Ready, Set, GO FAIR: Accelerating Convergence to an Internet of FAIR Data and Services

© Erik Schultes
Leiden University Medical Centre
GO FAIR International Support and Coordination Office
Poortgebouw N-01, Rijnsburgerweg 10, 2333 AA Leiden
The Netherlands
erik.schultes@go-fair.org

© George Strawn
Board Director
Board on Research Data and Information (BRDI)
US National Academies of Sciences, Engineering, and Medicine
USA
gstrawn@nas.edu

© Barend Mons
Leiden University Medical Centre
GO FAIR International Support and Coordination Office
Poortgebouw N-01, Rijnsburgerweg 10, 2333 AA Leiden
The Netherlands
barend.mons@go-fair.org

Abstract. As Moore's Law and associated technical advances continue to bulldoze their way through society, both exciting possibilities and severe challenges emerge. The upside is the explosive growth of data and compute resources that promise revolutionary modes of discovery and innovation not only within traditional knowledge disciplines, but especially between them. The challenge, however, is to build the large-scale, widely accessible, and automated infrastructures that will be necessary for navigating and managing the unprecedented complexity of exponentially increasing quantities of distributed and heterogenous data. This will require innovations in both the technical and social domains. Inspired by the successful development of the Internet and leveraging the FAIR Principles (for making data Findable, Accessible, Interoperable and Reusable by machines) the GO FAIR initiative works with voluntary stakeholders to accelerate convergence on minimal standards and working implementations leading to an Internet of FAIR Data and Services (IFDS).

Keywords: analytics and data management, data intensive domains, digital libraries, FAIR Data, GO FAIR Initiative, Internet of FAIR Data and Services (IFDS).

1 Introduction

Existing data stewardship practices are highly inefficient. Numerous studies indicate that data scientists both in academia and industry spend 70-80% of their time on mundane, manual procedures to locate, access, and format data for reuse [1,2]. Methodological legacies inherited from a pre-digital era (e.g., poor capture of metadata, broken links to various research assets) and outdated professional incentives (e.g., only rewarding publication of research articles rather than also datasets

and other research outputs) contribute to massive data loss and a well-documented reproducibility crisis [3-5]. Coupled with the exponential increases in data volumes (driven by, among other things, high through-put instrumentation and IoT data streams) the urgency for automated, commonly usable data infrastructures (i.e., an Internet for Machines) is increasingly recognised by numerous national and international organisations, science funders and industry [6-11]. Despite the urgent need, building a generalised, ubiquitous, data infrastructure that is widely used by diverse stakeholders is an inherently distributed and difficult process to direct. Knowing this to be the case, the GO FAIR initiative was

launched to accelerate data infrastructure development by leveraging general patterns of phased development described in other revolutionary infrastructures, including the Internet and the World Wide Web (WWW) [12].

2 Learning from previous Revolutionary Infrastructures

Revolutionary Infrastructures (for example, transportation, electrification, telecommunications, and computer networks) follow five phases of development [12,13]: (1) Vision: New discoveries and technologies lead to the anticipation of broad new application spaces; (2) Creolization: Inspired by the Vision, numerous experimental implementations are created, resulting in an uneven landscape of independently developed prototypes; (3) Attraction: Some solutions prove more viable, and are effectively generalised to achieve a simplified set of 'universal principles' that attract the attention of others working in the field; (4) Convergence: Various Attractors voluntarily decide to bridge otherwise isolated application solutions, and a compelling global infrastructure begins to emerge at the expense of the many other possibilities; (5) Exploitation: As widespread commitment to a particular implementation emerges, economy of scale kicks in, and what was hard and cost-prohibitive, now becomes easy and affordable. Users in the Exploitation phase might not even be aware of the infrastructure systems they routinely use (e.g., most users of the internet are blissfully ignorant of TCP/IP).

