Data Fabric Interest Group

DFIG

White Paper

Draft Version 0.5[[1]](#footnote-1)

***The task of DFIG is to design a flexible and dynamic framework of essential components and services, identifying those that enable efficient, cost-effective and reproducible data science and making these known and available to researchers and data scientists. The goal is to make it possible for scientific users to easily integrate their scientific algorithms into such a data fabric without needing to master the underlying details. Designing such a framework requires characterizing the landscape first. This builds onto and extends the efforts of other working groups in the RDA and of lighthouse projects with well-designed data organizations worldwide.***

**Main Editors**: Rob Pennington, Peter Wittenburg

**Editing Team**: Gary Berg-Cross, Keith Jeffery, Giridhar Manepalli, Rob Pennington, Beth Plale, Raphael Ritz, John Henry Scott, Yunqiang ZHU, Peter Wittenburg

# 1. Introduction

Five early working groups[[2]](#footnote-2) started up at the first plenary or just after it to address topics that the experts felt to be urgent barriers to be removed. The key people formulated the goals for the working groups they proposed without considering a wider picture at their start, since everything in RDA was new and the need to deliver concrete results within 18 months required hard work and short-term focus. However, during the work a close interaction between these groups was established and turned out to be crucial for progress and synchronization. The wish to continue this interaction as an Interest Group under the umbrella of the term "Data Fabric (DF)" [1] was formulated with three major tasks:

* put the work of these WGs and other relevant RDA groups into a broader picture
* use DFIG as a forum for maintaining the results of the terminating WGs
* indicate new, relevant work items with high priority that could be taken up by new or existing WGs working on overcoming concrete barriers

Three further sources of motivation can be identified:

* A large survey based on about 120 interviews and interactions [2] made it obvious that the way data is being created, managed and processed in the various institutes is far from being suitable for reproducible data science and is inefficient. Existing practices, mainly characterized by manual interventions and ad-hoc scripts, urgently need to be moved in the direction of automatic and self-documenting procedures (practical policies) that are suitable to meet the challenges of the data deluge. Many people agree with this need to change, but hesitate to invest in such effort since they lack directions, guidance and best practices.
* A DF session at Plenary 4 with engaged discussions clearly indicated that there are many experts that essentially agree with the label "data fabric" as a metaphor to describe the ecological landscape of components and services that are required to make our data practices more efficient and cost-effective.
* In a number of countries large scale national and international efforts are in the process of implementing data services in specific domains or across domains to simplify the tasks associated with data at all stages of the research process and improve scientific productivity and reproducibility. These projects need to be expanded and generalized from specific domains into the larger research arena in part by engaging with RDA in order to further the advancement of data interoperability.

The amount of data and its complexity, the number of tools and services, and the range and diversity of uses of the data by increasingly large groups of researchers will continue to grow very quickly in the coming years. This will become an ever larger problem, as some surveys [3] already indicate that only a small fraction of results emerging from analyzing data sets can be reproduced, thereby reducing the level of trust dramatically. This reduces the value of the data and unnecessarily increases the cost and effort required to conduct research and development. Thus we need to accept that we have a huge problem of scale that urgently requires new ways of thinking about and working with the data. The Data Fabric Interest Group (DFIG) has exactly this gap as its basic motivation: defining the ingredients that are necessary to professionally deal with large data sets based on well accepted concepts and mechanisms for the benefit of researchers and data professionals at all levels. Policy based automatic procedures that adhere to basic data organization principles are being worked out in RDA and making use of various components offering services is the most practical and feasible way forward.

DFIG has an important role in closing the gap associated with scale:

* consider and illustrate the possible directions,
* describe important common components and their services,
* describe principles of component and service interaction and
* indicate or otherwise identify best practices.

User communities have much to gain if and when DFIG is able to meet these goals: the impact would be huge and some have even argued that it can only be compared with the Internet revolution by defining the ingredients needed to allow a seamless connection of nodes worldwide. However, we must not forget that there is a lot of data and in particular software out there which researchers are using daily. It will be necessary to make this reality part of the articulated landscape.

We use the Data Life Cycle diagram (diagram 1) from the DataOne project[[3]](#footnote-3), to illustrate the way in which researchers use their data to produce their scientific results. Each of the steps shown in the diagram builds on the previous steps and should also have the capacity to integrate data and information from other sources as well as cross link to other steps. This view is necessarily a different perspective for working with data from that of the data scientist who is developing the tools and approaches that support this type of lifecycle work. The Data Fabric needs to describe the required set of components and their services that can be plugged into a flexible framework[[4]](#footnote-4) implementing such a lifecycle.

