Making FAIR Easy with FAIR Tools: From Creolization to Convergence

Mark Thompson[1][†], Kees Burger[1], Rajaram Kaliyaperumal[1], Marco Roos[1] & Luiz Olavo Bonino da Silva Santos[2]

[1]Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[2]GO FAIR international Support & Coordination Office (GFISCO), Leiden, The Netherlands

ABSTRACT

Since their publication in 2016 we have seen a rapid adoption of the FAIR principles in many scientific disciplines where the inherent value of research data and, therefore, the importance of good data management and data stewardship, is recognized. This has led to many communities asking "What is FAIR?" and "How FAIR are we currently?", questions which were addressed respectively by a publication revisiting the principles and the emergence of FAIR metrics. However, early adopters of the FAIR principles have already run into the next question: "How can we become (more) FAIR?" This question is more difficult to answer, as the principles do not prescribe any specific standard or implementation. Moreover, there does not yet exist a mature ecosystem of tools, platforms and standards to support human and machine agents to manage, produce, publish and consume FAIR data in a user-friendly and efficient (i.e., "easy") way. In this paper we will show, however, that there are already many emerging examples of FAIR tools under development. This paper puts forward the position that we are likely already in a creolization phase where FAIR tools and

† Corresponding author: Mark Thompson (E-mail: m.thompson@lumc.nl, ORCID: 0000-0002-7633-1442).

technologies are merging and combining, before converging in a subsequent phase to solutions that make FAIR feasible in daily practice.

1. INTRODUCTION

At a glance, the FAIR principles simply stipulate a number of "best practices" on how to deal with data and their associated metadata. However, a more careful reading of both the principles and their associated publications [1, 2] reveals some of the potential complexities when trying to implement FAIR [3]. These issues break down in at least three specific, orthogonal aspects. Firstly, a number of principles provide guidelines about the relationship between data, the representation of the data and the associated metadata that describes the data more fully (e.g., F1, F2, F3, I1, I2, I3, R1, R1.1, R1.2, R1.3). Even though it is clear what is required by these principles, it is not specified how it should be done, i.e., FAIR is not, in itself, a standard [2]. Secondly, there are a number of principles that require extensive infrastructural support like search engines, communication protocols and identifier resolution services (e.g., F4, A1, A2). Thirdly, there are a number of principles that refer to a community consensus or standard either explicitly (R1.3 and, by recursion, I2) or implicitly, concerning for example the definition of "rich", "shared" and "relevant" (F2, I1, R1). Moreover, the principles are open to interpretation with regard to the type of digital resource and its granularity. For example, when a principle talks about "data" does it refer to a data set as a whole, or could it refer to each individual data record (or item) contained in the data set? Finally, the principles need to be taken as guidelines that primarily aim to enable machines to (autonomously) interact with data [1], thus adding another possible layer of interpretation and implementation complexity.

In this paper we will consider which tools and technologies are currently available and which functionality, to the best of our knowledge, is still lacking to support stakeholders in each step from FAIR Data management planning to FAIR data creation, publication, evaluation and (re)use. As authors we have also developed such tools in recent years and we include them here in order to illustrate possible solutions and highlight open issues. A full and comprehensive review of relevant tools and technologies is out of scope for this paper, but references in this paper are available as a community-editable Wiki page [4] and we welcome contributions there in order to increase awareness of existing efforts and to facilitate technological creolization [5] and convergence.

2. FAIR DATA MANAGEMENT PLANNING

With the increase of data-driven research and the rising importance of digital research objects and other digital artifacts [6], e.g., for the purpose of reuse and reproducibility [7], there is more need than ever for researchers to follow proper data management procedures. Moreover, researchers are increasingly required to provide a Data Management Plan (DMP) that meets the requirements as set out by different funding organizations [8] and serves as an adaptable, guiding document of the data management process during the project. A large number of DMP tools have emerged to assist researchers to create and maintain DMPs.

