Report from Dagstuhl Perspectives Workshop 18472

Implementing FAIR Data Infrastructures

Edited by

Natalia Manola[1], Peter Mutschke[2], Guido Scherp[3], Klaus Tochtermann[4], and Peter Wittenburg[5]

1 **University of Athens, GR,** natalia@di.uoa.gr
2 **GESIS – Leibniz Institute for the Social Sciences – Cologne, DE,** peter.mutschke@gesis.org
3 **ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE,** g.scherp@zbw.eu
4 **ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE,** k.tochtermann@zbw.eu
5 **Max Planck Computing and Data Facility – Garching, DE,** peter.wittenburg@mpi.nl

Abstract

This report documents the programme and the outcomes of Dagstuhl Perspectives Workshop 18472 “Implementing FAIR Data Infrastructures”. The workshop aimed at bringing together computer scientists with digital infrastructure experts from different domains to discuss open issues implementing and adopting the FAIR principles in research data infrastructures and to shape the role that the field of computer science has to play.

Seminar November 18–21, 2018 – http://www.dagstuhl.de/18472
2012 ACM Subject Classification Information systems
Keywords and phrases fair principles, open data, open science, research data infrastructures
Digital Object Identifier 10.4230/DagRep.8.11.91

1 Executive Summary

Natalia Manola (University of Athens, GR)
Peter Mutschke (GESIS – Leibniz Institute for the Social Sciences – Cologne, DE)
Guido Scherp (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)
Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)
Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE)

The Dagstuhl Perspectives Workshop on “Implementing FAIR Data Infrastructure” aimed at bringing together computer scientists and digital infrastructure experts from different domains to discuss challenges, open issues, and technical approaches for implementing the so-called FAIR Data Principles in research data infrastructures. Moreover, the workshop aimed to shape the role of and to develop a vision for computer science for the next years in this field, and to work out the potentials of computer science in advancing Open Science practices.

In the context of Open Science, and the European Open Science Cloud (EOSC) in particular, the FAIR principles seem to become a common and widely accepted conceptual basis for future research data infrastructures. The principles consist of the four core facets that data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable in order to advance the discoverability, reuse and reproducibility of research results. However, the FAIR principles are neither a specific standard nor do they suggest specific technologies or implementations. They describe the core characteristics of data use. Thus, the FAIR principles cover a broad range of implementation solutions. This certainly incorporates the risk of having a highly fragmented set of solutions at the end of the day.

Given this, and in view of the "need for a fast track implementation initiative [of the EOSC]"[1], it is strongly needed to turn the principles into practice. Therefore, the workshop took the recommendations of the European Commission Expert Group on FAIR Data "Turning FAIR into reality" as a starting point and discussed what can be done next from the perspective of computer science to enable data providers to make their data FAIR.

The workshop started with three ignition talks on the wider background and context of the FAIR principles (given by Peter Wittenburg), the relationship of FAIR to Open Data (given by Natalia Manola) and the role of the principles within the EOSC (given by Klaus Tochtermann). Based on these talks as well as inputs from all participants in the forefront of the workshop, we have split the discussion into three working groups addressing, for each of the four principles, the main key challenges for implementing FAIR and the question what and how computer science can contribute to these key challenges. Based on the results of these three initial working groups we furthermore split into more focused groups addressing the problem of licenses w.r.t. data use, (self)improvement of FAIRification, and the relation of FAIR and data intensive science.

Finally, we identified three major areas to be addressed in the manifesto which we discussed in three further working groups:

1. Infrastructures & Services Aspects: This group focused on the question by which technical means research data infrastructures and data services can be advanced to better address and fulfil the FAIR principles.
2. Computer Science Research Topics: The working group discussed the relationship of research areas in computer science and topics relevant to implement FAIR data infrastructures.
3. FAIR Computer Science Research: While the other two groups mainly focused on the contribution of computer science to implement FAIR, this working group addressed the question how the FAIR principles are currently adopted by computer science research itself and what should be improved.

The participants will continue their work in the aforementioned issues, and a manifesto is foreseen to be ready by mid May 2019.

[1] https://www.dtls.nl/wp-content/uploads/2017/05/DE-NL-Joint-Paper-FINAL.pdf

2 Table of Contents

Executive Summary

Ignition Talks

Working Groups on Challenges of FAIR Principles for Computer Science

Working Groups on Specific Types of Challenges for FAIR

Working Groups on Identified Topics for the Dagstuhl Manifesto

3 Ignition Talks

3.1 Why FAIR and what is the context of it?

Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE)

Since about 2005, the special nature and value of data was commonly recognised. Since then intensive discussions took place in data science. OECD, a high level group of the EC, a workshop at ICRI 2012 that led to the establishment of the Research Data Alliance (RDA), the group of G8 Science Ministers, various RDA working groups, and many other places with the goal to identify ways to improve data sharing and reuse. These are currently hampered by a huge fragmentation and resulting in inefficiencies and costs. The FAIR principles finally formulated now widely accepted agreements on a minimal set of behaviours about the creation and management of digital objects in a convincing way. They formulate clear messages to change our current data practices which are characterised by 80% of waste of time of data professionals in data projects leading to consequences such as failed projects and exclusion of many experts and SMEs, etc. However, the FAIR principles are not blueprints to build the urgently needed infrastructures that will help changing practices. Initiatives such as RDA[2], GEDE[3], DONA[4] and GO FAIR[5] are now in agreement that the concept of "Digital Objects" (DO), which have a bit sequence encoding some content and are associated with a persistent and unique identifier and different kinds of metadata, is a way to implement interoperability at data organisation and modelling layer. The recently developed DO Interface Protocol (DOIP) therefore is a uniform interface to access all DOs in repositories independent of how these are set-up and the content the DOs are encoding. The FAIR principles and the DO implementation concept are therefore complementary and thus can be pillars of convergence towards improved efficiency in data management and reuse as requested by Wittenburg & Strawn [1].

