RDA Data Foundation and Terminology DFT: Adoption Note

Technical Editors: Gary Berg-Cross, Raphael Ritz, Peter Wittenburg

December 2014

Version 1.1

The documents produced by DFT are:

- DFT 1: Overview
- DFT 2: Analysis & Synthesis
- DFT 3: Term Snapshot
- DFT 4: Use Cases
- DFT 5: Term Tool Description

In this additional document we will describe the specific adoption measures. It is important to note for the interaction in the new phase that a 2-page flyer that has been created jointly and that this has already been disseminated and used in meetings (see Appendix).

1. Type of Output

DFT produced a set of definitions of terms that play a role in data organizations based on a conceptualization that emerged from analyzing a number of models and use cases. This output is not a specification that can be turned into a piece of software by developers or something similar, but it is a tool that can be used for improved communication. As a tool is supports has understanding the task of influencing the minds of data professionals with the intention to help move to better data practices which will reduce heterogeneity and thus improve interoperability, efficiency and cost-effectiveness.

2. Adoption History

The current conceptualization and thus the set of terms that have been defined already emerges from many intensive discussions

- with the communities and initiatives that contributed models and use cases
- within the DFT working group which includes many experts from the interested communities

Therefore we can claim that the interaction in the past 2 years has already had a bi-directional impact. On the one hand the communities and initiatives changed their way of thinking and acting. On the other hand the DFT conceptualization was influenced by these discussions. As an example we can refer to the EUDAT data infrastructure including its core communities, for example in the areas of language resources and technology (CLARIN), climate modeling (ENES) and seismology (EPOS). The EUDAT infrastructure has widely adopted the DFT model and built software and policies that are based on the current DFT model. Necessarily it also had a large impact on the core communities and a few other communities that EUDAT is federating data with. Another example we can refer to is the deep interaction between the Practical Policy group and the DFT group to synchronize

conceptualization and terminology. Due to recent developments some additional interactions need to take place to fully synchronize the latest versions, which is especially true as PP is still wrapping up some of its work.
In addition we can refer to the approximately 120 interviews and interactions which were held with data professionals from many different communities and organizations in Europe. In such interactions the RDA/EUDAT experts were not just listening, but also took the role of provoking and guiding. For many interactions slides were presented which include the basic elements of DFT's core model. This already changed a number of communities in their way of thinking so that we can conclude that most of them are now convinced that PID registration and MD creation is at the core of creating data that can participate in the open domain of data which can easily be shared.

3. Adoption Future

Now that DFT has reached a certain state of maturity which we call snapshot it is anticipated that ongoing discussions and new needs may influence the conceptualization again. That is, we expect that term definitions will need to be changed and extended in the coming phase. Nevertheless, we can and will again interact with as many communities as possible to move minds towards a common view on data organizations. The flyer that has been created is an excellent starting point to address community experts, since we do not expect that data professionals will read the 4-5 key documents which DFT created. They are expecting simple and clear messages that they can turn into action.

Here different strategies will apply to different regions. In Europe RDA has funds to take the following actions:

- interact with leading scientists which will be invited to the coming RDA Europe Science Workshops;
- interact with all EUDAT experts again including the communities EUDAT is federating with by submitting the flyer and addressing the issue at coming meetings;
- interact with all interested ESFRI research infrastructure projects (about 48 in all disciplines) by submitting the flyer, addressing the issue at coming meetings and organizing training courses;
- interact with those communities even more intensively that are willing to work on joint uptake projects;
- interact at policy level with more simple messages about the need to harmonize data organizations (a special flyer addressing policy level experts is in preparation).

In the US the following actions can be taken:

- discuss the synthesis document with relevant groups such as EarthCube and ESIP which have ongoing meetings
- interact with key RDA groups such as the various Metadata some of which will be invited to a Metadata/Semantics Summit Workshop to discuss how to add semantics to metadata;
- use the newly proposed DFT IG to arrange virtual meetings to identify areas of core terminology for use with such groups as Practical Policy
- interact more intensely with those communities, such as DataLink and Deep Carbon, that have expressed interest in adoption and are willing to work on joint uptake projects;
- interact at policy level with allied efforts such as National Data Service to create simple, common messages about the need to harmonize data organizations
-

Addressing just the European and US data professionals will not be enough. Therefore we are already in close interaction with Chinese, Russian and African researchers, but this will require different strategies and joint meetings. Also we need to see how the Australian, South African, Brazilian, Japanese communities can be contacted.

Appendix of the 2-page flyer

Data Foundation and Terminology Working Group

Responsible RDA Working Group Co-Chairs:
Gary Berg-Cross — Research Data Alliance Advisory Council, Washington D.C. USA
Raphael Ritz — Max Planck Institute for Plasma Physics, Germany
Peter Wittenburg — Max Planck Institute for Psycholinguistics, Germany

What is the Problem?

