Information Services & Use 37 (2017) 49–56
DOI 10.3233/ISU-170824
IOS Press

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud

Barend Mons [a,b,c,*], Cameron Neylon [d], Jan Velterop [e], Michel Dumontier [f], Luiz Olavo Bonino da Silva Santos [b,g] and Mark D. Wilkinson [h]

[a] *Leiden University Medical Centre, Leiden, The Netherlands*
E-mail: b.mons@lumc.nl
[b] *Dutch Techcentre for Life Sciences, Utrecht, The Netherlands*
[c] *Netherlands eScience Centre, Amsterdam, The Netherlands*
[d] *Centre for Culture and Technology, Curtin University, Perth, Western Australia*
[e] *Independent Open Access Publishing Consultant, Guildford, United Kingdom*
[f] *Institute for Data Science, Maastricht University, Maastricht, The Netherlands*
[g] *Vrije Universiteit Amsterdam, Amsterdam, The Netherlands*
[h] *Centre for Plant Biotechnology and Genomics U.P.M. – I.N.I.A., Madrid, Spain*

Abstract. The FAIR Data Principles propose that all scholarly output should be Findable, Accessible, Interoperable, and Reusable. As a set of guiding principles, expressing only the kinds of behaviours that researchers should expect from contemporary data resources, how the FAIR principles should manifest in reality was largely open to interpretation. As support for the Principles has spread, so has the breadth of these interpretations. In observing this creeping spread of interpretation, several of the original authors felt it was now appropriate to revisit the Principles, to clarify both what FAIRness is, and is not.

Keywords: FAIR Data, Open Science, interoperability, data integration, standards

1. Growing awareness of FAIRness

Open Science is a growing movement. The European Council adopted Open Science and the reusability of research data as a priority, as did the G7 at their summit in Japan [9]. This provided fertile ground for the rapid uptake of the FAIR Data Principles [25] since their recent publication [3]. The DG RTD (the Directorate General for Research and Innovation) of the European Commission took the lead [6], but in close collaboration with other directorates and the USA-based Big Data to Knowledge (BD2K) of the NIH (National Institutes of Health) [15]. Science Europe has adopted FAIR principles as the basis for sharing administrative data on funding [7]. The G20 went further in the 2016 Hangzhou summit by endorsing the FAIR Principles by name [8]. The Principles have also resonated in many discussions beyond their original scope of research data sharing, in domains as diverse as Archaeology [22], and

*Corresponding author: Barend Mons, Einthovenweg 20, 2333 ZC Leiden, P.O. Box 9600, 2300 RC Leiden, The Netherlands. Tel.: +31624879779; E-mail: b.mons@lumc.nl.

0167-5265/17/$35.00 © 2017 – IOS Press and the authors.

This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License (CC BY-NC 4.0).

environmental monitors for "smart cities" [12]. This wide embrace of the FAIR Principles by governments, governing bodies, and funding bodies, has led to a growing number of data resources attempting to demonstrate their FAIRness, for an example, see 'Being FAIR at UniProt' [10]. The UniProt example is spot-on, but there are also emerging indications that the original meanings of findable, accessible, interoperable, and reusable sometimes may be stretched; even, in some cases, in order to avoid change or improvement. In other cases, the proposed implementation of these principles, with the goal of an Internet of FAIR Data and Services, is beginning to raise concern and confusion. Therefore, with the broader community now forming independent, thoughtful opinions about the meaning and consequences of the FAIR Principles, it seems worthwhile to clarify their original intent and interpretation.

2. Becoming cloudy

Achieving the transition from the current closed and silo-based approaches to research towards more open and networked scholarship needs important changes in the science reward and methodological practice. But it also needs an increased support infrastructure of FAIR data-publishing, analytics, computational capacity, virtual machines and workflow systems.

These infrastructure needs have been – and are being – addressed intensively at the European Commission level, especially in the context of the 2016 Dutch EC Presidency [16] and the European Open Science Cloud (EOSC) [5], the e-IRG roadmap [16] and in the US through the NIH Data Commons projects. In Australia, ANDS [2] and AARnet [1] follow a very similar approach and recently, the East African Community has adopted the Dakar declaration on Open Science in Africa [23]. In South Africa, the African Data Intensive Research Cloud [21] is part of the roadmap for research infrastructures as well. Common to all these is the idea of building infrastructure based on rich metadata for the resources in the research environment, that support their optimal re-use. Provision of all such resources and services will necessarily involve a mix of players, including commercial and public ones. A group of early-adopter EU member states is preparing the GO FAIR initiative [13], which is a proposal for the fast-track implementation of the EOSC.