References

1 Peter Wittenburg and George Strawn. *Common Patterns in Revolutionary Infrastructures and Data*. 2018.
URL: http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

3.2 FAIR vs. Open

Natalia Manola (University of Athens, GR)

Open Scholarship and Open Science are becoming the modus operandi in research, but for data sharing to reach researchers a language of scholarly communication should be spoken. Excellent researchers do not have time to design infrastructure or temper technology, but

[2] http://rd-alliance.org
[3] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda/
[4] https://www.dona.net
[5] https://www.go-fair.org

they are eventually targeted towards communicating with peers through publishing any forms of research results, which contribute to their career development. FAIR and open data are now key aspects in scholarly communication, even from an early stage of research production, and our duty as infrastructure providers is to enable openness and FAIRness by design into our services and processes. Even though there is an unwarranted perception that FAIR and Open are overlapping terms, in reality "FAIR is not Open", and "Open does not imply FAIR" [1]. This adds some complexity in our communication to researchers, and we are often faced with questions of whether to aim for open or FAIR data, and in which situations would one or the other be enough? As a starting point, many of the FAIRness principles for data are prerequisites for their openness. However, data being FAIR does not directly imply that it is also open as there are "levels" of openness which are subjected to ownership, intellectual property rights, sensitivity issues, licensing etc. In practice, the FAIR principles are directed more towards technical aspects than towards moral and ethical aspects of data, especially as these address sharing by default for publicly funded research. Moreover, FAIR principles require clarity and transparency around the conditions governing access and reuse, and relevant services focus upon provisions to make data available for reuse under clearly-defined conditions and licenses, through well-defined processes, and with appropriate acknowledgement and citation. On the other hand, open does not directly mean FAIR. Open datasets without being FAIR, e.g., without proper metadata or software to access, without specifying the proper licenses are useless to their intended users. This is true for all fields, but it is more apparent in cases where we must have reproducible and accountable science, such as medical data where patient health history matters or humanities where working on already available data makes large part of the research. Furthermore, some key ethical issues still persist in openness: how do we code for machine readiness, with GDPR and data protection seen as examples; how do different domains perceive such issues and how do individuals react; where does big data and linked data come into play? In the end, costs put aside, we should aim for designing and operating infrastructures which produce "Open data in a FAIR way", considering above ethical issues.

References

1 Mons, B. and Neylon, C. and Velterop, J. and Dumontier, M. and Da Silva Santos, L. and Wilkinson, M. 2017. *Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud.* Information Services and Use. 37 (1): pp. 49-56. DOI: 10.3233/ISU-170824

3.3 FAIR Principles within the EOSC

Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE)

The idea of the European Open Science Cloud (EOSC)[6] is to leverage European research data management to the next level of excellence. The EOSC will connect existing and future research data centres with one another and will offer a free point of use, open, and seamless services for storage, management, analysis, and re-use of research data. The talk highlights

[6] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

three components related to FAIR: 1) data, 2) services, and 3) infrastructure. To establish the link between FAIR and EOSC, the talk analyses relevant EOSC documents, such as the reports of the EOSC High Level Expert Groups[7] or the Implementation Roadmap for the EOSC[8], events such as the EOSC summit, and developments such as new projects like FAIRsFAIR, FAIRPlus, and EOSC-Life, which have been awarded recently with EC funding to foster the FAIR principles. Within this context the Implementation Roadmap for the EOSC recommends to develop FAIR data tools, specifications, catalogues, and standards to best support scientists and innovators, and to stimulate the demand for FAIR data through consistent FAIR data mandates and incentives to open data by research funders and institutions across Europe. With respect to infrastructures the talk addresses the need for FAIR-compliant certification schemes for FAIR data infrastructures. And finally, the talk argues for the need of initial services that are required to gather and organise FAIR data and data-related research products and which should be made accessible via a service platform.

4 Working Groups on Challenges of FAIR Principles for Computer Science

The participants discussed in three parallel working groups each aspect of the FAIR principles with the following questions:

- What are the current challenges?
- What are possible solutions and how can computer science support?

The following sections are brief summaries from the respective working groups.

4.1 Working Group 1

Natalia Manola (University of Athens, GR / Moderator), Wilhelm Hasselbring (Universität Kiel, DE / Rapporteur), and Peter Mutschke (GESIS – Leibniz Institute for the Social Sciences – Cologne, DE / Contributor)

Findable

Challenges.

1. The data infrastructure landscape is characterised by a strong diversity and heterogeneity of data repositories and an over reliance on cataloguing. At the same time, stakeholder requirements are not really known. An overarching, one size fits all portal is missing.
2. Appropriate metadata standards and controlled vocabularies are missing as regards both common core as well as domain-specific standards.
3. An open issue is how to design identifiers and versioning (latest, history, releases), in particular how to precisely identify arbitrary subsets of data in a dynamic setting with

[7] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud-hleg

[8] https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

data being added, deleted, changed (see RDA WGDC Recommendation on Dynamic Data citation[9], with a slightly more extensive report[10].

4. Indexing is not sufficient. Field specific selections are also needed.

Possible solutions and support of computer science.

