Data Management Trends, Principles and Components (PARIS Paper) - Review of Comments so far

Peter, ??

20. August 2015

Until now we received about 20 comments (mostly written) of different nature on the “Paris Paper” and some are still in the pipe. We also expect comments from the 5 tracks being organized at 22.9.2015 - the day before the plenary starts and there is of course an ongoing discussion reflecting the term “Data Fabric” and what it could mean. After all these different types of comments it is not easy to respond, but in this document I will give it a first try without any claim to be comprehensive, since as some colleagues stated: there is no ONE answer to many of the issues. The procedure will be as follows:

1. this and perhaps other responses to be sent to all DFIG members begin of September^
2. asking the group of authors to complement my response until the plenary
3. aggregating additional comments such as being expected from the pre-plenary meetings
4. using the two DFIG sessions at the plenary to further elaborate on comments, interpretations and in particular to ways ahead
5. summarizing the state of discussions in form of a document in the autumn

A first general response is of course that ALL comments[[1]](#footnote-1) which we received so far are of enormous value to move our discussions ahead and sharpen our understanding. Some of the comments are about improving formulations and we will consider all of these when taking a next step. Some others have a more general view on the document as a whole. Since we wanted to raise discussions and not come up with a “scientific paper”, we need to take care to not invest too much time on creating a comprehensive document.

The best approach for responding now seems to be to

* first repeat the intentions of PP (as I see them),
* address
	+ terminology issues (and there were a lot of them),
	+ trends and principles issues,
	+ components issues,
	+ organizational issues and
* discuss some docs that were submitted separately
* last get to some general conclusions.

Let us clarify that the Paris document does not formulate RDA positions. It is written by a bunch of people mostly active as RDA members and others should feel free to also contribute.

# 1. Intentions of Paris Paper

In my view the main goal of the paper was and is to identify “common components” that need to be in place if we want to come to an efficiently operating global “data fabric”. The term “common” in this respect does not mean that every instantiation (concrete realization) of a data fabric[[2]](#footnote-2) needs to include all these common components, but the term should indicate that such components do have a common value that goes beyond a single project.

Another goal of this paper was to understand in how far we all agree on basic trends and principles which do have concrete implications for components. If we for example all agree on the need of “long-term accessibility” then we need to agree upon the existence of trustful repositories with long-term existence. Yet funders often do not take this step. Obviously we will not achieve agreement on specific terminologies which some have in their mind and which may be a result of years of thinking, therefore it is important that we try to overcome terminology aspects as far going as possible (see below).

Checking all comments with respect to these two main goals is probably the best way to move ahead. This implies that some comments may not be explicitly mentioned in this response; nevertheless they were and are important.

# 2. Terminology

Quite a number of comments address eminent terminology issues. How can we deal with them efficiently? In the DFT group we have seen that defining just a few basic terms in a very simple language costs us about 1.5 years. We cannot expect that a) people who are practitioners and want to solve a specific problem first need to learn a specific terminology that has grown over decades in some domains, b) people will adopt just one of these existing terminologies and c) people want to start terminology debates at length. The way out is probably to find a simple “interaction terminology” amongst the data professionals and then see how terminology in RDA will develop. As for every new community we must be willing to create our own “RDA terminology” where we feel it is needed, i.e. we need to dare to be disruptive since whatever “word” we will use it will have already been used/defined somewhere. It is excellent for each group in RDA, if experts are around that can immediately point to specific terms to address implications and to sharpen conceptualization. We have seen this in the DFIG discussions for example where experts suggested using “systems engineering terminology” and where the rest of the crowd was hesitant about the implications of adopting terms such as “architecture” since they introduce a certain unwanted bias. These issues are not new, but should be stated first in this response. In the following I will pick out a few terminology issues only.

**Data Types (Costantino)**

This is a completely overloaded term with lots of different meanings dependent on the context and it has a long history in CS. The Data Type Registry has now been introduced so that the term is living in the world of some data professionals. How to stop that? Experts with convincing messages need to make their points in the group sessions and suggest better terms people then may agree upon.