The main challenge for a DMP tool is to efficiently transfer knowledge regarding the many organizational, procedural and technical aspects of data management and data stewardship to an audience of researchers from different backgrounds and domains in order to produce an application- and domain-relevant DMP and to maximize opportunities for good data handling and reuse during and after the project. Many of these tools use the FAIR guiding principles for data management, but do so in a variety of ways. Here we take a look at two examples: DMPOnline [9] and the Data Stewardship Wizard (DSW, [10]), for a more complete discussion, please see [8]. DMPOnline has recently seen rapid adoption from researchers and organizations as the go-to tool to produce funder-compliant DMPs. It provides an online, collaborative environment with (mostly) open text forms divided into sections following a configurable funder's DMP template. For each section, DMPOnline embeds explanatory text from a configurable set of sources, which may be DMP guidelines from funding organizations or academic institutions and may (or may not) contain FAIR-specific guidance. In contrast, the DSW tool guides the user through a comprehensive, "FAIR-aware" data management knowledge model by asking a number of multiple-choice questions with embedded book excerpts for additional explanation [11]. This organization allows DSW to very efficiently point the user to the relevant data stewardship issues, tools and other resources by omitting the parts from the larger knowledge model that would only apply to other cases. DSW also facilitates automatic evaluation of the questions, for example in order to produce FAIRness metrics or other evaluation score. In the future we are likely going to see a continuation of efforts toward machine actionable DMPs and tooling[1], thus enabling DMP interoperability, exchange and (semi-)automatic evaluation of (parts of) the reported data management process. Interestingly, the FAIR metrics (see last section) share similar objectives, which suggest that DMP and FAIR metrics tools may be destined for co-evolution in the future.

3. FAIR DATA PRODUCTION

One of the main challenges following from the FAIR guidelines is that they propose a number of attributes to be associated with the data: unique identifiers [12], (qualified references to) rich metadata, use of vocabularies, provenance, etc. The value of these attributes to any downstream data consumer (be it a human or machine agent) is quite clear, but can also pose a burden on the data producer. We foresee the emergence of a category of tools that support data producers to make sure the data contain the required attributes. These "FAIRifier" tools may come in many different flavors: supporting either generic or domain-specific use cases, FAIRifying at the source or post-hoc, targeting different end-users (e.g., data scientists or data stewards), using different technologies (e.g., semantic Web technology) and supporting (semi)automated or manual workflows.

We have developed a general-purpose FAIRifier on the basis of the OpenRefine data cleaning and wrangling tool [13,14,15] and the RDF plugin[2]. This FAIRifier enables a post-hoc FAIRification workflow: load an existing data set (from a wide range of formats), (optionally) perform data wrangling tasks, add FAIR

---

[1] DMP Common Standards WG. Available at: https://www.rd-alliance.org/groups/dmp-common-standards-wg.

[2] OpenRefine RDF plugin. Available at: https://github.com/stkenny/grefine-rdf-extension.

(metadata) attributes to the data, generate a linked data version of the data and, finally, push the result to an online FAIR data infrastructure to make it accessible and discoverable. Literal values in a data set can be replaced by identifiers (URLs) either manually, by semi-automatic mapping to pre-loaded ontologies (using the OpenRefine reconciliation function) or by embedded, customizable script expressions. The interoperability of the data set can be improved by connecting these identifiers into a meaningful semantic graph-structure (model) of ontological classes and properties using the integrated RDF model editor. A provenance trail automatically keeps track of each modification and additionally enables "undo" operations and repetition of operations on similar data sets. A FAIR data export function opens up a metadata editor to provide information about the data set itself: title, publisher (author), license, and a range of additional optional metadata.