*Diagram 1: Data Life Cycle following tasks*

A lifecycle view allows us to address the question of how DFIG is anchored in the overall picture of RDA. To discuss this we like to use a slightly extended, nevertheless schematic and simple diagram which has been created recently to explain the role of persistent identifiers in the stages of data processing.

Diagram 2 schematically indicates the phases of data processing in scientific labs where new data is being created in cycles from combining existing data which can come from distributed repositories (relationships to the diagram above are indicated[[5]](#footnote-5)). Raw data is generated by sensors, simulations or human interaction. Such data will be registered (Persistent Identifier and Metadata assignment) and stored permanently to give access. Using discovery steps data collections will be selected, converted, contextualised for quality and relevance (the integrate step) then there follow steps for processing (management, analysis, derivation, etc.) with new collections being created[[6]](#footnote-6). In general these integrated collections should be registered again and stored permanently[[7]](#footnote-7).

Some data will be used to create scientific papers as a result of an intellectual step. Such papers will be stored in typical publication repositories which can also be subject to computations. It is important that e-Publications refer to the data collections that have been used to create the results. In RDA we already have quite a number of WGs and IGs that are working on optimizing the publication workflows and components, for example, focusing on the end-product of research work. Data Fabric IG looks at the data creation and consumption machinery as indicated by the diagram and wants to optimize methods. Of course a large number of WGs and IGs are dealing with various aspects of this Data Fabric, and success of the IG will require good communications with those groups (more in section 5).



*Diagram 2: Data creation and consumption machinery*

This diagram includes the possibility of course to define new collections from all stored data and re-purpose such collections in arbitrary ways. Collection building is often a complex, nonlinear step where methods of discovery are used and different types of digital entities (like research data, descriptions, pictures) are set into relation.

In order to bring these two complementary views of data (indicated by the two diagrams above) together, the different processes can be categorized in a number of ways. Currently we very often see that for a specific data collections a separate management and access solution is being developed by researchers who are often unaware of other approaches and mechanisms. Processes can have typical management tasks and also include all kinds of scientific operations. This results in unique but roughly parallel connections between the different types of data that are generated and the analytic tools and environments that they are used in. With increasing experience from working with different disciplines, it is becoming increasingly clear that commonalities exist that are valid for data objects (collections) across disciplines and discipline-specific processes. Diagram 3 indicates that this can be highly efficient when those processes which are common to all kinds of data objects can be bundled together into a well-developed trunk to connect the roots of science in the data generators to the growing canopy of highly sophisticated data discipline specific data analytics. It is one of the tasks from RDA to reduce the diameter of the trunk.

*Diagram 3: need for decreasing heterogeneity for data management solutions.*

DFIG needs to deal mainly with the infrastructure necessary to efficiently integrate the common processes that are useful and so be able to easily plug-in the scientific type of processes without requesting from the scientists that they understand how the overall machinery is working in detail. The researchers need to have a simple framework that allows integration of scientific service components without distorting the basic principles such as automatically documenting what is being done in the workflows (MD, provenance, etc.).

DFIG thus will provide input for the exploration and structuring discussion of the RDA landscape and will comment on activities relevant for DF, but it:

* does not have a formal role in the RDA processes, in contrast to the TAB
* recognizes that it only deals with certain aspects relevant for RDA, i.e. there are other aspects in RDA beyond the DFIG scope like the data publishing specifics, the legal and ethical questions and much more
* cannot have a claim to be the only group that deals with aspects of the scientific data process and
* is intended to provide a forum that embraces harmonization but does not tell others what is wrong and right.

DFIG concluded in its P4 session that we currently do not have a suitable language to describe what we are talking about and it was suggested that the group make use of the terminology that has been defined by systems engineering. In the interactions it was pointed out that this terminology could be a good starting point [11], but needs to be explained and made sufficiently clear so that all DF interactions reflect the same understanding. Chapter 2 elaborates the scope and nature of the DF we have in mind, chapter 3 elaborates the components and their services that may play an important role for shaping a DF compliant infrastructure. Chapter 4 summarizes related work. Chapter 5 explains why we need use cases and demos to move from purely theoretical debates to those grounded in concrete practices. Chapters 6 and 7 summarize the tasks of DFIG and the plan for the first months. The information on terminology is referenced in the appendix and is available through the Term Tool[[8]](#footnote-8) developed as part of DFT WG activity.

