Manifesto from Dagstuhl Perspectives Workshop 18472

Implementing FAIR Data Infrastructures

Edited by

Natalia Manola[*1], Peter Mutschke[*2], Guido Scherp[*3], Klaus Tochtermann[*4], Peter Wittenburg[*5], Kathleen Gregory[†6], Wilhelm Hasselbring[†7], Kees den Heijer[†8], Paolo Manghi[†9], and Dieter Van Uytvanck[†10]

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 Information Centre for Economics – Kiel, DE,** g.scherp@zbw.eu
4 **ZBW – Leibniz Information Centre for Economics – Kiel, DE,** k.tochtermann@zbw.eu
5 **Max Planck Computing and Data Facility – Garching, DE,** peter.wittenburg@mpi.nl
6 **Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, NL,** kathleen.gregory@dans.knaw.nl
7 **Universität Kiel, DE,** hasselbring@email.uni-kiel.de
8 **TU Delft, NL,** c.denHeijer@tudelft.nl
9 **ISTI-CNR – Pisa, IT,** paolo.manghi@isti.cnr.it
10 **CLARIN ERIC – Utrecht, NL,** dieter@clarin.eu

Abstract

The open science movement is gaining strength and momentum worldwide, signalling a fundamental shift in how scientific research is made accessible and reusable. In order to fulfill the promises of open science, reliable and sustainable research data infrastructures must be developed. While the FAIR data principles provide a promising conceptual basis for developing such data infrastructures, they do not provide technological guidance on how to do so.

Computer science is uniquely situated to fill this gap by researching and developing tools and technical specifications which can help to realize the creation of FAIR data infrastructures. To this end, this Dagstuhl Perspectives Workshop brought together computer scientists and digital infrastructure experts from across disciplinary domains to discuss key challenges and technical solutions to implementing and promoting the establishment of FAIR-compliant infrastructures for research data. This manifesto reports the findings from the workshop and provides recommendations along two lines: (1) how computer science can contribute to implementing FAIR data infrastructures and (2) how to make computer science research itself more FAIR.

Perspectives Workshop 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/DagMan.8.1.1

[*] Editor/Author
[†] Author

Except where otherwise noted, content of this manifesto is licensed under a Creative Commons BY 3.0 Unported license
Implementing FAIR Data Infrastructures, *Dagstuhl Manifestos*, Vol. 8, Issue 1, pp. 1–34
Editors: Natalia Manola, Peter Mutschke, Guido Scherp, Klaus Tochtermann, and Peter Wittenburg

Dagstuhl Manifestos
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Executive Summary

The FAIR Data Principles emphasize the importance of making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable [56]. They serve as a widely accepted conceptual basis for the development of research data infrastructures, as in the case of the European Open Science Cloud (EOSC)[1]. The FAIR principles do not, however, define technical specifications or suggest how they could be practically implemented within a research data infrastructure. As the principles do not provide a blueprint for their implementation, there is both a lack of specific reference models to guide the process of making data FAIR and a diversity of ways in which the principles are being implemented.

Computer science can greatly contribute to addressing this problem by developing technical approaches and through research. The Dagstuhl Perspectives Workshop "Implementing FAIR Data Infrastructures," which took place on November 18-21, 2018 with 29 participants, was convened to further explore these potential contributions. At the workshop, computer scientists and digital infrastructure experts from across disciplinary domains discussed the challenges, open issues, and technical approaches for implementing the FAIR Data Principles in research data infrastructures. The workshop aimed to identify the key elements required for the transition from scientific e-infrastructures and services to FAIR-compliant data infrastructures. In so doing, the workshop also provided an opportunity to further define the role of computer science in implementing FAIR-compliant infrastructures and in advancing open science practices as a whole.

This manifesto summarizes the key findings of the workshop[2] with the aim of stimulating future discussions and contributing to a deeper understanding of the core issues surrounding both the FAIR principles and their implementation within infrastructures. Recommendations are provided in the manifesto along two lines: how computer science can contribute to implementing FAIR data infrastructures (Section 2) and how to make computer science research itself more FAIR (Section 3).

Section 2 examines the technical means which can be used to implement the FAIR principles in research data infrastructures and in data services, as well as addressing how these challenges are related to areas of computer science research. Section 3 continues to investigate the role of research, although here the focus turns to questions such as how the FAIR principles are currently adopted within computer science as well as possible steps to increase the openness of research activities.

The recommendations proposed throughout the manifesto are briefly summarized below.

Recommendations: Implementing FAIR-compliant data infrastructures and services

These seven recommendations identify key steps for infrastructure and service providers, as well as for providers of innovative services resulting from computer science research.

1. Provide a user-friendly infrastructure that allows for the registration of persistent identifiers (PIDs) for all kinds of digital objects. This will help to establish interoperability at the data organization level.

[1] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[2] The manifesto represents the editors/authors' best endeavor to capture the key positions highlighted during the meeting by several participants, but may vary from individual opinions.

2. Develop registries of trustworthy repositories allowing users to quickly find services. To do this, a common certification scheme is needed.
3. Implement automatic processing for authentication & authorization in distributed scenarios. This should result in creating a "digital data passport."
4. Improve support for creating rich metadata. Possible ways to do this include automatic extraction of metadata and provenance information from workflows, as well as providing assistance for metadata validation and transformation.
5. Create schema and vocabulary registries which allow people to easily find and reuse these resources and to better support the use of semantic mappings, annotations or crosswalks.
6. Develop rich search interfaces for discovering (multidisciplinary) data.
7. Encourage widespread adoption of appropriate formal languages and techniques enabling interoperability and workflows.

These recommendations have a strong focus on improving automation and machine-actionability on the infrastructure level, incorporating common standards and vocabularies to provide easy-to-use and intelligent data services for researchers. In order to foster the implementation of these recommendations, and of FAIR data infrastructures in general, the following research areas were identified as being particularly relevant:

- Trustable exchange and tracking of digital objects
- New approaches for FAIR digital preservation
- Novel data management and storage concepts
- Intelligent data discovery
- Semantic aspects
- Distributed Computing

Recommendations: Increasing the FAIRness of computer science

Workshop participants found that the adoption of the FAIR principles, and of open science practices in general, is surprisingly not widespread within computer science.

Participants built on the work of [16], focusing on the "R dimensions" (i.e. repeatability, replicability, reproducibility, and reusability) of FAIR, to consider the FAIRness of the entire scientific process. This perspective requires that both research outputs and research workflows are reproducible and adhere to the FAIR principles.

The following four recommendations - targeted toward researchers, research communities and institutions, policy makers and funders - were made to advance the degree of FAIRness in computer science research.

1. Foster a sharing / publishing culture.
2. Foster (peer) review and scientific reward for published artifacts.
3. Promote the use of discipline-specific digital laboratories and publishing workflows to support the entire scientific process.
4. Foster cultural change.

Table of Contents

Executive Summary . . . 2

Motivation . . . 5

How computer science can contribute to implementing FAIR data infrastructures . . . 6

Introduction . . . 6

Status Quo: Implementation of FAIR . . . 7

Key Recommendations . . . 15

Related Computer Science Research Topics . . . 16

How to make computer science research more FAIR . . . 17

Introduction . . . 17

Status Quo: FAIR* Computer Science . . . 19

Key recommendations . . . 26

Participants . . . 28

References . . . 29

1 Motivation

The digitization of science and the increasing emphasis on sharing and reusing research data require the development of new infrastructures, tools, and methods for advancing open scholarship. Against this background, the open science movement is gaining strength and momentum worldwide, signalling a fundamental shift in how scientific research is made accessible and reusable - not only for the scientific community but also for industry and society as a whole. In order to fulfill the promises of open science, work to develop reliable and sustainable research data infrastructures is being undertaken, as evidenced by the emergence of the European Open Science Cloud (EOSC)[3]. While providing and using open access publications is becoming an increasingly accepted research practice, the same cannot be said for open data or open methodologies. Openly sharing research data and processes used during scientific production requires a mentality shift for many scientists. Reflecting this, there is also a great lack of available models, infrastructures, and services that enable researchers to cooperate on data, to share results, and to transform previously isolated research elements into an open, collaborative system.

In this context, the FAIR Data Principles [56] have become a common and widely accepted conceptual basis for planning and building research data infrastructures. The adoption of the term 'FAIR' mirrors the recognition that the not all data can be 'open,' as in the case of disciplines such as medicine or the social sciences, where sharing data openly is not always possible due to concerns about data privacy. The FAIR principles consist of four core facets: data must be **F**indable, **A**ccessible, **I**nteroperable, and **R**e-usable. However, the FAIR principles themselves are neither a specific standard nor do they suggest specific technologies or implementation pathways. They describe core characteristics of findable, accessible, interoperable and reusable data, but they are far away from providing practical guidelines that can assist data providers in implementing FAIR-compliant data management platforms. As a consequence, the FAIR principles allow a broad range of implementation solutions; this runs the risk of creating a highly fragmented data and service landscape.

Various European-wide initiatives address the concept of FAIR research data. The Expert Group on FAIR Data of the European Commission published their report [13] with recommendations for applying the FAIR principles to the development of the EOSC. The GO FAIR[4] initiative represents a bottom-up, stakeholder-driven initiative to implement the FAIR principles. Moreover, a number of projects like FAIRsFAIR[5], ENVRIFAIR[6], FAIRPlus[7] and EOSC-Life[8], have been recently awarded funding from the European Commission. Other discipline-specific initiatives, such as FAIR-DI[9], are in the process of investigating how to foster and implement the FAIR principles within particular research communities.

Despite the emergence of such efforts, the process of understanding how the FAIR principles can be fully implemented in research data infrastructures, as well as how such efforts can be measured (see the FAIR Data Maturity Model Working Group of RDA[10], is just beginning. In view of the "need for a fast track implementation initiative [of the

[3] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[4] https://www.go-fair.org
[5] https://www.fairsfair.eu
[6] https://envri.eu/envri-fair/
[7] https://fairplus-project.eu/
[8] https://www.eosc-portal.eu/eosc-life
[9] https://fairdi.eu/
[10] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

EOSC]" [25], the time is ripe for turning the FAIR principles into practice. Computer science is uniquely positioned to facilitate the implementation of the FAIR guidelines in research data infrastructures.

