolutely pursue this path, to incorporate data quality as a central goal at all levels in their own research strategies and, in this context, to further expand cooperation with the HEIs in particular. In the interest of comprehensive capacity building for the entire German science landscape, this should also – and especially in the research-related infrastructure sector – include a regular exchange of staff and joint commitment to data-related initial and continuing training. According to the RfII, the permeability proposed here offers great potential for all sides, particularly for establishing common binding standards and procedures (e.g. data management plans): By maintaining constant contact with research partners from HEIs, non-university institutions can protect their information infrastructures from a decoupling of innovations in university research. At the same time, individual researchers can learn from successful (data-) management procedures and practices, which can be better developed and implemented in the often more obligatory and collaborative organisational structure of non-university institutes.

4.7.4 RECOMMENDATIONS FOR THE FEDERAL GOVERNMENT AND THE *LÄNDER*

NFDI as an important starting point for joint action

As institutional sponsors of science in Germany, the Federal Government and the *Länder* play an important role with regard to the framework conditions in which researchers can work at the highest level of quality. By establishing the NFDI, the Federal Government and the *Länder* have shown that questions of data quality are of great importance to them in current and future science policy. In the opinion of the RfII, the NFDI must now be legally and organisationally structured in such a way that data quality in all research contexts can also be promoted from this network with a long term perspective.

Continuing to examine the prospects for the sustainability of project-based infrastructures and services

However, the NFDI is not an institution that can and should offer the numerous research and information infrastructures, which have been financed on a precarious – i.e. temporary – basis by project funding, a safe, stable port (at least not from its own resources). The question of a lasting development perspective for infrastructures successfully developing from projects at universities and university libraries is thus not solved. Many Collaborative Research Centres and Clusters within the framework of excellence funding are also developing valuable information infrastructures that need not lose their value after the

end of funding. Often, hardly any resources are available for their continuation or transfer to other infrastructure contexts, especially at the universities. The Federal Government and the *Länder* should therefore examine whether such services can be provided with longer-term development prospects – if they are proven to be of supra-regional importance and of structural relevance for the scientific system.

4.8 CONTINUING THE FAIR PROCESS

FAIR as starting point - not as an endpoint

In the European Research Area, the FAIR principles (Findable, Accessible, Interoperable, Re-usable) are currently being vigorously and somewhat successfully established as a benchmark for good research data management. The implementation of the FAIR principles is primarily aimed at creating usability and intensifying the use of data. The focus is on creating accessibility through machine readability as an essential enabling condition for data quality. In contrast, less attention is paid to the concrete operationalisation of the maxim essential for scientific research in the canon of FAIR principles: "(meta)data meet domain-relevant community standards" (cf. 1.2.5).

Connect FAIR with disciplinary quality discourses

The RfII sees an urgent need on the part of research to fill this leitmotif with life and to go beyond it. Discipline- and research field-specific quality criteria for data quality are necessary (cf. 4.2.3), and these must be actively incorporated into the curating and archiving processes on the infrastructure side. A link to subject-specific rules for (meta-)data documentation is urgently required, because good retrievability and shareability of subject-specific, non-quality assured or unchecked data would not support the FAIR intentions. The mere individualisation of the problem by shifting it to the responsibility of the individual scientist – both in the provision and in the course of valorisation – is not conducive to a FAIRe data culture. In the opinion of the RfII, science and infrastructure must always be considered together in the FAIR context.

4.8.1 LAUNCHING A SCIENTIFIC QUALITY OFFENSIVE

Raising awareness of content- and discipline specific quality standards

For this reason, the RfII recommends that the implementation of the FAIR principles be supplemented by a scientific quality offensive, which is committed to promoting appropriate descriptions of data for effective re-use and thus visibly qualifies the data for research practice. This calls for all actors at national and European level, who are significantly supporting and implementing the FAIR process. Especially in the context of the constitution of a new actor such as the NFDI, the RfII recommends that the scientific character of data should always be considered as a quality dimension of its own and henceforth be implemented parallel to the FAIR principles. As useful as it was at the beginning of the FAIR

process to focus on questions of data access in order to create a common basis on which various "Open" initiatives could build, it is just as important today to strive for a sharpened awareness of common quality standards not only in terms of form but also in terms of content. A quality initiative complementing FAIR can also give this struggle an obligatory claim to expand all efforts now and tomorrow from access to linkage and with linkage to connectivity and translatability of data in various scientific and social contexts.