Ensuring that in such globally dispersed infrastructures all provided resources are findable, accessible, interoperable and reusable, as well as ensuring that the qualities of a service (i.e. what it does, and how), as well as the quality of a service (i.e. the degree of excellence), are appropriate for the researchers' needs, requires widely shared and adopted standards and principles, In addition, there is a need for set of community-acceptable 'rules of engagement', that define how the resources within that community will/should function and promulgate themselves. These rules of engagement may vary depending on the needs or constraints within any given community, but in each case, the FAIR guidelines assist the interaction between those who want to use community resources and those who provide them. FAIR guiding principles provide a scaffold for building such rules of engagement within each community.

3. What FAIR is...

FAIR refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of both creator and user.

4. . . . and what FAIR is not

FAIR is not a standard: The FAIR guiding principles are sometimes incorrectly referred to as a 'standard', even though the original publication explicitly states they are not [25]. The guiding principles allow many different approaches to rendering data and services Findable, Accessible, Interoperable, to serve the ultimate goal: the reuse of valuable research objects. Standards are prescriptive, while guidelines are permissive. We suggest that a variety of valuable standards can and should be developed, each of which is guided by the FAIR Principles. FAIR simply describes the qualities or behaviours required of data resources to achieve – possibly incrementally – their optimal discovery and scholarly reuse.

FAIR is not equal to RDF, Linked Data, or the Semantic Web: The reference article in Scientific Data [25] emphasises the machine-actionability of data and metadata. This implies (in fact, requires) that resources that wish to maximally fulfil the FAIR guidelines must utilise a widely-accepted machine-readable framework for data and knowledge representation and exchange. While there are only a handful of standards and frameworks that could, today, fulfil this requirement, other potentially more powerful approaches may appear in the future. As such, the FAIR Principles explicitly do not prescribe the use of RDF or any other Semantic Web framework or technology. That said, RDF, together with formal ontologies, are currently a popular solution to the knowledge-sharing problem that also fulfil the requirements of FAIR. As such, RDF and widely adopted ontologies or vocabularies figure prominently in many of the early FAIR examples [4,17,27]. We would, however, like to emphasise that, as with any technology, RDF has its range of suitable applications, but is unsuitable for others. Therefore, it is very likely that applications in the Internet of Data and Services, and the Internet of Things, will use a variety of data formats that allow specific and scalable manipulation of data for pattern recognition and knowledge discovery, and these representations may or may not be FAIR. RDF plus proper ontologies are very effective for the purpose of interoperability and information-sharing, particularly at the level of metadata; however, any other format may also be used in a FAIR context, including size-efficient formats aimed at high performance analytics applications. Data (or portions of data) should only be exposed using FAIR formats if this clearly increases their findability, accessibility, or reusability.

FAIR is not just about humans being able to find, access, reformat and finally reuse data: The official press release following the publication of the FAIR Principles states the authors' position clearly: "The recognition that computers must be capable of accessing a data publication autonomously, unaided by their human operators, is core to the FAIR Principles. Computers are now an inseparable companion in every research endeavour". In recent surveys, the time reportedly spent by PhD students and other researchers in projects dealing with discovering and reusing multiple data sources – so called 'data munging' – has been pegged at 80% [19]. Were these colleagues and their machine-assistants only having to deal with FAIR data and services, this wasted time would be reduced to a fraction of what it is today. The avoidance of time-wasting would be a first return on investment in good data stewardship. To serve this potentially enormous cost reduction, FAIR compliant (meta)data and services should be actionable by machines without human supervision whenever and wherever possible.

FAIR is not equal to Open: The 'A' in FAIR stands for 'Accessible under well defined conditions'. There may be legitimate reasons to shield data and services generated with public funding from public access. These include personal privacy, national security, and competitiveness. The FAIR principles, although inspired by Open Science, explicitly and deliberately do not address moral and ethical issues pertaining to the openness of data. In the envisioned Internet of FAIR Data and Services, the degree to which any piece of data is available, or even advertised as being available (via its metadata) is entirely at the discretion of the data owner. FAIR only speaks to the need to describe a process – mechanised or

manual – for accessing discovered data; a requirement to openly and richly describe the context within which those data were generated, to enable evaluation of its utility; to explicitly define the conditions under which they may be reused; and to provide clear instructions on how they should be cited when reused [11]. None of these principles necessitate data being "open" or "free". They do, however, require clarity and transparency around the conditions governing access and reuse. As such, while FAIR data does not need to be open, in order to comply with the condition of reusability, FAIR data are required to have a clear, preferably machine readable, license. The transparent but controlled accessibility of data and services, as opposed to the ambiguous blanket-concept of "open", allows the participation of a broad range of sectors – public and private – as well as genuine equal partnership with stakeholders in all societies around the world.

