Report from the DFIG Meetings

Bridget, Jianhui, Peter

The Data Fabric IG was involved in 5 sessions at the Denver plenary:

- **Breakout 1: Discussion about Guidelines/Recommendations**
- **Breakout 2: configuration building and Minimal PID Types**
- **Breakout 4: DFIG Core Session**
- **Breakout 5: Joint session with Brokering Group**
- **Breakout 7: Joint meeting with Publishing Data Workflows**
- **We were also invited to participate in a BoF organised by the GEO people.**

In this report we want to highlight major results. For details we refer to the slides which are all uploaded to the DFIG web-site.

Major Results

Turning to practical steps was and is always very important for DFIG, therefore the major actions to be taken are summarised here.

- **Breakout 1: Discussion about Guidelines/Recommendations**
  - approach has been confirmed
  - emails with concrete steps to start interactions will follow asap
- **Breakout 2: configuration building and Minimal PID Types**
  - persons to participate in the analysis to come to minimal PID information types (metadata) have been identified
  - the concrete work will be started asap
- **Breakout 4: DFIG Core Session**
  - 2 new co-chairs
  - action lines and current foci are confirmed
  - work on Global Digital Object Cloud will be intensified
  - Repository Registry case statement to be renewed and campaign needs to be started to find interested parties worldwide
  - link to brokering to be intensified
- **Breakout 5: Joint session with Brokering Group**
  - workshop to be planned with active communities to identify areas of heterogeneity
  - a short-term action towards a small start-up project will be sketched
- **Breakout 7: Joint meeting with Publishing Data Workflows**
  - creation of joint statements to continue bridge building
- **Joint BoF with Geo**
  - planning a joint workshop about the value chain (data -> knowledge), usefulness of current RDA outputs and focus of a possible RDA group

1. DFIG Core Session

Alan, Zhu, Bridget, Jianhui, Peter

Co-Chairs

Two new co-chairs were elected:

- Bridget Almas from Tufts University US
- Li Jianhui from CAS China

The two former co-chairs Alan and Zhu were thanked for their many contributions during the last two years in which DFIG was very active and showed a dynamic development. A new European co-chair will be elected in Barcelona to replace Peter.

Activity Overview and Discussion

An overview was presented about the action items DFIG dealt with and focused on. It was clear that DFIG has done a considerable amount of work over the two-year period, and that the focus has shifted to deal with new emerging issues and on some near-term efforts.

- A White Paper was written to define what the Data Fabric is. An update is needed.
- DFIG gathered many use cases, including cases collected by other groups.
- We placed existing WG/IG work on the data cycle in the labs.
- We then started identifying components that are relevant to improve the efficiency of work.
- As one concrete example „repository registries" were discussed by a group of interested people which led to a WG case statement.
- DFIG then started the discussion about testing which led to concrete activities for funding adoption projects in US and EU.
- Recently DFIG pushed the discussion about „guidelines and clear messages" that could be given to practitioners and funders to overcome the barriers for taking investment decisions. This was discussed in a special session and in EU the large GEDE group was set up to include the many infrastructure builders in Europe.
- To make the step from „talking" to „doing" DFIG stimulated concrete composition work. Here two approaches can be identified:
    - A short term oriented goal to put existing results of RDA groups (DFT, PIT, DTR) and use of Handle to build a concrete configuration solving some problems.
    - A long term oriented vision labelled „Global Digital Object Cloud" with a focus on working with colleagues from a few interested communities on concrete efforts.

Obviously much time of the co-chairs during the last few months went into the last two action points (guidelines and compositions) leaving the other topics slightly out of focus. The discussion about the overview resulted in two major topics that were discussed:

Repository Registry

There are basically two categories of registries of information about repositories. 1) One for human consumption mainly and here we have with re3data a very good service already. 2) Another one for machine consumption including lots of details about the services of a repository allowing large federations to operate smoothly. A WG Case Statement has been submitted to the secretariat[1].

The third category discussed (collection registry) was more of an initiative to make the RDA web-site more attractive for users, but efforts on this track has not yet started.