4.8.2 COMMUNICATING DATA QUALITY – BUILDING TRUST

Developing a joint media strategy

An essential component of the proposed quality offensive is frank and precise communication. The RfII recommends discussing a common media strategy throughout the scientific community, in which all actors of the scientific system can participate and which is implemented with commitment by the science organisations (see also recommendation 4.6.4). In particular, the RfII suggests that data-related activities be attractively prepared for purposes of journalism and given a stronger presence in the mass media: It is not uncommon for the "real" nature of thorough research to become tangible here. For numerous scientific breakthroughs are based on high-quality prepared and analysed data – especially in the area of interdisciplinary cross-sectional topics, with which science provides answers to major social challenges (demographic change, climate protection, widespread diseases, etc.). Similarly, keywords such as "AI" or "Industry 4.0" are incomplete for journalistic enlightenment without the topic of "good data". Data quality is a fundamental prerequisite for such future topics.

Bringing data-intensive research to the public

Scientific institutions, but also individual researchers, must give more weight to the infrastructure and data dimension of their research, both in their own scientific communication and reporting, in the assessment and recruitment of staff and in teaching as a relevant and necessary driver of scientific progress. The fact that this can be achieved is shown by regular reporting on the findings on income and wealth inequalities or the development of poverty risks for different population groups or the results of astrophysics and research into the elementary particles of matter – i.e. research in which the infrastructure reference is easily recognisable through the use of large social science population surveys or large-scale scientific equipment (radio telescopes, particle accelerators). For other scientific fields, the representation of data sources in the media is more difficult to achieve, but is certainly possible with targeted efforts.

4.8.3 PROMOTING A CULTURE OF TRANSPARENT EXPLICATION

Incorporating the culture of explication in research practice and institutional models

In the course of the quality initiative, the RfII encourages all actors in the scientific system to adopt a culture of "transparent explication" within the scope of their responsibilities – from the supervisor of a qualification project to the management of a scientific institution, from the information infrastructure to the publishing house and research funding. Information on the quality assurance process should be made available as a matter of course, as should, for example, the method and process of scientific knowledge production. In addition, scientific communities should agree on an adapted culture of referencing or citing data (stocks), which is both quality assuring with regard to research and reputation-promoting for the data-producing side and the participating researchers. The RfII considers such a culture of transparent data dissemination to be an essential complement to the FAIR principles and an important contribution to good scientific practice. The RfII recommends that this be reflected in institutional models for good scientific practice and actively promoted in the day-to-day research activities of HEIs and non-university research institutes. In these matters the Council expects important impulses from the NFDI and its consortia, which should trigger resonance not only in the national framework but also in the European research area.

BIBLIOGRAPHY

Amann, Rudolf I. et al. (2019): Toward Unrestricted Use of Public Genomic Data, in: Science 363, Nr. 6425, p. 350-352, available at: https://science.sciencemag.org/content/363/6425/350.full.pdf, DOI: 10.1126/science.aaw1280, (accessed: 30-08-2019).

Baker, Karen S./Duerr, Ruth E./Parsons, Mark A. (2015): Scientific Knowledge Mobilization. Co-evolution of Data Products and Designated Communities, in: International Journal of Digital Curation 10, Nr. 2, p. 110-135, available at: http://www.ijdc.net/article/view/10.2.110/411, DOI: 10.2218/ijdc.v10i2.346, (accessed: 30-08-2019).

Carpenter, Todd (2017): What Constitutes Peer Review of Data – A survey of published peer review guidelines, available at: https://arxiv.org/ftp/arxiv/papers/1704/1704.02236.pdf, (accessed: 30-08-2019).

CCSDS- Consultative Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Practices. CCSDS 650.0-M-2, Washington, available at: https://public.ccsds.org/pubs/650x0m2.pdf, (accessed: 30-08-2019).

Daston, Lorraine/Galison, Peter (2007): Objektivität, 1. Aufl., Frankfurt am Main: Suhrkamp, 530 p.

DBV- Deutscher Bibliotheksverband- Sektion 4 (2018): Wissenschaftliche Bibliotheken 2025, dbv:, available at: http://www.bibliotheksverband.de/fileadmin/user_upload/Sektionen/sektion4/Publikationen/WB2025_Endfassung_endg.pdf, (accessed: 30-08-2019).