**Enabling Technologies (Costantino, Donatella, Andrew)**

I agree with the comments in so far as we indeed speak about a set (stacks) of services that must be in place to allow users to carry out certain activities. These services are thus “enablers” and the term “technologies” makes sense to indicate that the concrete implementations vary over time. I found the term used quite convincing. The term “set of services” may be correct from an abstract computer science perspective, but it is not practical enough to transmit clear messages. We had similar discussions about metadata where from CS perspective we all speak about relations at the end, but what does this tell me with my practitioner hat on.

**DM Commons (Costantino)**

Costantino a.o. is questioning whether the terms “external and internal characteristics” are appropriate. Indeed I am not sure myself if such a distinction can be made and is useful at the end. These questions were the reason that we did not include them in the DFT core set. I am still looking for a good description of those properties of digital objects that typical “data managers” do not have to look at and which are meant for “data scientists” that need to interpret the content. A property such as “size” is relevant almost only for managers. If a digital object, however, contains the tag “temperature” it will be important for those re-using the data since they need to figure out how to interpret the corresponding number. Heterogeneity of data is of course much higher when including the “internal” properties such as syntax and semantics, requiring meta-level descriptions to cope with this increased heterogeneity.

There are digital objects such as metadata objects that are relevant for managers and scientists for various known reasons. I would like to see a separation into harmonized packages that clearly identify those parts that are only relevant for managers making it easy across borders to interpret the used categories. This seems to be a task for the metadata groups in RDA and I was happy when the MD groups recently expressed their interest of to tackle the packages aspect.

In this realm Costantino speaks about new database concepts such as array databases. Indeed this is a very interesting development. Yet it needs to find its way into practice.

**PID Issue (Costantino)**

In the DFT and DF groups we analysed practices from many groups and people use PIDs in various ways. There is a clear trend to use PID records to bind various characteristics and computer science needs to accept that. Identifiers are stored together with information (for sure it is a kind of metadata in the general sense, but it fulfils certain functions) that allows us for example to carry out checks, to find related information such as locations where to find the bit-sequences, etc. It can all be read in the DFT documents. This means that we need to deviate in our terminology from standard comp science terms to be able to talk about these needs specific enough. The PIT group called this “PID Information” being of specific types.

**Registered Data and Trusted Repositories (Costantino)**

I know the classification that is mentioned by Costantino and we used it in the discussion when preparing the “Riding the Wave” report. In practical lab scenarios this classification does not help too much, since when data is being created no one knows what kind of status it will have at the end. People are start re-using, re-purposing and combining data when they know where to find it and what it contains. The lab situation must be dynamic where a classification of the sort that is mentioned does not play a role in most instances in the first lifecycle phases. Of course in some communities there are curated reference data sets such as the human reference genome being used to normalize sequence data. These are examples where the classification applies.

**Logical & Physical Levels (Costantino)**

Several colleagues commented on the usage of these terms. Indeed they have been defined by the database community and their definitions are mentioned in the comments. We had the same discussions in the DFT group and since we still had questions about their use we did not include them in the core set. Still we could live with any other useful terms that would distinguish “data” from all kinds of metadata. Keith suggests using the term “metadata” since according to him “everything is metadata”. From an IT point of view this is correct, but then data people have the problem that they cannot be specific enough. What is necessary to separate the “bit-sequence” carrying the content from PID, metadata (here in the narrow sense[[3]](#footnote-3)), rights and relations describing different characteristics for different purposes. If someone has a better suggestion for useful terms it would be excellent.

**Data Model Issue (Costantino)**

The kind of model that is referred to is the DFT model used to derive useful terminology. So it was used for conceptualizing purposes, however, what DFT always claimed is that the model is derived from data practices. Personally I found it important to reach consensus in the DFT group.

**Schema Registry System (Costantino)**

Fully agree with the comments about data service registry and indeed appropriate metadata models are missing as far as I know. Some earlier attempts failed and in some communities a new start has been made.

**Open Data (Justin)**

Open Data is meant in the sense that data that is visible but access is restricted to a certain degree should not be excluded. This is the reason that we need indeed authentication and authorization mechanisms.