A big challenge is to get enough use cases from large federations to understand how to optimally describe repository services and characteristics so that those descriptions can be exposed allowing federations and service providers to grasp what they need for their intentions.

Composition Building

Building compositions of components in different contexts and for different purposes will require to talk about „connectors" as well. It will most probably be some brokering layer that will enable this

[1] In the meantime we got reviews about the case statement and where asked to check possible overlap with another WG.

kind of flexible integration needed. DFIG should take this up as well (see joint sessions with the brokering groups below).

2. Guidelines/Recommendations

Rebecca, Peter, Alan

This new action line was presented also in the realm of the new opening towards different types of outputs from RDA. It was stated as motivation that we urgently need joint messages to the communities with the signal to reduce the solution space and thus to make solution maintenance possible, to include software industry and to increase interoperability of course. Here the worldwide adoption for TCP/IP and later HTTP as common grounds can be seen as key examples. Finding such messages can only be done inclusively, i.e. bringing all relevant people together. RDA has shown to be a good place to achieve this level of inclusion.

One aspect is of course how to integrate the many research infrastructure builders. Is the way to set up a parallel European platform under the umbrella of RDA (see GEDE) the way to go? We need to further explore this and be ready to adapt to address potential issues and problems.

The session stimulated an open discussion:

- This is a considerable coordination effort and it will only work when there is proper coordination and leadership.
- Not all of the participants were convinced about the processes and the success. There are some risks included.
- The charettes were seen indeed as the way to go (short periods for aggregating relevant assertions and then identifying the areas of consensus).
- Some believe that more time for research on bundles would be necessary to make sure that all relevant players are on the radar. The DFIG agreed that it will be important to invite all relevant infrastructures in Europe and make sure that a broad spectrum of opinions, solutions and knowledge would be included in the effort.
- There was a question about how other regions would be integrated – there is no clear answer and depends on initiatives. Initiatives such as GEDE should be open to delegates from other serious infrastructure building groups.

The next steps will be as follows:

- A first email will ask for participants to discuss suggestions for „bundles“ and in parallel the GEDE group will also be asked for bundles.
- A second email will ask for people who want to participate in a PID bundle discussion and in parallel the GEDE group will also be asked the same question.
- A time plan will be worked out for both discussion tracks.

3. Configuration Building

Beth, Tobias

PID Profiles: summary of P8 session (Beth Plale)

DFIG held a session at P8 on PID profiles, or otherwise known as minimal metadata associated with PIDs. The motivation for the work is to enable new forms of discovery and management of data objects. Two specific examples were given: provenance and rapid internet-speed filtering/routing of data objects. Data provenance, despite being in existence since 2005, and having standard description languages, remains siloed in systems that create the provenance. Attaching a minimal object oriented provenance record directly to a Handle, opens the opportunity for new tools that remove the silos thus fully realizing the capabilty of data provenance. Rapid decision making on PIDs:

suppose a client tool is handed a list of 100,000,000 PIDs and needs to take action on the items in the list quickly. The only action it is going to be able to do quick enough is consult the PID minimal metadata. How does this capability enable a new ecosystem of tools?

The meeting accomplished two things: 1) we agreed to restrict our focus to Handles in this activity. Other PID types may be follow on work. 2) we formed four small subgroups that will define a profile per group between now and P9. The groups and their membership is given below. Slides from the session are attached.

Data provider - Digital Humanities

1. Bridget Almas, Tufts
2. Ulrich Schwarzmann, GWDG, Germany
3. Beth Plale, Data To Insight Center, Indiana University

Data consumer - Digital Humanities

1. Daan Broeder, MPI
2. Mike Jones, Mendeley
3. Beth Plale, Data To Insight Center, Indiana University

Data provider - natural/physical science

1. Stuart Chalk, Univ North Florida
2. Alex Thompson, iDigBio
3. Yumiang Zhu, Institute for Geographic Sciences and Natural Resources, CAS, China
4. Cyndy Chandler, Woods Hole
5. Stuart Rhea, AgConnections
6. Mario Silva, Institute for Systems and Computer Engineering, Portugal
7. Beth Plale, Data To Insight Center, Indiana University
8. Tobias Weigel, DKRZ, Germany

Data consumer - natural/physical science

1. Stuart Chalk, Univ North Florida
2. Alex Thompson, iDigBio
3. Kei Kurakawa, Nat'l Institute of Informatics, Japan
4. Sharef Youssef, NIST
5. Jim Duncan, Vermont Monitoring Cooperative
6. Stuart Rhea, Ag Connections
7. Beth Plale, Data To Insight Center, Indiana University
8. Tobias Weigel, DKRZ, Germany

4. Joint session with Brokering Group

Stefano, Jay, Peter, Larry
An elaborative note from Jay is added as attachment.