In the specific case of the Internet, there had been early **Visions** of interlinked computers throughout the 1950s and 1960s. By 1969, ARPAnet had initiated the phases of **Creolization** (and later **Attraction**) with the co-existence of multiple, specialised solutions, e.g., X25, Ethernet, ARCNET, and others. This work demonstrated the feasibility of computer networks and drew the attention of large investors (e.g., IBM, DEC). But this investment resulted in numerous incompatible standards that in some ways slowed progress. **Convergence** was eventually triggered with TCP/IP protocols (early 1970s) and the 7-layer ISO/OSI reference model (early 1980s). This was because these minimal standards allowed various networks to interoperate while at the same time maintaining maximum freedom to engineer solutions at the implementation layer 'below' and application layer 'above' (creating the so-called "hourglass" architecture of the Internet, with TCP/IP at the narrow waist). It was working implementations (however embryonic) and the simplicity of the hourglass approach that motivated influential decision makers "to move towards using TCP/IP as universal for implementing global computer networking" [13]. With a stabilized universal in place, **Exploitation** soon followed, with rapid investment in both hardware and software, that is the now familiar story of the Internet. By 1992, the Internet Society was set up to coordinate further develop TCP/IP approaches to networking.

Findable:

F1 (meta)data are assigned a globally unique and persistent identifier;

F2 data are described with rich metadata;

F3 metadata clearly and explicitly include the identifier of the data it describes;

F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

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

I2 (meta)data use vocabularies that follow FAIR principles;

I3 (meta)data include qualified references to other (meta)data;

Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2 the protocol allows for an authentication and authorisation procedure, where necessary;

A2 metadata are accessible, even when the data are no longer available;

Reusable:

R1 meta(data) are richly described with a plurality of accurate and relevant attributes;

R1.1 (meta)data are released with a clear and accessible data usage license;

R1.2 (meta)data are associated with detailed provenance;

R1.3 (meta)data meet domain-relevant community standards;

Figure 1 The 15 FAIR Principles ensuring machine Findability, Accessibility, Interoperation and Re-use of digital resources [18][19]

It is important to note that the use of TCP/IP has always been voluntary, and at no time was its use ever required. Indeed, top-down enforcement policies would likely have killed its effectiveness as an attractor. Instead, once a 'critical mass' of influential users had adopted TCP/IP, the larger community followed, driving convergence. An analogous pattern of development (voluntary use, attractor effect in the community) occurred soon after with the formation of the WWW, in this case with HTTP playing the role of TCP/IP. The significance of this historical insight can not be understated. It enables some degree of control in the development of new infrastructures, because only a relatively few (albeit influential) users need be convinced to invest in a particular technology. Once the 'critical mass' is assembled, the 'long tail' of community stakeholders will likely follow.

Even before the 2000's, visionaries had already anticipated the need for a general-purpose data infrastructure. Digital Object Architectures (DOA), systems supporting Persistent Identifiers (PIDs) and the Semantic Web (a framework for knowledge representation built on top of existing Internet and WWW infrastructures) appeared as an important components, ensuring both data interoperation and machine readability. Since then, difficult problems in this space have been investigated resulting in a plenum of new, co-existing methods, languages, software and specialised hardware, producing by now, a protracted period of Creolization. By 2012 the Attraction phase was underway with public discussions about component specifications, principles and procedures for semantically enabled data infrastructures [14-16]. By early 2014, in a workshop hosted by the Lorentz Center (Leiden), this discussion culminated in the generalised and broadly applicable FAIR Principles for data reuse [17]. In a now widely cited commentary (indicative of the Attraction phase) [18], the FAIR approach had been defined as "Data and services that are findable, accessible, interoperable, and re-usable both for machines and for people" and 15 high-level Principles had been articulated, Figure 1. Immediately following

their publication (April 2016), the FAIR Principles (and later, the corresponding FAIR Metrics [20]) have been acting as a powerful attractor in the emerging data infrastructure.

Following the previous examples, the Convergence phase of the data infrastructure will commence once a 'critical mass' of users commits to particular, minimal specification for automatic routing of FAIR data and services (see for example the continuing discussions around Digital Object Architecture [14,15,21]). This globally distributed data infrastructure will likely be substantially more complex than its predecessors in that an Internet of FAIR Data and Services (IFDS) necessitates elaborate semantically enabled metadata descriptions. The 'FAIRification' of digital resources is not trivial, and widespread application will require an ecosystem of methods, tooling, services and training that help communities of diverse stakeholders to create and use FAIR resources. GO FAIR supports and coordinates bottom-up community initiatives that aim to 'Make FAIR easy" [11, 22].

3 GO FAIR

3.1 Accelerating Convergence toward a FAIR data infrastructure

Given that many different combinations of technology choices and use of standards could conceivably implement the FAIR Principles, the GO FAIR initiative was launched in late 2017 by the Dutch, German and French governments as a means to pragmatically accelerate community Convergence. The initial vehicle for GO FAIR is the International Support and Coordination Office (GFISCO) [11]. Following the examples of the Internet and WWW, the GFISCO operates through voluntary stakeholder participation attempting to reach a 'critical mass' of users committed to a set of absolute minimal technology specifications. Beyond these minimal specifications, there is unrestricted room to innovate.

