FAIR and RDA DFT: sharing the same key messages

Peter, Larry, Raphael, Gary, Beth
February 2015

Summary

It is agreed that we urgently need to overcome the lack of interoperability and the inefficiencies of working with data. Therefore, it is important to identify key messages that emerge across disciplines, countries and initiatives and turn them to widely agreed messages independent of differences in packaging and style. In this report we show that FAIR principles widely overlap with RDA results which is excellent, since it will allow us to come close to the ideal of widely accepted messages which then have the potential to be accepted as guidelines from funders which in turn will reduce the huge solutions space, increase interoperability, and thus reduce costs.

Background

From broad surveys [1], publications [2,3] and activities such as FAIR and Research Data Alliance (RDA) (see annexes) a number of statements can be made about the inadequate approach we currently have for dealing with data:

- Working with data is too inefficient, too expensive and prevents many scientists and companies from finding new insights and creating new businesses.
- There are a huge number of information infrastructure initiatives that all work on solutions for creating integrated and interoperable data and tool spaces within various domains.
- This approach of letting 1000 flowers blossom was useful in including broad groups of researchers and technologists, but has had two negative side effects
    - despite many interactions these initiatives work within their own agenda and create silo solutions
    - the policy makers in government and science are confronted with a huge solution space hampering decision taking and thus fast progress on overcoming barriers
- There are many sociological and technical barriers preventing efficient data sharing and re-use.

As a consequence a number of global initiatives started looking at common data principles such as G8, FAIR/FORCE and the Nairobi Initiative. Stehouwer and Wittenburg[4] compared these principles and came to the conclusion that we basically have the same messages at an abstract level despite small differences in scope and wording. According to these principles we all share the view that data must be searchable/findable, accessible, understandable/interoperable/re-usable and manageable/ persistent. It is excellent that at this abstract level we share the core messages giving us hope that these will help in changing the culture of sharing.

We see an urgent need to stress the agreements and not to point to remaining disagreements in small nuances, since this would confuse most of the scientists who are not dealing with these aspects in a daily manner and who are waiting for clear messages to feel safe in making investments of their time and resources.

Recommendations

However, we cannot stay at that level of abstraction and need to turn these more abstract recommendations into more concrete recommendations. In particular the FAIR initiative and the Research Data Alliance are active and relevant, although coming from different starting points:

- FAIR as an initiative within the world-wide bioinformatics domain where the huge investments and the need to use globally shared data and tools require fast action in converging on domain-wide agreements.
- RDA as a global and cross-disciplinary initiative to overcome the many sociological and technological barriers that hamper improved data sharing and re-use.

Both share obviously similar goals although the approaches differ. RDA working and interest groups, in general, start from a large number of use cases from different disciplines and countries and try to generalize from these individual practices. Overarching groups such as the Data Fabric Interest Group and also the Technical Advisory Board bring the different outputs into a bigger picture and it is now agreed that much testing and adoption of RDA group outputs needs to take place to improve the outputs where necessary and to draw conclusions. After roughly 3 years we can observe quite a number of agreements emerging within the RDA community based on the intensive cross-disciplinary and global interactions.

FAIR is a domain-based approach and the intensive discussions within the domain, but also the interactions with many other initiatives resulted in increasingly detailed agreements. {needs to be worked out by Barend, Michel, etc.}

This paper aims to show the relationships between the statements of these two initiatives. For RDA we primarily refer to the results of the Data Foundation & Terminology (see annex B) and Data Fabric Groups, but also note other relevant RDA groups as appropriate. We take the FAIR statements as the starting point, as it is not easy to do it the other way round since RDA has various results outside of the scope of the FAIR statements.

FAIR Statement	RDA Statement	Comment
(meta) data[1] are assigned a globally unique and eternally persistent identifier.	A Digital Object (DO) is referenced and identified by a persistent identifier. Metadata descriptions and collections are DOs as well.	FAIR speaks about “unique” PIDs while RDA assumes PIDs to be unqiue. RDA does not make a difference between various types of DOs. But since metadata descriptions are DOs the statements are overlapping.
data are described with rich metadata.	A Digital Object has properties that are described by metadata.	FAIR is adding the adjective “rich” which is hoped for by everyone in particular since otherwise machines will not be able to interpret them.
(meta) data are registered or indexed in a searchable resource.	DOs have bitstreams that are stored in repositories. Metadata descriptions are stored in metadata repositories.	RDA makes a difference between data and metadata repositories, since often data and metadata are treated differently. Often metadata and data are maintained by the same repository. Registries are built up by services providers aggregating metadata and providing search facilities.
metadata specify the data identifier.	Metadata minimally needs to contain the PID of a DO.	RDA is fairly elaborate on this issue, since PIDs are a special type of metadata, a PID record can include other properties of DOs allowing for example the inclusion of the PID of the metadata description.