Future development plans include features to make the FAIRifier easier to use for non-technical users. This includes functionality to suggest transformations and (semi)automatic application of graph models based on libraries of ontologies and graph models created by other (expert) users. Many other tools have demonstrated FAIRification capabilities with different benefits and limitations. To name a few: Karma③ offers a user-friendly interface and automatic model selection capability that are not available in the OpenRefine-based FAIRifier, but lacks some of its other features. RightField [16] and Ontomaton [17] transparently integrate FAIRification to end-users by pre-configuring spreadsheet applications with a semantic data model. The different concepts and functionalities offered by these tools are all worth further evaluation and development in the context of creating a rich ecosystem of FAIRifier tools. Finally, note that the tools mentioned in this section and the next, adopt ontologies [18] and linked data [19]. These technologies align very well with a number of FAIR principles "out of the box", but other tools may choose a different core technology for their implementation.

4. PUBLISHING FAIR DATA

Data coming from a FAIRifier can still not be considered fully FAIR and machine actionable, unless they have been published to, or otherwise made available via the Internet. Here we focus mainly on the principles collected under the "A" and related infrastructural aspects, for issues regarding Findability of FAIR data sets, please see the last section. Arguably, the main challenge regarding Accessibility is to make every part of the access process machine actionable, so that machines are enabled to automatically negotiate access (based on conditions set by the data owner) and to retrieve data and metadata in order to (semi) automatically evaluate their fitness for purpose. Part of this problem relates to the representation of accessibility conditions and their organizational, regulatory or legal framework [20, 21, 22]. Another part requires specific support from the infrastructure, i.e., if conditions permit access, the infrastructure should allow data consumers to get to the data in a straightforward, predictable way. This means choosing between a large number of protocols and APIs and their respective standards and conventions.

---

③ Karma: A data integration tool. Available at: http://usc-isi-i2.github.io/karma/.

We have developed the concept of a FAIR Data Point (FDP) [23] with a dual, ongoing goal: 1) to demonstrate comprehensive compliance to the FAIR principles and metrics and 2) as a light-weight infrastructural component and standard that may be used by existing repositories and infrastructures. Primary design objectives to support these goals were to require only minimal (but extensible) semantic descriptions and to adopt a light-weight interface. An FDP serves relevant, FAIR metadata as RDF over a simple RESTful API [24] on five different hierarchical layers starting with metadata about the FDP itself, followed by Catalogs, Data sets, Distributions and, finally, record-level metadata. Its metadata is mainly based on the widely used DCAT[④] and Dublin Core[⑤] standards, with minor extensions to comply with FAIR principles (detailed in the FDP specification document[⑥]). Given a FDP URL, a DCAT-aware REST client can automatically traverse the FDP hierarchy down to the level of actual data records. Traversal may be directed by the client's evaluation of the metadata (e.g., for relevance) or may be halted by the FDP if access restrictions for that level apply. We intend to use the FDP in combination with more refined, currently emerging semantic models to describe access conditions (e.g., based on consent and GDPR regulations) and integration with an Authorization and Authentication Infrastructure for applications in the health domain [25]. There are a number of other standards (most notably Linked Data API[⑦], Hydra[⑧] and Linked Data Platform[⑨]) that provide more sophisticated descriptions to the client about API state transitions and additional API functionality such as querying. We consider these efforts complementary to the FDP and combinations are likely possible. We are currently evaluating in which scenarios such combinations would offer additional benefit before extending the FDP core functionality accordingly.

5. EVALUATING THE FAIRNESS OF A RESOURCE

An emerging consideration for the different stakeholders involved in FAIR activities is the assessment of the FAIRness level of resources. It is often useful to assess to which extent a resource (data or metadata) follows the FAIR principles. This assessment can help evaluate if initial goals for the resource have been achieved and also can help identify desirable points for improvements. A number of different initiatives are currently working on defining frameworks, methods and criteria for evaluating FAIRness. Initiatives include the FAIR Metrics Group[⑩], the RDA FAIR Data Maturity Model Working Group[⑪], the NIH Data Commons Pilot Phase Consortium[⑫] and others and they are mostly ongoing efforts. Nevertheless, a number of online evaluation tools and forms have become available [26, 27, 28, 29, 30], which illustrates the perceived importance of helping users to measure theirs or other people's FAIRness in all phases of the data life cycle.