GFISCO is stakeholder governed, and includes researchers from specialized knowledge domains (e.g., earth sciences [23], chemistry [24]) but also policy bodies (e.g., CODATA, RDA, FORCE11), publishers (e.g., Elsevier, Springer-Nature), repositories (e.g., Figshare), and funding agencies (e.g., The American NSF and NIH, the Health Research Board of Ireland, and the Dutch ZonMW). GFISCO brokers among stakeholders, the choice of standards implementing the functions of the FAIR Principles and emerging best practices leading to the Internet of FAIR Data and Services. GFISCO operates via supporting and coordinating Implementation Networks (INs), which are voluntary international consortia that self-organize to implement elements of the IFDS. GO FAIR INs belong to 3 broad topical pillars: GO BUILD, GO TRAIN and GO CHANGE.

3.2 GO BUILD

GO BUILD focuses on the technological aspects of the IFDS, including the design and building of reference implementations for elements composing the IFDS such as FAIR Metrics [20], FAIR Data Points [25,26], FAIRification tools and other FAIR-compliant services. Furthermore, via ongoing "Metadata for Machines" workshops and "Community Challenges", GO BUILD supports and coordinates communities who aim to achieve adoption of globally unique and persistent identifiers, agree on common metadata representation formats, agree on a minimal set of generic metadata content and define domain-relevant community standards. Currently, there are 8 INs under the GO BUILD pillar.

3.3 GO TRAIN

The overall objective of the GO TRAIN pillar is to create a scalable framework that is used in higher education programs and throughout industry to train large numbers of certified data stewards (estimated to be 500,000 for Europe [27], millions more world wide). GO TRAIN supports and coordinates two activities: 1) The development of canonical training curricula focused on FAIR Data Stewardship; 2) The development of certification schema for competencies in FAIR Data Stewardship (providing professional career trajectories, that in turn, are intended to drive rapid uptake of FAIR practices among diverse stakeholders). Currently there are two GO TRAIN INs. The first is the Training Frameworks IN which aims to develop schema for FAIR Data Stewardship education (including train-the-trainer curricula and endorsement specifications), with lenses for Managers, Principal Investigators and Data Stewards themselves. Secondly, The FAIR Curriculum IN will re-use the Carpentries Open, community based curriculum development model [28] to develop novel modular lessons for FAIR data stewardship.

3.4 GO CHANGE

The overall purpose of the GO CHANGE pillar to support and coordinate systemic culture change that transforms existing data management practices into the respected profession of data stewardship. This includes the development of new funding schema, sustainability strategies, and business models. GO CHANGE stakeholders range from international policy makers and national governments to organisation managers and front-line data producers and data stewards. A key IN for GO CHANGE is a FAIR resource hub that aggregates multiple resources for FAIR data stewardship planning, compliance, and assessment.

4 Participating in GO FAIR

4.1 Implementation Networks

GO FAIR INs foster a collaborative community of harmonized practice which leads to Convergence and allows members to 'speak with one voice' on critical

issues regarding FAIR data infrastructures. Anyone (i.e., a person, an institution or a network organisation) can join an existing or create a new GO FAIR IN [29]. The list of current GO FAIR INs can be found at the GO FAIR website [30].

4.2 Launching an IN

The requirements to become an IN are minimal: 1) have a plan to implement an element of the IFDS (including adequate resourcing to accomplish the proposed goals); 2) comply with the GO FAIR Rules of Engagement (essentially, commitment to the FAIR Principles and 'no vendor lock-in' [https://www.go-fair.org/implementation-networks/rules-of-engagement/]); 3) have sufficient critical mass to be regarded as thought leaders in the field of expertise. Moreover, IN leaders will compose a 'manifesto' describing the goals and mode of operation of the IN [31]. Drafting the manifesto can be done in assistance with the GFISCO as part of ongoing, periodic, 1-day Manifesto Workshops [32]. Interested parties can initiate the application process by completing an online intake form [33].