FAIR is not a Life Sciences hobby: The first definition of the FAIR principles was formulated by a group that was perceived as primarily coming from a life sciences background. However, the principles may be equally applied to any data, or any service, in any discipline. The problems that hinder data reuse in the Life Sciences – including ambiguity of symbols; too many persistent identifiers for the same concept; semantic drift; linguistic barriers; descriptions of analytical methodologies; tools and their capabilities; and the need for adequate and accurate citation – are issues affecting many other scholarly and professional domains, such as the humanities or law.

5. Is FAIR fair?

The actual meaning of the term 'fair' in everyday life is in some ways also confusing. People have different perceptions and connotations associated with it. One major criticism (relating to the machine-actionability aspect of the principles) is the perception that non-machine-readable data would be considered in some way 'unfair'. We must point out, again, that we explicitly describe FAIR as a spectrum, and a continuum; that there is no such thing as 'unfair' being associated with the FAIR principles, except maybe the specific case of data that are not even findable. As we noted above, not all data can, or should, be machine-actionable. There are numerous circumstances where making data machine-actionable would reduce its utility (e.g. due to the lack of tools capable of efficiently processing the machine-actionable format). We emphasise that as long as such data are clearly associated with FAIR metadata, we would consider them fully participating in the FAIR ecosystem.

A very positive connotation of FAIR is that the acronym carries the ring of general 'fairness'. On the one hand the 'A' allows fair shielding or protection of data that cannot be open for good reasons of various kinds, so that citizens and medical researchers, but also for instance industry, are assured of proper data protection. On the other hand, from the basic principle that FAIRness is maximised when data are open, maximising 'A' implies maximising openness. This includes addressing, to the greatest extent reasonable, the machine-actionability aspect of FAIR. The 'fair' connotation should therefore not be underestimated either. Data that are not open will simply participate less in the Open-Science-driven Social Machines that will dominate science in the near future.

6. Partly FAIR may be fair enough

Figure 1 shows how data can become increasingly FAIR digital objects: Panel A represents the (unfortunate) situation of more than 80% of the datasets in current practice effectively being unavailable for reuse. Almost as many are simply unusable [20], which is why we coin the term 'reuseless' for those data