DFG- Deutsche Forschungsgemeinschaft (2017): Replizierbarkeit von Forschungsergebnissen. Eine Stellungnahme der Deutschen Forschungsgemeinschaft, Bonn, available at: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2017/170425_stellungnahme_replizierbarkeit_forschungsergebnisse_de.pdf, (accessed: 30-08-2019).

DFG- Deutsche Forschungsgemeinschaft (2019): Leitlinien zur Sicherung guter wissenschaftlicher Praxis (Kodex), Bonn, available at: https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf, (accessed: 30-08-2019).

DINI- Deutsche Initiative für Netzwerkinformationen e.V. (2018): Thesen zur Informations- und Kommunikationsinfrastruktur der Zukunft, available at: https://edoc.hu-berlin.de/bitstream/handle/18452/19876/DINI-Thesen_2018_2.pdf?sequence=1&isAllowed=y, DOI: 10.18452/19126, (accessed: 30-08-2019).

Europäisches Parlament/Rat der Europäischen Union (2019): Richtlinie (EU) 2019/1024 des Europäischen Parlaments und des Rates vom 20. Juni 2019 über offene Daten und die Weiterverwendung von Informationen des öffentlichen Sektors (Neufassung), in: Amtsblatt der Europäischen Union, 28, L 172/56- L 172/83, online verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32019L1024&from=EN, (accessed: 30-08-2019).

Fecher, Benedikt/Friesike, Sascha/Hebing, Marcel (2015): What Drives Academic Data Sharing?, in: PLoS ONE 10, Nr. 2, p. 1-25, DOI: 10.1371/journal.pone.0118053.

Field, Laurence et al. (2013): Realising the Full Potential of Research Data. Common Challenges in Data Management, Sharing and Integration Across Scientific Disciplines. Version 3, available at: http://orca.cf.ac.uk/66034/1/ESFRI_Common_Challenges_v1.pdf, (accessed: 30-08-2019).

Hodson, Simon et al. (2018): Fair Data Action Plan. Interim Recommendations and Actions From The European Commission Expert Group On Fair Data, DOI: 10.5281/ZENODO.1285290, (accessed: 30-08-2019).

KE- Knowledge Exchange (2014): Sowing the Seed. Incentives and Motivations for Sharing Research Data, a Researcher's Perspective, Kopenhagen, available at: http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf, (accessed: 30-08-2019).

King, Gary/Persily, Nate (2019): A New Model for Industry-Academic Partnerships, available at: https://gking.harvard.edu/files/gking/files/partnerships.pdf, DOI: 10.1017/S1049096519001021, (accessed: 30-08-2019).

Kleiner, Matthias (2010): "Qualität statt Quantität"- Neue Regeln für Publikationsangaben in Förderanträgen und Abschlussberichten. Pressekonferenz, DFG- Deutsche Forschungsgemeinschaft, Berlin, available at: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2010/statement_qualitaet_statt_quantitaet_mk_100223.pdf, (accessed: 30-08-2019).

Klimpel, Paul (2015): Eigentum an Metadaten? Urheberrechtliche Aspekte von Bestandsinformationen und ihre Freigabe, in: Euler, Ellen et al. (ed.): Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft, Berlin, Boston, available at: https://irights.info/wp-content/uploads/2016/01/Klimpel-2015-Eigentum-an-Metadaten.pdf, (accessed: 30-08-2019).

Koepler, Oliver et al. (2018): Thesenpapier Nationale Forschungsdateninfrastruktur für die Chemie (NFDI4Chem), DOI: 10.5281/ZENODO.1404201, (accessed: 30-08-2019).

Lauber-Rönsberg, Anne/Krahn, Philipp/Baumann, Paul (2018): Gutachten zu den rechtlichen Rahmenbedingungen des Forschungsdatenmanagements, available at: https://tu-dresden.de/gsw/jura/igetem/jfbimd13/ressourcen/dateien/publikationen/DataJus_Zusammenfassung_Gutachten_12-07-18.pdf?lang=de, (accessed: 30-08-2019).

Lazer, David/Kennedy, Ryan/Vespignani, Alessandro (2014): The Parable of Google Flu: Traps in Big Data Analysis, in: Science, Nr. 343, p. 1203-1205, available at: https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf, (accessed: 30-08-2019).

Liggesmeyer, Peter (2009): Software-Qualität. Testen, Analysieren und Verifizieren von Software, 2. Aufl., Heidelberg: Spektrum Akademischer Verlag, online verfügbar unter: http://dx.doi.org/10.1007/978-3-8274-2203-3, accessed: 30.08.2019.