**Digital Object and Data Sets (Leonardo)**

Please have a look at the DFT model for the use of the term Digital Object. Instead of speaking about Data Sets we used the term Collection.

**Reference Architecture (Leonardo, Franco)**

Well - we had intensive discussions about this term “architecture” in the Data Fabric group (DFIG) and decided to not use it since it is biased by current practices. So in fact the enterprise of the DFIG group is to describe various possible instantiations where well-defined components will be re-used in different configurations. It remains to be discussed whether some components will be mandatory. Everyone is invited to participate in this discussion. Guess we fully agree that “one size fits all” has its limitations and as you indicate there are different PID solutions for different tasks: some want to point to semantic categories, others to bit sequences contained in digital objects or collections, etc.

**Data Management (Leonardo, Donatella)**

First, we should indicate that the terms mentioned under principles mostly come from other documents. Let’s follow the good traditions of some logicians and agree that it is sometimes very good for human interaction if the terms are defined in a fuzzy way. The term “data management” is used differently and the document gives in appendix some fuzzy definitions of the actors including data managers. In the context of the other principles it is obvious to me what is meant with the term. It is even meant in the sense of including curation/stewardship.

Donatella is right of course when she states that different solutions for DM are partly unavoidable. I guess what the tree diagram wants to express is that we need to reduce heterogeneity, since a) it hampers interoperability and b) software solutions will not be sustainable.

**Term “Component” (Leonardo, Donatella)**

I fully see the point of critique and different “things” are mentioned under in the components chapter. If there is a better term that easily describes what is expressed it would be fine for me.

**Users, Actors, Views**

There are various comments on the view we take and how actors are included. Here short comments:

* Indeed often actors are humans and machines - we should stress this where applicable.
* We don’t take an explicit role when we wrote the paper. We know about Comp Sci methods that offer frameworks (such as RM-ODP) to describe distributed processing systems from various perspectives. Doing this multi-view analysis is a lot of work and colleagues in the European ENVRI project have used this approach. This paper is not the place to discuss this approach in detail, but for sure no one of the authors has the time to do such a study and it needs to be shown by others that more comprehensive statements can be generated.

# 3. Trends and Principles

**Data Management Plans (Thorsten)**

We did not mention DMPs in the document. They could be said to be a “principle or trend” all seem to agree upon. Recently there were a couple of discussions about DMPs which can be summarized as follows: (1) Asking for DMPs creates a lot of awareness raising about data treatment. (2) DMPs at this moment are not really productive for researchers, i.e. paper is filled to satisfy call criteria, but they disappear in a box after the evaluation. (3) Researchers now see them increasingly often as burocratic acts and they just look for usable templates they can copy&paste for their case. (4) DMPs should become productive, i.e. being usable as a tool throughout the project and help researchers in doing things. There is an Interest Group in RDA which wants to work on these issues. Don’t know yet whether the group just wants to add burocracy or look for strategies to make DMPs useful. It is an important topic and RDA Europe will organize a workshop on this topic and invite experts and researchers.

So indeed we should mention DMPs under Trends & Principles.

**Outsourcing Data Management (Thorsten)**

This is an increasingly important issue and I agree with the basic statement. In DFT WG we spoke about “registered data” that has a status relevant to others for re-use[[4]](#footnote-4) and thus should be stored at a trustful repository, having MD and a PID. Data with such a status needs to be managed and partly curated by professionals. In fact the principles that are mentioned in the document will lead to the kind of outsourcing Thorsten is asking for. The use of “components” needed are “open repositories” that fulfil certain criteria, i.e. repositories that can be used by everyone easily. Such repositories need to maintain a suite of tools to help in DM/DC. The maintainability of such tools will depend in how far repository systems (software stacks) can be harmonized, currently most of the software being used will probably not survive since it is made for too narrow.

I guess that CLARIN is on a very good way to have such a system available for language resources. CLARIN experts should bring in their knowledge into the new Repository Software WG. Perhaps we also need a group working on “management and curation plans” that can then be turned into execution with the help of “practical policies” where RDA already has a group.