1. Meta search engines, supporting both multi-disciplinary and disciplinary search, and standardized search and content harvesting APIs are needed ((a) Verborgh & Dumontier (2018): A Web API Ecosystem through Feature-Based Reuse [1], (b) combination of metadata and API example: [2] from the life science community). A further help is seen in search query and metadata auto completion, e.g., enabled via machine learning in "data lakes".
2. A common representation of metadata (see EDMI from EOSCpilot[11] and schema.org) as well as intelligent assistance for metadata creation (see Ted Nelson's vision of a literary machine for science[12] is urgently required.
3. Computer science could help by providing standard components and engineering support.
4. Leveraging industry-scale markup like schema.org (e.g. bioschema.org) would alleviate this problem.

Accessible

Challenges.

1. A major problem from user perspective is seen in dead links and the lack of reliable mechanisms to deal with dead links in a sustainable way.
2. A further problem is seen in the great amount of heterogeneity of authentication, authorisation and identification processes.

Possible solutions and support of computer science.

1. Audit trail management services are required that provide evidence and quality control of access data and links and by this a greater transparency of data accessibility t the user.
2. A standardised protocol for authentication, authorisation, and identification (AAI) is strongly needed (e.g. a data passport containing a decision tree providing a meaningful response to users on what to do to get access).

Interoperable

Challenges.

1. The variety of data formats and data encoding methods (e.g. different time/date formats, time zones) as well as the lack of format validation is seen as the major problem.
2. An infrastructure to annotate data for improving interoperability is missing.
3. Dealing with different vocabularies and languages is a major obstacle as regards interoperability.
4. Ensuring portability of data as well as tools between different platforms is still a challenge.

[9] https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
[10] http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
[11] https://eosc-edmi.github.io
[12] https://en.wikipedia.org/wiki/Literary_Machines

Possible solutions and support of computer science.

1. Common standards, such as W3C / schema.org, should be reused consequently. Data transformation methods (by use of semantic technologies and common markup languages, e.g. YAML, extract transform load) as well as advanced format validation tools are strongly required.
2. Tools for semantic annotations and an ontologies lookup service as a gatekeeper are needed.
3. Ontology crosswalks (LOV=linked open vocabularies[13]) and smart ontology mapping methods could alleviate this problem.
4. Container technologies to enable portability might help.

Reusable

Challenges.

1. Different data protection regulations, intellectual property rights and licensing models make reusability challenging .
2. Provenance data that capture the entire lifecycle of data, i.e. the social process of "making" data (incl non-digital interactions), is often missed.
3. Peer review of data sets and data curation is needed.

Possible solutions and support of computer science.

1. Computer actionable licences are required to address this problem. Identifiers and versions should be FAIR as well.
2. Services that record workflow-generated provenance metadata are strongly needed.
3. Semi-automated data curation tools (workflow based) and tools for data management plans are required.

References

1 Ruben Verborgh and Michel Dumontier. *A Web API Ecosystem through Feature-Based Reuse*. IEEE Internet Computing, Volume 22 , Issue 3 , May./Jun. 2018. DOI: 10.1109/MIC.2018.032501515

2 Carlos Horro Marcos1, John M. Hancock, Wiktor Jurkowski, Annemarie Eckes. *Brassica Information Portal and Elixir: MIAPPE & BrAPI.*Presented at Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017. URL: https://doi.org/10.7490/f1000research.1114610.1

4.2 Working Group 2

Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Daniel Mietchen (University of Virginia, US / Rapporteur)

Findable

Challenges. Rich metadata that includes context and provenance information described in different languages and annotations from different views will be essential to facilitate

[13] https://lov.linkeddata.es/dataset/lov/

broad findability and finally interpretation and reuse of digital objects by researchers across communities. An extension of search to content aspects would be helpful, however, we need to have simple ways to define content patterns and an infrastructure supporting content search is required. In some analyses metadata is treated as "data", therefore it makes sense to treat metadata as separate Digital Objects of special types allowing machines to carry out suitable operations. It was argued that many metadata assertions are made at different phases, but that we do not yet have suitable means to collect these assertions to come to rich metadata. These challenges pose requirements on infrastructure development and computer science.

Possible solutions and support of computer science.

- An infrastructure based on the concept of Digital Objects such as worked out in RDA is urgently required. This infrastructure needs to be based on an identifier system that is available for everyone and supports the large number of stable links needed.
- Repositories and portals need to offer metadata annotation frameworks such that these extensions are kept as separate digital objects, but nevertheless can be made part of the findability context.
- An increased application of automatic workflows is required to automatically generate the rich metadata required by machines. However, this step will only be done if the domain researchers get access to easy to use and flexible workflow orchestration frameworks.

Accessible

Challenges. Automatic content negotiation including variants and versions is missing and creating inefficiencies for users. The rules for dealing with dynamic metadata and data have been specified clearly within RDA and the repositories need to make clear which policies with respect to versioning they are using. The current authentication mechanisms lack the support of detailed credentials such as specific "roles" which are needed for fine graded access control. Although basic technology exists, we miss an efficiently working authorisation system for several different distributed scenarios. Also, these challenges pose requirements on infrastructure development and computer science.

Possible solutions and support of computer science.

- Infrastructure components such as repositories need to support content negotiation and the application of the RDA rules on dynamic data.
- The currently used authentication systems need to be extended to support the needed detail of credentials.
- Effort needs to be taken urgently to design and develop practically usable authorisation solutions for distributed data scenarios such that also machines can easily find the correct information.