Lipton, Zachary C. (2016): The Mythos of Model Interpretability, available at: https://arxiv.org/pdf/1606.03490v3.pdf, (accessed: 30-08-2019).

Neuroth, Heike et al. (ed.) (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, Boizenburg/Göttingen: Hülsbusch/Universitätsverlag, available at: https://publiscologne.th-koeln.de/files/428/Publikation_Osswald_Langzeitarchivierung_Bestandsaufnahme.pdf, (accessed: 30-08-2019).

Parsons, M. A./Fox, P. A. (2013): Is Data Publication the Right Metaphor?, in: Data Science Journal 12, 32-46, DOI: 10.2481/dsj.WDS-042, (accessed: 30-08-2019).

Peer, Limor/Green, Ann/Stephenson, Elizabeth (2014): Committing to Data Quality Review, in: IJDC- International Journal of Digital Curation 9, Nr. 1, p. 263-291, DOI: 10.2218/ijdc.v9i1.317, (accessed: 30-08-2019).

Pfaffenberger, Fabian (ed.) (2016): Twitter als Basis wissenschaftlicher Studien: Eine Bewertung gängiger Erhebungs- und Analysemethoden der Twitter-Forschung, Wiesbaden: Springer, 146 p.

RatSWD-German Data Forum (2018): Activities Report 2017 of the Research Data Centres (RDCs) accredited by the German Data Forum (RatSWD), Berlin, DOI: 10.17620/02671.36, (accessed: 30-08-2019).

RfII – German Council for Scientific Information Infrastructures (2019): Digital Competencies – Urgently Needed! Recommendations on Career and Training Prospects for the Scientific Labour Market, Göttingen, available at: www.rfii.de/?wpdmdl=4015, (accessed: 30-08-2019).

RfII – German Council for Scientific Information Infrastructures (2016): Enhancing Research Data Management: Performance through Diversity. Recommendations Regarding Structures, Processes, and Financing for Research in Data Management in Germany, Göttingen, available at: https://d-nb.info/1121685978/34, (accessed: 30-08-2019).

RfII – German Council for Scientific Information Infrastructures (2017): An International Comparison of the Development of Research Data Infrastructures: Report and Suggestions, Göttingen, available at: http://d-nb.info/1143738454, (accessed: 30-08-2019).

RfII – German Council for Scientific Information Infrastructures (2019): Statement of the Council for Scientific Information Infrastructures (RfII) on current developments concerning Open Data and Open Access, Göttingen, available at: http://www.rfii.de/?p=3814, (accessed: 30-08-2019).

Samek, Wojciech/Wiegand, Thomas/Müller, Klaus-Robert (2017): Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, available at: https://arxiv.org/pdf/1708.08296.pdf, (accessed: 30-08-2019).

Stuart, David et al. (2018): Practical Challenges for Researchers in Data Sharing (Whitepaper), DOI: 10.6084/M9.FIGSHARE.5975011, (accessed: 30-08-2019).

Swan, Alma/Brown, Sheridan (2008): The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the Jisc, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf, (accessed: 30-08-2019).

Wang, Richard Y. (1998): A Product Perspective on Total Data Quality Management, in: CACM- Communications of the ACM 41, Nr. 2, p. 58-65, DOI: 10.1145/269012.269022, (accessed: 30-08-2019).

Wang, Richard Y./Strong, Diane M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers, in: Journal of Management Information Systems 12, Nr. 4, p. 5-33, available at: http://www.jstor.org/stable/40398176, (accessed: 30-08-2019).

Whyte, Angus/Pryor, Graham (2011): Open Science in Practice: Researcher Perspectives and Participation, in: IJDC-International Journal of Digital Curation 6, Nr. 1, p. 199-213, DOI: 10.2218/ijdc.v6i1.182, (accessed: 30-08-2019).

Wilkinson, Mark D. et al. (2016): The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: Scientific data 3, p. 1-9, available at: https://www.nature.com/articles/sdata201618.pdf, DOI: 10.1038/sdata.2016.18, (accessed: 30-08-2019).

Wouters, Paul/ Wouter Haak (2017): Open Data. The Researcher Perspective. Hg. v. CWTS- Leiden University's Centre for Science and Technology Studies und Elsevier, available at: https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf, (accessed: 30-08-2019).