**Capturing metadata (Justin)**

Guess that we all agree that capturing metadata at an as early stage as possible is what we should do. Often the tools are not (yet) available to let us enter such metadata efficiently. I guess that we should add this to a paper on recommendations as some communities already do.

**Trend Identification (Leonardo, Franco)**

We found that it is not so relevant how “trends” have been identified and we also admit that they are of different nature. The paper wanted to raise discussions and look where we agree and disagree. So there could be many more trends and these should be mentioned. You refer to a trend that “big science” should be differentiated from “long-tail science” for example. If you could provide for the cases you are raising a short motivation, people could start arguing about your statement. To me this difference between “big science” and “long-tail science” is not so clear. Let me give an example: our ethnologists revolutionized the way to find evidence for the evolution of languages. They extract a large number of features from a large set of digital objects worldwide (each of them being a small lexicon or annotation for example), aggregate them in a large feature matrix which is then fed into a phylogenetic algorithm to calculate interpretable dependency trees. What is this? Another example is from the EUDAT experience: the B2SHARE service allows users to upload files - this is a typical mechanism for long-tail data. The B2SAFE service is meant for large data volumes where other procedures need to be applied. Here the difference is made for practical data management reasons but it does not indicate how the uploaded data will be used for solving scientific questions.

As said: the authors of the Paris paper would be happy to get short descriptions of other trends that people see.

**Self-Documenting Issue (Leonardo)**

It is obvious that we need to improve reproducibility and PROV, PIDs that may help in this have already been mentioned. In practice these techniques are not yet widely used, however there is a clear trend that increasingly often experts agree that we should change practices. Self-documenting here means simply to implement default mechanisms in workflows that do exactly what has has been indicated: assign PIDs, create PROV compliant descriptions, etc.

**Description Approaches** (such as RM-ODP)

Given the disputes between people preferring the agile approach and those preferring more systematic approaches such as ODP I cannot speak about a trend. This may change in future. There was an attempt to start an RDA IG on systematic description methods such as ODP, but that was not put into action finally as far as I know.

# 4. Components

**Provenance (Chris)**

I absolutely agree with the comment on the “data fabric” where we continuously repurpose/combine existing data to derive new data and that provenance plays a key role. W3C’s PROV-O should indeed be indeed an important component in our considerations. We need to embed this in our workflows and need to see whether we can do more to help researchers in using this.

**Discoverability & Ontologies (Costantino)**

PIDs as applied by quite a number of repositories allow users to do more than just to pinpoint to the location of data objects and collections (data sets). It is up to the community to decide which granularity should be chosen. We discussed this in the language resource area for a decade and learned to be pragmatic - important is that decisions must be made explicit. PIDs as they are used could be used for discovery, in general metadata is used. The comment about new methods being used to extract information from the content is correct. In general this extracted information is added to the metadata object, but again: there are different practices.

And yes we are all using some types of knowledge sources to be able to interpret metadata and to map them onto other metadata information. The question to me is how we can come to improved agreements in the semantic domain to make semantic work much more efficient and machine actionable. Experience across domains shows that semantic methods and technologies are heavily under-used in daily practices. The reasons are to a large extent well known, yet we are looking for ways to overcome the barriers and RDA could play a role.

**Collection Processing (Tobias&Michael)**

T&M argue for a PID-based data organization approach and that is what many labs start doing to exactly be able to trace later what has been done (reproducibility). This comes close to what Chris mentioned when pointing to PROV. Both using PIDs and PROV is probably the way many are taking and then using PIDs to identify all kinds of digital objects (data, software, knowledge, configurations, etc.) in a way that allows checking identity and integrity. T&M argue that it should be possible to process collections sharing certain attributes to become efficient. In one of the repositories we built we indeed implemented such a system that is based on data types which allowed us to do management based on types such as “open all files with property X to the world”. An open question is still where repositories would store this type of specific metadata information (Mime types, metadata, PID record, etc.) - there is still no agreement. Guess that the Practical Policy group should address these issues.