Interoperable

Challenges. It was stated that there are many different layers of interoperability starting with the structural and semantic specification of protocols enabling the exchange of, for example, digital objects up to the encoding of phenomena in data and metadata that are close to the research topics being studied. While the first can be specified in all detail, the others are subject of changes and dispute. In addition, semantic interpretation varies substantially with the semantic distance to the original source and the vocabularies used. It was concluded that at this level, semantic precision is an illusion. The minimum that is expected is that everyone defines and registers the schemas and concepts being used, allowing others to interpret and refer to them. Another important topic raised was the lack of means to link digital objects with operations that are suitable for their type. This led to some conclusions about the needed infrastructures and computer science actions. Especially with respect to interoperability, a gap between computer science research and urgently needed infrastructure building was considered.These challenges are huge and conceptual support from computer science is urgently required.

Possible solutions and support of computer science.

- Obviously a systematic and systemic approach for registries of schemas, concepts and vocabularies is required to overcome the current fragmentation. Such a solution would support users to reuse existing best practices and to optimise the implementation of "annotation by stealth" systems.
- A systemic solution to support crosswalks between different semantic spaces and to share these with other users would increase efficiency of work.
- Also, the broad availability and use of mechanisms to link types of digital objects with operations as suggested by RDA's data type registry specifications would increase efficiency for human users and open the path towards automation.

Reusable

Challenges. Detailed provenance information is required in particular to make workflow orchestration and execution possible. Yet there are no best practices how to link extensive provenance information allowing machines to find it. Improved mechanisms for provenance and containers would also improve the conditions for reproducibility as a subset of reusability and to access earlier versions of digital objects. Making licenses machine readable and actionable is a challenging topic, the way smart contracts are defined needs to be standardised. Increasing the quality of data sets is another big challenge and applying checkpoints in workflows to control quality is recommended. The review of data sets is difficult given the huge amounts of data and given that only a small percentage of data will be published officially. These challenges pose requirements on infrastructure development and computer science.

Possible solutions and support of computer science.

- More efforts need to be taken to have mechanisms to create and link provenance information that can be broadly used.
- Improve reproducibility of research results is an urgent requirement. Here we can refer to other Dagstuhl workshops, such as Reproducibility of Data-Oriented Experiments in e-Science[14]
- Also, in the area of machine actionable licenses more conceptual effort needs to be invested to come to a limited set of best practices.
- Improving data quality is an urgent requirement, since reuse is based on trust in the content. Ways to improve and test quality need to be worked out.

[14] https://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=16041

4.3 Working Group 3

Klaus Tochtermann (ZBW – Leibniz-Informationszentrum Wirtschaft – Kiel, DE / Moderator), Kathleen Gregory (Data Archiving and Networked Services, NL / Rapporteur), and Luiz Olavo Bonino da Silva Santos (GO FAIR – Leiden, NL / Contributor)

Findable

Challenges.

1. Searchability does not directly equate to findability, particularly for human actors, who engage in "user journeys" of browsing, linking to related content, serendipitous discovery, and direct searching. A variety of data, not just data produced through research, including open or governmental data, are also of interest.
2. Metadata creation is problematic. It is often generated by humans in unstandardized ways. Even when standards exist within a discipline, there are various interpretations and variations in how metadata fields are populated. At the same time, humans also provide a necessary level of control.
3. Sometimes a PID resolves to a metadata landing page; sometimes it redirects the user/agent directly to the data itself. This discrepancy is especially problematic for machine agents.
4. Private entities (e.g. Google DataSearch, publishers) are often in charge of discovery platforms. These entities could decide to restrict or cut off access to search platforms, which would limit the discoverability of data.

Possible solutions and support of computer science.

1. Semantic techniques could help both human and machine agents to locate distributed data.
2. Machine-generated metadata which is subsequently annotated by humans or semi-automated metadata generation could alleviate this problem.
3. The solution to this problem is not purely technical, but also falls within the policy realm. Policies need to be made that standardize where PIDs direct agents. Developing PID information types, metadata describing the information type where a PID points, could also be a possible solution.
4. Platforms should be open source to ensure that they remain open and accessible.

Accessible

Challenges.

1. If new communication protocols are required, how can it be demonstrated that they are open, free, universally implementable and allow for authentication and authorisation when necessary?
2. How long should metadata be retained for data that are no longer available?

Possible solutions and support of computer science.

1. Issuing certificates to individuals/agents, creating registries of access protocols, and certifying the openness of these protocols could help to automate and streamline access.
2. These metadata could be kept for a time period defined by a timestamp; at the defined time, archivists could review the metadata and decide to retain or delete the metadata.

Interoperable

Challenges.

1. The variety of metadata standards and representation languages between and within domains makes interoperability challenging.

Possible solutions and support of computer science.

1. Domain specific use cases exploring interoperability could help to better understand these differences. Semantic technologies, e.g., ontology alignment and applying translations between representation languages, could help to make these data interoperable.

Reusable

Challenges.

1. Provenance information is vital for reuse. Computer-actionable provenance schemas exist, but are perhaps underused.

Possible solutions and support of computer science.

1. The use of schemas for provenance information should be further investigated and encouraged. The code used to generate the data should also be included with the data or added to the metadata. The reuse of algorithms, software, and standards necessary for interpretation also needs to be supported.

5 Working Groups on Specific Types of Challenges for FAIR

5.1 Legal Tech: Licenses and Software to Ensure Trust

Marie Farge (ENS – Paris and CNRS, FR / Moderator) and Ron Dekker (CESSDA ERIC, NO / Rapporteur)

This working group discussed the complexity with legal issues regarding the access and use of research data and how technology and software in the sense of “Legal Tech” can help.