---

④ https://www.w3.org/TR/vocab-dcat/.
⑤ http://dublincore.org/.
⑥ https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/development/spec.md.
⑦ https://github.com/UKGovLD/linked-data-api/blob/wiki/Specification.md.
⑧ https://www.hydra-cg.com/spec/latest/core/.
⑨ https://www.w3.org/TR/ldp/.
⑩ http://www.fairmetrics.org/.
⑪ https://rd-alliance.org/groups/fair-data-maturity-model-wg.
⑫ https://commonfund.nih.gov/commons/awardees.

For instance, the aforementioned Data Stewardship Wizard incorporated in its knowledge model metrics from the FAIR Metrics Group so that the user can have an indication of the FAIRness level that is expected from the yet to be created data. After data creation, another evaluation can be performed to measure the achieved FAIRness level, and if necessary, a review of the plan can be made to mitigate any problems [31].

6. FINDING AND (RE)USING FAIR DATA

Arguably, efficient use and reuse of data is a major objective of the FAIR guiding principles. Consider an ideal digital world where all data are FAIR: machine agents should then be able to (autonomously) execute a process or workflow to find (principles F) and access (A) any available, relevant data sources and automatically integrate, query and reason over the interoperable (I) data toward a useful result to a problem formulated by either human users or indeed other machine agents. It may seem therefore that reusability (R) is trivially solved if resources fully comply to F, A and I principles and infrastructure exists to support it. However, we would argue that without due consideration of the principles under R, the data would still not be very (re)usable and that the effects and requirements of the R principles permeate through to all the other principles, all steps in the data life cycle, as well as any FAIR supporting infrastructures and tools. Let's for example look at the step of finding relevant data, a problem for which many technical solutions exist, even those exhibiting certain FAIR characteristics. This includes for example the FAIR data search engine prototype, which harvests FDP metadata, indexes it and offers a search UI and API for human and machine searches, respectively [32]. An alternative approach uses structured embedded metadata which may be crawled and indexed by existing online search services: for example, a Web page related to a data set could contain structured "Data set" metadata⊗ and would allow the data set to show up in the Google Data set search® service. Hybrid approaches are also possible: for example the FDP includes a simple UI that embeds schema.org metadata. Even as there appears to be sufficient infrastructure to support "Findability", the data that are found will not actually be usable if the metadata does not specify the legal conditions under which it may be used (R1.1), if the origin, relevance and trustworthiness of the data is not clear (R1.2) or if it does not follow standards relevant for a given domain (R1.3). The main challenge regarding the reusability of the data is therefore to make sure that any FAIR resource includes such a "plurality of accurate and relevant attributes" (R1) to support data reuse. In the findability use case, these attributes could furthermore be used to improve search results by automatically prioritizing relevant, trustable results that the requester is legally able to use for his specific purpose. We note that non-technical developments are of influence as well: a positive example is the recent adoption of the GDPR [33], which is increasingly cited as motivation for works capturing and modeling data usage conditions and constraints [20]. Such works are important precursors for convergence toward broadly accepted and generically applicable metadata standards for data use and access constraints that have yet to emerge. Finally, communities themselves need to identify, develop and promote the required metadata standards and metadata registry services play an important role toward convergence within and across domain boundaries.

⊗ https://schema.org/.

® https://toolbox.google.com/data setsearch.

Registries may range from full-featured, generic solutions like FAIRsharing®, to relatively simple community recommendation lists [34, 35, 36].