# 2. Scope and Nature of Data Fabric

## 2.1 Common View

As has been indicated already, the task of DFIG is (1) to characterize the landscape of essential components and services that enable efficient, cost-effective and reproducible data science, (2) to identify the needed common components by the services they are offering and their capabilities and limitations, (3) to define the principles of their interaction (interfaces, protocols) and (4) to indicate how scientific users who are not interested in the details of the underlying data fabric can easily integrate their scientific algorithms.

The framework we are aiming at is a new platform for data science that is easy enough to use by the scientific users yet flexible enough to cope with different interests and new emerging technologies. In the sense of the terms discussed in in the Term Tool it is a framework that enables different configurations as well as adhering to basic agreed upon principles.

Diagram 4 indicates the nature of the flexible framework which is our goal. On the one hand a set of common core components will be offered with clear specifications of services and interfaces which will be extended over time and which will be subject to continuous renewal based on technological advancement and improved insight. Some of these core services may be expected to be mandatory services needed to keep the resulting system functioning, others may be optional. On the other hand there will be an unlimited number of component/services offered by researchers that will continuously change in all dimensions due to scientific innovation. However, in an increasing number of cases the techniques and services may – with benefit – be used by researchers from a different domain.

What will make it all work together is the interaction layer determined by interface and protocol specifications. In addition it must be specified how a researcher can register his digital objects that can contain data and code to make them part of the flexible and configurable framework.



*Diagram 4 indicates a Data Fabric as a set of selected components and their services interacting to fulfill defined tasks.*

A core part of the framework will be how it consumes existing data objects and creates new data objects as part of collections as indicated in diagram 2. Diagram 5 illustrates possible processing at the atomic level. According to the DFT snapshot model a processing unit that wants to carry out some operation on a specific digital object will make use of the existing metadata description, the existing PID record[[9]](#footnote-9) and of course the bit stream encoding some meaningful content (with references amongst them as illustrated). The processing unit[[10]](#footnote-10) will create a new digital object that is characterized by a new metadata object including some provenance information about the operation and a new PID record containing new state information all containing properties of the new digital object. 

*Diagram 5 indicates possible atomic modules consuming and creating specific information as part of the data machinery.*

Diagram 6 [9] is complementary to the diagrams above in so far as it indicates the steps needed to make use of a digital object:

* A user has a collection to be made available for use. This requires registration of a PID and associated state information (such as access path), a metadata description, as well as the schema, and defined key terms being used so that it can be accessed and understood.
* These registration steps will then allow others to (1) to find the collection by using and interpreting metadata, (2) to access it and check its identity and integrity by resolving the PID into state information and (3) to make use of it by interpreting the metadata, perhaps consulting a Data Type Registry and perhaps contextual descriptions.
* To make this work a number of openly accessible registries (components) offering appropriate services need to be available to allow seamless processing.

*Diagram 6 indicates the steps necessary to make use of a data object stored in some repository.*

* In this canonical access process the following components are essential: (1) metadata registries, registries containing their schemas and concepts so as to be able to search for objects and to interpret them; (2) registries containing PID and associated state information in order for example to locate instances of the bitstreams or to check identity/integrity; (3) registries for distributed authentication and authorization in order to grant access and (4) trusted and persistent repositories so that requests can be made of the bitstream to carry out the intended operations.

The above explanations are just one example and the list of components and services are only indicative. It is the task of DFIG to study such typical stepwise processes and to analyse well thought out use cases to better understand the requirements and possibilities and thus to stepwise improve the description of the flexible framework DFIG has in mind.

We can summarize that DFIG is about defining

* basic and flexible machinery that helps make data science reproducible, i.e. a machinery in which all steps are guided by actionable policies, and in which all steps adhere to basic organizational principles and all steps are self-documenting (i.e. provide some provenance metadata about the activities);
* the common components, their services and interfaces;
* a number of facilitating components, their services and interfaces (such as DTR).
* general principles for the interaction between services (interface, protocol);
* the methods to allow users to integrate their digital objects into the machinery;
* the methods to allow users[[11]](#footnote-11) to select data and scientific services they want to combine to create new data and results.