[1] FAIR combines “(meta) data” as a short hand indicating that the statements hold for data and metadata.

(meta) data are retrievable by their identifier using a standardized communications protocol. The protocol is open, free, and universally implementable. The protocol allows for an authentication and authorization.	A PID record contains a set of attributes describing DO state properties. State information is metadata information that describes those current properties of the DO that are relevant for proper management and access.	Within the various communities PIDs are being used differently, therefore RDA needs to use these definitions. Location information to retrieve/access a DO's content is part of the state information. In RDA a group started working now on an API for repositories that allows accessing DOs. Indeed DAA methods need to be considered since some data will be protected.
metadata are accessible even when the data are no longer available.	PID records are stored persistently and PIDs need to be resolvable persistently.	RDA does not make a statement on the persistency of metadata except for the metadata (DO state properties) that are stored in the PID record. Here FAIR is more explicit.
(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	In RDA the Metadata Groups are working on metadata issues. MD groups talk about packages now to distinguish different functions of MD. There is common agreement, although not result of a specific WG/IG output, that metadata schemas need to be registered and the concepts used need to be defined and registered in open registries. Yet given the practices for in particular metadata there is no agreement to use "a formal ... language for knowledge representation". It seems to be an excellent objective to agree on export formats independent of how people organize their metadata. Yet RDA does not have a group that formulates general requirements for data. Here FAIR is more explicit and RDA can support the their formulations.	There is a wide agreement in RDA that the schemas and semantics for DOs must be openly registered and defined. The schema implies a structured description and it is expected that semantics are stored in suitable vocabulary registries. There is a new RDA group to deal about vocabulary registries. Registries must be open.
(meta)data use vocabularies that follow FAIR principles.		
(meta)data include qualified references to other (meta)data.		RDA speaks about binding information in the PID record. This is crucial in order to connect each DO with its bitstream, metadata, landing pages, rights records, etc. Repositories may implement this, however, in different ways. It is agreed that metadata is the place that should include references to relevant context and history.
meta(data) have a plurality of accurate and relevant attributes.		RDA recognizes that there are different types of metadata created at different moments by different stakeholders such as descriptive md, state information (systems md), rights md, references to contexts and provenance.
(meta)data are associated with their provenance.		
(meta)data are released with a clear and accessible data usage license.		RDA agrees that descriptive metadata should be openly accessible. Data will be protected as needed and metadata should include clear statements about licenses.
(meta)data meet domain-relevant community standards.		This is a rather vague formulation with which RDA can of course agree. However, there are so many md standards out there, that RD as a first output in the md domain created a registry of standards.

We can identify a few differences.

- FAIR statements about metadata and data content are much more detailed and specific at this point than are the statements from RDA groups. However, we do not see any difference in basic positions.
- FAIR requires persistence of metadata. Currently there is no agreement on this in RDA. It is widely agreed that relevant property information in the PID records must be persistent.
- RDA differentiates between types of metadata, which will be crucial in the way we want to standardize and handle metadata. PID records, for example, should at least include the typical

'passport' information (ID, fingerprint, dates, locations) so that amongst others at every moment identity and integrity can be assessed.

- RDA defines the term "collection". This is an important concept since most of the management and analytics processes will be executed on collections and, further, collections are logically one type of DO.

Given these small differences and the large overlap we believe that FAIR and RDA are in agreement on the same basic messages despite the differences in wording and packaging. RDA is busy formulating common agreements that are beyond the outputs of the RDA working and interest groups and we should closely synchronize with the FAIR initiative and also FAIR should closely synchronize with RDA.

Conclusions

Most important for the data scientists and data managers in the thousands of labs doing data-driven science are unified messages where possible to reduce the amount of confusion and to create an atmosphere of innovation.

RDA will continue finding agreements on data principles and also components within RDA, but also look for agreements with other initiatives. Therefore, the FAIR principles will be brought in into the discussions of RDA groups to seek support where RDA, until now, is not explicit. RDA will continue to invite key people behind FAIR principles to do cross-fertilization and it will refer to the FAIR principles where applicable.

To overcome the huge fragmentation and to turn results and statements of initiatives into recommendations to funders to finally increase interoperability and reduce costs of infrastructure building it is obvious that we need to take all chances to identify agreements across initiatives.

[1] RDA Europe: Data Practices Analysis: http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f
[2] Riding the Wave: https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data
[3] Data Harvest Report: https://rd-alliance.org/data-harvest-report-sharing-data-knowledge-jobs-and-growth.html
[4] Herman Stehouwer, Peter Wittenburg, Principles for Data Sharing and Re-use: are they all the same?, http://hdl.handle.net/11304/1aab3df4-f3ce-11e4-ac7e-860aa0063d1f

Annex A: FAIR Principles

https://www.force11.org/group/fairgroup/fairprinciples

Preamble

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows. Here, we describe **FAIR** - a set of guiding principles to make data **Findable, Accessible, Interoperable, and Re-usable**.