7. CONCLUSIONS

In this paper we have shown that there are many ongoing efforts that directly or indirectly contribute to the objective of making FAIR a reality. We have shown that these tools contribute to an ecosystem of FAIR tooling that covers everything from FAIR data management planning, to production, publication, evaluation, finding and (re)using FAIR data. Some of these tools contribute to the design and development of (components of) FAIR infrastructures and platforms, while others address a solution to a very specific FAIR challenge. In most cases there are a number of alternative solutions with some overlapping, but also many complementary features. Moreover, almost all of these efforts have dependencies on, or reach full potential only in combination with other FAIR tools and resources. e.g., FAIRifiers are typically more effective with the availability of registries of (community adopted) FAIR data models and metadata standards, FAIR search and accessibility services cannot work without descriptions of usage and license conditions, etc. In our opinion this signals a creolization phase [4] of FAIR tool development. In the near future we will likely see an increase in the number of available FAIR tools, while simultaneously these tools will evolve, converge and merge in ways that cannot currently be foreseen. Periodically checking alignment with the original aim and intention of the FAIR principles will help to converge such efforts toward the realization of mature FAIR tool ecosystems and infrastructures, FAIR-based domain-specific applications like the Personal Health Train® [37] and the generic Internet of FAIR data and Services [38].

AUTHOR CONTRIBUTIONS

M. Thompson (m.thompson@lumc.nl) has drafted the first version of this paper; R. Kaliyaperumal (R. Kaliyaperumal@lumc.nl), L.O. Bonino da Silva Santos (luiz.bonino@go-fair.org) and K. Burger (c.a.burger@lumc.nl) have proof-read and contributed improvements to the text; all authors have contributed to the design and implementation of the FAIRifier and FAIR Data Point software and specifications described in the paper.

ACKNOWLEDGEMENTS

® https://fairsharing.org/.
® https://www.dtls.nl/fair-data/personal-health-train/.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[2] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O. Bonino da Silva Santos & M.D. Wilkinson. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37(1)(2017), 49-56. doi: 10.3233/ISU-170824.

[3] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, ... & E. Schultes. FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(2020), 10–29. doi: 10.1162/dint_r_00024.

[4] M. Thompson, K. Burger, R. Kaliyaperumal & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: Community editable Wiki page. Available at: https://osf.io/x2h3t/wiki/home/.

[5] P. Wittenburg & G. Strawn. Common patterns in revolutionary infrastructures and data.

[6] J. Borycz & B. Carroll. Managing digital research objects in an expanding science ecosystem: 2017 conference summary. Data Science Journal 17(2018), 16. doi: 10.5334/dsj-2018-016.

[7] S. Bechhofer, D. De Roure, M. Gamble, C. Goble & I. Buchan. Research objects: Towards exchange and reuse of digital knowledge. Nature Precedings (2010), 1–6. doi : 10.1038/npre.2010.4626.1.

[8] S. Jones, R. Pergl, R. Hooft, T. Miksa, R. Samors, J. Ungvari, R.I. Davis & T. Lee. Data management planning: How requirements and solutions are beginning to converge. Data Intelligence 2(2020), 208–219. doi: 10.1162/dint_a_00043.

[9] M. Donnelly, S. Jones & J.W. Pattenden-Fail. DMP online: The digital curation centre's Web-based tool for creating, maintaining and exporting data management plans. In: Research and Advanced Technology for Digital Libraries (ECDL 2010), 2010, pp 530–533. doi: 10.1007/978-3-642-15464-5_74.

[10] M. Suchánek & R. Pergl. Data stewardship wizard for open science. Available at: https://www.researchgate.net/publication/331357542.

[11] B. Mons. Data stewardship for open science: Implementing FAIR principles. Boca Raton: CRC Press, 2018.

[12] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(2020), 30–39. doi: 10.1162/dint_a_00025.

[13] OpenRefine, a free, open source, powerful tool for working with messy data. Available at: http://openrefine.org/.

14] OpenRefine RDF plugin. Available at: https://github.com/stkenny/grefine-rdf-extension.

[15] Karma: A data integration tool. Available at: http://usc-isi-i2.github.io/karma/.