It is agreed that whatever DFIG does it must be

* looking at flexible open frameworks;
* technology independent as far as possible to prevent lock-ins;
* dedicated to the creation of a robust framework designed to avoid single point failures or overall system failures
* open to replace “old” components by new ones or to allow for competitive services as long as they adhere to defined agreements;
* focusing on cross-disciplinary common pillars of an infrastructure everyone can use;
* grounded on use case examples and demos as far as possible.

In starting to define the scope of DFIG’s work, we recognise that

* in the domain of data publication - the end of data lifecycle - important and knowledgeable groups have been formed and are working towards efficient solutions;
* that there are many “meta”-topics that are relevant for a distributed data fabric framework, such as legal & ethical aspects, certification methods, user engagement, etc.

DFIG will need to interact with these groups to ensure maximal compliance and as a matter of principle in RDA everyone can join the discussions.

## 2.2 Extended Views

Beyond these common views a couple of suggestions were made in the DFIG discussions that need further elaborations. They may be seen as further points enriching the discussion on DF:

* In addition to provenance metadata, “log-files” that result from self-documentation should be sent to provenance consolidators periodically;
* An RDA-agreed upon list of service types would be useful;
* An interface to a publish & subscribe system which serves as the DF control mechanism would be useful;
* A DF is a set of software and hardware infrastructure components that are used to manage data, information and knowledge[[12]](#footnote-12);
* When an enterprise implements a data management solution, one of multiple types of DF infrastructures is typically chosen to enable the processes[[13]](#footnote-13);
* An “Open and Trusted Forum” for all kinds of services (data, analytics, etc.) should be established across disciplines allowing researchers to easily find useful components/services without having to test out all kinds of offered possibilities. It is not obvious how to organize such a forum and it must be based on what has already been developed by communities, but it must work efficiently. It can have a strong impact when it can be combined with workflow frameworks. An analogy for this might be an apps store type of environment or approach.

More aspects may come up that may inspire DF discussions.

# 3. Components and Services

As indicated above we are still in a phase where we need to explore the landscape that Data Fabric will cover. It seems that one aspect indicated by diagram 3 is widely agreed: (1) There will be an increasing number of ways sensors and algorithms are built that create scientific data and there are an increasing number of variants of analytical and visualization software services all designed to discover specific aspects of natural and societal phenomena, i.e. these will be widely discipline-specific specific (although services in one research domain may with benefit be used in another). (2) All these sensors and software systems create data objects and collections that are subject to management, stewardship and access tasks which are widely discipline-independent .

Therefore we assume that there are common and discipline-specific data services and that the number of common data services will be limited. We still need to sort out whether there are mandatory common services needed to create the open and accessible data domain which we are all looking for to make data more visible, accessible, reusable and interoperable.

Thus, the key question is whether in the data domain we also need such common key services to deal with data efficiently. A deeper understanding will be essential to get a more complete landscape of components/services and as in other RDA groups DFIG should also follow a bottom-up strategy and look at interesting and far-ranging use cases and do test demos.

Independent of answers to this question a number of common components/services have been mentioned already:

* services to register digital objects (DO) incl. data, code and knowledge requiring a worldwide system to register persistent identifiers and resolve them to useful state information such as access paths, fingerprint information, etc.;
* services to register actors (ORCID [10], etc.) allowing the unique identification of people[[14]](#footnote-14);
* registry service of trusted repositories for registered DOs guaranteeing long-term access and proper stewardship;
* services offering assessed policy principles (proper organizing mechanisms, self-documenting, certified, etc.);
* services of trusted registries that are meant for humans and in particular for machines (types & concepts, metadata and provenance schemas, metadata instances, repositories, PIDs, policies, terminology etc.);
* services that include semantics but how to do this remains a challenge. There are repositories for ontologies, for example, that may contribute to the knowledge aspects of the infrastructure as well as many semantic brokers to do such things as bridge terminology differences and convert data for valid integration. There is agreement that this is an important but challenging area! One semantic interoperability BoF was held at P3 to discuss this topic.

Components and their services need to optimally support

* data management – for building data repositories, manage information/MD catalogues, and enforce management policies;
* data analysis – to enable (workflow) analysis of data collections and the automating of processing pipelines for packaged results;
* data preservation –to support building and maintaining reference collections of citable data;
* data publication and linkage as part of citation;
* data discovery and access to data collections;
* data sharing – including controlled sharing of data collections, shared analysis workflows, and MD/information catalogues.