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be 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 authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be 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.

To be Re-usable:

R1. meta(data) have 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 their provenance.
R1.3. (meta)data meet domain-relevant community standards.

Annex B: RDA Data Foundation and Terminology - DFT: Results RFC

The goals of the DFT Working Group were:

- Moving the discussion in the data community towards an agreed, upon suite of terms relevant to RDA groups with a focus on basic core model of related terms along with some basic principles that will harmonize the data organization solutions.
- Fostering an effective RDA community culture by converging on essential terminology arising from agreed upon reference models.

Based on a variety of data models and use cases presented by experts coming from different disciplines and about 120 interviews and interactions with different scientists and scientific departments, the DFT WG has composed a number of simple definitions for digital data in a registered2 domain based on group conceptualization and synthesis.

1. DFT Core Term Definitions

1.1 Digital Object (DO)

Definition

A digital object (DO) is represented by a bitstream, is referenced and identified[3] by a persistent identifier and has properties that are described by metadata.

Note: As indicated we only talk about registered DOs in the context of this document.
Note: Properties included in metadata include discovery, contextual, schema, rights, curation and provenance information.
Note: A DO is said to be a dynamic DO when the information content represented in a DO is changing for some period of time or even for an indefinite duration.

1.2 Persistent Identifier (PID)

Definition

A persistent identifier is a long-lasting ID represented by a string that uniquely identifies a DO and that is intended to be persistently resolved to meaningful state information about the identified DO[4].
Note: We use the term Persistent Resolvable Identifier as a synonym.

1.3 PID Record

Definition

A PID record contains a set of attributes stored with a PID describing DO properties.

1.4 PID Resolver (aka Resolution System)

Definition

A PID resolution system is a globally available infrastructure system that has the capability to resolve a PID into useful, current state information describing the properties of a DO[5].

---

[2] There will always exist data in private, temporary stores, which will not be organized and made accessible in a standard way.
[3] Various repositories include passport-like metadata information along with the PID which goes beyond pure referencing.
[4] This can be information such as checksum, access paths, references to additional information, etc. Some repositories call this administrative or system metadata, but some think a minimal set of attributes to be included has not been well defined yet. In cases where the digital objects do not exist anymore, due to finite lifetime for example, the PID is expected to continue to exist and can still be resolved into useful information.
[5] There are a couple of comparisons such as http://www.clarin.eu/content/comparison-pid-systems

1.5 Metadata

Definition

Metadata contains descriptive, contextual and provenance assertions about the properties of a DO.
Note: Such metadata will make the DO for example discoverable, accessible and usable/interpretable.
Note: To make metadata referable it needs to be associated with a PID and thus is a DO.
Note: Metadata minimally needs to contain the PID of the DO.

1.6 Aggregation

Definition

A digital aggregation is a bundle of digital entities.
Note: The term "aggregation" as a base concept does not add substantially to our understanding of the intuitive idea of collections as resulting from some aggregation process and thus is not used as a separately defined concept.

1.7 Digital Collection

Definition

A digital collection is an aggregation which contains DOs and DEs. The collection is identified by a PID and described by metadata.
Note: A digital collection is a (complex) DO.
Note: A digital collection is an aggregation in so far as there are other types of aggregations.

1.8 Digital Entity

Definition

A digital entity is anything that can be represented by a bitstream.

1.9 Repository

Definition

A digital repository is an infrastructure component that is able to store, manage and curate DOs and return their bitstreams when a request is being issued.

1.10 Bitstream

Definition

A bitstream is a sequence of bits that encodes a specific content, either stored on some media or being transferred under control of protocols.
Note: The term "bit-sequence" is seen as a synonym in the context of DFT.

1.11 State Information

Definition

State information is “metadata” information that describes those current properties of the DO that are relevant for proper management and access.

1.12Property

Definition

A property of a digital object specifies one of its characteristics as digital data.

1.13 Metadata Repository

Definition

A digital metadata repository is a digital repository that is able to store, manage and curate metadata.
Note: A metadata repository is a digital repository.
Note: Metadata can be aggregated by service providers to registries or catalogues.

1.14 Checksum

Definition

A checksum is metadata and an important property of a digital object to allow verifying identity and integrity.

2. Basic Data Organization model

The data organization model which is the basis for the above term definitions has been drawn from the models that have been suggested to overcome the current deficits in dealing with data. It should be noted that this model only applies to the domain of registered data and that there may always be special requirements in the area of very large data sets for example.