**Smart Resolvers (Tobias&Michael)**

Fully agree that we can think of many added value functions that PID resolvers could offer or invoke and I see that people start doing it. The problem is that we yet do not have agreements about where and how to store certain metadata information. Developing software now means taking a risk, since it might become outdated very quickly. This even indicated the importance for clarifications in the metadata area.

**PIDs and Dynamic Data (Christian et al)**

This is indeed an interesting issue and the (Dynamic) Data Citation Group worked on this topic and came up with a convincing solution. It seems that there is no reason to not assign a PID to dynamic data, but then one needs to invent other mechanisms such as specifying a date/time to identify exactly where one is referring to. In Handles this could be a fragment identifier that is being resolved by the local database for example. In the repositories I participated in setting up mostly “slowly changing” data was handled by giving specific registered versions a PID, i.e. a number of PIDs were assigned through lifetime of a “digital work”. PID records could then be used to refer to older versions if wanted. Guess that the Data Citation group has come up with a proper solution for this topic. Obviously we miss tolls that allow us to take the necessary steps in daily practice.

I will not further comment on the DOI/Handle issue. Just to say that it is good to have DOIs (Handles with prefix 10 and some rules behind it) for published and quality controlled data and use other Handles for all kinds of data or even tools/things etc. one want to refer to. The worldwide Handle system is reality - it requires some simple form of agreement for communities to make use of the registration and resolution services. For the Handle system increasingly more countries are willing to set up national services in addition to the existing services.

**CORBEL (Christian et al)**

I hope that with establishing CORBEL we (RDA) will be in close touch with the medical folks about components in a bi-directional interaction.

**Prefabricated PP Modules (Andrew)**

The practical policies WG/IG certainly would be an excellent place to discuss this issue. If we all believe that “self-documentation” as we called it in the document (writing metadata in PROV style, registering PIDs, etc) will be the step we all should do, then it should be possible to come up with some generic software code that carries out these default tasks. Registering a PID according to the PIT API is not something that everyone has to re-invent. Adding new metadata information to existing is also something that could be shared to a certain extent. I still believe that we can simplify life of developers everywhere by providing tested software snippets and sharing it via registries.

# 5. Organizational Issues

At this stage I will not comment on the organizational issues that were mentioned in the wiki discussions. They are important but should be addressed at a plenary session.

# 6. Separate Documents

Here we will discuss a few documents that have been sent to us as a reaction to the Paris documents, but that describe domain/project approaches. We will only mention those aspects that either show disagreement to the principles and trends[[5]](#footnote-5) or that add to the components discussion. These descriptions are in addition to all use case descriptions which were submitted to Data Foundation &Terminology WG and Data Fabric IG which can be seen in the wikis.

## 6.1 BBMRI

BBMRI is focusing on improving the work with the different biobanks about samples and data series in various dimensions. To fulfil this purpose BBMRI has drawn an architecture of components that partly need to be built in the coming months/years. The components mentioned in the architecture diagram are arranged according to three levels:

**Applications**: **Metadata Directory** of Biobanks, Biobank networks, samples, data collections, etc.

 **Sample Broker** - an extended search and interaction tool for sample selection

 **Sample Locator** - an advanced tool that can help to overcome the severe privacy

constraints for Biobank sample data

 **Sensitive data processing platform** - obviously a data service environment for the

secure processing of sensitive data

**Clinical Record Extraction** - a framework that helps extracting supplementary

information from sensitive clinical records

**Distributed Secure Operation** - a tested tool suite that can be installed on servers in

clinics to aggregate useful data without hurting data privacy issues

Middleware: **Ontology Translator** - a tool that helps to translate between the semantic spaces of

the various biobanks

**Database interoperability** - set of federating and semantic mapping tools

**Anonymization tools** - set of tools to help in anonymization

**Distributed AAI** - the usual technology to enable distributed and secure access

**Logging and Auditing** - some software to keep track of all access actions

Except for the specific biobank tools (applications 2 to 5) there is much overlap with what other communities are doing. So there must be a gain by coming to common components.