Brokering has two major characteristics: 1) mediating between heterogeneous services, 2) handing over mediating services to 3rd parties. The last aspect is very much related with TRUST and SECURITY. One colleague expressed the close relation between these two characteristics: Reduction of complexity in services leads to an increase in governance (e.g. trust and security) complexity.

In focus of the session was how brokering can play a role in the actual composition building plans/processes of DFIG. In particular, in near demonstrations and in the long term plan for a Global Digital Object Cloud (GDOC). There are actually quite a number of opportunities:

- DFIG talks about different compositions of RDA and other components. It is agreed that where possible we should have common components shared by many if not everyone to reduce overall complexity and not end up in the tower of Babel. But there will be so many more specific components and links with common components that need to be combined through a flexible approach. Putting flexibility into all components is probably a very bad solution since it cannot be maintained. Therefore, smart brokering will be the only solution to go, but it will require trust and security.
- Another area of brokering is to be found between the layers of the GDOC (e.g. to virtualize the present and future information storage systems).

Investing in Common Connectors (brokers) in a domain of common or widely common components seems to be natural. The question was raised whether there are generic patterns for the construction of the set of different brokers (middleware components) that will be needed.

The session ended with the clear recommendation to turn abstract discussions into practices through the implementation of a focused demonstration. The definition of such a demonstration needs to include both information experts and users. It was recommended that a workshop be held in October or Novemver where communities who are highly interested in working on the GDOC (we have already two large communities) will come together with a few experts to discuss where heterogeneity can be expected and how this can be dealt with. There may be another short term action based on small funds to carry out a project. It is hoped that a summary of the workshop can be provided to the next RDA co-chairs meeting for further discussion.

5. Joint meeting with Publishing Data Workflows

Amy, Larry, Peter

The purpose of the session was to build bridges between the two until now fairly independent operating RDA communities Data Fabric (DFIG) and Publication Workflows (PWWG), since an investigation from the covenors clearly show that there is much overlap despite all terminology and cultural differences. This intention was welcomed by the audience. For the different views we refer to the slides.

A number of specific points were discussed:

- A bit of terminology clarification was necessary:
    - DFIG talks about **Digital Objects** using the definition by DFT. They include any kind of digital entities (data, metadata, software, workflows, configurations, etc.) including also metadata objects describing physical entitities.
    - **Data Objects** are thus one type of Digital Object and can be a file, a query into a database etc. For some, software code can also be data of course.
    - **Research Objects** are digital objects that include all kinds of relationships that describe the context of a research work.
    - **Publication Objects** are obviously a specific kind of research object. There was a fair amount of discussion about the definition of a „publication“. Some of the points:
        - Peter Fox suggested the following definition: „a publication object is a contextualized assembly with explicitly stated relationships among research objects.“

        - We discussed the idea of a publication as simply being a special type of collection of objects
        - Amy suggested that a publication:
            - is registered (i.e. has a PID)
            - is distributed
            - is referenceable
            - has longevity
        - The idea of „published“ being simply an attribute on a data object had some traction
        - Larry suggested that „publication“ may no longer be a finely grained enough term (analogous to the term „copy“ which is no longer sufficent to describe an operation)