Definitions.

- Data: Any identifiable object is a data.
- License: It is a subset of contracts.
- FAIR: It qualifies the processes and protocols necessary to authorise a human or a machine to access data (its content and/or its metadata).
- Legal Tech: Use of computer science methods and software to help stakeholders solve
- legal issues on data production and use, e.g., on privacy and security questions, copyright and IPR (intellectual property rights), etc.
- reproducibility issues on all publications and processes, e.g., on the European GDPR (General Data Protection Regulation) issues, etc.

Present situation. Before internet and electronic publishing the interaction between a data owner and a data user was between two humans (who usually knew each others), or was mediated by a third human, usually a librarian. This interaction was reciprocal and based on mutual trust, without or with very little legal control.

Today, the interaction is between a data owner and a very large number of data users. They no more know each others, while today their interaction is mediated by networks, service providers, institutions, platforms (machines and software), and humans acting as brokers. Moreover, there are increasing requirements concerning privacy, security, economic or scientific value, and idiosyncratic risk (where each case requires new negotiations, new contracts or agreements, etc.).

Needed evolution. Key issues are trust and compliance, and today machines together with software step in. We need frameworks and infrastructures that can do verification using software. We would like to transform licenses into computational models.

Today access to confidential data is cumbersome. There is lack of trust between owner and user: on access, on use, on applications. There are many – too many – contracts, agreements, etc. Institutions act as brokers between owners and users, but this is labour intensive and complex. If machines become brokers, this will simplify offer better performance on compliance, and enable to do post-compliance (on new papers). If machines instead of humans are allowed to access data (i. e., queries via algorithms instead of access to individual data) and process them, one will no longer need to anonymise data to comply with privacy issues. Indeed, sometimes anonymisation is not possible; moreover, most of the time it induces loss of information, and it is not necessarily a guarantee for compliance.

One also needs to simplify the types or numbers of contracts/agreements and the legal and soft conditions for approval to use (e.g., on confidential data), to meet pre-conditions, post-approval compliance, etc. To achieve this, one could replace the current broker, doing checks manually, by machines and software. They would provide even more than a broker could ever do, hence provide more trust into the system.

We need to engage computer science with legal, economic, and technical communities. We need contract-language that is machine-parsable on transactions. For this one should specify the rules for audit and use distributed machine-learning. Digital Objects should also have the license-metadata with them.

Regarding the use of legal tech for reproducibility issues we left with the following open questions:

- Is there a method for reproducibility checking on confidential data done by a machine?
- Is it possible to do all reproducibility checks by a machine?

Use cases.

- NHS Research Passport[15]
- Image Processing Online (IPOL)[16]: Open Access publishing platform which offers the possibility to test one's own data on a given algorithm (implemented in open source) to see if it is useful for such data. Each article contains a text, an algorithm, and its source code, with an online demonstration facility and an archive of experiments. Text and source code are peer-reviewed and the demonstration is controlled. Moreover, the software of the publishing platform is open source and available on GitHub.

[15] https://www.nihr.ac.uk/about-us/CCF/policy-and-standards/research-passports.htm

[16] https://www.ipol.im

5.2 (Self-)Improvement of FAIRification

Carole Goble (University of Manchester, GB / Moderator & Rapporteur) and Michel Dumontier (Maastricht University, NL / Contributor)

The working group addressed the question: How "FAIRification" can be usefully assessed? A central focus of the discussion was about concerns in FAIR assessments and certifications.

The FAIR principles are aspirational – they articulate a set of desirable properties in digital objects to increase their potential to be discovered and reused by others. Achieving the vision of an Internet of FAIR digital objects will pose a substantial challenge to create it in a sustainable manner. Some aspects of the FAIR principles are readily achievable, while others may entail substantial and sustained effort. The likelihood that any given resource will completely fulfil the FAIR principles out of the box or at any given time is low – but that is to be expected and not a negative situation, and offers the opportunity to improve its value to a wider community of (re)users. Therefore, the role of any assessment tool should not be to "judge" a repository, but to provide indications of what can be expected from a resource.

Towards obtaining a picture of the state of FAIRness in digital objects, initiatives so far range from the development of questionnaires to elicit self reflection to gather largely qualitative assessments to metric-based software aim to gather evidence of quantitative adherence. Indications of what can be assessed as span this quantitative/qualitative spectrum. The language ranges from "metrics", understood as numerical (or ordinal) measures of quantitative assessment for comparison and compliance, to "indicators" as an attempt to embrace a range of signals beyond those that can be readily counted, and to incorporate non-mechanistic means of assessment that take into account the costs and return on investment of FAIRification of datasets by data providers. The working group highlighted the challenges of communicating FAIR by overly simplistic methods such as star ratings, as exemplified in early experiments with the DANS FAIR Assessment tool[17], for example.

Others, such as GO-FAIR, are examining the feasibility of FAIR certification via nationally accredited third-parties that could apply to datasets, repositories, software, services (such as training), organisations, and people (such as FAIR data stewards). Certification involves the confirmation of certain characteristics of an object, person, or organisation. Certifications aim to establish trust, set expectations in terms of quality and utility, offer choice, encourage criticism and roadmaps for improvement. An example of certification is The Core Trust Seal (CTS), which offers certification for online repositories, highlighted by the European Commission's High Level Expert Group on Turning FAIR into Reality. The working group cautioned against premature certification against assessment criteria that has yet to be fully understood or have buy-in from communities. Dangers highlighted included: favouring one community over another (for example digital librarianship and funder compliance over domain specialist data providers) and one actor over another (for example dataset consumers over dataset providers).