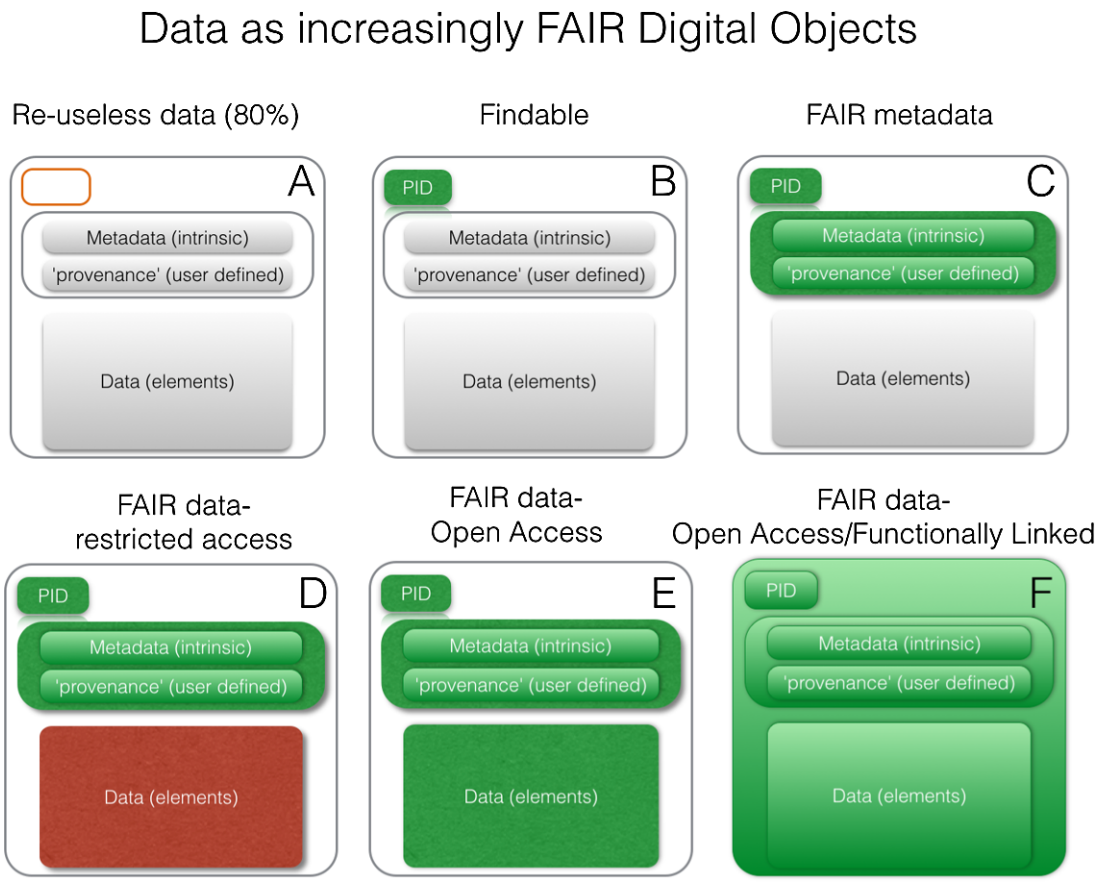

Fig. 1. Varying degrees of FAIRness. As elements become coloured, they have become FAIR. For example, adding a persistent identifier (PID) increases the fairness of that component. Coloured elements in green are FAIR and open, coloured elements in red are FAIR and closed. In the final panel, the mechanism for expressing the relationship between the ID, the metadata, and the data, is also FAIR (i.e. follows a widely accepted and machine-readable standard, such as DCAT or NanoPublications) and interlinked with other related FAIR data or analytical tools on the Internet of FAIR Data and Services.

and services, rather than the term ‘unfair’. Reuseless data are, for instance, those published as obscure and unstable links to supplemental data in narrative articles, not even (as a set) having a proper, machine-resolvable, Persistent, Unique Identifier (PID) which renders both the data elements themselves, as well as their metadata, non-machine-readable.

A minimal step towards FAIRness is to provide the data set, as an entity in its own right, with a PID that is not only intrinsically persistent, but also persistently linked to the data set (research object) it identifies (panel B). However, without machine-readable metadata it will still be difficult to find the data, unless one knows the PID. So a PID is necessary, but not sufficient.

We distinguish ‘intrinsic metadata’ and ‘user-defined’ metadata. The former category (albeit with the boundaries sometimes blurred) are the metadata that should be constructed at data capture. In other words, they describe the metadata that is often automatically added to the data by the machine or workflow that generated the data (e.g. DICOM data for biomedical images, file format, time date stamps, and other features that are intrinsic to the data). Such metadata can be anticipated by the creator and added in order to be useful to Find, Access, Interoperate, and thus reuse, the research object.

As it is very burdensome to peer review the quality of data at the time they are first published, the ongoing and extended review and annotation of data sources during the period of their existence and reuse is a crucial process in Open Science, an approach addressed for instance in the CEDAR project [14]. We argue that both intrinsic and user-defined provenance (e.g. contextual) metadata should be added, and made FAIR whenever possible (panel C).

Not all data lend themselves to be machine-actionable without human intervention (some raw data, but also images for example). However, many data that have a relational or an assertional character can be captured perfectly correctly in a machine-processable semantic syntax. Nevertheless, even if data are technically FAIR, it may be necessary to restrict access to them for reasons discussed above (panel D). That said, the default for maximal FAIRness should be that the data themselves are made available under well-defined conditions for reuse by others (panel E).

We argue here that even the step from A to B would already have a profound effect on the actual reuse of research objects, because at least they can be consistently located by those who know the identifier, and thus can be shared via that identifier. However, thereafter, the addition of rich, FAIR metadata is the first major step towards becoming maximally FAIR. When the data elements themselves can also be made FAIR and made open for reuse by anyone, we have reached the highest degree of FAIRness. When all of these are linked with other FAIR data, we will have achieved the Internet of (FAIR) Data. Once an increasing number of applications and services can link and process FAIR data we will finally achieve the Internet of FAIR Data and Services (panel F). However, when data are not FAIR (at least at level C) they simply cannot truly participate in this future scenario.