- Often in lab practices the wish to publish results comes late and introduces additional requirements (long-term accessibility of data, etc.). If this step is not already prepared in the early steps then it costs an enormous effort to fulfill the criteria. On the other hand the lab processes have their own rules which can not be overloaded with additional bureaucratic requirements which cannot be fulfilled at too early a stage.
- The notion of „publishing data“ needed to be spelled out:
    - Amy presented a view about the criteria to be met to talk about a (data) publication.
    - Many labs have already started to associate PIDs to Digital Objects and upload it to a trustworthy repository at an early stage to allow proper referencing. Is this already a „publication“, since data is accessible and referenceable from that moment on? Some labs already use this as a kind of data publication in particular when metadata is also being associated with the DO.
    - Some demand some quality control in relation with publications. In this realm it was suggested that curators should do part of the work (metadata creation) instead of the reearcher.
    - It was noted that scientists probably most often „publish“ collections and not the large amount of individual objects they want to refer to in stable ways.
- „Publishing releases“ led to an intensive discussion with following points made:
    - A common statement seems to be „don’t publish releases“
    - In many sciences, however, it is already good practice to make early versions accessible to allow others to contribute. Often Digital Objects are never finished, such as a lexicon in linguistics.
- It seems to be widely agreed that Digital Objects independent whether they are being used for referencing or citing should be uploaded to a trustworthy repository which also takes care of assigning PIDs and organising the data and metadata.

The result of this meeting is that the chairs agreed to create a set of joint statements which will be discussed in the two communities with the hope of wide agreement across the communities. It may be necessary to organise a virtual meeting where all session participants are invited.

6. GEO BoF

Ari, Peter, Jay, Stefano

(Moving from Observations to information and knowledge)

This BoF was obviously meant

- to discuss strategies within GEO to establish the value chain towards knowledge
- to understand what RDA and in particular DFIG can contribute and
- to see whether it makes sense to work towards an RDA group to work on the chain.

Ari provided an introduction and overview of the session objectives including the points made above. Barb Ryan, Director of the GEO Secretariat, presented an overview of GEO and its initiatives. DFIG (Peter) presented its data cycle and, to make it concrete enough to understand it, a use case from the linguistics field was presented that also stressed the cross-disciplinary interest in such a value chain. It was discussed which RDA outputs and DFIG activities could play a role and it is also obvious that for some outputs there needs to be a mediator to bring over the essential messages. The long-term view of a Global Digital Object Cloud seems to be a very interesting vision to continue interaction.

The session participants expressed interest in an interest group for coordinating the value chain developments in the RDA, but it seemed that the idea still needs a lot of work to refine details

It was agreed to organise a joint workshop with some experts to go into more detail and understand the potential on the one hand and on the other hand understand the detailed needs within the GEO community and benefits of using the RDA process. In particular, there were recommendations that examination of the value chain in the context of data sharing impacts at various places along the chain would provide insight into ways in which research data sharing and the support Data Fabric can be improved. The initiative for a workshop should come from GEO, but RDA Europe certainly has the potential to contribute substantially. The potential of a co-location with RDA or GEO meetings was discussed. For this reason, a joint virtual meeting on the GEO/RDA collaboration and use of RDA to discuss value chain was suggested. As soon as the participant list is received, this will be organized.

Joint Meeting of RDA IG Brokering, IG Data Fabric

Session 6 Sept 16, 2016

Organized by Jay Pearlman, Stefano Nativi, Peter Wittenburg and Larry Lannom

A list of participants is provided in the appendix of these minutes.

The session was introduced by Jay Pearlman. He noted that there is synergy between data fabric and mediation approaches in complex environments. Even with the objectives of the data fabric to minimize complexity, some will remain. Thus a discussion of opportunities for collaboration is timely as we move to demonstrations.