It is also clear that the integration of ideas, components and services from a number of working groups will intersect in the domain of policy definition and implementation. A concept policy graph that flows from the purpose through the actual operation will have to be jointly worked out as part of the DFIG effort.

# 4. Related Work

The Data Fabric IG needs to view its challenges in the context of other work that has been done already. We distinguish between work being done outside and inside RDA.

## 4.1 Work outside of RDA (more should be added here)

* DFIG will make use of SE terminology [11] as far as possible and elaborate on it to make sure that we all understand the terms in the same way. But for each term this needs to be checked for usefulness to RDA groups. We should use the Term-Tool created by DFT for this discussion. The above elaboration can only be a start;
* It is claimed that within the Grid and Cloud community documents are available that may foster our discussions[[15]](#footnote-15);
* Tony Hey’s book about the 4th Paradigm [12] might also be a good inspiration for ideas about the Data Fabric;
* It is claimed that there are many tools/services available that could be integrated in DF. These should be described in the Wiki;
* It is claimed that much of what DF needs to consider has already been discussed in computer science. All contributions that are relevant need to be deposited in the Wiki.
* There is some interoperability work within TAMBIS [13];
* There is relevant workflow work within TAVERNA [14].
* There is a body of work on semantics to support data sharing and interoperability across heterogeneous domains [for example 15]

## 4.2 Work within RDA

Quite a number of groups are working on topics that will play a very important role within the Data Fabric. We list them here in two groups with a short first description of their possible contribution.

**Working Groups**

* **Brokering Governance**: brokering among services is essential part of a landscape with many services from very different groups that need to interact with each other;
* **Data Citation**: starting at the very beginning of the life of data objects registration (PID assignment) and metadata creation is important to enable proper citation;
* **Data Foundation and Terminology**: the Data Fabric needs to adhere to the basics of data modelling as described by DFT;
* **Data Type Registries**: DTR’s will be essential components in the data fabric landscape to allow both machines and humans to interpret types and to allow dissemination of new types;
* **Metadata Standards Directory**: metadata registries will be essential components in the data fabric landscape to enable interpretation of metadata descriptions;
* **PID Information Types**: PIT results (API, core term agreements) are important to harmonize access to PID registration and resolution systems and thus to allow seamless machine access;
* **Practical Policy**: the results of the PP group are at the core of the data fabric landscape since actionable policies (chains of procedures) will carry out the automatic processing that needs to be achieved;
* **Repository Audit and Certification**: this group is essential to the DF landscape as it will specify how we can have trusted repositories as essential components.

**Interest Groups**

* **Active Data Management Plans**: active management plans are at the beginning of specifying the data life cycle and in that way are relevant for the DF landscape;
* **Big Data Analytics**: BDA will be an essential part of the data processing cycle within the DF landscape, so discussions of this group will be crucial for DF;
* **Brokering**: see above;
* **Metadata**: this group is the umbrella group for all metadata discussions and as such crucial for the DF landscape;
* **Data in Context**: this is one of the metadata groups;
* **Research Data Provenance**: this is another group working in the realm of metadata, here focusing on documenting the operations being carried out on the digital objects. Provenance is crucial for a functioning DF landscape;
* **Reproducibility IG**: all aspects of mechanisms in the DF landscape need to be focusing on making data science reproducible, i.e. there is a high thematic overlap requiring a close interaction;
* **Domain Repositories:** domain repositories will be part of the trusted repository landscape which will be essential for DF;
* **Ethics and Social Aspects of Data:** this group will give relevant contributions to the discussion about trust building mechanisms in a DF landscape;
* **Federated Identity Management**: the DF landscape will be highly distributed in all dimensions requiring strong and reliable FIM mechanisms, therefore is FIM crucial for DF;
* **Libraries for Research Data**: libraries want to play a more active role in the domain of data provisioning, therefore they need to know what is being discussed as being relevant in the DF landscape at an early moment in time and contribute based on their experience;
* **Long tail of research data**: in the DF landscape all kinds of data will be processed and managed, i.e. LT data must be part of the DF landscape discussions;
* **PID Interest Group**: PIDs will help identifying digital objects in the DF landscape and help resolving them to useful state information, therefore the discussions are crucial for the DF landscape;
* **RDA/CODATA Legal Interoperability**: this group will have relevant contributions to the discussion about trust building mechanisms in a DF landscape;
* **RDA/WDS Certification of Digital Repositories**: see above under WG;
* **Service Management**: understanding distributed and federated service landscapes will be crucial for a functioning DF landscape requiring a close interaction with this group.