WR- Wissenschaftsrat (2012): Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Drs. 2359-12, Berlin, available at: http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf, (accessed: 30-08-2019).

WR- Wissenschaftsrat (2015): Empfehlungen zu wissenschaftlicher Integrität. Positionspapier. Drs. 4609-15, Köln, available at: http://www.wissenschaftsrat.de/download/archiv/4609-15.pdf, (accessed: 30-08-2019).

WR- Wissenschaftsrat (2018): Empfehlungen zur Internationalisierung von Hochschulen. Drs. 7118-18, WR-Wissenschaftsrat, München, available at: https://www.wissenschaftsrat.de/download/archiv/7118-18.pdf, (accessed: 30-08-2019).

ONLINE RESOURCES

Cochrane Germany
https://www.cochrane.de/de/cochrane

COPDESS – Coalition for Publishing Data in the Earth and Space Sciences
http://www.copdess.org

The German Association of University Professors and Lecturers, Barometer
https://www.hochschulverband.de

The German Museums Association
https://www.museumsbund.de

German Text Archive
http://www.deutschestextarchiv.de/

ESA Sentinel Online
https://sentinel.esa.int/web/sentinel/home

Force 11
https://www.force11.org

Forschungsdaten.org
https://www.forschungsdaten.org

GEO-Wiki
https://www.geo-wiki.org

GFBio – German Federation for Biological Data
https://www.gfbio.org/about

GFZ Data Services
http://dataservices.gfz-potsdam.de/portal/about.html

Go FAIR Initiative
https://www.go-fair.org

ICSU- World Data System
http://www.icsu-wds.org

The Institute for Documentology and Scholarly Editing
https://www.i-d-e.de

Schema.org – Community Website
https://schema.org

Social Science One
https://socialscience.one

SpringerNature
https://www.springernature.com

Research Data Centres of official statistics
https://www.forschungsdatenzentrum.de/en

The Open Definition
https://opendefinition.org

UAG Datenmanagementpläne der DINI-nestor AG Forschungsdaten
https://www.forschungsdaten.org/index.php/UAG_Datenmanagementpläne

Visual6502-Projekt
http://www.visual6502.org

W3C Standards
https://www.w3.org

Wikidata
https://www.wikidata.org

APPENDIX

A. DEFINITIONS

In the course of its work, the Preparatory Committee on Data Quality also revised two of the definitions of the RfII, which were published in 2016 in the position paper Enhancing Research data management: Performance through diversity. The plenary session adopted the following new wording at the 10th Council meeting in November 2017.

DATA QUALITY

The term data quality refers both to general, typical properties of the data itself, which are required from a methodological point of view, as well as their suitability for further use, which may be additionally created by quality assurance measures.

The evaluation of data quality is based on the requirements to be defined for the data, which depend on the respective research question and thus on the use to develop a research result. These include, for example, the accuracy of measured values, the reliability of an empirically obtained result, the completeness or timeliness of data and the documentation of data collection and data storage.

In addition, sustainability aspects cannot be separated from the assessment of the specific quality of data. Such aspects include the nature of the data, such as, for example, the exchange of the data or the durability of data carriers. In particular, they relate to the forward-looking storage of research data for later, ideally manifold and possibly yet unknown, forms of scientific, economic and societal use.

From the point of view of further use ("subsequent use"), data quality is determined by the fact that data sets and collections are easy to research/find and that they contain sufficient additional information. This should be in the form of technical and professional metadata on quality aspects that are as standardised as possible and provide information on data generation, further processing and the instruments and methods used. A prerequisite for the traceability and, if possible, subsequent use of digital research results is that the data they contain are comprehensively documented with regard to the data models on which they are based (applied vocabularies, formats, etc.) and the methods used (such as measuring instruments, surveys, algorithms, etc.). Wherever possible, not only metadata, but also further, possibly special documentation, should follow acknowledged standards.

The availability, accessibility and citation of research data including their metadata – also in the long term – are in turn aspects of the quality of information infrastructures and services, which enable secure storage, targeted retrieval, access to the data and their subsequent use (also in the context of long-term archiving). The clarification of the legal framework of a possible data use is also part of data quality in connection with information infrastructure services.

Sources, position papiers

On good scientific practice: Allianz der Wissenschaftsorganisationen (2003) – Berliner Erklärung und DFG (2013) – Sicherung guter wissenschaftlicher Praxis, p. 21–22; DFG (2019)– Leitlinien zur Sicherung guter wissenschaftlicher Praxis; on data quality: OECD (2007) – Access to Research Data.