Unlike the domain of computer networks where the TCP/IP and ISO/OSI models serve as a common reference point for everyone, there is no common model for data organisation, which leads to the fragmentation we are currently seeing everywhere in the data domain. Not having a common language between data communities, means that working with data is very inefficient and costly, especially when integrating cross-disciplinary data. As Bob Kahn, one of the Fathers of the Internet, has said, "Before you can harmonise things, you first need to understand what you are talking about."

When talking about data or designing data systems, we speak different languages and follow different organization principles, which in the end, result in enormous inefficiencies and costs. We urgently need to overcome these barriers to reduce costs when federating data.

For the physical layer of data organisations, there is a clear trend towards convergence to simpler interfaces (from file systems to SWIFT-like interfaces[1]). For the virtual layer information, which includes persistent identifiers, metadata of different types including provenance information, rights information, relations between digital objects, etc., there are endless solutions that create enormous hurdles when federating. To give an idea of the scale of the problem, almost every new data project designs yet more new data organisations and management solutions.

We are witnessing increasing awareness of the fact that at a certain level of abstraction, the organisation and management of data is independent of its content. Thus, we need to seriously change the way we are creating and dealing with data to increase efficiency and cost effectiveness.

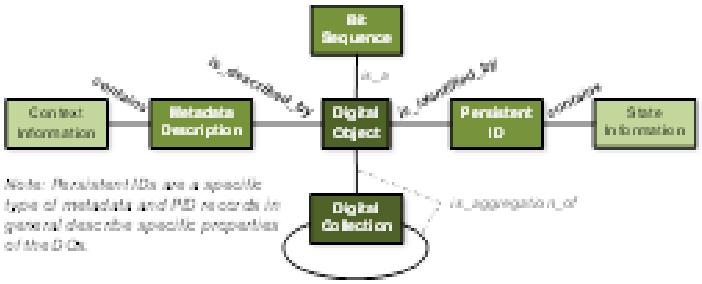
What were the goals?

The goals of this Working Group (WG) were:

- Pushing the discussion in the data community towards an agreed basic core model and some basic principles that will harmonize the data organization solutions.
- Fostering an RDA community culture by agreeing on basic terminology arising from agreed upon reference models.

What is the solution?

Based on 21 data models presented by experts coming from different disciplines and about 120 interviews and interactions with different scientists and scientific departments, the DFT WG has defined



Note: Persistent IDs are a specific type of metadata and PID records in general describe specific properties of the DOs.

This diagram describes the essentials of the basic data model that the DFT group worked out. In a simplified way. Agreeing on some basic principles and terms would already make a lot of difference in data practices.

[1] https://wiki.openstack.org/wiki/Swift

a number of simple definitions for digital data in a registered [1] domain based on an agreed conceptualisation.

These definitions include for example:

- **Digital Object** is a sequence of bits that is identified by a persistent identifier and described by metadata.
- **Persistent Identifier** is a long-lasting string that uniquely identifies a Digital Object and that can be persistently resolved to meaningful state information about the identified digital object (such as checksum, multiple access paths, references to contextual information etc.).
- A **Metadata description** contains contextual and provenance information about a Digital Object that is important to find, access and interpret it.
- A **Digital Collection** is an aggregation of digital objects that is identified by a persistent identifier and described by metadata. A Digital Collection is a (complex) Digital Object.

A number of such basic terms have been defined and put into relation with each other in a way that can be seen as spanning a reference model of the core of the data organisations.

What is the impact?

The following benefits will come from wide adoption of a harmonized terminology which will be expanded stepwise:

- Members of the data community from different disciplines can interact more easily with each other and come to a common understanding more rapidly.
- Developers can design data management and processing software systems enabling much easier exchange and integration of data from their colleagues in particular in a cross-disciplinary setting (full data replication for example could be efficiently done if we can agree on basic organization principles for data).
- It will be easier to specify simple and standard APIs to request useful and relevant information related to a specific Digital Object. Software developers would be motivated to integrate APIs from the beginning and thus facilitate data re-use, which currently is almost impossible without using information that is exchanged between people.
- It will bring us a step closer to automating data processing where we can all rely on self-documenting data manipulation processes and thus on reproducible data science.

When can we use this?

The definitions have been discussed at RDA Plenary 4 meeting (Sept 2014) and will become available as a document and on a semantic wiki to invite comments and usage at January 2015. RDA and the group members will take care of proper maintenance of the definitions. For more information see

https://rd-alliance.org/group/data-foundation-and-terminology-wg.html and

http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

In the next phase of the work, more terms will be defined and interested individuals will have the opportunity to comment via the semantic wiki.

What is RDA?

The Research Data Alliance (RDA) was planned and launched in March 2013 by an international group of collaborating data professionals with a vision of researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society. Members of the RDA voluntarily work together in self-formed Working groups or exploratory interest groups to create deliverables that will directly enable data sharing, exchange, or interoperability. RDA is supported by the European Commission, the United States National Science Foundation and the Australian Government. Information can be found on www.rd-alliance.org

Produced by RDA Europe
rda-outputs@europe.rd-alliance.org

[1] *There will always exist data in private, temporary stores, which will not be made accessible in a standard way.*