# 5. Use Cases & Demos (Need more input on this)

Use cases and concrete demos will help us to define the scope and nature of DF by defining the different components, interfaces, landscapes and a framework. Quite a number of well thought through solutions have been developed by various communities and institutes to come closer to automated and self-documenting processes for various tasks (management, analysis, aggregation, etc.) and to improve interoperability. Often these initiatives are working within one discipline and within national boundaries and when it comes to federations across discipline or country boundaries again barriers can be found.

Nevertheless, we need to collect such use cases and analyse them carefully to see where they give directions, where they have disadvantages or might lead to lock-ins. Abstractions from concrete use cases and implementations will help to determine which components are being thought of already, what kind of characteristics they have, which services they offer and thus which role they could play in the intended DF landscape. We need to define a template for "use cases" to get comparable descriptions or re-use an existing template[[16]](#footnote-16).

**5.1 Gravitational Physics**

One relevant example comes from the US NDS [16] and illustrates one possible complete path from the acquisition of data, analysis, publication and possible reuse. This path is hypothetical at this stage but is useful as an example of the scope and range of what is being undertaken in some of the research domains. Some aspects of this process are coming together as a framework for testing as part of the NDS Labs.

“In 2021 the LIGO gravitational wave observatory detects a strong “transient” burst event with an unknown source; an alert is issued. Across the US, physicists and astronomers (who have never worked directly together) engage NDS discovery services to find relevant data from other instruments, leading them to detections from the IceCube neutrino observatory, further isolating the originating portion of the sky. NDS discovery services connect the researchers to the federated discovery tools of the Virtual Observatory to collect data by sky position from large surveys like DES and LSST to look for electromagnetic precursors. Through literature searches, they find publications describing characteristics of similar detections; recent publications and an arXiv preprint supporting NDS data linking lead them to the data underlying the analyses. They use NDS data transfer services to migrate previous detection data as well as simulation data held at the Blue Waters supercomputing system, containing previously unpublished neutrino emission predictions, to DataScope, a specialized computing platform to compare observations with theoretical models. From this analysis, a crucial insight suggests a new class of stellar object. Using NDS transfer tools, they pull together the LIGO data, corresponding IceCube detections, image cutouts from LSST, and analyses of simulation data into their private space in an NDS repository. NDS metadata generation tools help them organize a new collection. Soon, a paper is submitted to a journal, including identifiers for the new data collection. Once the paper is accepted, the NDS data collection is sent to a campus archive for longer-term curated management. With the new publication, readers have direct access to the underlying data, enabling them to verify and extend the results. Results and data are further available to educators, who bring the discovery to a broad audience by updating astronomy e-textbooks.”

**5.2 EUDAT** (not a use case description yet)

[Use cases to come]

**5.3 CLARIN** (not a use case description yet)

[Use cases to come]

**5.4 EPOS** (not a use case description yet)

* A geoscientist working on seismology wants to research correlations between seismic events and volcanological events in Sicily.
* She accesses the EPOS-IP ICS-C, the ICS-C confirms authentication / authorisation for the request
* The EPOS-ICS-C assembles from the catalog metadata on
	+ Relevant datasets, location, quality and relevance legalistics, financials
	+ Relevant software, location, quality and relevance, legalistics, financials
	+ Relevant resources (computers, data storage, detectors), location, quality and relevance, legalistics, financials
* And creates a proposed deployment script with distributed parallelism
* The EPOS-ICS-C confirms with the end-user that the assembled proposed workflowed deployment is correct
	+ Appropriate assembled resources (relevance, quality)
* EPOS ICS-C then
	+ Sets up a screen graph of the workflow for the end-user (to keep track of processing)
	+ Dispatches selection software to each ICS-d with appropriate datasets to send results to analytical ICS-d
	+ Dispatches analytical software to appropriate ICS-d ready for processing;
	+ Initiates the execution on analytical ICS-d when selected datasets assembled there;
	+ Receives results for end-user and displays to her
	+ Closes and writes away the detailed log of the processing for future optimisation and audit

**5.5 DataOne** (not a use case description yet)

DataONE is an excellent example of a well worked out set of use cases and implementations that is already associated with NDS. We plan to work with DataOne to develop more information from their use cases.

**6.6 Virtual Observatory Domain (no description yet)**

In the VO domain it is claimed that they have a system of interacting services in place and provide concrete guidelines about services such as for example "how to get data from an archive" etc.

