

Principles for Data Sharing and Re-Use: are they all the same?

Version 3, 23.4.2015

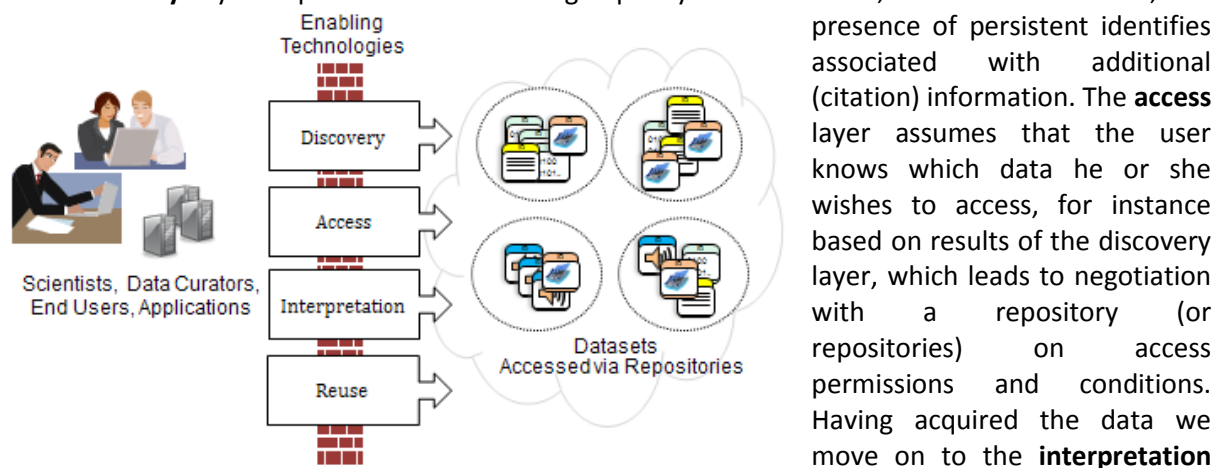
Herman Stehouwer, Peter Wittenburg

Several principles for sharing and reuse of research data have been proposed to change our inefficient data practices. The *Data Practises* report from RDA Europe identifies and described data practices based on many interviews and interactions with different scientific communities [1]. Having a plethora of principles may seem confusing for the general researcher. In discussions at the RDA Europe science workshop [2] three questions were raised:

1. Does a researcher have to be familiar with all these sets of principles?
2. Do people writing Data Management Plans (e.g. for respond to calls funded in programmes that require them, e.g. NSF, H2020) have to keep all the principles in mind?
3. Which principles should be used as basis for further activity (e.g. to guide RDA activities).

We have analysed a number of these principles and in this note (*Request for Comments*) we describe them in a systematic way that allow possible comparison. We reached a provisional conclusion that the principles convey a coherent message and are aligned.

Many of these principles are based on the thoughts presented by Lannom at the DAITF ICRI workshop in 2012 [3]. Lannom formulated a layered approach to data access and reuse technology. In the diagram below these layers are sketched. It is clear that these actions imply several things; The **discovery** layer implies the need for high quality metadata and, for certain activities, the



layer, which implies that the structure and semantics of the acquired data are described adequately (for instance in the metadata). Finally the **reuse** data implies yet more additional information on provenance, semantics, and structure as the data is reused outside of its original context.

We have mapped the following principle sets:¹

- G8 Principles for an Open Data Infrastructure [5]
- G8 London Science Minister Statements [6]
- Nairobi Data Sharing Principles [7]
- FORCE11 Data Citation Principles [8]

¹ We are not aiming to be complete in terms of principle sets, there are many more such as the Panton Principles [4].

- FAIR Data Principles [9]

Establishing agreeable principles is useful as communities and regions need to make them "owned" and "adopted" by their members. The Nairobi principles, for example, emerged in a specific context and an utterly important cooperation interaction between developing countries and UNESCO. However, to avoid confusion they should build on each other. As we reach maturity in this discussion and RDA would be ready to turn principles in practical behaviours, we believe that it is now time to prevent proliferation of additional principles. Funders worldwide start asking for Data Management Plans and publishers requesting research data to be stored in "trustworthy" repositories² which adhere to policy rules as they have been established for example by the Data Seal of Approval [10], World Data System [11], Nestor Seal (DIN 31644) [13], and ISO 16363 [14] initiatives³.

In all four analysed sets of principles we found the basic elements indicated by the diagram above with some specific focus and wording and a few extensions depending on the view angle and context. For example, we think that FAIR principles are more specific and directed towards machine processing which requires agreements on more precise terminology etc., the FORCE11 principles come from the specific view of improving scholarly communication and including data in this communication, while the other two are more generic. The Nairobi principles add a few statements which are important from the view of developing countries in particular. The G8 Principles emerge from an international policy orientation statement adopted at very high level (G8 Ministerial Meeting) [5] and [6]. Its usefulness, besides the high-level policy legitimacy, relates with their relevance in triggering the Open Science policies being developed round the world by funders of research. They may expect the research communities and RDA to come-up with the necessary infrastructure, services and advocacy to make them adopted.