With this aim, this Dagstuhl Perspectives Workshop took the recommendations of the European Comission Expert Group on FAIR Data ("Turning FAIR into reality" [13]) as a starting point to consider how computer science can contribute, both through technical solutions and through research, to enabling data providers to make their data FAIR. The central goal of the workshop was to bring together computer scientists, digital infrastructure experts and researchers from across domains to discuss the key elements required for the transition from scientific e-infrastructures and services to FAIR-compliant data infrastructures. The workshop also provided an opportunity to develop a vision for computer science in this space for the coming years and to explore the role of computer science in advancing open science practices as a whole. In this context, this document highlights the findings of the workshop and provides recommendations along two lines: how computer science can contribute to implementing FAIR data infrastructures (Section 2) and how to make computer science research itself more FAIR (Section 3).

Since the initial submission of this manifesto, a number of papers have been published which address FAIR from different perspectives. The special issue of the journal Data Intelligence on emergent FAIR practices[11], especially the two comprehensive introductory articles on the interpretation of the FAIR principles and their implementation in practice [40, 31], are of particular relevance. Also of note is recent work exploring FAIR Digital Objects [18], "FAIRsemantics" [36] and maturity models for FAIR [45]. While these articles represent a major step forward in the discussion of implementing the FAIR principles, the points they raise strengthen rather than weaken the considerations and recommendations discussed in this manifesto.

2 How computer science can contribute to implementing FAIR data infrastructures

2.1 Introduction

Data-intensive science presents unique challenges, in particular (1) extracting knowledge from increasingly vast amounts of distributed, complex data and (2) integrating and exploiting the knowledge from various sources. Infrastructures providing stable and persistent identifiers and access points to (meta)data, whose specific implementation may change over time, are needed to facilitate these two major challenges. Such infrastructures should form and contribute to an ecosystem of essential components and services at various layers indicated above [13].

Computer science can help to overcome the many sources of inefficiencies caused by current gaps in this infrastructural ecosystem by suggesting better methods for working with data and designing generalised interfaces. This chapter focuses on how computer science can address problems involved at the infrastructure and service level to make data infrastructures more FAIR-compliant (Section 2.2). Based on the recommendations of the EC report on FAIR Data [13], we identified a number of requirements and fields of actions for computer

[11] https://www.mitpressjournals.org/toc/dint/2/1-2

science (Section 2.3), in particular from the perspective of data-intensive science. A further focus of the chapter is to identify areas of computer science research which can be used to foster the implementation of FAIR data infrastructures (Section 2.4).

Services which provide FAIR data are important in overcoming recent hurdles regarding data sharing and reuse. The concept of Digital Objects (DO) was discussed as one promising way of achieving interoperability between repositories and registries to address these challenges. (The term "DO" was used in the describtion of the core model developed by the RDA Data Foundation & Terminology Group; this term was then extended to FAIR Digital Objects in the report of EC's expert group on FAIR implementation [13] and by experts in the RDA GEDE Collaboration) [32, 6, 58, 50] .

A related approach to be taken into account is the concept of Research Objects[12] [5], which forms part of the IEEE Standardisation activity for BioCompute Objects[13]. Trustworthy repositories and registries providing data and metadata services are essential pillars in a stable data domain. Improved functionality, more focus on "unpublished data" [41] and certification according to CoreTrustSeal[14] and emerging FAIR Maturity Indication models [57] (cp. FAIR Metrics[15], FAIR data assessment tool[16], 5 Star Data Rating tool[17,18], FAIR self-assessment tool[19]) are important to developing trust.

Distributed Authentication and Authorisation (AAI) methods facilitating access to repositories need to be enhanced to meet researchers' needs in distributed scenarios. Improved approaches to create rich metadata, to represent metadata in ways that facilitate reuse, and to develop search interfaces are urgently needed. There is no doubt that a higher degree of automation will need to be achieved to make use of large volumes of data more efficiently. Better support for the creation, management and exchange of workflows by researchers who are not IT experts is required. This needs to be accomplished by creating methods to easily save and exchange the contexts of data processing procedures to allow easy replication. Finally, better and easier-to-use frameworks for semantic operations such as supporting cross-walks were discussed.

2.2 Status Quo: Implementation of FAIR

Digital Objects and Persistent Identifiers

There are many different ways for implementing FAIR data infrastructures. To achieve a higher degree of convergence it is urgent to develop interoperability solutions at various layers. The concept of Digital Objects (DOs) addresses the interoperability challenge at the data organisation level. According to recent surveys [42], 80% of the time of data scientists is wasted with "data wrangling." To a large extent this is due to bad data organisations, the fact that the necessary information to access and interpret data is inexplicit, and overall low data quality. Currently, repositories organize data differently; for example, they store

[12] http://www.researchobject.org
[13] https://osf.io/h59uh/
[14] https://www.coretrustseal.org
[15] http://fairmetrics.org
[16] http://blog.ukdataservice.ac.uk/fair-data-assessment-tool
[17] https://research.csiro.au/oznome/tools/oznome-5-star-data
[18] http://oznome.csiro.au/5star/
[19] https://www.ands-nectar-rds.org.au/fair-tool

data and types of metadata in different ways by applying different models (i.e. files, clouds, databases, spreadsheets, XML schemas, etc.).Basic essential information is often not provided by the creators.

The DO specifies that it has a bit sequence that can be stored in various repositories; this bit sequence is associated with a Persistent Identifier (PID) and different types of metadata. These metadata descriptions may be DOs themselves, as it is typical, for example, for repositories. A separate type of metadata is the PID Kernel Information [55], which is stored together with the PID in the identifier resolution system. This metadata affords autonomous decisions taken by software agents at the e-infrastructure level. According to the recommendation of the RDA Data Foundation & Terminology group, the concept of DO specifies that these different entities need to form an inseparable bundle if FAIRness is to be achieved.

The DO concept allows the DO Interface Protocol (DOIP)[20] to be defined. This protocol can be seen as a "gold standard" to normalize organisational differences, just as TCP/IP was used to interconnect the many different network types decades ago. If a client has a PID, it should be possible for the client to request either the bit sequence or the checksum or to get the pointers to the descriptive metadata, the access rights record, the transaction records stored in a blockchain etc. Interoperability in this respect is defined by the ITU X.1255[21] standard. DOIP, in conjunction with the protocol to resolve PIDs associated with DOs, does exactly that, independent of how repositories and scientific communities organise and model their data and metadata. Currently, there is a broad discussion about the second version of DOIP; this is particularly important given the impact of the DO concept.

Thus, for the benefit of users, the concept of DO abstracts away from all implementation details. The RDA Data Type Registry group (DTR)[22] defined a standard (now an ISO process) to link "data types" with "operations" which is already being tested in some communities. Therefore, a DO's PID can also be used to access a DTR and find suitable operations to carry out processes on the DO's bit sequence. In doing so the DO also enables encapsulation, which is known to be a powerful concept in managing complex systems.

From the above, it is obvious that the DO concept heavily relies on a stable system for PIDs which can be resolved to useful PID Kernel Information. With the Handle system[23], such a globally available infrastructure is ready to be used. Many ESFRI (European Strategy Forum on Research Infrastructures) communities agreed already on the use of persistent, unique and globally resolvable identifiers (PIDs) for all data entities. This is particularly important for all data which is being created in the labs, being dealt with in collaborations, referred to from, i.e. workflows, metadata, etc. and which will either not be published at all or only published at a later stage. This data covers far more than 80% of the total amount of data that is being created and stored, and is increasingly often associated with Handles issued by different service providers such as ePIC (European PID Consortium)[24]. Published collections of data are often associated with DOIs[25] which are also Handles with prefix 10. In doing so, many repositories implement the FAIR principles in this respect. These communities are also making use of the possibility to add "passport"-type information (i.e. checksum, creation date, version, etc.) and important references (types of metadata) as

[20] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
[21] https://www.itu.int/rec/T-REC-X.1255-201309-I
[22] https://rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries
[23] https://en.wikipedia.org/wiki/Handle_System
[24] https://www.pidconsortium.eu
[25] https://en.wikipedia.org/wiki/Digital_object_identifier

attributes to the PID record. The RDA PID Kernel Information working group [21] started to standardize such attributes in open type registries to improve interoperability at this layer. Yet only few communities make use of these agreed upon kernel types and, for example, also follow policy recommendations such as adding a "tombstone" mark in the PID record when the bit DO's sequence is being deleted.

It is time to establish an easy-to-use infrastructure that allows the registration of PIDs for all kinds of DO types by every researcher and which acts as a support for all kinds of tools to effectively implement the FAIR guidelines. Such a system would allow us to also realize the FAIR-compliant domain of DOs, to establish interoperability at the data organisation level systematically, to motivate repositories to develop suitable adapters to DOIP and to pave the way towards automation of data processing. Here, a particular challenge is seen in dynamic data citation [43, 44] which has been addressed by the RDA Working Group on Dynamic Data Citation[26]. The working group issued recommendations for identifying arbitrary subsets of (potentially highly dynamic) data by (1) timestamping and versioning changes to evolving data and by (2) assigning PIDs to the queries selecting the according subsets.

Trustworthy Repositories and Registries

A research infrastructure landscape is populated by repositories - which store, manage, curate and offer access to data, metadata and other types of digital objects - and by registries aggregating various types of metadata (i.e. descriptive, scientific, access rights, transactions, etc.). All these repositories and registries are core pillars of such infrastructures and we rely on their services and their trustworthiness. With respect to their services, a list of wishes was formulated: 1) format recognition and validation tools should be provided to extract and enrich metadata, 2) content negotiation which makes use of the PID infrastructure should be supported, 3) there should be clear references about which schemas and semantic spaces are being used for all digital objects, 4) there should be options to let users add annotations and tags and to offer them to new users in addition, and 5) there should be registries of trustworthy repositories and registries allowing users to quickly find services.

It is well-known that about 80% of the data that is being created and stored in repositories has never been used for published studies or will only be officially published after long periods. Nevertheless, many of this "dark data" [41] is being reused in collaborations and referenced by workflows etc. It is an urgent recommendation therefore to assign PIDs to almost all digital objects being created in order to have stable references and thus increase reproducibility. This "dark data" needs to be made part of the visible data domain.

Tracking reuse of data and software is an interesting topic for various purposes and has not yet been solved satisfyingly. The systematic use of PIDs would help greatly, but also tools that extract references to digital objects in full text papers and relating them (for first approaches see InFoLIS[27] and OpenMinTeD[28]) would be very valuable in case that PIDs are not used. These methods could also be used to relate statements made in a paper with statements in earlier papers or with original data.