While well intentioned, current FAIR assessments and certification schemes only provide a narrow picture of features that may be needed to fully realize the FAIR vision. At NETTAB 2018, Christine Durinx, Executive Director SIB & co-lead of ELIXIR Data Platform, presented

[17] https://www.surveymonkey.com/r/fairdat

an expanded set of indicators that better captures the goals of provisioning of high quality database service (http://www.igst.it/nettab/2018/files/2018/10/NETTAB2018_Durinx.pdf). This included the use of persistent and unique identifiers, number and growth of entries in the repository, technical performance of system, use of community-recognised standards for (meta)data, documentation of provenance, availability of data, and customer service. In this way, the FAIR principles are an important, but underspecified set of requirements. Moreover, there is serious concern that a resource that does not adhere to the FAIR assessment or certification could be seen as having lower quality or value than a resource that ticks all the boxes. Additionally, fully meeting the FAIR principles may be prohibitively expensive for individuals and particular organisations. This is particularly of concern for organisations that are under pressure from funding agencies to be FAIR, while other well justified concerns prioritise their efforts.

Another promising option is to consider FAIR as a contract. By this we mean that FAIR is used as a mechanism for reporting, for expectation management of consumers, and as a roadmap. Contracts in every aspect of FAIR would enable an agent to reliably assess what can be expected from a resource and to decide if it should use the resource. Roadmaps and frameworks are helpful for systematic reviews. Review helps decision making and direction setting. For these reasons, the working group was in favour of developing "FAIRification Roadmaps" that would foster positive discussions with stakeholders on improving FAIRness with a sensitivity to the value-returning activities. The return on investment made by all stakeholders (repository owners, data consumers, funding agencies) is a key part of a "FAIRification maturity model". Established computer science principles can help to implement corresponding models, e.g., based on CMMI[18] and related assessments. See also the FAIR Capability Maturation Model which is developed in the FAIRplus project[19] as example.

The debate regarding FAIR assessment highlighted the spectrum of agendas, viewpoints, and contexts that need to be taken into account:

- from data consumers to data providers
- from objective automated FAIR assessments to manually mediated assessments incorporating subjectivity
- from focus on supporting automated analytics to supporting human interaction and decision making
- from FAIR assessments confined to technical aspects to those that incorporate social, political, and economic aspects (particularly for data providers)
- from a set of rules and tick lists to a roadmap or contract

Examples and use cases.

- FAIRmetrics[20]
- FAIRshake tool[21]
- ANDS FAIR Data Self-Assessment Tool[22]
- CMMI[23]
- DANS FAIR Assessment Tools[24]

[18] https://en.wikipedia.org/wiki/Capability_Maturity_Model_Integration
[19] https://fairplus-project.eu
[20] http://fairmetrics.org
[21] http://fairshake.cloud
[22] https://www.ands.org.au/working-with-data/fairdata/fair-data-self-assessment-tool
[23] https://en.wikipedia.org/wiki/Capability_Maturity_Model_Integration
[24] https://dans.knaw.nl/en/projects/projects

- Science Europe Guidelines[25]
- ELIXIR CDR criteria[26]
- Core Trust Seal[27]

References

1 Jerry Z. Muller. *The Tyranny of Metrics*. Princeton, 220 pp, February 2018. ISBN: 978 0 691 17495 2

2 James Wilsdon et al. *The Metric Tide Sage*. 168 pp, February 2016. ISBN: 978 1 4739 7306 0

3 Christine Durinx, Jo McEntyre, Ron Appel, Rolf Apweiler, Mary Barlow, Niklas Blomberg, Chuck Cook, Elisabeth Gasteiger, Jee-Hyub Kim, Rodrigo Lopez, Nicole Redaschi, Heinz Stockinger, Daniel Teixeira, Alfonso Valencia. *Identifying ELIXIR Core Data Resources*. URL: https://f1000research.com/articles/5-2422/v2

4 Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos & Michel Dumontier. *A design framework and exemplar metrics for FAIRness*. URL: https://www.nature.com/articles/sdata2018118

5 *Turning FAIR into reality*. Final report and action plan from the European Commission expert group on FAIR data. Nov 2018. ISBN: 978-92-79-96546-3 DOI: 10.2777/1524

6 Alastair Dunning, Madeleine de Smaele, Jasmin Böhme. *Are the FAIR Data Principles Fair?* URL: http://www.ijdc.net/article/view/567

5.3 FAIR Principles & Data Intensive Science

Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Kees den Heijer (TU Delft, NL / Rapporteur)

The big question that was addressed in this working group was whether data intensive science (DIS) poses different requirements with respect to the FAIR principles as other types of data usage. Since definitions of terms such as DIS and Big Data are fuzzy, the group, consisting of experts working on large data sets and computer scientists, first briefly characterised what is meant by DIS. Attributes were mentioned such as the 4 Vs (Volume, Variety, Velocity, Veracity), large data sets with numerical data generated by sensors and simulations, an increased obligation to use automatic workflows which need to be sufficiently flexible, an area of numerical transformation to integrate data from different sources, application of complex statistical AI methods to first extract knowledge before it can be exploited using semantic technologies, and existence of the iceberg phenomenon where much data is being created and reused from collaborators far before subset collections will be published.