Summary

Basically no fundamental disagreement can be identified between the essential messages of the compared sets; in contrast they are overlapping widely which may help reducing confusion. For the Research Data Alliance [12] work for example we believe that we can make use of the assertions made and feed them into the discussion processes in the groups and also to evaluate the results of RDA working groups. This combination of principle sets paints a clear way forward for possible DMP guidelines as well.

References

- [1] Herman Stehouwer & Peter Wittenburg, 2014, *Second year report on RDA Europe analysis programme*, <http://europe.rd-alliance.org/sites/default/files/RDA-Europe-D2.5-Second-Year-Report-RDA-Europe-Forum-Analysis-Programme.pdf>
- [2] to come soon
- [3] Larry Lannom, 2012, *DAITF: Enabling Technologies*, DAITF ICRI 2012 workshop; <http://www.icri2012.dk/www.ereg.me/ehome/index06e1.html>
- [4] Panton Principles: <http://pantonprinciples.org/>
- [5] G8 Principles for an Open Data Infrastructure, <http://purl.org/net/epubs/work/12236702> (*White Paper: 5 Principles for an Open Data Infrastructure*, editors on behalf of the Data Working Group of the G8+O6 Group of Senior Officials on Research Infrastructures, Dr A Blatecky (NSF) (Ed.), Dr J Bicarregui (STFC Rutherford Appleton Lab.) (Ed.), Dr C Morais Pires (EC) (Ed.) – May 2013

² This term "trusted repository" is now widely being used, although "trust" is a relationship that can only be defined by users. Assessment procedures can help users in improving the trust relationship.

³ It should be noted here that these two initiatives are aligning their rules within the RDA process as far as the areas covered overlap.

- [6] G8 London Science Minister Statements, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf
- [7] Nairobi Data Sharing Principles: <http://www.jkuat.ac.ke/wp-content/uploads/2014/08/Nairobi-Data-Guidelines-20142.doc>
- [8] FORCE11 Data Citation Principles: <https://www.force11.org/datacitation>
- [9] FAIR Data Principles: <https://www.force11.org/group/fairgroup>
- [10] Data Seal of Approval: <http://datasealofapproval.org/en/>
- [11] World Data System: <https://www.icsu-wds.org/>
- [12] Research Data Alliance: <http://rd-alliance.org>
- [13] *DIN 31644: Information and documentation - Criteria for trustworthy digital archives.*
- [14] *ISO/TC 20/SC 13, ISO 16363:2012: Space data and information transfer systems -- Audit and certification of trustworthy digital repositories.*

In this document we take the G8 principles as basis and map the other principles where it makes sense. We allow ourselves to do abbreviations where this seems to be opportune – for details we refer to the documents about these principle sets.

G8 Principles	Fair Principles	Nairobi Principles	FORCE11 Principles
Discoverability of data sets with the help of metadata and persistent identifiers to easily find suitable ones.	Data should be findable at any point in time and thus needs to be uniquely and persistently identifiable. To make it re-findable metadata is important that allows users/machines to distinguish from other data.	Data should be easy to find and access	
Accessibility of data sets to make as many sets openly available.	Data should be accessible by humans and machines based on appropriate authorization and well-defined protocols.	Data should be open and unrestricted . Data may be restricted, if adequately justified.	Data citations should facilitate access to data and to associated metadata, documentation, code and other materials.
		Data sharing should be timely .	Data should be cited and citable .
		Data access should be equitable .	A data citation should include a persistent method for identification that is machine actionable, globally unique and widely used.
Understandability of data sets by providing context, structural and semantic information to make them effectively re-usable.	Data should be Interoperable by providing machine readable metadata, data formats that utilize shared vocabularies and ontologies and by guaranteeing the object being syntactically parseable and semantically machine accessible.	Data should be informative and quality controlled.	Data citations should facilitate access to data and to associated metadata, documentation, code and other materials.
	Data should be re-usable by adhering to all earlier criteria, by providing sufficiently well-described and rich data.		
Manageability by providing policies and common technology for		Data should be sustainable .	Unique identifiers and metadata describing the data and its

maintaining availability and curation over long periods.			disposition should persist .
People need to be educated that enable data intensive science.			
Global Research Infrastructures to be built to facilitate international cooperation.			
Openness of Research Data to be fostered to make widest use of data for science and societies.			
Access to Scientific Research Results to be free and rapid to foster the work on Societal Challenges in particular.			
			Data should be considered legitimate, citable products of research.
		Data generation should be flexible and relevant.	
			Data citations should facilitate identification of, access to, and verification of the specific data that support a claim.
		Data Contributors should be given credit .	Data citations should facilitate giving scholarly credit to all contributors.
			Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation across communities.