Documents need to be uploaded to the wiki under use cases.

**6.7 ESGF (Earth System Grid Federation) in climate research**

ESGF has been developed as data infrastructure for sharing results from global climate modelling. The most recent application example is the CMIP5 data management. CMIP5 (Climate Model Intercomparison Project phase 5) is globally coordinated climate modelling effort that produced Earth system numerical model results that have been used as data basis for scientific publications on climate change to be used in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR5). Central data that are replicated across three ESGF core data nodes sum up to 2 PB with about 5 Mio individual data instances (NetCDF data files). The total amount of data that has been calculated under CMIP5 is a factor of five larger. First projections of the next generation of this modelling effort provides for CMIP6 an increase in volume and number of data entities in the order of 20 – 50.

ESGf data management challenges are not only related to the volume of data but also to the number of individual data entities:

* Data replication across a globally distributed data federation
* Keep track of millions of data entities at different geographic locations
* Realisation of data access transparently from actual storage location
* Capture provenance information in a structured way
* Grouping metadata (provenance information, data documentation, context information) and scientific data entities to collections with cross referencing between related entities
* Integration of network based, transparent data processing across the federation

ESGF data management challenges might of wider interest for other scientific domains and data federations. DFIG developments could be entering future ESGF data infrastructure improvement.

# 6. Results

The following topics can be results of DFIG:

* act as umbrella for the maintenance of a number of existing WG outputs (PIT, DTR, DFT, PP,?)
* update the landscape description for DF (components, services, interfaces/protocols)
* collection and analysis of use cases
* design of demos incl. willing implementers and report on them
* determining new WGs to work on specific aspects

# 7. Planning (12 months – from P5 to P7)

As part of the process, we need to define some urgent tasks and responsibilities for the next 12 months. Here are a few issues which are meant to stimulate discussion that are not intended to be complete or comprehensive:

* push terminology section to term tool (already done)
* check correctness of DFT terms and suggest how to update them (Gary, Raphael, Peter)
* check correctness of DTR terms and suggest how to update them (Larry, Daan, Giridhar)
* check correctness of PIT terms and suggest how to update them (Tobias, Tim)
* when the PP is ready, check compliance with their terms (Rainer)
* interact with MD, Provenance, Brokering and other groups and exchange the White Paper versions
* collect use cases and analyse them in a system engineering context (all)
* design demos, work with implementers and analyse them (all)
* identify gaps to be addressed by WGs
* categorize and summarize the results as they are developed

# References

[1] DFIG Case Statement: <https://rd-alliance.org/group/data-fabric-ig.html>

[2] RDA Europe Survey: <https://europe.rd-alliance.org/Content/NewsRoom.aspx?id=221&Page=1&Cat=0!3>

[3] Begley, C. Glenn, and Lee M. Ellis. "Drug development: Raise standards for preclinical cancer research." *Nature* 483.7391 (2012): 531-533

[4] Data Type Registries: <https://rd-alliance.org/groups/data-type-registries-wg.html>

[5] Data Foundation & Terminology: <https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html>

[6] PID Information Types: <https://rd-alliance.org/groups/pid-information-types-wg.html>

[7] Practical Policies: <https://rd-alliance.org/groups/practical-policy-wg.html>

[8] Metadata[[17]](#footnote-17): <https://rd-alliance.org/groups/metadata-standards-directory-working-group.html>

[9] Larry Lannom: DAITF: Enabling Technologies; DAITF-ICRI Workshop on Data, Copenhagen, March 2012

[10] ORCID: <http://orcid.org/>

[11] Systems Engineering: Systems and Software Engineering -- Vocabulary, ISO/IEC/IEEE 24765-2010(E), <http://pascal.computer.org/sev_display/index.action>

[12] Tony Hey: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

[13] <http://www.cs.man.ac.uk/~stevensr/tambis/>

[14] <http://www.taverna.org.uk/>

[15] [21] Chen, Liming, et al. "Managing Semantic Metadata for Web Grid Services." *International Journal of Web Services Research* 3.4 (2006): 73-94.

[16] NDS: <http://www.nationaldataservice.org/>

[17] EUDAT: <http://www.eudat.eu>

[18] CLARIN: <http://www.clarin.eu>

[19] EPOS: <http://www.epos-eu.org/>

[20] DataOne: <https://www.dataone.org/>

[21] VO: //

# Appendix

# Basic Terminology

A number of terms should be further elaborated on to set a basis for our discussions[[18]](#footnote-18). We should make use of DFT and System Engineering definitions as far as possible and where needed extend or explain the definitions.