## 6.2 Human Brain Project

(to come)

## 6.3 Perseus Project

(to come)

# 7. Conclusions

Some comments and documents provided have not been analysed yet, i.e. the conclusions may change.

**General**

A general observation is that the knowledge about the activities of the many groups active in RDA is not clear to many people. We obviously need to do more to spread messages in simple terms that are easy to understand by the practitioners. This is a task for the secretariat and helping groups such as the Senior/junior team in RDA Europe. Without these simple to understand messages we will not have sufficient impact.

**Terminology**

The comments made clear that we face a huge terminology challenge in RDA. Different terminologies are suggested for re-use that grew up in the last decades in some communities, but the discussions in the groups indicate that not all of these terminologies are easily accepted and that RDA needs to dare to develop its own terminology. At this moment I cannot see how far this will go, but I see it as essential that the data practitioners find a level to interact without having first to study extensive terminologies and accepting some fuzziness at the beginning. On the other hand it is very useful if experts trained in existing terminologies are assisting to accelerate the process of finding useful definitions. We have an interest group on terminology which could aggregate useful terms as they emerge in the other groups and we should include young experts to help.

***Finding a proper way on this should be discussed in plenary sessions.***

***We need to clarify the terms data management, data curation and data stewardship within RDA so that we better know what is meant.***

**Trends and Principles**

There seems to be wide agreement on the trends and principles that have been briefly described in the document. Yet the implications have not thought through and need to be made clear by RDA as well as the many infrastructure initiatives. A few additional trends were indicated by commenters:

* We should mention Data Management Plans since they are increasingly often requested and are a “trend” despite all questions about them.
* It is indicated there there seems to be a trend to outsource data management (in the broad sense) to professional repositories. This is certainly true for certain types of data in some communities and perhaps it should become a recommendation.
* The principle of capturing metadata (in the broad sense) as early as possible is widely agreed and should be taken serious by all equipment and software builders.
* Doing self-documentation with workflows is important to increase reproducibility and we should point explicitly to PROV.

**Components**

It is obvious that more discussions are requested to identify “common components” and to discuss their characteristics, services and interfaces. There is a growing view that it is good to have the various WGs in RDA that produce results, but that we can only make steps ahead when we broadly test them and based on these test improve RDA work and come to further results. A few additional topics were mentioned:

* A huge area where more clarification is needed is metadata. We have different types of metadata with which we associate different functions (PID info types, rights, relations, etc.) and we have metadata that is being created at different times (contextual, descriptive, system, provenance, etc.). It is urgently needed to have a proper terminology so that we know where we are talking about and for sure we need to identify “packages” that can be re-used easily. Even if the chosen classification will not be optimal, we need soon steps. It is partly terminology, but to a large extent we are looking for usable “components”.
* The other huge area full of questions is “semantics”. How to reduce complexity and enable easy user fine tuning to make “semantic” processing a daily practice? Can RDA do something about it?
1. We should explicitly thank Costantino, Thomas, Chris, Andrew, Peter B, Thorsten, Christian, Wolfgang and Steve, Leonardo, Donatella, Justin, Tobias and Michael, Franco, Petr and Jan-Eric, Larry, Susanna, David and Sean for their input. [↑](#footnote-ref-1)
2. Just to mention again that the term „data fabric“ does not refer to a specifically designed architecture, but wants to denote a “flexible system” of interacting components that offer interoperable services. Perhaps the term “data fabric instantiations” can describe better that perhaps each concrete approach realization or implementation will be a “data fabric” adhering to some general principles. [↑](#footnote-ref-2)
3. Here we are also waiting on an agreed term set from the metadata groups since there are overlapping definitions around for descriptive, system, etc. metadata. We need to sort this out asap. [↑](#footnote-ref-3)
4. Guess that we need to distinguish data being made available to others and data that have been published. Most of the highly interesting data will be re-used by others than the creators before some form of “publication” in a journal or so will occur. [↑](#footnote-ref-4)
5. For reasons of simplicity we assume agreement in those cases where no disagreement could be found. [↑](#footnote-ref-5)