References

[1] Stehouwer H & Wittenburg P. RDA Europe: Data Practices Analysis. (Jan 11, 2018) http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f

[2] Data Scientist Report, Crowdflower. (2017) https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html

[3] Schloss PD. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. mBio, 9(3), e00525–18. (2018) http://doi.org/10.1128/mBio.00525-18

[4] Gorgolewski KJ & Poldrack RA. A Practical Guide for Improving Transparency and Reproducibility in Neuroimaging Research. PLOS Biology 14(7): e1002506. (2016) https://doi.org/10.1371/journal.pbio.1002506

[5] Barend Mons. Data Stewardship for Open Science: Implementing FAIR Principles, 1st Edition. Chapman and Hall/CRC. (2018)

[6] Research Data Alliance https://www.rd-alliance.org

[7] Implementation Roadmap for the European Open Science Cloud (14 March 2018) http://www.esfri.eu/ri-world-news/implementation-roadmap-european-open-science-cloud

[8] New Models of Data Stewardship, NIH Data Commons https://commonfund.nih.gov/commons

[9] How expensive is FAIR compliance and how expensive is it to not be FAIR compliant. RDA 11th Plenary BoF meeting (2018) https://rd-alliance.org/how-expensive-fair-compliance-and-how-expensive-it-not-be-fair-compliant-rda-11th-plenary-bof

[10] G7 SCIENCE MINISTERS' COMMUNIQUÉ. Turin, 27 – 28 September http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqué.pdf

[11] Progress Towards the European Open Science Cloud: GO FAIR Office Established, Global ActionPlatform (2017) http://globalactionplatform.org/post/progress-towards-the-european-open-science-cloud-go-fair-office-established

[12] Thomas P Hughes. Networks of Power: Electrification in Western Society, 1880–1930 . Baltimore: Johns Hopkins University Press. (1983).

[13] Wittenburg P & Strawn G. Common Patterns in Revolutionary Infrastructures and Data. US National Academy of Sciences (February, 2018) https://www.rd-alliance.org/sites/default/files/Common_Patterns_in_Revolutionising_Infrastructures-final.pdf

[14] International DAITF Workshop at the ICRI 2012 Conference http://www.icri2012.dk/www.ereg.me/ehome/index06e1.html

[15] Research Data Alliance, Data Foundation & Terminology Group, Core Terms and Model http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

[16] The FAIR Data Principles, FORCE11 https://www.force11.org/group/fairgroup/fairprinciples

[17] Jointly designing a data FAIRPORT (13-16 January 2014), Lorentz Center faculty of Science of Leiden University, Leiden The Netherlands https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602

[18] Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (2016), 160018. doi:10.1038/sdata.2016.18

[19] GO FAIR, FAIR Principles Explained. https://www.go-fair.org/fair-principles/

[20] Wilkinson MD, et al. A design framework and exemplar metrics for FAIRness. Sci. Data 5:180118 doi: 10.1038/sdata.2018.118 (2018)

[21] Research Data Alliance, Data Type Registries Recommendations (Endorsed) https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

[22] GO FAIR International Support and Coordination Office (GFISCO), http://go-fair.org

[23] American Geophysical Union's Enabling FAIR

Data Project http://www.copdess.org/enabling-fair-data-project/

[24]   GO FAIR Chemistry Implementation Network (ChIN), Supporting FAIR Exchange of Chemical Data Through Standards Development https://iupac.org/event/supporting-fair-exchange-chemical-data-standards-development/

[25]   Wilkinson MD, et al. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Computer Science 3:e110 https://doi.org/10.7717/peerj-cs.110 (2017)

[26]   FAIR Data Point Specification https://github.com/DTL-FAIRData/FAIRDataPoint/wiki/FAIR-Data-Point-Specification

[27]   500,000 data scientists needed in European open research data, JoinUp Platform, European Commission (2016) https://joinup.ec.europa.eu/news/500000-data-scientists-need

[28]   The Carpentries https://carpentries.org

[29]   GO FAIR Implementation Networks https://www.go-fair.org/implementation-networks/

[30]   GO FAIR Current Implementation Networks https://www.go-fair.org/implementation-networks/overview/

[31]   GO FAIR Implementation Network manifesto template https://www.go-fair.org/manifesto-template/

[32]   GO FAIR Manifesto Workshops https://www.go-fair.org/implementation-networks/starting-new-implementation-network/manifesto-workshop

[33]   GO FAIR Implementation Network intake form https://www.go-fair.org/implementation-networks/starting-new-implementation-network/implementation-network-application-form/