Scientific Data Management and Making Danish ESFRI[1] data more FAIR[2]

ESFRI case AnaEE

1 Introduction

DeIC, within the auspice of its *Research Data Management Forum*, has decided to initiate a pilot project aiming at analysing how larger research communities are faring in respect to *FAIRifying* their research data – i.e. in applying the FAIR data principles. The overall aim is twofold, in gathering knowledge and experience as to:

1) how university and national research support functions are best equipped and maned, in the pursuit of more *FAIR* data.
2) how universities, possibly nationally coordinated, best can design a curriculum aimed at producing Data Stewards.

In so doing DeIC wishes to establish cooperation with three ESFRI projects:

a. AnaEE (Infrastructure for Analysis and Experimentation on Ecosystems)
b. WindScanner (European WindScanner Facility)
c. ICOS ERIC (Integrated Carbon Observation System)

These represent larger well established international communities that are part of the European research framework. Hence they are subject to both external expectations as to FAIRifying their research data, as well as assumed to have a self-interest in so doing.

2 Datatype, structure, geographical, organisation state

Initially we aim to better understand the nature of the data in question. I.e. we need to map the variety, type and structure of data, as well as how the data is currently stored, accessed and used, within which organisational settings, with which economic and geographical attributes. Subsequently prioritization is done such that data, which would be most worthwhile to FAIRify (i.e. giving best scientific impact), is selected.

The most obvious low hanging fruit best suitable for *FAIRifying* in AnaEE Denmark is the meteorological data (time series). These can be described in a DataCite Metadata Schema: https://schema.datacite.org/meta/kernel-4.2/

The total volume of such data is manageable, estimated to be lower than a few TB.

The data originates from several geographically dispersed sites and each site data set consists of sets of homogeneous flat ACII-files, possibly divided into annual time series. In total, at most 10 sets of data are situated at 3 AnaEE Denmark partners (at KU, AU and DTU, respectively).

. The 3 partners all control the data in question in an autonomous manner, hence all work on the data must be based on interest and consensus must be achived in terms of common standards, quality control etc.

It was agreed that one, maximum two, dataset from each of the three partners (3-4 in all), covering one year, would be chosen to start with.

3 Optimal attack vector aimed at the elaborated FAIR principles

We need to analyse the status of most the important datasets against the FAIR principles, using the Go-FAIR topology[3], and decide where FAIRifying actions have best research impact.

In choosing suitable datasets, the approach is to select them from the specific Research Platforms run by the partner sites, consisting of one, maximum two datasets (possibly in more than one file). For each, it will be decided to which extent the dataset can be FAIRifyed, i.e. the extent of the FAIRification

[1] Danish ESFRI projects: http://roadmap2018.esfri.eu/projects-and-landmarks/browse-the-catalogue/?members=DK

[2] FAIR Principles: https://www.go-fair.org/fair-principles

[3] FAIR Principles: https://www.go-fair.org/fair-principles

process must be estimated (https://www.go-fair.org/fair-principles) for each dataset.

4 FAIRification effort output product

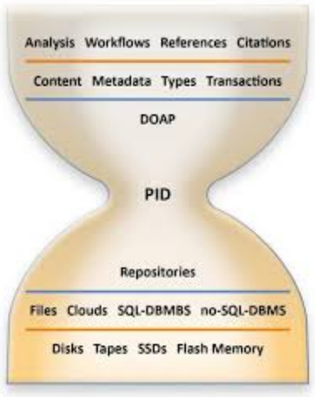
Analysis Workflows References Citations
Content Metadata Types Transactions
DOAP
PID
Repositories
Files Clouds SQL-DBMBS no-SQL-DBMS
Disks Tapes SSDs Flash Memory

Figure 1: The hour glass model used to describe the key message for the Internet transformed to the data domain with PID numbers and the Digital Object Access Protocol in its centre (Source: RDA & Go-FAIR).

In the pursuit of highest possible research impact, with a minimal FAIRification effort, the point of departure is the assumption that high impact can best be obtained at the "*belt of the FAIR hour glass*" (Figure 1). I.e. it is assumed that as a minimal effort one needs to embed data in a Persistent Identifier (PID) Infrastructure, that is machine readable and accessible, knowing that the above analysis 3 (*optimal attack vector*) might influence such prioritisation. Based on this understanding, a product output based on a minimal FAIRification effort is defined. This could be any change making the research data more FAIR.ing the research data more FAIR

A minimum output (project requirements) is thought to be:

- agreed initial quality control of datasets;
- making the dataset Findable, via persistent identifiers (PIDs);
- making the dataset Accessible via a Landing Page.
- Preferably some degree of Interoperable.

In total, we seek to archive at least one level of *FAIRifying* from each of the four letters in FAIR[4]

A plan to handle versioning must be developed, if not implemented.

5 Suggested Actions

Based on the above aimed for FAIRification effort and output product, a project is defined describing, not only the task to be done, but also how it is done, by whom at which cost, within which timeline.

- Katrine Korsgaard Vendelboe <katrine.vendelboe@plen.ku.dk> will lead the initiative, and be single contact point from AnaEE. Rene Belsø <Rene.Belso@deic.dk> will be contact point from DeIC.
- Katrine will find data-site managers from each of the three sites; will introduce this initiative and write a more elaborated plan for the work to be done, based on this agreed point of departure.
- The plan will consist of a time schedule.
- The plan will consist of a budget, in categories as needed, for DeIC to consider and possible contribute to.
- The three data sites will together organizes a Workshop, in Denmark, financed by DeIC.
  - ➢ The workshop will for each dataset aim *to* FAIRifying each dataset, as well as a plan to FAIRify all AnaEE meteorological data
  - ➢ As a minimum conclude DOI's (Digital Object Identifiers) for included datasets
- The workshop date, venue and budget will be presented to DeIC by end of May beginning of June, aiming to hold the workshop before the summer holidays.
- The results of the project will be presented to DM Forum at one of its meetings in 2019.

[4] https://www.go-fair.org/fair-principles