With the help of the diagram we would like to stimulate the discussions based on the idea of using system engineering terminology and show possible relationships between the different terms. At the left edge of the figure, the researcher sees a coherent system that functions in clear, well defined ways to meet a set of specific needs and expectations. It is important to note that the choices, options and possibilities for interactions increases to the right in the figure, illustrated by the increasing width.

All terms being discussed until now can be found in the Term Tool from DFT group[[19]](#footnote-19):

<http://smw-rda.esc.rzg.mpg.de/index.php?title=Special:SearchByProperty&property=Scope&value=RDA+Data+Fabric+Interest+Group>

To participate in the Term Tool elaborations you first need to register:

<http://smw-rda.esc.rzg.mpg.de/index.php?title=Special:UserLogin&returnto=Main+Page>

1. We see the versions 0.x as early drafts of the White Paper. [↑](#footnote-ref-1)
2. Data Type Registry WG [4], Data Foundation & Terminology WG [5], PID Information Types WG [6], Practical Policies WG [7], Metadata Standards Directory WG [8] [↑](#footnote-ref-2)
3. <https://www.dataone.org/best-practices> [↑](#footnote-ref-3)
4. A framework is seen as a reusable design that can be refined (specialized) and extended to provide some portion of the overall functionality of many applications (for details see attachment A). [↑](#footnote-ref-4)
5. Some steps such as "integration" are in this cycle diagram processes that for example do conversion. [↑](#footnote-ref-5)
6. Sometimes newly generated data will be stored in temporary stores only since they are results of test runs etc. In this paper this is not further debated. [↑](#footnote-ref-6)
7. It is debated to which moment in time exactly "registration" should take place. It is important to mention that PID and MD assignment should be done as early as possible in the highly automated Data Fabric, since extensions at later moment will often not happen or are very costly. But the notion of DF does not include any particular moment in time when this should be done. [↑](#footnote-ref-7)
8. <http://smw-rda.esc.rzg.mpg.de/index.php?title=Special:SearchByProperty&property=Scope&value=RDA+Data+Fabric+Interest+Group> [↑](#footnote-ref-8)
9. It should be noted here that properties stored in the PID record are also metadata in the generic sense. However it is important to make differences between different types of metadata – a classification into packages being worked out within the metadata groups in RDA. [↑](#footnote-ref-9)
10. In this paper we do not make differences between the kinds of processes being carried out. As indicated in the Practical Policy group's overview they can be of rather different nature fulfilling different purposes all resulting in some form of transformations. [↑](#footnote-ref-10)
11. This could also be done automatically by using some operations acting on behalf of the user especially with increasing scale. [↑](#footnote-ref-11)
12. This definition seems to be compliant with the descriptions in this document, but it is at a different level of abstraction. [↑](#footnote-ref-12)
13. This expression is compliant with what is written in the text as a requirement for DF. People must be able to configure their solution within certain boundaries that are obviously defined by the need to use certain common services and to adhere to specifications. [↑](#footnote-ref-13)
14. There is an ongoing discussion whether also such a component needs to be able to turn an identifier to state information that allows for example verification which ORCID currently does not. . [↑](#footnote-ref-14)
15. <http://cordis.europa.eu/fp7/ict/ssai/docs/future-cc-2may-finalreport-experts.pdf> ;

<http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf>;

<http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-expert-group/roadmap-dec2012-vfinal.pdf>

ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg\_eg\_final.pdf

ftp://ftp.cordis.europa.eu/pub/ist/docs/ngg2\_eg\_final.pdf

ftp://ftp.cordis.europa.eu/pub/ist/docs/grids/ngg3\_eg\_final.pdf [↑](#footnote-ref-15)
16. This needs to be done as next step to finish the White Paper. [↑](#footnote-ref-16)
17. This is seen as standing for the various metadata WG/IGs. [↑](#footnote-ref-17)
18. We should not have the mission to bring out proper definitions in the sense of terminological completeness, but to allow us to interact on the DF ideas. [↑](#footnote-ref-18)
19. Elaboration notes should be made by all participants to make sure we all understand the same. Some terms are not take from SE, but quickly noted to have a start for elaborations. [↑](#footnote-ref-19)