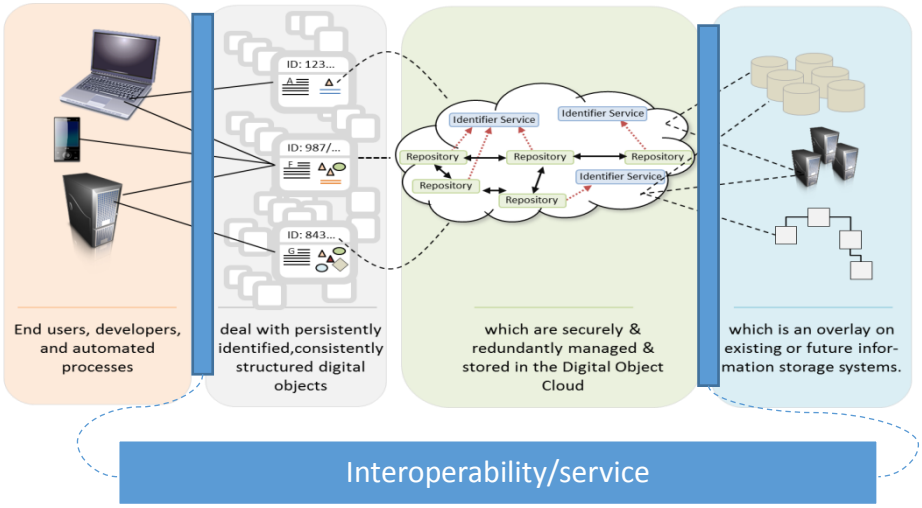
Stefano Nativi

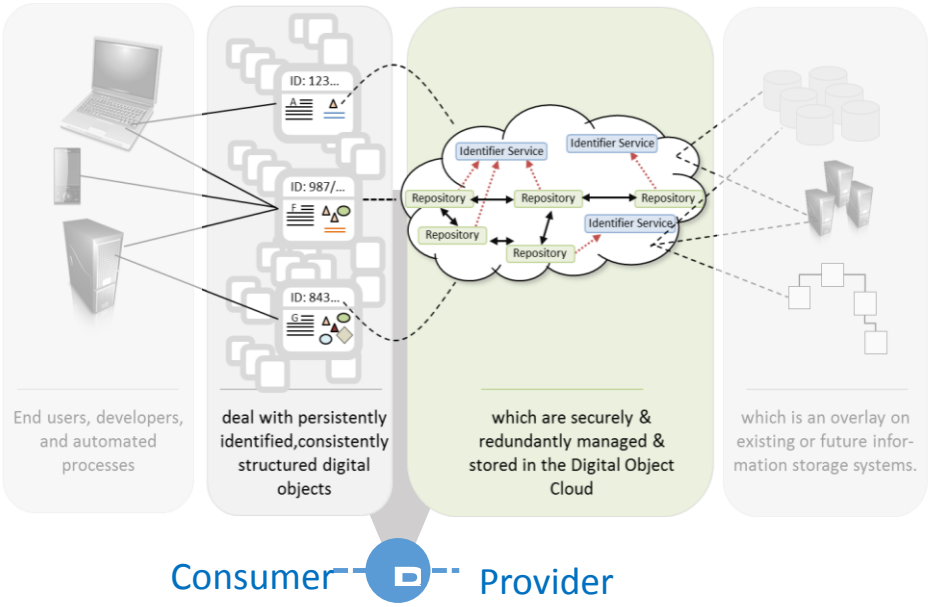
Stefano provided a brief summery of brokering history. He noted attributes of a broker framework include the broker as a third party service which is flexible and extensible; also he commented that the broker can operate in a PID centric data management environment.

The Broker, which works in a heterogeneous environment containing multiple remote objects that interact synchronously or asynchronously, typically demonstrates the ability to:

1. Finalize requests on behalf of its clients against a vast supply system –e.g. by transforming different interoperability protocols;
2. Support many clients at the same time in a dynamic way;
3. Access large, distributed, and heterogeneous supply systems in a dynamic way;
4. Be fully autonomous from its clients and accessed supply systems;
5. Be flexible, configurable (even at run-time) and extensible.

Stefano identified several areas of the PID Central Data Management and Access (CDMA) where brokering can provide support including interoperability service interfaces (digital object cloud in the middle) and connections between the PID CDMA system and users or suppliers (see figures below). For example, the broker could decouple the relationship between the digital object cloud (in the middle), users and suppliers. It could also interconnect component or common components in a data fabric composition, thereby allowing the mapping/mediation of different digital objects.