[16] K. Wolstencroft, S. Owen, M. Horridge, O. Krebs, W. Mueller, J.L. Snoep, F. du Preez & C. Goble. RightField: Embedding ontology annotation in spreadsheets. Bioinformatics, 27(14)(2011), 2021–2022. doi: 10.1093/bioinformatics/btr312.

[17] E. Maguire, A. González-Beltrán, P.L. Whetzel, S.A. Sansone & P. Rocca-Serra. OntoMaton: A Bioportal powered ontology widget for Google Spreadsheets. Bioinformatics 29(4)(2013), 525–527. doi: 10.1093/bioinformatics/bts718.

[18] G. Guizzardi. Ontology, ontologies and the "I" of FAIR. Data Intelligence 2(2020), 181–191. doi: 10.1162/dint_a_00040.

[19] C. Bizer, T. Heath & T. Berners-Lee. Linked data: The story so far. In: A. Sheth (ed). Semantic Services, Interoperability and Web Applications: Emerging Concepts. Hershey, PA: IGI Global, 2011, pp. 205–227.

[20] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The "A" of FAIR – as open as possible, as closed as necessary. Data Intelligence 2(2020), 47–55. doi: 10.1162/dint_a_00027.

[21] I. Labastida & T. Margoni. Licensing FAIR data for reuse. Data Intelligence 2(2020), 199–207. doi: 10.1162/dint_a_00042.

[22] C. Brewster, B. Nouwt, S. Raaijmakers & J. Verhoosel. Ontology-based access control for FAIR data. Data Intelligence 2(2020), 66–77. doi: 10.1162/dint_a_00029.

[23] L.O. Bonino da Silva Santos, M.D. Wilkinson, A. Kuzniar, R. Kaliyaperumal, M. Thompson, M. Dumontier & K. Burger. FAIR data points supporting big data interoperability. In: M. Zelm, G. Doumeingts & J.P. Mendonça (eds) Enterprise Interoperability in the Digitized and Networked Factory of the Future. London: ISTE Press, 2016, pp. 270–279.

[24] R.T. Fielding. Architectural styles and the design of network-based software architectures. PhD Thesis, University of California, Irvine, 2000.

[25] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The "A" of FAIR – as open as possible, as closed as necessary. Data Intelligence 2(2020), 47–55. doi: 10.1162/dint_a_00027.

[26] FAIR assessment tool. Available at: https://www.ands-nectar-rds.org.au/fair-tool.

[27] FAIR data assessment tool. Available at: https://www.surveymonkey.com/r/fairdat.

[28] FAIRshake. Available: https://fairshake.cloud/.

[29] Fair maturity indicators and tests. Available at: https://linkeddata.systems:3000/FAIR_Evaluator//. [Accessed: 10-Apr-2019].

[30] Fairbear services. Available at: https://fairbearservices.com/.

[31] R. de Miranda Azevedo & M. Dumontier. Considerations for the conduction and interpretation of FAIRness evaluations. Data Intelligence 2(2020), 285–292. doi: 10.1162/dint_a_00051.

[32] FAIR search engine prototype. Available at: https://github.com/FAIRDataTeam/FAIRSearchEngine.

[33] Council of the European Union and European Parliament. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, Apr. 2016.

[34] Open Metadata Registry (formerly the NSDL Registry). Available at: http://metadataregistry.org.

[35] Research Data Alliance Metadata Directory. Available at: http://rd-alliance.github.io/metadata-directory/standards/.

[36] Metadata Standards Catalog. Available at: https://rdamsc.bath.ac.uk/.

[37] O. Beyan, A. Choudhury, J van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, Md. R. Karim, M. Dumontier, S. Decker, L.O. Bonino da Silva Santos & A. Dekker. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence 2(2020), 96–107. doi: 10.1162/dint_a_00032.

[38] M. van Reisen, M. Stokmans, M. Basajja, A. Ong'ayo, C. Kirkpatrick & B. Mons. Towards the tipping point of FAIR implementation. Data Intelligence 2(2020), 264–275. doi: 10.1162/dint_a_00049.