It is commonly agreed that certification of repositories and registries is urgently required to indicate levels of trustworthiness, i.e., implying that there cannot be a black/white statement about the compliance with the CoreTrustSeal requirements and the compliance

[26] https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html
[27] https://www.gesis.org/?id=8948
[28] http://openminted.eu

to the FAIR Principles. Currently, a globally used certification scheme is used to audit repositories on the quality of their procedures. The CoreTrustSeal (CTS)[29], for example, which was developed within RDA and which is already widely used, is based on years of experience. This certification promotes a self-assessment approach, i.e. the assessment only checks whether the claims a repository is making are indeed true. This allows users to identify whether they find a repository sufficiently trustworthy for their goals.

CTS, however, does not assess whether the digital objects stored are FAIR. Many initiatives are now working on developing FAIR Maturity Indication models that can make statements about the degree of FAIRness of a data set in a repository or registry. It is agreed that a self-mandated organisation assessing FAIRness will not be accepted. Gradational maturity models are therefore needed to assess the degree to which data are FAIR; these models should take into account both domain-specific research data as well as, and in particular, cross-discipline research data. The aim of the maturity model should not be only to indicate the degree of the maturity of the FAIRness of data under study, according to the 15 criteria for FAIR (which may evolve over time), but also to point to specific needs for actions depending on the level of FAIRness of the data in question.

In practice, however, not all criteria will be fulfilled in the same way by real-world data infrastructures. Therefore, gradational maturity levels would better allow indicating to which extent the data under study are FAIR or not. A gradational maturity model would allow providing clear indications about which criteria for which single FAIR principle are to which degree fulfilled and which are not. By this, data providers are enabled to clearly identify concrete action points to increase the FAIRness of their data and the FAIR compliance of their data management systems respectively. A gradational maturity model should be provided as an adaptable tool for assessing FAIR. The model should be validated in a research-centered way to ensure that the theoretically developed maturity model takes researchers' needs into account in a proper way. A still open question is who defines the criteria. To achieve high acceptance community, involvement is highly needed. It was also agreed that the certification schemes will need to evolve over time, and that CoreTrustSeal and FAIR Maturity Indication Models may merge over time given this overlap.

Authentication and Authorisation

Current authentication and authorisation infrastructures are not appropriate for a number of distributed scenarios and for the automation of data processing which will be a requirement for the future. The bit sequences of digital objects are often stored in several repositories, for example, for preservation reasons or access efficiency, but the access permissions and licences should in most cases be independent of the specific repository delivering the content. Yet there is no suitable authorisation infrastructure to support such a distributed scenario, with the consequence that copies are often simple backup copies that cannot be accessed. With respect to licenses, attempts have been made in Finland[30] to have a central license registry where users can sign agreements once that then are being used for all content that share the same license type. Such an infrastructure is urgently needed and needs to support high security requirements.

With respect to authentication, the current set of credentials is not sufficient for scientific use. The introduction of detailed roles and groups is needed. The codification of ethical rules and their integration into access procedures would be a great step ahead as well.

[29] https://www.coretrustseal.org/about/

[30] https://www.csc.fi/en/consultation-and-tailored-solutions/-/asset_publisher/KP8P2hmf5Vk2/content/fiona-tilastokeskuksen-etakayttojarjestelma

The trend towards automatic processing requires rethinking authentication & authorisation allowing software agents to act on behalf of users and to interpret license conditions; "smart contracts"[31] as they have been introduced in the blockchain world[32] need to be applied. These "smart contracts" lack the generality and standardization which is required to facilitate data science. Thus, we need frameworks and infrastructures that perform verification of access and processing conditions at all levels of interaction with digital objects to improve the transparency of data access and use. For this, mechanisms to transform licenses into formal, computational models (i.e. a machine-parsable contract-language) are needed. These mechanisms should automatically match user requests to preconditions of use and provide machine-actionable information about access rights, as well as a sort of post-approval compliance monitoring. Those frameworks should return a digital "data passport" of sorts (computer actionable licences) for each dataset that includes certificates for openness.

Metadata Representation and Searching

From linking to searching and enhanced data use. Creating and using links between (meta)data is key to overcoming issues of findability and interoperability, particularly given the ever increasing amounts of both research data and metadata. In order to increase both findability and interoperability, all data items should ideally be represented according to the Digital Object concept and be marked by a PID. All metadata should be represented according to common representation schemas and vocabularies. This will make metadata exposures findable by crawlers, as well as enable the creation of Linked Open Data (LOD), resulting in an ecosystem where datasets can be treated as a single web of knowledge. Although a number of Web standards exist, data are often represented by proprietary metadata and vocabularies; this makes finding datasets a challenging task.

This situation suggests the creation of assistance services (i.e. a knowledge-based system) that either provide guidance or transform metadata into common representation schemas using broadly accepted vocabularies (DCAT[33], DCMI[34], RDA Schema Catalogue[35] or schema.org[36]), as well as conventions for using schema.org in scientific domains, for example Bioschema.org[37]. Transformation of metadata into common standards is important, as datasets should be allowed to have multiple descriptions, using different domain-specific metadata standards. This would allow for dynamic and fit-for-purpose enrichment of metadata and inks to other objects, both during and after the data publishing stage, by people creating and using the dataset. LOD could be used to express relations between heterogeneous descriptions and entities.

There is a need for improved support for activities such as creating and exposing rich metadata, searching and mapping, including provenance information, that is facilitated by registered schemas and semantics. A limited set of formats should be accepted for metadata transformation (such as XML, JSON, RDF, etc.).A possible area of investigation is to see how (or if) these standards should be extended for advanced data retrieval needs and to better meet the needs of particular communities.

[31] https://en.wikipedia.org/wiki/Smart_contract
[32] https://www.nist.gov/publications/blockchain-technology-overview
[33] https://www.w3.org/TR/vocab-dcat/
[34] http://dublincore.org
[35] https://www.rd-alliance.org/metadata-standards-directory
[36] https://schema.org
[37] https://www.rd-alliance.org/bof-using-schemaorg-and-enriched-metadata-enableboost-fairness-research-resources-rda-13th-plenary

Data transformations should rely on semantic technologies, in particular LOD, and common markup languages (e.g. YAML[38]). It is important to embed steps of metadata creation into existing workflows to better capture provenance (see Section 2.2). Metadata descriptions from different scientific communities do exist, although they use a broad variety of schemas and serialisations. Within most communities, there is an awareness of the need to be FAIR-compliant. How to achieve FAIRness when aggregating metadata from several sources remains a challenging question. Many metadata aggregators (e.g. EUDAT's B2FIND[39], Europeana[40], HumaNum's isidore[41], CLARIN's VLO[42], OpenAIRE's EXPLORE[43]) have been confronted with this question and have addressed it in several ways. As there is no "golden bullet" solution, we mention a few possible strategies below:

- Linked Open Data (based on augmented RDF) provides a popular framework to express relations between heterogeneous descriptions and entities. For this, services providing semantic mappings and crosswalks (see also Section sec:CS4FAIR.SA) are strongly required.
- Common metadata schemas (e.g. DCMI, EDM, EDMI, DCAT, DataCite, Schema.org, OpenAIRE guidelines[44]) are another strategy. This could either be translated into the requirement to provide metadata descriptions in these formats directly, or by a conversion towards the common format. Such conversions pose the risk of being lossy.
- Metadata based on registered schemas and concepts is a third way towards acheiving semantic interoperability between various sources of metadata. This approach can be combined with the two strategies listed above, and will be discussed in detail in the next section.

Systematic registration of schemas and semantic definitions. We need a systematic and systemic approach to schema, concept and vocabulary registration that allows people to easily register, find and reuse these resources, thus making them also FAIR. RDA offers such a schema registry[45]. There are also community-specific registries available, such as the CLARIN component registry. In general, the availability of such registries can be seen as a basis for improved support of semantic crosswalks (see Section 2.2).

Validation & metadata extraction services. Both interoperability and reusability can benefit significantly from machine-supported digestion and inspection of data sets. First, by applying automatic validation, a guarantee can be given that the data is not corrupted and is in line with the format specifications. If these requirements are not met, the person or program responsible for the ingestion can be informed in a timely manner, so that the data can be curated and metadata can contain correct type specifications etc.

Second, in many cases, the often costly process of metadata generation can be made more efficient by extracting information from the data that is described. In many cases there is embedded metadata available, e.g., in form of headers embedded in the data. This presupposes that the format specification is respected and that the embedded metadata is correctly generated. This might require additional attention at the moment of the data generation (e.g. calibration of sensors).

[38] https://en.wikipedia.org/wiki/YAML
[39] https://eudat.eu/services/b2find
[40] https://www.europeana.eu/portal/de
[41] https://isidore.science
[42] https://www.clarin.eu/content/virtual-language-observatory-vlo
[43] https://explore.openaire.eu
[44] https://guidelines.openaire.eu
[45] http://rd-alliance.github.io/metadata-directory/

Search interfaces. Eventually, better and FAIRer metadata, using the various strategies mentioned above, can lead to more effective search facilities. In any case, searching in a collection of diverse metadata sources can be a fairly complex operation that requires insight into the heterogeneity of the metadata available. To overcome this, assistance services are needed that help users to represent a query in a format that meet common data representation formats - e.g. translating a GUI-supported query into SPARQL to search in a LOD set.

Accordingly, standardized FAIR-compliant search APIs for data are needed to support querying distributed metadata and data collections. Some examples in use are SRU-CQL[46](for metadata in the library and museum sector) and Federated Content Search (for language resource data).

Finally, it should be noted that data discovery in a multidisciplinary scenario through such rich search interfaces remains challenging. While the recall naturally increases, striking the right balance with a good precision – finding the proverbial needle in the haystack of millions of records in a search portal – is not at all trivial. Realizing this, and therefore being transparent about the indexing and ranking mechanisms used, is an important step to turn the FAIR principles into reality. At the end we assume a landscape of smart and specialised services tuned to specific multidisciplinary scenarios.

Workflow Frameworks

There is growing agreement that only an increased automation of data processing, with its implicit possibility of systematically including code that implements the FAIR principles and documenting actions, will help to overcome current inefficiencies and be able to facilitate reproducibility/reusability. However, the unpredictability in scientific work makes automation difficult to achieve.