Peter Wittenburg

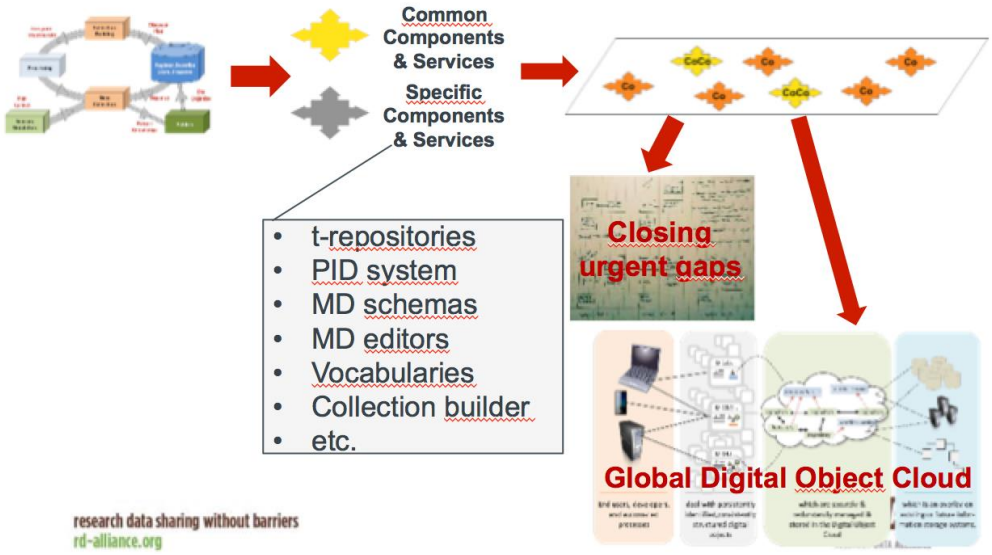
Peter started the discussion saying that the broker is about 2 capabilities: the broker can provide connectivity where there is an mismatch between component interfaces (improper fitting); second, for certain operations such a single user sign-on, the broker could provide a “trusted” third party environment.

Peter introduced the Data Fabric cycle.

This data fabric cycle is inefficient; the steps in implementation are to identify components that could improve efficiency (stepwise), specify them and then motivate people to build and test them. There will be common and specific components. Questions then arise on the fabric composition.

For example, do we create one standard that everyone will accept or do we use brokering? How do we address Composition – put together components into compositions; different types of compositions; how flexible do we need to be? He continued that we are probably talking about common components that are essential to reduce complexity and asked if they can be served by a third party? Peter went through a description of the history of the Handle system and its current operations by DONA.

We are entering the period where we need to move from abstract fabrics to solutions. Peter showed a data fabric flow:



This chart was the basis of further discussions in the session.

The following conclusions were offered:

- Component compositions probably require flexible solutions
- Common components will be required where possible to reduce complexity – built in convergence trends
- Where to build in flexibility – separate layer?
- Many examples for 3rd party services
- Trust is the major issue
- Abstract discussions don‘t move us ahead. How to move on in concrete building/testing projects

Discussion with Participants:

Reagan Moore offered, as a general comment, that broker and data fabric are both middleware. They serve with different characteristics – fabrics maintain names and assignments; brokers manage remote operations.

Mario suggested that we need a detailed discussion. Bridget asked if the development of a more detailed concept should be done in an RDA working group.

Peter and Stefano discussed various options with an emphasis on a near term demonstration. This could be done with one or two use cases that would allow examination of interfaces and identification of gaps. Use cases such as climate observations/information and the work in natural history museum catalogs were mentioned as examples. The trade between flexibility, efficiency and complexity was again discussed.

Recommendations

Organize a small workshop of experts this fall (before December) to look at the options for a near and longer term data fabric implementation including brokering which addresses key questions associated with the digital objects cloud, composition implementations and between users, the fabric and the repositories. Look at the work that can be done with 2 communities e.g. the climate community (working with Tobias Weigel of the German Climate Computing Center) and the natural history museum community (working with Dimitris Koureas of the Natural History Museum, London, UK). The small workshop (no more than a dozen participants) would be a working meeting of one or two days with representatives from the data fabric IG, brokering IG, climate community and Museum community. Results of the workshop would be reported in the RDA co-chair meeting in December 2016 and could provide the basis for further discussions at the next plenary and a demonstration. A working group proposal to RDA could be considered.

A proposal to RDA Europe may be used for funding of the initial workshop.