7. FAIR and closed could support FAIR and open

In a recent press interview article, Barend Mons proposed a new business model to allow ‘closed’ to pay for ‘open’ [24]. The basis of this proposed business model is that cloud services be free at the point-of-use in the situation when, and if, the user (not just the originator or creator of the data) contributes fully to Open Science. In other words, all user queries, annotations, analytical results and subsequent publications would be fully Open Access and therefore contribute to the public good of open data and services. Those users, however, who wish to keep any of these actions and results private or secret would need to pay. This is perceived as just (fair) by both academics and colleagues in the private sector. Researchers in hospitals, companies, national security agencies and other secrecy-prone players use Open Public Good data as much as all others, so it seems only fair that they contribute to the sustainability of the open data when they use these services for their private or proprietary goals. The fair use of FAIR data is a critical asset in the toolbox of further (and hopefully realised) sustainable development goals.

8. Early adoption

A Skunkworks-like group of coders spontaneously formed at the original Lorentz workshop and, after attracting additional experts from various fields, this group has recently published an exemplar implementation for Web-based discovery and interoperability that is fully compliant with the FAIR principles [26]. This exemplar is not intended to be prescriptive, or even a recommendation; its sole purpose is to describe a novel interoperability infrastructure that naturally leads to adherence to every aspect of FAIR – thus answering the question “what does FAIR look like?”

Many other implementations will no-doubt be necessary, and will be welcomed, to solve a broader range of problems currently precluding effective sharing and reuse of data and services. Obviously, the FAIR Principles are not magic, nor are they presenting a panacea, but they guide the development of infrastructure and tooling to make all research objects optimally reusable for machines and people alike, which is a crucial step forward. It is very important that the community continues to discuss, challenge and refine their own implementation choices, within the 'behavioural' guidelines established by the principles.

9. In conclusion

The FAIR Principles have further propelled the global debate about better data stewardship in data-driven and open science, and they have triggered funding bodies to discuss their requirements for implementation of the FAIR principles; some of these are very embryonic, while others have matured to actual guidelines [7] and there are already attempts to implement supporting prototypes [18]. We strongly believe that FAIR data and services are a key substrate for evidence-based decisions; allow the exposure of research and intellectual property malpractices of multiple kinds; the full participation of citizens and citizen-scientists (i.e. not only professional scientists) from developed and developing countries alike. While intentionally demanding certain qualities and properties from data resources, the FAIR principles nevertheless allow a great deal of freedom with respect to implementation. We hope that this revisiting of the FAIR principles will serve to remove some misperceptions. Nevertheless, we welcome feedback about any concerns that may emerge from the community in the future.

References

[1] AARNet, Welcome to AARNET [Internet], [cited 2016 Dec 15]. Available at https://www.aarnet.edu.au/.
[2] ANDS, Home [Internet], [cited 2016 Dec 15]. Available at http://www.ands.org.au/.
[3] Article Metrics, The FAIR guiding principles for scientific data management and stewardship, *Scientific Data* [Internet], [cited 2016 Dec 15]. Available at http://www.nature.com/articles/sdata201618/metrics.
[4] L.B. Da Silva Santos, M.D. Wilkinson, A. Kuzniar, R. Kaliyaperumal, M. Thompson, M. Dumontier et al., FAIR data points supporting big data interoperability, in: *Enterprise Interoperability in the Digitized and Networked Factory of the Future* [Internet], M. Zelm, G. Doumeingts and J.P. Mendonça, eds, iSTE Press, 2016, pp. 270–279. Available at http://www.iste.co.uk/index.php?f=a&ACTION=View&id=1073.
[5] Eudat, Liber, OpenAIRE, Egi, Geant, European Open Science Cloud for research [Internet], Position Paper, 2015 Oct. Available at http://libereurope.eu/wp-content/uploads/2015/11/OSC_Position_Paper-final-30.10.15.pdf.
[6] European Commission, European Open Science Cloud | Open Science – Research & Innovation [Internet], [cited 2016 Dec 15]. Available at http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
[7] European Commission, Directorate-General for Research, H2020 programme guidelines on FAIR data management in horizon 2020, 2016 Jul.
[8] G, G20 Leaders' Communique Hangzhou Summit [Internet], 6 September, 2016. Available at http://www.consilium.europa.eu/en/meetings/international-summit/2016/09/Leaders-CommuniqueHangzhouSummit-final_pdf/.
[9] G7 Science and Tsukuba Communiqué, *G7 Science and Technology Ministers' Meeting*, 2016 May.
[10] Inside UniProt, Being FAIR at UniProt [Internet], 2016 [cited 2016 Dec 15]. Available at http://insideuniprot.blogspot.com.es/2016/11/being-fair-at-uniprot.html.
[11] Joint Data Citation Synthesis Group, Declaration of data citation principles – final, in: *FORCE11, San Diego, CA*, M. Martone, ed., 2014.
[12] A.V. López, Environmental FAIR data: jugar limpio con la información ambiental [Internet], *Madrid: Productor de Sostenibilidad* (2016), 1–16. Available at http://www.conama2016.org/web/generico.php?idpaginas=&lang=es&menu=257&id=1426&op=view.
[13] B. Mons, GO-FAIR initiative – Dutch Techcentre for Life Sciences [Internet], GO-FAIR Initiative, [cited 2017 Feb 6]. Available at http://www.dtls.nl/go-fair/.