In general, analytic steps in research need to be redone frequently, using different methods and adjusted parameter sets, until evidence for an interpretable scientific output can be achieved. Suitable workflows need to concatenate modular parts addressing specific analytic steps that can be automated. They need to be easily modifiable by scientists and offer the possibility of reusing existing code, thus addressing issues of both FAIRness and documentation. These workflows also need to be parameter-driven.

Electronic notebooks, such as Jupyter notebooks[47], and workflow frameworks such as Galaxy [1], KNIME [7], and Nextflow[48] offer a way forward toward automation . Over 230 workflow management systems are in use in research today[49]. The field of computer science has a long history of workflow research [3, 12], including the discoverability of workflows [24], workflow registries such as myExperiment.org [17], and workflow provenance [8].

The Common Workflow Language (CWL)[50] [2] is becoming a standard for describing analysis workflows and tools in a common way that makes them discoverable, portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. CWL is designed to meet the needs of data-intensive science, such as in bioinformatics, medical imaging, astronomy, physics, and chemistry, and forms part of the technology stack of BioCompute Objects[51], the GA4GH Workflow Execution Service[52] and the EOSC Life workflow collaboratory[53].

[46] https://www.loc.gov/standards/sru/cql/
[47] https://jupyter.org
[48] https://www.nextflow.io
[49] https://s.apache.org/existing-workflow-systems
[50] http://commonwl.org
[51] https://biocomputeobject.org
[52] https://ga4gh.github.io/workflow-execution-service-schemas/
[53] https://www.eosc-life.eu

To facilitate automation, libraries of code snippets should be made available that are specified in CWL and implement generic approaches for data management, access and use. An example is the Digital Object approach proposed by RDA, i.e. using code snippets to create a checksum, to register a handle, to create new metadata based on the existing one and add provenance information. Workflows amended with such reusable code snippets would be self-documenting in terms of provenance and versioning and would automatically create new, full-fledged and FAIR-compliant Digital Objects.

Aspects such as defining suitable packaging and container formats (Research Objects[54], Docker[55]) are also of great relevance to facilitate reproducibility and exchange of workflows. Increasingly often, code will have to be moved to the place where the data is being hosted and being executed at the corresponding machines. In this "bring your code to the data" scenario, a researcher needs to be able to execute (arbitrary) code close to the data. This is not possible in most data centres today. This is a particular problem in data-intensive science, where a solution is urgently needed.. An even more challenging scenario exists when workflows run in a distributed setting, i.e. when parts of a workflow are being executed at different locations exchanging messages. Such a scenario is of great interest in cases where sensitive data located in different repositories need to be virtually integrated to solve a specific analytic problem, but cannot be moved for ethical or juridical reasons. Making such scenarios common practice requires much better organizational and administrational solutions; for example, governments will need to overcome current restrictions that are hampering cross-border data intensive science.

Semantic Aspects

Discussions at the workshop clearly indicated a gap in services to facilitate semantic interoperability. Three major aspects were mentioned: 1) lack of a systematic and systemic approach to the definition and registration of schemas, semantic categories and vocabularies, 2) better help for semantic crosswalks and 3) the possibility to add user tags to existing metadata of data and services.

Currently it is almost impossible to find schemas or the definitions of semantic categories and vocabularies for people who are not experts in the field; once found, the formats are so different that machines can hardly make use of them. In collaboration with the Digital Curation Center, RDA has worked out a registry for schemas for metadata[56] as they are being used in various research communities; FAIRsharing[57] is a widely used metadata standards registry [52]. The schema.org initiative from Google is often referred to as focusing on schemas associated with web pages. The Bioschemas.org initiative has developed conventions for using schema.org markup for Life Science resources. Many research communities have done extensive work in the area of registries of semantic categories and vocabularies, extending to creating comprehensive ontologies. Ontologies are often underused in current research practice. A more systematic and systemic approach to the registration of schemas and semantic categories/vocabularies is strongly recommended as a basis for facilitating the work at semantic level such as crosswalks.

[54] http://www.researchobject.org
[55] https://en.wikipedia.org/wiki/Docker_(software)
[56] https://www.rd-alliance.org/metadata-standards-directory
[57] http://www.fairsharing.org

Researchers increasingly need to combine metadata and data from different sources that include different semantic spaces [27] An example of this is the project SoRa[58], funded by the German Research Foundation, which aims at linking data from the social sciences with data from spatial sciences to enable research addressing the relationship between health, well-being, or attitudes and local spatial condition. Thus, there is an urgent requirement to facilitate semantic crosswalks. One aspect of facilitating crosswalks would be to offer quick access to FAIR vocabularies and ontologies at all steps when creating or reusing metadata or data, as this would facilitate the understanding and the identification of fine semantic differences. Another aspect is to have improved tools to exploit aggregated assertions (c.f. RDF triples, nano-publications), as is indicated by mapping knowlets [39] in a domain of Linked Open Data and identifying core concepts, their scope and their relationships and by using Linked Open Vocabulary approaches[59]. A third point is to create easy-to-use tools that allow the creation of semantic mappings between concepts, as well as the ability to store and share them. Semantic mappings are often dependent on the specific type of task a researcher wants to carry out; complex ontologies are often too unwieldy to adapt to the concrete needs.

Another point that was flagged as being urgent is the need to have better support for users to add annotations or tags to data or metadata as is possible, for example, in repositories. These commentaries, which need to be stored separately from the original data and metadata, would help new users in understanding the overall topic of the (meta)data. This is often problematic, as data creators tend to use specific terminologies not known to everyone.

2.3 Key Recommendations

The FAIR principles do not form a blueprint for building infrastructures. There are many different ways to implement them. The following recommendations were identified as being particularly crucial for implementing FAIR-compliant data infrastructures and services. These recommendations target infrastructure and service providers as well as providers of innovative services resulting from computer science research:

1. Provide a user-friendly infrastructure that allows for the registration of **persistent identifiers (PIDs) for all kinds of digital objects**. This will help to establish interoperability at the data organization level.
2. Develop registries of **trustworthy repositories** allowing users to quickly find services. To do this, a common certification scheme is needed.
3. Implement **automatic processing of authentication & authorization** in distributed scenarios. This should result in creating a digital "data passport"
4. Improve **support for creating rich metadata**. Possible ways to do this include automatic extraction of metadata and provenance information from workflows, assistance services for metadata validation and transformation.
5. Create **schema and vocabulary registries** which allow people to easily find and reuse these resources and to better support the use of semantic mappings, annotations or crosswalks
6. Develop **rich search interfaces** for discovering (multidisciplinary) data
7. Encourage widespread adoption of appropriate **formal languages and techniques** enabling interoperability and workflows.

[58] http://www.sora-projekt.de
[59] https://lov.linkeddata.es/dataset/lov/

2.4 Related Computer Science Research Topics

The recommendations mentioned in Section 2.3 have a strong focus on improving automation, machine-readability, and computer-actionability on the infrastructure level to provide easy-to-use and intelligent services for researchers. This generally requires that data and metadata can be read and interpreted by machines based on corresponding standards and vocabularies. The following research aspects should be further addressed to foster the implementation of the recommendations and of FAIR data infrastructures in general.

- **Trustable exchange and tracking of digital objects** (contributes to Rec 2 & 3): In the context of the trustworthiness of repositories, it is vital that access to and use of digital objects is applied according to defined permissions which are trackable and tamper-proof. Important research aspects here are security and privacy frameworks to provide FAIRness for sensitive data. Furthermore, research should examine novel provenance concepts such as the use of blockchain technology, e.g., to store transaction records or to encapsulate authorisation decisions in smart contracts, taking scalability aspects into account. In the context of automation, further research aspects include software agents for automated informed consent negotiation and inference as well as computer actionable licences.
- **New approaches for FAIR digital preservation** (contributes to Rec 1 & 2): Research in this area is needed to ensure that FAIR data (and metadata) remain FAIR over a long period of time and that the used PIDs are stable. This aspect should also be part of repository certification, which is related to digital preservation problems in general.
- **Novel data management and storage concepts** (contributes to Rec 1 & 4): Data intensive science often relies on large-scale research data across evolving versions that are aggregated over long time scales. It is important that different data versions are accessible and that aggregated data is either persistent itself or that the computation including PIDs to all used data is preserved. In this context, the persistent provision of such data and the automatic generation of PIDs should be further examined.
- **Intelligent data discovery** (contributes to Rec 6): Data discovery is a significant challenge which becomes even more complex in multidisciplinary scenarios. Here, we see different possible research aspects. Rich search interfaces are an important prerequisite for data discovery; here, semantic aspects of intelligent FAIR services (based on ontologies and vocabulary crosswalking) should be further examined. Given cross disciplinary search scenarios, where heterogeneous data sources with different underlying structures come into play, an imporant challenge is finding evidence of significant correlations and improving interoperability between various data sources. Hence, the role of AI technologies, especially the deployment of machine and deep learning agents and their interaction with discoverability services, becomes critical. These role of these technologies need to be further examined to assist discovery workflows and to improve the quality of search results and recommendation systems. As these technologies are complex, performance aspects should also be considered. Finally, text-based search has limitations for data discovery; novel visual analytics technologies should also be examined to allow visual data discovery.
- **Semantic aspects** (contributes to Rec 4 & 5): Lack of practices by researchers in their daily work with semantic technologies along with inadequate understanding for structural and semantic specification for research data within the scientific community is an explicit obstacle in achieving the interoperability aspect of the FAIR data principles. An active area of research to deal with this problem is to work to establish a fully operational semantic layer, built upon the principles of semantic web and linked data

technologies, to ensure the consistent application and usage of machine-readable common standards (W3C / RDA) in harmonizing multidisciplinary content representation. How this semantic layer can support services, applications, metadata and data interoperability for distributed, federated resources still needs to be investigated. Another area of research is the establishment of universal registry infrastructures and underlying mechanisms for semantically rich knowledge graphs which, regardless of specific disciplines, can provide lookup services for the reuse of available schema, concept, vocabulary / ontologies, schema crosswalks and annotation services. In this context, rich metadata are the core for intelligent FAIR services. Whenever possible, this data should be automatically generated. One important research aspect is capturing provenance in workflow frameworks. Another aspect is reasoning on such provenance data to enable automated data identification and integration across heterogeneous domains.