RESEARCH DATA, RESEARCH DATA MANAGEMENT

Research data are not only the (final) results of research. Rather, it is any data that is generated in the course of scientific work, for example through observations, experiments, simulations, surveys, questioning, source analysis, records, digitalisation, evaluations. Research data also include such data which are not acquired by science itself, but which science accesses for research purposes in order to use them as a methodologically necessary basis for the concrete research process. This is the case, for example, when official statistics or other official data or products of non-scientific service providers are scientifically processed. The fact that the research tools used as well as the traces of scientific work – i.e. process data, which are often automatically produced by digital research – are also included in research data is important wherever the documentation and archiving of research processes and research data is part of their quality assurance or is required for sustainability aspects and historical research. Pragmatically, although not always clearly separated, research data can be distinguished from metadata. Metadata document and contextualize the process of research data creation. In the research process, metadata can themselves itself become the subject of further research, which is important for the research data life cycle.

Research data management includes all measures – beyond research activities in the narrower sense also organisational measures – that have to be taken in order to obtain high-quality data, to adhere to good scientific practice in the data life cycle, to make results reproducible and to take into account any existing documentation obligations (for example in the health care system). The availability of data for re-use (possibly across domains) is also an important point. Data management plans are increasingly used in scientific institutions. Data management plans, which are developed and laid down at the beginning of a project or are the result of a research project, should describe the data to be used and generated and the necessary documentation, metadata and standards, identify possible legal restrictions (for example data protection) early in the process, plan required storage resources and define criteria for making data available to external parties. At the organisational level, research institutions (e.g. universities) must provide access to appropriate infrastructure services within the institution (e.g. by establishing and expanding appropriate capacities) or in cooperation with external partners (through cooperation agreements, etc.). In doing so, efforts should also be made to achive the overarching objective of cross-domain, science-wide data use.

Sources, position papiers

Allianz-Initiative Digitale Information – AG Forschungsdaten (2015) – Research data at your fingertips; on research data = basis of argumentation and calculation: EC (2013) – Guidelines on Open Access, p. 3; on research data = primary source of scientific activity: OECD (2007) – Access to Research Data, p. 13; on research data from the perspective of the social sciences: RatSWD (2010) – Kriterien Forschungsdaten-Infrastruktur, p. 4; on research data as data from the research process: Allianz-Initiative (2012) – Leitbild 2013–2017, p. 7; WR (2012) – Empfehlungen zu Informationsinfrastrukturen, p. 53–57; DCC – Data Management Plans, http://www.dcc.ac.uk/resources/data-management-plans (accessed: 30.08.2019); DFG (2015) – Leitlinien Forschungsdaten; HRK (2014) – Management von Forschungsdaten; HRK (2012) – Hochschule im digitalen Zeitalter.

B.1. COUNCIL, MEMBERS, AND GUESTS

The German Council for Scientific Information Infrastructures has 24 members and is composed as follows to ensure equal participation:

- 8 representatives of scientific users from a broad spectrum of scientific disciplines
- 8 representatives of information facilities who cover the entire range of the German scientific system
- 4 representatives of the German Federal Government and the *Länder*
- 4 representatives of the public

The first 16 representatives are appointed in a procedure similar to that for members of the German Council of Science and Humanities. The other 8 representatives are nominated by the Federal Government and the governments of the *Länder* in the Joint Science Conference (GWK). All members are appointed by the chair of the Joint Science Conference for a term of four years. Guests can be invited to council meetings or parts thereof when there is a corresponding need.

"The composition of the Council reflects our conception that the future of scientific information infrastructures is a joint task of the providing institutions, the scientific users, the funders, and related national and international stakeholders."

