Data Fabric IG

Rob Pennington (NCSA), Yunqiang Zhu (CAS-IGSNRR), Peter Wittenburg (MPS)
Purpose of this plenary session

- Information on the Data Fabric IG and its intentions
- Synchronize on the “understanding” and scope with everyone in RDA
  - Are we starting from the same points and headed for the same goals?
- Agenda for this session
  - Rob: Introduction (what is DF, history, where we are)
  - Peter: Analysis of first use cases
  - Zhu: Use case template and participation
  - Q&A and discussion
- If you are interested – join
  - DFIG Core Session (Tuesday 4 pm breakout 6)
  - BoF on Repository Registry (Wednesday 11 am breakout 7)
What is the Data Fabric?

- “Data Fabric” is ambiguous if you look across-cultures
- Isn’t this at the core of RDA? We like this ambiguity and its possible connotations
- Can we describe what DF is?
- A short history
Data Fabric is a Bottom-up Effort

- At the first plenaries the first WGs started in a more or less isolated way under pressure to deliver artifacts.
- It was understood that we all work on a larger picture of integration – call it a framework for processing our data.
- Realization that the WG outputs and also the topics of other groups (WGs/IGs) are working on are components and their services with a place in this landscape.
- DF is a place to discuss such components and understand how they all will fit together.
- So together with a number of chairs we started DF IG.
this represents our scientific data creation & consumption machine in the scientific labs with various steps and various components

Not an architecture diagram!
Some of the Big Questions for RDA:

How can we maximally support this machinery

- unload researchers from unnecessary details,
- make science reproducible,
- How to identify the essential components and services
- let people configure them according to their needs
- etc.
Thinking about How People Work with Data in their Research

all phases must be considered in DF IG

Originally from DataONE
Data Fabric where we are

- 2nd WG Chairs meeting
- Draft White Paper
- 1st DFIG Session at P4 in Amsterdam
- updated draft version of WP
- several meetings where DFIG was presented & discussed
- lots of commenting in DFIG wiki
- first real WP version
- start of collecting Use Cases
- 2nd DFIG Session at P5 in San Diego
Data Fabric first analysis

- **goals**
  - understand components/services infrastructures are using
  - extract common components/services and their characteristics

- **two strands of input for analysis**
  - current data practices
  - Use cases

- now also analysis of Large Scale Data Infrastructures
Data Fabric first analysis

how to come to essential components & services?

Analyze Data Practices
Data Management Conclusion

Management of data objects is widely type and discipline independent.
PID system is core

Internet Domain nodes with IP numbers packages being exchanged
standardized protocols

Data Domain objects with PID numbers objects being exchanged
standardized protocols

Value Added Services

Internet Protocol Suite

Network Technology

IP

domain

PID

Digital Objects

Data Sets RDBMS Files

Local Storage Cloud Computed

Analysis Persistent Reference Citation

Apps Custom Clients Plug-Ins

Resolution System Typing

Metadata attributes

point to each other
Data Fabric first analysis

how to come to essential components & services?

Analyze Use Cases
10 (+5) Use Cases so far (2 in development, others mature)

- Environmental science
- Natural science
- Life science
- Humanities, soc. sciences
- IT, various

All indicated nodes are centers of national, regional and even worldwide federations.
Issues of Relevance

- Data fabric domain
  - documented processing
  - new collection
  - new metadata
  - temp store
  - management, analytics, conversion
  - provenance – reproducibility
  - workflows, policies, deployment

- Collection
  - highly distributed in federations
  - AAI/FIM
  - virtual collection builder
  - FS, Cloud, DB Repository System

- Raw data
  - PID, Metadata
  - Rights
  - Syntax, Types
  - Semantics
  - Relations

- Publishing domain
  - citing
  - data & paper publications

- Sensors
  - Simulations
  - Crowd
  - etc.
How do WG/IGs fit?

Data fabric domain

collection

DFT
Data Foundation & Terminology
simple model for digital data in a registered domain

MD
Metadata
community curated standards catalogue for metadata interoperability

PIT
PID Information Types
common interface for providers and users of persistent ID services worldwide with harmonized categories

DTR
Data Type Registries
allowing humans and machines to act on unknown data types

FIM
Practical Policy
defining best practice workflows how to deal with data automatically and in a documented way

documented processing

processing

registration

permanent store

citing

raw data

DMP

CITDD

CERT

BROK

BDA

PP

Prov

REPRO
Components I

- domain of registered digital objects (DO) incl. basic organization principles (data, code, knowledge) -> **worldwide PID system (Handles/DOI)**

- domain of registered actors -> **worldwide ID system (ORCID)**

- domain of trusted repositories for DOs -> **worldwide Rep Registry**
  - proper DFT/DSA/WDS compliant **repository systems**

- accepted policy commons (proper organization support, self-documenting, tested/certified, etc.) -> **policy component registry**

- policy/services -> **service registry**

- authentication system -> **various in place (ORCID just number)**

- authorization system -> **authorization registry**
Components II

- MD components/schemas -> metadata schema registry
- data types /schemas/formats -> data type registry
- semantic categories -> category registry
- vocabularies -> vocabulary registry

... much already out there but ...
... why does it cost months
  - to federate and integrate data
  - to make data interoperable
... need to harmonize, raise trust & value
... make it ready for machines
In order to compare different use cases and extract common characteristics of components and services of use cases, DFIG made a use case description template.

1. Scientific Motivation and Outcomes
2. Functional Description
3. Describe essential Components and their Services
4. Describe optional/discipline specific Components and their Services
5. Describe essentials of the underlying Data Organization
6. Indicate the type of APIs being used
7. Achieved Results
Use cases template II

1. Scientific Motivation and Outcomes (max. 0.5 pages)
   Provide a short summary of the scientific or technical motivation for the use case. What would be the best possible outcome and why?

2. Functional Description (max. 1 page)
   Give at least one diagram that indicates the overall structure/architecture of the data creation and consumption machinery that is being used in the lab/infrastructure. Describe in simple words the functioning of the machinery.

3. Describe essential Components and their Services (max. 1 page)
   Describe the most essential infrastructural components of the machinery and the kind of services they offer. These descriptions don't have to be comprehensive.
4. Describe optional/discipline specific Components and their Services (max. 1 page)

Describe the optional/discipline specific infrastructural components of the machinery and the kind of services they offer. These descriptions don't have to be comprehensive.

5. Describe essentials of the underlying Data Organization (max. 1 page)

Describe the most important aspects of the underlying data organization and compare it with the model outlined by DFT.

6. Indicate the type of APIs being used (max. 1 page)

Describe the most relevant APIs and whether they are open for being used.

7. Achieved Results (max. 0.5 pages)

Describe the results (if applicable) that have been achieved compared to the original motivation.
Summary

- DFIG as a platform for WG/IG chair interaction about all kinds of components/services that are essential to make data work more efficient, cost-effective and reproducible.
- The idea is to do Use Case studies to identify such components/services based on what people are doing.
- The method is thus learning from examples and from there to do abstractions to common components.

Please provide your Use Cases and join discussions on their essentials.
Thanks for your attention.
some answers !?

- lack of broad conviction in science – missing guidance, thus too risky to invest (thus no broad uptake and lack of quality)
- lack of widely trusted, stable and accessible services
- lack of explicitness of structures and semantics
- lack of agreed common interfaces
- brokering versus harmonization