- **Distributed Computing** (contributes to Rec 1, 3, 4 & 7): Concepts such as (FAIR) Digital Objects provide an abstraction layer for the basic management of any form of research data. The concept of Research Objects fills the gap to computational aspects for a better connection of code and data related to intensive operations like processing and analysis. Research Objects, ranging from simple scripts to large workflows, should be ready to be executed in a highly distributed computing and storage environment. This is still an important field for computer research. One possible research direction is to investigate further advancing the portability of executable code by encapsulating containers while also archiving the FAIRness of Research Objects with respect to reproducibility. Ultimately, this is the basis for building research environments such as Virtual Research Environments (VRE).

3 How to make computer science research more FAIR

3.1 Introduction

Computer science plays an important role in the implementation of the FAIR principles, both through performing research and delivering tools to support FAIRness within other disciplines. Computer science research, which relies on reusing digital objects such as research data and software to perform experiments, should also be FAIR itself. In particular, computer science should seek solutions to address what has come to be known as "FAIR*-ness," where efforts focus on what [16] terms the "R dimensions."

The scientific process behind digital computer science thinking should be inspired by these dimensions, namely[60] *repeatability* ("same research activity, same laboratory"), *replicability* ("same research activity, different laboratory"), *reproducibility* ("same research activity, different input parameters"), and *reusability* ("using a product of a research activity into another research activity") [16]. The FAIR*-ness of science applies to the entire scientific process, requiring that both results (i.e. FAIR data, FAIR software) and the processes used to create those results are reproducible. The following research ecosystem components play a key role in achieving FAIR*-ness in computer science:

- Research Literature: digital documents describing a research activity, e.g., articles, theses, slides, blogs, posters, etc. Literature is the essential and minimal scientific product required to share the output of a digital experiment. The online publication of research

[60] The twelve Rs were first presented in a blog post in 2010 which is no longer accessible: http://www.scilogs.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/

literature announces and represents the research activity itself; describes its motivations and methodology; illustrates and justifies the experimental steps (possibly including errors committed or wrong assumptions); and presents alternative solutions. The research literature can be enhanced when it is properly linked to the digital products used to perform the experiment or those generated by the experiment, i.e. other literature, research data and software, or experiments. This is not relevant for some sub-disciplines of computer science that are not data-driven or software-driven, such as the theory of computing.

- Research data: digital objects used as evidence of phenomena for the purpose of research. In computer science, research data may include a diverse range of objects, i.e. corpora of texts used for text mining; corpora of images or videos used for feature extraction and matching; sample data; training sets for machine-learning or bag of words, etc. Some communities also use source code as data. Mining Software Repositories (MSR)[61], for example, analyzes the rich data available in software repositories to uncover interesting and actionable information about software systems and projects. The variety of computer science disciplines is so broad that it is difficult to define a cross-discipline classification of research data.
- Research software: code to be compiled or interpreted. Any digital product intended for execution to become a process that is the concrete result of a research activity, i.e. implementations of algorithms in programming languages such as R scripts and Java, can be seen as research software.
- Research experiments: representations of digital experiments, that encapsulate repeatable sequence of actions (e.g. workflows in KNIME, Galaxy, CWL) or aggregation of digital objects to be combined into an experiment (e.g. Research Object[62] docker containers). Research experiments are typically related to the digital laboratory assets required to execute them; these include scientific services such as Taverna and lab notebooks, as well as possibly their input and output data, and the research software used in the experiment.
- Digital Laboratories (or virtual research environment (VRE)). Digital laboratories are compounds of assets required to technically operate a digital scientific life-cycle. These include ICT tools (services, software, frameworks, workflow engines, etc.), paradigms, methodologies, and standards, to enable the reuse of digital products and to generate new ones. Research literature, data, and software are all part of a digital laboratory. Research software can be deployed in such an environment so that it is accessible online via APIs or web interfaces.

In the following, we analyse the state of the art of FAIR*-ness for the research activities in computer science in Section 3.2. We do this to assess current trends in publishing digital scientific products in a way that guarantees FAIR*-ness of the individual products (i.e. FAIRness of data and software) and FAIR*-ness of the overall research activity. Based on this analysis, we propose recommendations for making computer science research more FAIR* in Section 3.3.

[61] http://www.msrconf.org
[62] http://www.researchobject.org

3.2 Status Quo: FAIR* Computer Science

Here, the FAIR*-ness of computer science refers to the FAIRness of the complete research cycle, including all digital objects that are used or created. For the scientific experiment to be FAIR*, best practices typically require these objects to be neatly described by metadata, to be uniquely identified and semantically interlinked, and to be (possibly openly) accessible on the web. In advanced scenarios, the digital objects resulting from an experiment encompass the digital and executable representation of the experiment itself (e.g. a workflow, a process), not solely the digital outcome of the experiment. The same object may be described by different kinds of metadata to support a variety of consumption options, from findability to re-use to cite and to credit. In the following, we analyse the status quo of FAIR*-ness for research literature, research data, research software, and digital laboratories.

Research literature

Research activities, by which we mean the entire scientific process, are typically published via research literature, e.g., scientific articles or research papers, which describe the underlying motivation, thesis, process, and conclusions. The scientific article is typically the main channel of scientific communication, while research data, research software, and in some cases the digital representation of the experiments, are additionally provided to achieve better FAIR*-ness. Vice versa, in some cases the "data paper" or "software paper" approach is adopted. Researchers publish an article with the explicit purpose of describing research data or software and make such products re-usable by other scientists [9, 35].[63]

Executable paper. The general idea of an executable paper is to enrich classical print-publications with executable elements, so that underlying computational elements can be directly explored and repeated. A while ago, Elsevier conducted the "Executable Paper Grand Challenge"[64] [23] during the ICCS 2011 (International Conference on Computational Science). In addition to addressing computational science at a general level, this effort also specifically addressed computer science research. Several projects presented their concepts of executable papers, which are published in the corresponding conference proceedings [46]. The winner of the challenge was invited to implement the "Collage" platform which was at one time used by the "Computers & Graphics" journal in the "Special Issue on executable papers for 3D object retrieval" [49]. Afterwards, this project was not continued, and the concept was rarely applied in individual cases. An example of where this conecpt was applied is the paper [37], which contains a literate Curry programme that is directly executable.

Artifact evaluation and badging. Many journals have begun to allow the addition of supplementary material for articles such as data and software. Artifacts, for example, can be software systems, scripts used to run experiments, input datasets, raw data collected in the experiment, or scripts used to analyze results. Several conferences and journals have established so-called artifact evaluation, in which the review process is extended to the supplementary materials that must be provided by the author.

Computer science disciplines have been experimenting with artifact evaluation since 2011. Known examples are the ESEC/FSE conference in Software Engineering and the ECOOP conference in Programming Languages (www.artifact-eval.org), as well as the Empirical

[63] https://www.nature.com/search?subject=mathematics-and-computing&journal=sdata
[64] https://www.iccs-meeting.org/iccs2011/

Software Engineering journal [38]. ACM started with "Artifact Review and Badging"[65] a quite popular approach applied in some sub-disciplines of computer science. Examples include Software Engineering (ACM SIGSOFT and ACM ESEC/FSE), Programming Languages (ACM SIGPLAN), and Databases (ACM SIGMOD) see [34], as well as the IEEE International Requirements Engineering Conference[66]. Some sub-disciplines of computer science such as Information Retrieval (ACM SIGIR)[22] are still discussing whether they should adopt the artifact evaluation process. Complementary to this, the "CENTRE reproducibility evaluation of CLEF/NTCIR/TREC tasks"[67] and "The Open-Source IR Replicability Challenge" (OSIRRC 2019)[68] propose a discipline-specific approach for Information Retrieval.

Artifact evaluation has also been a topic of interest at Schloss Dagstuhl with the Dagstuhl Perspectives Workshop on "Artifact Evaluation for Publications" [11], conducted in 2015. Since 2015 Schloss Dagstuhl provides the Dagstuhl Artifacts Series (DARTS)[69] to complement its own conference proceedings series.

Research data

Computer science follows trends and standards of formats to manage and exchange data, developed by the community in order to provide innovative, optimal, and efficient technologies. Database management systems vary today from SQL technologies (MySQL, Postgres, HIVE), NOSQL technologies (MongoDB) to column stores (Cassandra, HBASE, Spark) and graph databases (Neo4j, Virtuoso, GraphX), which can support different kinds of processing, whether local, parallel or distributed. On the side of exchange formats and protocols, examples include the Resource Description Framework (RDF), the eXtensible Markup Language (XML), JSON, AVRO, Protocol Buffers, CSV, etc., flanked by SPARQL, OpenSearch, etc. At a higher level of abstraction, tables are also extremely common across computer science disciplines and are shared via any standard format or file type.

Although data practices are solid and aligned across different computer science fields, supported by consolidated open source practices, the same cannot be said about publishing data according to the FAIR principles. To our knowledge, different sub-disciplines of computer science adopt different metadata data models, make use of different technologies, and follow different data acquisition, pre-processing, processing, and preservation practices. Their ability to converge on common methodologies and tools depends again on their maturity, where it exists, in their networks and governance, and sometimes by industry (e.g. MPEG-7 is widely used in industry and research [47]), which governs standards via ISO, W3C, and other similar entities.

In general, computer science disciplines do not reconcile their data publishing practices under research infrastructures or similar initiatives. Accordingly, computer scientists do not follow best practices or conventions of research data publishing and, despite the "de facto" digital science features, they fall under the umbrella and pitfalls of the "long tail of science" [54]. Table 1 classifies a number of known practices for sharing /publishing research data, highlighting if these practices can guarantee basic properties of data FAIRness: presence of persistent identifiers (PIDs), web discoverability means, preservation, and attribution metadata for citation.

[65] https://www.acm.org/publications/policies/artifact-review-badging
[66] https://re20.org/index.php/artifacts/
[67] http://www.centre-eval.org
[68] https://osirrc.github.io/osirrc2019/
[69] https://www.dagstuhl.de/en/publications/darts/artifacts/

Some computer scientists, perhaps pushed by funder mandates or because they work in fields sensitive to sharing practices (e.g. information science), have started to follow FAIR practices by publishing data in research data repositories and publishing "data papers" describing their research data. These practices help to ensure visibility, reusability, and scientific reward for the necessary, skills-enabled, often tedious job of generating research data [9]. Overall, however, the lack of practices and dedicated repositories makes it very difficult for computer scientists to correctly share, discover, and reuse research data.

Data publishing and preservation. Computer scientists (or scientists in general) often publish research data via websites. These websites are not intended for this purpose, as they do not give support for PIDs, preservation and structured citation metadata.