– Joint Science Conference, November 2014 –

Representatives of scientific users

Prof. Dr. Marion Albers
Faculty of Law, University of Hamburg

Prof. Dr. Lars Bernard
Faculty of Environmental Sciences, Technical University of Dresden

Prof. Dr. Stefan Decker
FIT – Fraunhofer Institute for Applied Information Technology and RWTH Aachen

Prof. Dr. Petra Gehring (Chair)
Department of History and Social Sciences, Technical University of Darmstadt

Prof. Dr. Kurt Kremer
MPI – Max Planck Institute for Polymer Research, Mainz

Prof. Dr. Wolfgang Marquardt
Forschungszentrum Jülich GmbH

Prof. Dr. Joachim Wambsganß
Centre for Astronomy of Heidelberg University (ZAH)

Prof. Dr. Doris Wedlich
KIT – Karlsruhe Institute of Technology – Division I: Biology, Chemistry, and Process Engineering

Representatives of the Federal Government and the Länder

Rüdiger Eichel
Ministry of Science and Culture of Lower Saxony

Dr. Hans-Josef Linkens
Federal Ministry of Education and Research

Dr. Dietrich Nelle
Federal Ministry of Education and Research

Annette Storsberg
Ministry of Culture and Science of North Rhine-Westphalia

Representatives of information infrastructure facilities

Sabine Brünger-Weilandt
FIZ Karlsruhe- Leibniz Institute for Information Infrastructure GmbH

Prof. Dr. Dr. h.c. Friederike Fless
DAI – German Archaeological Institute and Free University, Berlin

Prof. Dr. Michael Jäckel
Trier University

Prof. Dr. Stefan Liebig (Deputy Chair)
DIW – German Institute for Economic Research, Berlin

Prof. Dr. Sandra Richter
German Literature Archive

Katrin Stump
University Library of Braunschweig

Prof. Dr. Klaus Tochtermann
ZBW – Leibniz Information Centre for Economics and Kiel University

Prof. Dr. Ramin Yahyapour
GWDG – IT in science and University of Göttingen

Representatives of the public

Dr. Anke Beck
IntechOpen publishing

Marit Hansen
Data Protection Commissioner of Schleswig-Holstein

Dr. Nicola Jentzsch
SNV – Stiftung Neue Verantwortung (Foundation for New Responsibility, until 03/2019)

Dr. Harald Schöning
Software AG

B.2. PROJECT DATA QUALITY

COMMITTEE ON DATA QUALITY (2017)

Prof. Dr. Frank Oliver Glöckner (Moderation), Prof. Dr. Lars Bernard, Prof. Dr. Petra Gehring, Dr. Margit Ksoll-Marcon, Prof. Dr. Stefan Liebig

WORKING GROUP DATA QUALITY (2017–2019)

Prof. Dr. Dr. h.c. Friederike Fless (Chair), Prof. Dr. Marion Albers, Prof. Dr. Lars Bernard, Prof. Dr. Petra Gehring, Prof. Dr. Frank Oliver Glöckner (Guest), Prof. Dr. Stefan Liebig, Dr. Hans-Joseph Linkens (represented by Dr. Lena Maerten), Prof. Dr. Doris Wedlich

EDITORIAL GROUP (2019)

Prof. Dr. Petra Gehring (Chair), Prof. Dr. Lars Bernard, Prof. Dr. Dr. h.c. Friederike Fless, Prof. Dr. Kurt Kremer, Prof. Dr. Stefan Liebig, Dr. Hans-Josef Linkens (represented by Dr. Lena Maerten)

The committees received substantive and organisational support of the RfII Head Office from Dr. Stefan Lange, Dr. Kirsten Gerland, Dr. Ilja Zeitlin, Dr. Sven Rank, Dr. Maximilian Räthel.

B.3. ACKNOWLEDGEMENT

The RfII would like to thank all experts who participated in the work of the Working Group on Data quality.

The experts of the discussions in May 2018 and September 2018 were:

Dr. Michael Diepenbroek
Prof. Dr. Gerd Gigerenzer
Prof. Dr. Uwe Hasebrink
Prof. Dr. Richard Lenz
Prof. Dr. Thomas Ludwig
Prof. Dr. Eva Schlotheuber
Dr. Markus Schmalzl
Dr. Nico Siegel
Dr. Cornelia Weber

For expert advice:

Dr. Kai Denker

IMPRINT

Courtesy translation, published February 2020
The original text was adopted by the RfII in September 2019

German Council for Scientific Information Infrastructures (RfII)
Head Office
Papendiek 16
37073 Göttingen, Germany

Phone +49 551 392 70 50
Email info@rfii.de
Web www.rfii.de

DESIGN, TYPESETTING, AND PRINTING
NEFFO DESIGN (Buchholz), Klartext GmbH (Göttingen)

SUGGESTED CITATION
RfII – German Council for Scientific Information Infrastructures: The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn, Göttingen 2020, 120 p.

The German National Library lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.