RDA Data Fabric IG Proposal

# Data Fabric (DF) IG

Several RDA groups are working on core components supporting a basis for a Data Infrastructure for reproducible science. These include: PID IG, PIT WG, PP WG, MD WG and IG, Provenance IG, Brokering IG, DFT WG and others. Some of these groups are Working Groups and will finish in 2014. Yet we miss an overarching concept & discussion framework for most of these groups to relate the components with the overall data lifecycle landscape and identify gaps, obstacles and possible incompatibilities. The loose concept of the Data Fabric is a more nuanced view of data, data management and data relations along with the supporting set of software and hardware infrastructure components that are used to manage this data, along with associated information, and knowledge. This vision offers the possibility to formulate more coherent visions for integrating some RDA work and results. The Data Fabric discussions will also contribute to discussions in TAB where aspects of the Data Fabric can be utilized & integrated into the broader context of RDA.

# Background

Excellent, leading research in many areas such as human brain analysis, health care, climate change, etc. are based on smart computations on large aggregated data collections. These collections exhibit a high complexity/heterogeneity in a number of dimensions such as spatial and time resolution, with the type of multidisciplinary data extending from linear time series to array data with that data’s knowledge and information context expressed as relations between descriptive metadata which may be formalized as ontologies. Barrier breaking results that give new insights of how to deal with the big societal challenges need to rely on the availability of more efficient and cost-effective ways to make use of and combine all relevant data and software services which have been created. It becomes increasingly obvious that we need to turn to automatic workflows adhering to some criteria when we want to cope with these volumes and the complexity such that reproducible and traceable science is guaranteed. These workflows are guided by what we call “practical policies”, i.e. self-documenting procedural statements at an abstract level that are specified by data managers, data scientists and other persons involved.

*This diagram indicates possible actions within a Data Fabric. Primary data, stored in repositories is in a continuous process of being enriched and analyzed creating new derived data. Some of this data (testing) will be temporary, but most of the data will be part of some workflows and thus should be referable. In some cases when the results have been used and quality checked for a publication they may become citable. Also publications are part of this Data Fabric since they are often used for data mining and other analysis.*



# Characteristics of a Data Fabric

We characterize the Data Fabric concept by the following observations:

* The Data Fabric covers a domain of registered software components (workflows, services) that are in fact a special class of digital objects (DOs).
* Data in the fabric includes a means of registering DOs using an authorized registration site
* DOs are stored and managed in persistent and accessible repositories.
* DOs have metadata describing their creation, context and history (provenance).
* DOs are registered with a PID at authorized registration sites.
* Actions on DOs may be guided by abstract policies that are explicit and thus auditable.

There can be multiple data fabric implementations and to advance work these should be highly interoperable. These data fabrics are a critical component of infrastructures paving the way to reproducible science. In a data fabric we see how the separate components, developed separately, can be made to work together, this means that for different sets of components the data fabric will be different. We note, strongly, that it is meant as a descriptive way to deal with the interrelation between many components, rather than prescriptive (like you would have with an architecture). We note also that a data fabric is in a constant state of evolution and adaptation. The technologies in use for any component today, will be replaced with a cheaper and higher performing technology in the future.

The suggested Data Fabric IG is therefore planned as a forum to discuss:

* Alternate views, components and aspects of the DF concept.
* How the outputs from the RDA working groups fit in the DF concept and how they relate to each other and to various related WGs and IGs within the RDA.
* Which further activities are required to advance the data fabric concept.
* Continuation and initialization of working group activities related to the DF.
* Improving the uptake of various WG outputs by connecting and relating them as a coherent whole within the DF concept.

The goal of the discussions is to:

* Write a concept paper that defines a vision of data fabrics, including within RDA IG &WGs, to drive the discussions and work program.
* Identify the need for new WGs that remove concrete barriers on the way towards a functioning DF.
* Define and initiate these WGs.
* Promote DF WG results & demonstrations as they are produced, integrate the work of the WGs into an overall DF concept and maintain them in order to evaluate their effectiveness.
* Promote the adoption of DF WGs outputs within other RDA efforts

We believe that this will provide a useful and productive addition to the discussion in RDA.

# Intentions

* The first action of this IG will be to agree on a description of what is meant by the concept of DF. We will propose to organize a BoF session at P4 in Amsterdam. Various people/groups are asked to work out short concept notes beforehand and exchange them. The group of initiators will work on these notes to structure the presentation part and the discussion.
* At the P4 meeting the chairs of this IG will be identified. Currently Peter, Gary, Reagan and Keith offer to push the discussion and chair the interactions.
* The second action is to formulate a White Paper based on the notes and the discussion results and offer it for open discussions in the RDA forums. This White Paper should be ready in December 2014.
* The IG will give the current WG chairs collaborations a formal RDA basis as an IG which is being treated the same way as all other IG. The DF topics as described above will be an issue of the regular IG meetings.

# Initiating Members

The idea for this IG emerged from the discussions amongst the chairs of the WGs who are the initiating members. Some more experts who showed interest joined this group.

Rebecca Koskela, Keith Jefferey, Jane Greenberg, Reagan Moore, Rainer Stotzka, Tim Delauro, Tobias Weigel, Raphael Ritz, Gary Berg-Cross, Peter Wittenburg, Daan Broeder, Larry Lannom, Juan Bicarregui, Herman Stehouwer