Data citation and identification. Research data is typically referred to via URLs from the article, not vice versa, and is not discoverable via search engines dedicated to research or computer science. The same holds true for computer scientists relying on GitHub, which offers known value-added functions for collaborative work, but is not intended to serve only scientists nor address research data publishing.

Data peer review and scientific reward . Data in computer science is still not regarded as a core scientific product, but rather as supplementary material, in support of the article. This is reflected by the author guidelines in conferences and journals in the field, which only recently have started to mention research data and data papers (e.g. IRCDL2019 conference (Italian Research Conference on Digital Libraries)[70], and in general do not bias reviewers to mandate their acquisition for review.

Research software

Sharing software as open source [53] is an established practice in computer science. Developers rely on software repositories and management tools (e.g. GitHub, Bitbucket) which support collaborative programming as well as maintenance and versioning of open source projects [30]. However, sharing and publishing research software are different practices and such tools do not support publishing. Publishing has implications of scientific reward, such as metadata for attribution and PIDs for citation (of different software versions), of findability, accessibility, and preservation, which such tools are not intended to target. To fill the gap, and to meet Open Science publishing expectations, we are currently witnessing a large number of initiatives attempting to bridge the two worlds of software development and research software publishing [51]. Table 2 classifies a number of known practices for sharing /publishing research software, highlighting whether basic properties of FAIRness can be guaranteed: presence of persistent identifiers (PIDs), web discoverability means, preservation, and attribution metadata for citation.

Software publishing and preservation. Research software in computer science is usually not published and archived in a FAIR* way, for example by using a common vocabulary to describe these artifacts with metadata and in a citable way with a persistent identifier. GitHub is not a platform for scholarly publishing; the common practice to link to GitHub repositories from the literature (PDF) as footnotes is far from acheiving FAIR* software publishing to facilitate the R* of science. Repositories like Zenodo and FigShare support publishing software as effective research products, which are assigned a DOI, attribution metadata, and descriptions, and which can be properly cited from literature [29]. (Zenodo, for example, supports direct linking with GitHub).

[70] https://ircdl2019.isti.cnr.it

Table 1 Practices for sharing /publishing research data in computer science.

	PID	Discoverability	Preservation	Metadata for citation
Data published on web sites, e.g., ClueWeb12 for information retrieval benchmarks[71] and 3Dscanrep 3D graphics repository[72].	URL	- Web search engines (e.g. Google)	Not addressed	Not addressed
Data sharing platforms for programming challenges, e.g., Kaggle datasets for computer science[73].	URL	- Web search engines (e.g. Google) - Data search engines when schema.org compatible (e.g. Google Data Search, OpenAIRE, GeRDI)	Not addressed	Not addressed
Programming practices, e.g., GitHub[74].	URL	- Web search engines (e.g. Google)	Not addressed	Not addressed
Research data repositories, e.g., general-purpose data repositories (e.g. Zenodo[75], figshare[76]).	DOI	- Data search engines (e.g. Google Data Search, DataCite, OpenAIRE)	Guaranteed by repository SLAs	Enabled
Data papers, e.g., International Journal of Robotics Research[77], Scientific Data[78], SpringerPlus[79], Applied Informatics[80].	DOI	- Data search engines (e.g. Google Data Search, DataCite, OpenAIRE) - Publication search engines (e.g. Google Scholar, OpenAIRE, Scopus, etc.)	Guaranteed by repository SLAs	Enabled

Research software can also be published in the sense of "software papers"[81]. The "Journal of Open Source Software"(JOSS)[82], for instance, is a general journal for open source software publishing. Research software can also be published in the "Research Ideas and Outcomes" (RIO)[83] journal, which generally allows the publication of all artifacts around the research cycle. However, the majority of journals in this context are "software metapapers", in which research software is described in a paper. In these cases, the associated software is not "published" (with DOI etc.) but rather is attached as supplemental material or stored and linked via an internal or external repository. In computer science, the "Journal of Machine Learning Research" (JMLR)[84], for example, provides its own internal software repository[85] which is linked with the journal[86]. However, software publishing is seldom adopted in computer science.

Preservation is the key to enable software re-use and replicability from a long-term perspective. It remains a great challenge, however, to collect, preserve, and share all the software source code. Software Heritage [19] is an initiative of INRIA which aims at building the largest software archive world-wide, thus enabling preservation of software repositories beyond their lifetime. Software Heritage keeps an in-sync copy of all versions of all repositories forever and for those generates a unique (hash-based) identifier which ensure unambiguity and retrieval.

Software citation and identification. Several works related with software citation and identification[87,88,89,90] [33, 48, 14, 15] exist today. The FORCE11 Software Citation Working Group Software Initiative provides metadata standards for software citation[91]. The Software Heritage archive[92] addresses software preservation.

Software peer review and scientific reward. Peer review is currently the core mechanism for quality assurance in the scientific publishing system and an import quality indicator for journals. Thus, software journals such as JOSS have a peer review process, which usually includes code review, re-execution tests and so forth. ACM established a review process to check different degrees of R* for artifacts, including software. Passed checks are associated with a corresponding badge, see the previous discussion on "Research literature" on page 19.

Research experiments

As mentioned above, digital science is not just about producing digital scientific products, but also about adopting digital services in a digital laboratory, and possibly combining them to build complex experiments. Depending on the technical homogeneity of the digital laboratory, the combination of such services, as well as their configuration, to obtain a final result, can be itself expressed as a digital object. As such, this object can be published, attributed

[81] https://www.software.ac.uk/which-journals-should-i-publish-my-software
[82] https://joss.theoj.org
[83] https://riojournal.com
[84] http://www.jmlr.org
[85] https://mloss.org/software/
[86] http://jmlr.csail.mit.edu/mloss/
[87] https://www.software.ac.uk/how-cite-software
[88] https://www.force11.org/software-citation-principles
[89] https://guidelines.openaire.eu/en/latest/software
[90] http://www.citethisforme.com/guides/ieee-with-url/how-to-cite-a-software
[91] https://www.force11.org/group/software-citation-working-group
[92] https://www.softwareheritage.org/archive/

Table 2 Practices for sharing/publishing research software in computer science.

	PID	Discoverability	Preservation	Metadata for citation
Software published on web sites and FTP sites	URL	- Web search engines (e.g. Google)	Not addressed	Not addressed
Programming practices, e.g., GitHub, Bitbucket.	URL	- Web search engines (e.g. Google)	Addressed indirectly by Software Heritage[93].	Not addressed
Research software repositories, e.g., general-purpose data repositories (e.g. Zenodo[94], figshare[95]).	DOI	- Software search engines (e.g. OpenAIRE's EXPLORE for all software and, swMATH for sotware in mathematics)	Guaranteed by repository SLAs	Enabled
Software papers, e.g., Joss Journal[96].	DOI	- Data search engines (e.g. Google Data Search, DataCite, OpenAIRE) - Publication search engines (e.g. Google Scholar, OpenAIRE, Scopus, etc.)	Guaranteed by repository SLAs	Enabled

to its authors, be provided with a DOI and metadata, and be accessed and retrieved by other scientists for re-use. The optimal scenario is one where the digital experiment can be consumed by a service to automatically reproduce or repeat the experiment; examples are workflow management systems such as Galaxy, Taverna, and KNIME. Other scenarios are possible, where the digital object represents a manual sequence of steps - Standard Operating Procedures or Lab Protocols (e.g. protocols.io[97]), or describe hybrid contexts where only part of the workflow can be executed [10].

Scientific workflows. Scientific workflows encode sequences of actions, typically resulting from the execution of services that obey the standards and practices imposed by a particular language and can therefore be combined into a pipeline. As highlighted in Section 2.2, workflow languages, management systems, and related functionalities constitute a specific, mature, and still innovative field in computer science [4, 26]. Although much research has produced widely adopted products which are used to model processing workflows in several disciplines (e.g. BPEL[98], Common Worklow Language (CWL)[2], Taverna [59], Galaxy [1], myexperiment.org [17]) , only a few sub-disciplines of computer science heavily rely on such tools to actually perform and share experiments. KNIME [7], used in the context of data mining and analytics, is one such exception. In general, computer science does not follow or promote best practices that rely on workflow-oriented approaches to model processing flows or eventually publish them as research products.

Research objects. Unlike scientific workflows, the aggregation of objects does not necessarily target sequences of steps. Here are two common examples:

- Aggregation of objects: Some approaches define representations of object aggregations, in the form of labelled graphs whose node entities can describe the elements of an experiment and possible their composition. Such representations enable a degree of interoperability, come with out-of-the-box tools, and unburden scientists from the definition of ad-hoc encodings. Examples include work on workflow research objects for reproducibility and preservation, e.g., Research Object[99] [5] and RMap Disco [28].
- Virtual Machines: Some approaches explore the notion of generating virtual machines, which are by definition born to facilitate the sharing of code, data, and processes (e.g. EGI Application Database). This approach makes the research described in an article fully executable, as the authors share and publish the digital representation of the experiment, thereby minimizing the efforts required by readers to re-assemble individual components and re-build the necessary execution environment. These approaches are based on development practices or on more advanced scholarly communication practices.

Digital laboratories

Digital laboratories are the digital twin of traditional scientific laboratories. Computer scientists make use of digital resources to perform their experiments. Such resources range from personal laptops, to the internet itself, encompassing cloud resources, shared on-line services, or software as a service solutions. In order to maximize FAIR*ness, the digital laboratory assets used in an experiment should also be shared, together with the experimental

[97] https://www.protocols.io
[98] http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html
[99] http://www.researchobject.org

conditions (e.g. configurations, software versions). In the last decade in Europe, efforts have been made to construct research infrastructures and e-infrastructures; these initiatives allow research communities to identify a governance body to plan policies, best practices, and services for researchers to better leverage scientific workflows.

Specific sub-disciplines of computer science have exploited such opportunities to agree on common digital laboratories and are today putting efforts into building services and governance. Examples of this can be seen in SoBigData[100] for social mining and OpenMinTed[101] for text mining. In general, however, computer scientists seem convinced that their current efforts are more than enough to perform FAIR* science [20, 21] and do not see the benefit of investing time and effort into building common digital laboratories.