[14] M.A. Musen, C.A. Bean, K.-H. Cheung, M. Dumontier, K.A. Durante, O. Gevaert et al., The center for expanded data annotation and retrieval, *J Am Med Inform Assoc* [Internet] **22**(6) (2015), 1148–1152. Available at http://jamia.oxfordjournals.org/content/22/6/1148.abstract.
[15] NIH, Commons Home Page, Data Science at NIH [Internet], [cited 2016 Dec 15]. Available at https://datascience.nih.gov/commons.
[16] NLU, Amsterdam call for action on open science [Internet], 2016 Jul. Available at file:///home/markw/Downloads/amsterdam-call-for-action-on-open-science.pdf.
[17] L. Olavo Bonino, A. Gavai, A. Kuzniar, R. Kaliyaperumal and K. Burger, FAIR data point software specification – FAIR Data point – Confluence [Internet], [cited 2017 Feb 8]. Available at https://dtl-fair.atlassian.net/wiki/display/FDP/FAIR+Data+Point+Software+Specification.
[18] R. Pergl and M. Suchánek, DS wizard [Internet], ELIXIR data stewardship wizard, [cited 2017 Feb 6]. Available at http://dmp.fairdata.solutions/.
[19] G. Press, Cleaning Big Data: Most time-consuming, least enjoyable data science task, survey says, *Forbes* [Internet], 2016 Mar 23. Available at http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3643728a7f75.
[20] D.G. Roche, L.E.B. Kruuk, R. Lanfear and S.A. Binning, Public data archiving in ecology and evolution: How well are we doing?, *PLoS Biol* [Internet] **13**(11) (2015), e1002295. Available at http://dx.plos.org/10.1371/journal.pbio.1002295.
[21] R. Simmonds, R. Taylor, J. Horrell, B. Fanaroff, H. Sithole, S.J. van Rensburg et al., The African data intensive research cloud, in: *2016 IST – Africa Week Conference* [Internet], IEEE, pp. 1–8. Available at http://ieeexplore.ieee.org/document/7530650/.
[22] Standing Committee on Archaeology, Harvard University. Critical Perspectives on the Practice of Digital Archaeology | Standing Committee on Archaeology [Internet], Critical perspectives on the practice of digital archaeology, [cited 2017 Feb 6]. Available at http://archaeology.harvard.edu/critical-perspectives-practice-digital-archaeology.
[23] The Sci-GaIA Consortium, The Dakar declaration on open science in Africa [Internet], Sci-GaIA Open Access Repository, 2016. Available at http://dx.doi.org/10.15169/sci-gaia:1457961379.87. doi:10.15169/sci-gaia:1457961379.87.
[24] When privacy-bound research pays for open science [Internet], *EuroScientist* **27** (2016), [cited 2016 Dec 15]. Available at http://www.euroscientist.com/privacy-bound-research-pays-open-science/.
[25] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al., The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* [Internet] **3** (2016), 160018. Available at http://www.nature.com/articles/sdata201618.
[26] M.D. Wilkinson, R. Verborgh, L.O.B. da Silva Santos, T. Clark, M.A. Swertz, F.D.L. Kelpin et al., Interoperability and FAIRness through a novel combination of web technologies, *PeerJ Preprints* [Internet] **4** (2016), e2522v1. doi:10.7287/peerj.preprints.2522v1.
[27] M.D. Wilkinson, R. Verborgh, L.O.B. da Silva Santos, T. Clark, M.A. Swertz, F.D.L. Kelpin et al., Interoperability and FAIRness through a novel combination of web technologies, *PeerJ Preprints* [Internet], Report No. e2522v2, 2017 Jan [cited 2017 Feb 8]. Available at https://peerj.com/preprints/2522/.