In different scenarios, the terms used in the FAIR principles need to be interpreted in the respective contexts:

- Finding suitable data sets, for example, for training models by machine learning requires very rich metadata including provenance information. If search would be extended to

[25] https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

[26] https://www.elixir-europe.org/platforms/data/core-data-resources

[27] https://www.coretrustseal.org

content search, a broadly available HPC infrastructure would be required which does not exist.
- The need for increased richness of metadata leads to more complex search interfaces which researchers in general do not like, i.e., interface engineering to camouflage complexity is not trivial.
- The interpretation of the term "interoperability" as introduced in the FAIR principles needs[28] to be amended since the integration of typical data in DIS from different sources mostly requires extensive numerical transformations and normalisations due to differences between the underlying models and the calculations of missing data. At that stage one cannot speak of the application of "knowledge representation languages", instead we speak about structures and formats with headers informing about the variables.
- In DIS there is an urgent need to turn to automatic workflows with repeating patterns and extensions to create the necessary documentation. These, however, need to be flexible and parameterised, and to allow accessing subsets of data, demand specific knowledge to be created which is beyond the knowledge of the domain specialists.
- A reliable and highly performant system to register and resolve PIDs that will be available for many decades, that will be independent of protocol specifications which may change over time and that is free of misleading semantics such as included in URLs.
- Professional mechanisms to allow code to be moved to data and to be executed remotely are a must posing complex organisational and administrative challenges.
- The availability of smart VREs supporting the various functions would reduce the complexity of the tasks for the researchers.

Summarising, it was agreed that DIS requires specific interpretations of a few terms used in the FAIR principles and are stressing the needed infrastructures and services due to the scale of data and calculations. Computer science would have to collaborate closely with domain scientists to work out solutions for the various challenges indicated.

6 Working Groups on Identified Topics for the Dagstuhl Manifesto

6.1 Infrastructures & Services Aspects

Peter Wittenburg (Max Planck Computing and Data Facility – Garching, DE / Moderator) and Dieter Van Uytvanck (CLARIN ERIC – Utrecht, NL / Rapporteur)

This working group went through all notes to indicate relevant topics for building and extending infrastructures and the service landscape. Although building large infrastructures includes politics, socio-economics, and technology, the discussions focused mainly on the latter. Many points were addressed in the different sessions, here we can only mention a few:
- The need for a PID registration and resolution infrastructure available for every researcher (and beyond) and fulfilling several requirements (powerful, protocol-independence, stability of references over decades, support of standardised attributes, support of binding of digital object components to facilitate interoperability, inclusion of passport information, etc.).

[28] I1. (meta)data use a *formal, accessible, shared, and broadly applicable language* for knowledge representation

A PID system needs to be based on interoperability standards such as ITU X.1255 and a standardisation of attributes as done in RDA.

- A set of ready-made services to continuously monitor the availability and state of URI and PIDs and make use of its possibilities would be extremely helpful (link checking, cross-referencing, integrity checking, etc.). Since these checks can be resource-intensive (e.g. computing, network) some degree of centralization, as to prevent duplication, can provide a higher efficiency.
- It is important to have improved support for rich metadata creation, exposure, searching, and mapping, including provenance facilitated by registered schemas and semantics. A limited set of serialisation formats should be accepted, such as XML, JSON, RDF, etc.
- Since the creation of metadata is a demanding task, automatic extraction of metadata on the basis of data files should be considered. A landscape analysis of such available technologies would be valuable.
- We need a systematic and systemic approach to schema, concept, and vocabulary registration allowing people to easily register, find, and reuse them. RDA offers such a schema registry. This can be seen as a basis for much better support of semantic crosswalks.
- Changes to data and metadata must be versioned and traceable to understand the state of the data at their (various) time(s) of use.
- A systematic solution allowing registering types of digital objects and link sets of operations with it, such as developed in RDA, is highly needed to foster automation.
- A much more improved authentication and authorization infrastructure addressing different distributed scenarios is highly required. This must be amended by computer-readable license consent (smart contracts).
- Much more support for creating flexible workflows is required. These should also facilitate reproducibility and should make use of state of the art packaging formats, as suggested, for example, by Research Objects[29].
- A move towards self-explanatory APIs that include semantic annotations is needed to facilitate machine action. They are digital objects having a PID and descriptions.
- Much better support for the scenario where code is transferred to the data to be executed is required.
- Finally it was agreed that suitable VREs bundling the access to infrastructures and its many services would offer great opportunities for more easy access to all features by researchers.

6.2 Computer Science Research Topics

Achim Streit (KIT – Karlsruher Institut für Technologie, DE / Moderator) and Tobias Weigel (DKRZ Hamburg, DE / Rapporteur)

This working group discussed research topics in computer science related to implement FAIR data infrastructures. This covers a broad spectrum of subfields from mathematical foundations, algorithms and data structures, security, AI, software engineering, applied

[29] http://www.researchobject.org

computer science up to theory of computation (cf. Wikipedia page on computer science[30]). The working group put a focus of its discussion in defining future computer science research topics in the form of PhD topics. These were:

- Tamper-proof exchange and tracking of digital objects across distributed infrastructures
- Provenance capturing and reasoning on provenance data to enable automated data identification and integration across heterogeneous domains
- Privacy-preserving analytics across distributed data sets
- Demonstration of a fully automated closed loop research cycle system, from hypothesis generation, data identification, analysis, hypothesis verification to novel hypothesis derivation
- Automated informed consent negotiation and inference
- Intelligent content-based searching using AI and performance benchmarking with state-of-the-art metadata-powered search techniques
- Visual analytics in support of data finding – new forms of Human-Computer-Interaction based on interdisciplinary research with Arts and Social Sciences
- Quantifiable FAIR-ability of architectural frameworks of data infrastructures
- Representation of ethics and moral in technical solutions in FAIR data infrastructures
- Semantic* in support of intelligent FAIR services and based on ontologies and vocabulary crosswalking
- Security & Privacy frameworks for making data FAIR
- Framework to analyse the impact of FAIR metrics
- Novel data management/storage concepts enabling persistent provisioning of large-scale research data across evolving versions aggregated over long time scales
- New approaches for digital preservation to ensure FAIR data remains FAIR over long periods of time
- Social Machines and FAIR: Crowdsourcing FAIR
- Trust and identity in the context of FAIR data infrastructures