3.3 Key recommendations

It is ironic that although ICT and applied computer science form the pillars of e-infrastructures, the adoption of open science practices in (data-driven) computer science research is lacking. The FAIR principles, particularly the interpretation of FAIR*-ness as we discuss it above are not widely adopted within research in the field. While some branches of computer science have recently started to create research infrastructures and to push for common practices, standards and services, there is a general lack of adoption of these practices across computer science and related disciplines.

In order to advance FAIR*ness within computer science, we propose the following four recommendations targeted toward researchers, research communities and institutions, policy makers and funders.

1. **Foster sharing / publishing culture**: Publishing on GitHub is not sufficient to acheive FAIR*ness. Identification, metadata, versioning, and preservation support provide a solid basis, but this is not enough to reach minimal FAIR requirements. Researchers and their communities in computer science should advocate for extending publishing workflows in support of research data, research software, and research experiments. They should aim to preserve these objects with generic metadata (e.g. DataCite, Software Citation metadata) and use DOIs/PIDs to link them to theliterature. Existing solutions such as Zenodo or Software Heritage (for software) may be used; alternatively community-specific repositories with FAIR* in mind could be realized.
2. **Foster (peer) review and scientific reward for published artifacts**: Research communities in computer science should foster and establish the concept of artifact publishing (i.e. data, software, experiments) and review to ensure the FAIR*-ness of the published research which these artifacts support. Additionally, publishing data, software, and experiments should be seen as having a stand-alone value; these objects can later be described for proper re-use in a dedicated article. The consolidation of scientific reward criteria at the community level is key for open science practices to take hold. Communities, disciplinary sub-communities and institutions should introduce evaluation practices reflecting this vision, where evaluation is based on the entirety of researcher's scientific output, rather than just on published literature. Discipline-specific quality measures reflect this vision need to be identified.

[100] http://sobigdata.eu
[101] http://openminted.eu

3. **Promote the use of discipline-specific digital laboratories and publishing workflows to support the entire scientific process.**: Computer science and its sub-disciplines should devise agreed-upon, community-driven processes for the definition, governance, and maintenance of digital laboratories (e.g. research infrastructures, scholarly communication services). As proven by research infrastructure experiences, such a collaborative approach helps to create the community endorsement required to drive the development of common understandings of the scientific process and its subsequent publishing. In such contexts, sub-disciplinary communities and engaged institutions can organize/endorse the way they produce and publish research data, software, and experiments (i.e. using common APIs, common descriptors, discipline ontologies, and software portability, etc.). They can agree upon common policies, digital laboratories, methodologies, and workflows and publish research objects to ensure R* within the field. As a consequence, researchers would be able to find common ground for scientific development and debate. In the ideal scenario, scientific assessment would be as automated as possible, scientific fraud would be minimized, and redundancy (and hence the cost of science) would be mitigated by re-use. Finally, the usage of tools to digitally describe non-digital scientific processes (in non data-driven or software-driven sub-disciplines) could also be adopted to formalize the methodologies.
4. **Foster cultural change**: To accelerate this cultural shift, policy makers and institutions with strong publishing mandates should also be involved. This will enable linking funding and policies to support FAIR*ness and open science. In this respect, it is crucial that these principles and goals are also implemented at the educational level. During their training, students and PhD researchers should encounter and acquire knowledge and skills about the elements required to perform FAIR* computer science research. Such skills include learning about open source software development practices, software engineering, data curation and sharing practices, adoption of standard tools, and the establishment of digital laboratories. Institutions, departments, and universities play a key role in planning and providing courses[102,103,104] for training and support, and in identifying and supporting new roles such as "data stewards"[105,106]. "Train the trainer" approaches offer a promising way for teachers and professors to become fluent with issues related to making computer science FAIR*, as well as the related issues of ethics, trust, reward, and cost within science.

[102] https://www.datasciencedegreeprograms.net
[103] https://www.ipp.mpg.de/4532371/10_18
[104] https://www.mu-ds.de/
[105] https://www.mastersindatascience.org/careers/
[106] https://www.openaire.eu/item/the-role-and-value-of-data-stewards-in-universities-a-tu-delft-case-study-on-data-stewardship

4 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
- 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
Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, NL
- Gregor Hagedorn
Museum für Naturkunde – Berlin, DE
- Wilhelm Hasselbring
Universität Kiel, DE
- Kees den Heijer
TU Delft, NL
- 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 Information Centre for Economics – Kiel, DE
- Birgit Schmidt
SuB – Göttingen, DE
- Achim Streit
KIT – Karlsruher Institut für Technologie, DE
- Klaus Tochtermann
ZBW – Leibniz Information Centre for Economics – 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

Acknowledgements. The participants wish to thank Schloss Dagstuhl for their strong support of this workshop. The editors especially wish to thank all contributing authors for their input to this manifesto and all participants for their quality check. Thanks also to Atif Latif (ZBW – Leibniz Information Centre for Economics) for his valuable input for the related computer science research aspects.

The workshop was supported by funding of the Leibniz Research Alliance Open Science[107] within the funding line strategic networks of the Leibniz Association.

References

1 Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544, July 2018. URL: https://academic.oup.com/nar/article/46/W1/W537/5001157, doi:10.1093/nar/gky379.

2 Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, and Luka Stojanovic. Common Workflow Language, v1.0. page 5921760 Bytes, 2016. URL: https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156/2, doi:10.6084/M9.FIGSHARE.3115156.V2.

3 Malcolm Atkinson, Sandra Gesing, Johan Montagnat, and Ian Taylor. Scientific workflows: Past, present and future. *Future Generation Computer Systems*, 75:216–227, October 2017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X17311202, doi:10.1016/j.future.2017.05.041.

4 Adam Barker and Jano van Hemert. Scientific Workflow: A Survey and Research Directions. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Roman Wyrzykowski, Jack Dongarra, Konrad Karczewski, and Jerzy Wasniewski, editors, *Parallel Processing and Applied Mathematics*, volume 4967, pages 746–753. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. URL: http://link.springer.com/10.1007/978-3-540-68111-3_78, doi:10.1007/978-3-540-68111-3_78.

5 Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, February 2013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X11001439, doi:10.1016/j.future.2011.08.004.

6 Gary Berg-Cross, Raphael Ritz, and Peter Wittenburg. RDA DFT Core Terms and Model, 2016. URL: https://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318.

7 Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 319–326. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. URL: http://link.springer.com/10.1007/978-3-540-78246-9_38, doi:10.1007/978-3-540-78246-9_38.

[107] https://www.leibniz-openscience.de/en/

8 Shawn Bowers. Scientific Workflow, Provenance, and Data Modeling Challenges and Approaches. *Journal on Data Semantics*, 1(1):19–30, May 2012. URL: http://link.springer.com/10.1007/s13740-012-0004-y, doi:10.1007/s13740-012-0004-y.

9 Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani. Data journals: A survey: Data Journals: A Survey. *Journal of the Association for Information Science and Technology*, 66(9):1747–1762, September 2015. URL: http://doi.wiley.com/10.1002/asi.23358, doi:10.1002/asi.23358.

10 Leonardo Candela, Paolo Manghi, Fosca Gianotti, Valerio Grossi, and Roberto Trasarti. HyWare: a HYbrid Workflow lAnguage for Research E-infrastructures. *D-Lib Magazine*, 23(1/2), January 2017. URL: http://www.dlib.org/dlib/january17/candela/01candela.html, doi:10.1045/january2017-candela.

11 Bruce R. Childers, Grigori Fursin, Shriram Krishnamurthi, and Andreas Zeller. Artifact Evaluation for Publications (Dagstuhl Perspectives Workshop 15452). *Dagstuhl Reports*, 5(11):29–35, 2016. URL: http://drops.dagstuhl.de/opus/volltexte/2016/5762, doi:10.4230/DagRep.5.11.29.

12 Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, and Christophe Blanchet. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75:284–298, October 2017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X17300316, doi:10.1016/j.future.2017.01.012.

13 Sandra Collins, Françoise Genova, Natalie Harrower, Simon Hodson, Sarah Jones, Leif Laaksonen, Daniel Mietchen, Rūta Petrauskaitė, and Peter Wittenburg. Turning fair into reality: Final report and action plan from the european commission expert group on fair data, 2018. doi:https://doi.org/10.2777/54599.

14 Roberto Di Cosmo, Morane Gruenpeter, and Stefano Zacchiroli. 204.4 Identifiers for Digital Objects: The case of software source code preservation. September 2018. URL: https://osf.io/kde56/, doi:10.17605/OSF.IO/KDE56.

15 Roberto Di Cosmo, Morane Gruenpeter, and Stefano Zacchiroli. Referencing Source Code Artifacts: A Separate Concern in Software Citation. *Computing in Science & Engineering*, 22(2):33–43, March 2020. URL: https://ieeexplore.ieee.org/document/8946737/, doi:10.1109/MCSE.2019.2963148.

16 David De Roure. The future of scholarly communications: Based on a paper presented at the 37th UKSG Conference, Harrogate, April 2014. *Insights: the UKSG journal*, 27(3):233–238, November 2014. URL: http://insights.uksg.org/articles/10.1629/2048-7754.171, doi:10.1629/2048-7754.171.

17 David De Roure, Carole Goble, and Robert Stevens. The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, May 2009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X08000939, doi:10.1016/j.future.2008.06.010.

18 Koenraad De Smedt, Dimitris Koureas, and Peter Wittenburg. FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. *Publications*, 8(2):21, April 2020. URL: https://www.mdpi.com/2304-6775/8/2/21, doi:10.3390/publications8020021.

19 Roberto Di Cosmo. Software Heritage: Why and How We Collect, Preserve and Share All the Software Source Code. In *2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 2–2, May 2018.

20 Nicola Ferro. Reproducibility Challenges in Information Retrieval Evaluation. *Journal of Data and Information Quality*, 8(2):1–4, February 2017. URL: https://dl.acm.org/doi/10.1145/3020206, doi:10.1145/3020206.

21 Nicola Ferro, Norbert Fuhr, Kalervo Järvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *ACM SIGIR Forum*, 50(1):68–82, June 2016. URL: https://dl.acm.org/doi/10.1145/2964797.2964808, doi:10.1145/2964797.2964808.

22 Nicola Ferro and Diane Kelly. SIGIR Initiative to Implement ACM Artifact Review and Badging. *ACM SIGIR Forum*, 52(1):4–10, August 2018. URL: https://dl.acm.org/doi/10.1145/3274784.3274786, doi:10.1145/3274784.3274786.

23 Ann Gabriel and Rebecca Capone. Executable Paper Grand Challenge Workshop. *Procedia Computer Science*, 4:577–578, 2011. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877050911001189, doi:10.1016/j.procs.2011.04.060.

24 Daniel Garijo, Yolanda Gil, and Oscar Corcho. Abstract, link, publish, exploit: An end to end framework for workflow sharing. *Future Generation Computer Systems*, 75:271–283, October 2017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X17300274, doi:10.1016/j.future.2017.01.008.

25 Joint Position Paper on the European Open Science Cloud, 2017. Germany and the Netherlands. URL: https://www.dtls.nl/wp-content/uploads/2017/05/DE-NL-Joint-Paper-FINAL.pdf.

26 Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the Challenges of Scientific Workflows. *Computer*, 40(12):24–32, December 2007. URL: http://ieeexplore.ieee.org/document/4404805/, doi:10.1109/MC.2007.421.

27 Kathleen M Gregory, Helena Cousijn, Paul Groth, Andrea Scharnhorst, and Sally Wyatt. Understanding data search as a socio-technical practice. *Journal of Information Science*, 46(4):459–475, August 2020. URL: http://journals.sagepub.com/doi/10.1177/0165551519837182, doi:10.1177/0165551519837182.

28 Karen L. Hanson, Tim DiLauro, and Mark Donoghue. The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CE on Joint Conference on Digital Libraries - JCDL '15*, pages 281–282, Knoxville, Tennessee, USA, 2015. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2756406.2756952, doi:10.1145/2756406.2756952.

29 Wilhelm Hasselbring, Leslie Carr, Simon Hettrick, Heather Packer, and Thanassis Tiropanis. From FAIR research data toward FAIR and open research software. *it - Information Technology*, 62(1):39–47, February 2020. doi:10.1515/itit-2019-0040.

30 Wilhelm Hasselbring, Leslie Carr, Simon Hettrick, Heather Packer, and Thanassis Tiropanis. Open source research software. *Computer*, 53(8):84–88, August 2020. doi:10.1109/MC.2020.2998235.

31 Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Philippe Rocca-Serra, Robert Pergl, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Schneider, George Strawn, Mark Thompson, Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighagen, Peter Wittenburg, Marco Roos, Barend Mons, and Erik Schultes. FAIR Principles: Interpretations and Implementation Considerations. *Data Intelligence*, 2(1-2):10–29, January 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/dint_r_00024, doi:10.1162/dint_r_00024.

32 Robert Kahn and Robert Wilensky. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2):115–123, April 2006. URL: http://link.springer.com/10.1007/s00799-005-0128-x, doi:10.1007/s00799-005-0128-x.

33 Daniel S. Katz and Neil P. Chue Hong. Software Citation in Theory and Practice. In James H. Davenport, Manuel Kauers, George Labahn, and Josef Urban, editors, *Mathematical Software – ICMS 2018*, volume 10931, pages 289–296. Springer International Publishing, Cham, 2018. URL: http://link.springer.com/10.1007/978-3-319-96418-8_34, doi:10.1007/978-3-319-96418-8_34.

34 Shriram Krishnamurthi and Jan Vitek. The real software crisis: repeatability as a core value. *Communications of the ACM*, 58(3):34–36, February 2015. URL: https://dl.acm.org/doi/10.1145/2658987, doi:10.1145/2658987.

35 Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci. OpenAIRE's DOIBoost - Boosting Crossref for Research. In Paolo Manghi, Leonardo Candela, and Gianmaria Silvello, editors, *Digital Libraries: Supporting Open Science*, volume 988, pages 133–143. Springer International Publishing, Cham, 2019. URL: http://link.springer.com/10.1007/978-3-030-11226-4_11, doi:10.1007/978-3-030-11226-4_11.

36 Yann Le Franc, Jessica Parland-von Essen, Luiz Bonino, Heikki Lehväslaiho, Gerard Coen, and Christine Staiger. D2.2 fair semantics: First recommendations, March 2020. doi:10.5281/zenodo.3707985.

37 Steffen Mazanek and Michael Hanus. Constructing a bidirectional transformation between BPMN and BPEL with a functional logic programming language. *Journal of Visual Languages & Computing*, 22(1):66–89, February 2011. URL: https://linkinghub.elsevier.com/retrieve/pii/S1045926X10000741, doi:10.1016/j.jvlc.2010.11.005.

38 Daniel Méndez Fernández, Martin Monperrus, Robert Feldt, and Thomas Zimmermann. The open science initiative of the Empirical Software Engineering journal. *Empirical Software Engineering*, 24(3):1057–1060, June 2019. URL: https://github.com/emsejournal/openscience/, doi:10.1007/s10664-019-09712-x.

39 Barend Mons. FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services. *Data Intelligence*, 1(1):22–42, March 2019. URL: https://www.mitpressjournals.org/doi/abs/10.1162/dint_a_00002, doi:10.1162/dint_a_00002.

40 Barend Mons, Erik Schultes, Fenghong Liu, and Annika Jacobsen. The FAIR Principles: First Generation Implementation Choices and Challenges. *Data Intelligence*, 2(1-2):1–9, January 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/dint_e_00023, doi:10.1162/dint_e_00023.

41 P. Bryan Heidorn. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2):280–299, 2008. URL: http://muse.jhu.edu/content/crossref/journals/library_trends/v057/57.2.heidorn.html, doi:10.1353/lib.0.0036.

42 George Strawn Peter Wittenburg. Common Patterns in Revolutionary Infrastructures and Data. 2018. URL: https://b2share.eudat.eu/records/4e8ac36c0dd343da81fd9e83e72805a0, doi:10.23728/B2SHARE.4E8AC36C0DD343DA81FD9E83E72805A0.

43 Stefan Pröll and Andreas Rauber. Scalable data citation in dynamic, large databases: Model and reference implementation. In *2013 IEEE International Conference on Big Data*, pages 307–312, October 2013. doi:10.1109/BigData.2013.6691588.

44 Andreas Rauber, Ari Asmi, Dieter Van Uytvanck, and Stefan Proell. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of IEEE Technical Committee on Digital Libraries, Special Issue on Data Citation*, 12(1):6–15, May 2016. URL: https://www.force11.org/sites/default/files/d7/project/81/ieee-tcdl-dc-2016_paper_1.pdf.

45 Research Data Alliance FAIR Data Maturity Model Working Group. FAIR Data Maturity Model: specification and guidelines - draft. 2020. URL:

https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines, doi:10.15497/RDA00045.

46 Mitsuhisa Sato, Satoshi Matsuoka, Peter M. Sloot, G. Dick van Albada, and Jack Dongarra, editors. *Proceedings of the International Conference on Computational Science, ICCS 2011*. ScienceDirect, 2011. URL: https://www.sciencedirect.com/journal/procedia-computer-science/vol/4/suppl/C.

47 Shih-Fu Chang, T. Sikora, and A. Purl. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001. URL: http://ieeexplore.ieee.org/document/927421/, doi:10.1109/76.927421.

48 Arfon M. Smith, Daniel S. Katz, Kyle E. Niemeyer, and FORCE11 Software Citation Working Group. Software citation principles. *PeerJ Computer Science*, 2:e86, September 2016. URL: https://peerj.com/articles/cs-86, doi:10.7717/peerj-cs.86.

49 Michela Spagnuolo and Remco Veltkamp. Special issue on executable papers for 3D object retrieval. *Computers & Graphics*, 37(5):A7–A8, August 2013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0097849313000587, doi:10.1016/j.cag.2013.04.006.

50 George Strawn. Open Science, Business Analytics, and FAIR Digital Objects, 2019. doi:http://doi.org/10.23728/b2share.6ceeed13eb6340fcb132bcb5b5e3d69a.

51 Jonathan Tennant, Ritwik Agarwal, Ksenija Baždarić, David Brassard, Tom Crick, Daniel J. Dunleavy, Thomas Rhys Evans, Nicholas Gardner, Monica Gonzalez-Marquez, Daniel Graziotin, Bastian Greshake Tzovaras, Daniel Gunnarsson, Johanna Havemann, Mohammad Hosseini, Daniel S. Katz, Marcel Knöchelmann, Christopher R Madan, Paolo Manghi, Alberto Marocchino, Paola Masuzzo, Peter Murray-Rust, Sanjay Narayanaswamy, Gustav Nilsonne, Josmel Pacheco-Mendoza, Bart Penders, Olivier Pourret, Michael Rera, John Samuel, Tobias Steiner, Jadranka Stojanovski, Alejandro Uribe-Tirado, Rutger Vos, Simon Worthington, and Tal Yarkoni. A tale of two 'opens': intersections between Free and Open Source Software and Open Scholarship. preprint, SocArXiv, March 2020. URL: https://osf.io/2kxq8, doi:10.31235/osf.io/2kxq8.

52 the FAIRsharing Community, Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L. Lister, and Milo Thurston. FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4):358–367, April 2019. URL: http://www.nature.com/articles/s41587-019-0080-8, doi:10.1038/s41587-019-0080-8.

53 Georg von Krogh and Eric von Hippel. The Promise of Research on Open Source Software. *Management Science*, 52(7):975–983, July 2006. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1060.0560, doi:10.1287/mnsc.1060.0560.

54 Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, 8(7):e67332, July 2013. URL: https://dx.plos.org/10.1371/journal.pone.0067332, doi:10.1371/journal.pone.0067332.

55 Tobias Weigel, Beth Plale, Mark Parsons, Gabriel Zhou, Yu Luo, Ulrich Schwardmann, Robert Quick, Margareta Hellström, and Kei Kurakawa. RDA Recommendation on PID Kernel Information (version 1), 2018. doi:https://doi.org/10.15497/RDA00031.

56 Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry

Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, December 2016. URL: http://www.nature.com/articles/sdata201618, doi:10.1038/sdata.2016.18.

57 Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5(1):180118, December 2018. URL: http://www.nature.com/articles/sdata2018118, doi:10.1038/sdata.2018.118.

58 Peter Wittenburg, George Strawn, Barend Mons, Luiz Boninho, and Erik Schultes. Digital Objects as Drivers towards Convergence in Data Infrastructures, 2019. doi:https://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11.

59 Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P. Balcazar Vargas, Shoaib Sufi, and Carole Goble. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(W1):W557–W561, July 2013. URL: http://academic.oup.com/nar/article/41/W1/W557/1094153/The-Taverna-workflow-suite-designing-and-executing, doi:10.1093/nar/gkt328.