6.3 FAIR Computer Science Research

Wilhelm Hasselbring (Universität Kiel, DE / Moderator & Rapporteur) and Paolo Manghi (ISTI-CNR – Pisa, IT / Contributor)

After focussing on how computer science can contribute to implement the FAIR principles in several working groups, this working group addressed the question how are the FAIR principles are currently adopted by computer science itself. The group started to discuss examples from software engineering.

In software engineering everything you can store as file is called an artifact. Most major software engineering conferences meanwhile offer artifact evaluations for papers accepted to the conference main programme. ACM SIGMOD, for instance, calls its artifact evaluation as reproducibility evaluation. Papers that also have artifact evaluation have more citations [1]. In some cases, computer science experiments are not reproducible and only repeatable (e.g. performance research, you need the same hardware to reproduce results). See also [2].

[30] https://en.wikipedia.org/wiki/Outline_of_computer_science

Based on this input the group started to collect aspects and issues to make computer science research FAIR.

Findable.

- Publish as much as possible. Artifacts are: software, data, employed methods, workflows, protocols, services, virtual machines/containers, documents, etc.
- Assign PIDs to everything (not necessarily DOIs).
- (Pre-)Publication of scientific processes/workflows (e.g. protocols.io).
- Use repositories like DockerHub.
- Software metadata remains a great challenge: software citation (SSI, OpenAIRE, CodeMeta), software identification (RDA Group).
- Executable papers and echanced publications [3].

Accessible.

- Research artifacts should be published with preservation in mind. GitHub, for example, does not do that. Publishing involves citation and preservation (e.g. Zenodo.org).

Interoperable

- Portable software is required (subproblem of software preservation).

Reusable.

- Use (and describe in the metadata) standard tools for sharing scientific thinking. Containers (Docker etc) and virtualisation help. Same for Jupyter notebooks. Use workflow languages such as the common workflow language (CWL)[31].
- Allowing repeatability by offering cloud-based services, such as VREs for example. Distinguish between Software-as-Code (e.g., via GitHub) and Software-as-a-Service (e.g., via some cloud service).
- Follow standards for APIs and metadata. Documentation is essential.
- Conference artifact evaluation processes already help to check quality via peer review.

References

1 Bruce R. Childers; Panos K. Chrysanthis.*Artifact Evaluation: Is It a Real Incentive?*. 2017 IEEE 13th International Conference on e-Science (e-Science). DOI: 10.1109/eScience.2017.79

2 Shriram Krishnamurthi and Jan Vitek. *The real software crisis: repeatability as a core value.* Commun. ACM 58, 3 (February 2015), 34-36. DOI: https://doi.org/10.1145/2658987

3 Bardi, A. and Manghi, P. *Enhanced Publications: Data Models and Information Systems.* LIBER Quarterly, 23(4), pp.240–273. DOI: http://doi.org/10.18352/lq.8445

7 Acknowledgements

We thank all participants of the workshop for their valuable contributions and the Dagstuhl team who have made this a successful event.

[31] https://www.commonwl.org

Participants

- Marcel R. Ackermann
LZI Schloss Dagstuhl & dblp
Trier, DE
- Luiz Dlavo Bonino da Silva Santos
GO FAIR – Leiden, NL
- Timothy W. Clark
University of Virginia, US
- Ron Dekker
CESSDA ERIC, NO
- Kees den Heijer
TU Delft, NL
- Michel Dumontier
Maastricht University, NL
- Marie Farge
ENS – Paris and CNRS, FR
- Sascha Friesike
VU University of Amsterdam, NL
- Carole Goble
University of Manchester, GB
- Kathleen Gregory
NL
- Gregor Hagedorn
Museum für Naturkunde – Berlin, DE
- Wilhelm Hasselbring
Universität Kiel, DE
- Oliver Kohlbacher
Universität Tübingen, DE
- Paolo Manghi
ISTI-CNR – Pisa, IT
- Natalia Manola
University of Athens, GR
- Daniel Mietchen
University of Virginia, US
- Peter Mutschke
GESIS – Leibniz Institute for the Social Sciences – Cologne, DE
- Heike Neuroth
FH Potsdam, DE
- Andreas Rauber
TU Wien, AT
- Marc Rittberger
DIPF – Frankfurt am Main, DE
- Raphael Ritz
Max Planck Computing and Data Facility – Garching, DE
- Guido Scherp
ZBW-Leibniz-Informationszentrum Wirtschaft – Kiel, DE
- Birgit Schmidt
SuB – Göttingen, DE
- Achim Streit
KIT – Karlsruher Institut für Technologie, DE
- Klaus Tochtermann
ZBW-Leibniz-Informationszentrum Wirtschaft – Kiel, DE
- Dieter Van Uytvanck
CLARIN ERIC – Utrecht, NL
- Tobias Weigel
DKRZ Hamburg, DE
- Mark D. Wilkinson
Polytechnic University of Madrid, ES
- Peter Wittenburg
Max Planck Computing and Data Facility – Garching, DE