FAIR Technology Matrix
Phase 2 Impressions

Kristina Hettne (Leiden University Libraries, CDS), Peter Wittenburg (GEDE, RDA)
September 2019

We received 20 Questionnaires. 2 were doubled and 3 empty, i.e. 15 could be analysed. 15 are not sufficient to draw final conclusions, but they are sufficient to get first impressions. Thus, all statements in this document need to be interpreted with care.

General

- Questionnaires are not liked and people don't want to spend time on them. Only via funding engagement detailed answers can be expected. A possible solution is to link the process of filling out the questionnaire with events such as workshops and conferences, so that it can be done together with colleagues. This would add an element of collaboration and a feeling of being in this together. We might also consider a more direct coupling with what others have filled in, so that if you fill in for example a metadata standard you can choose to immediately see what others have filled in, in for example a graph representation or in a matrix format.
- Obviously, the length of the questionnaire and the detail of the questions shocked a few colleagues, so that they stopped answering. A possible solution to this would be to automize what can be automized, by making machines find the answer to some of the questions. This work is already underway by us and others. This would be facilitated by having repositories expose relevant information via an API. For more loosely defined communities with no central point of information this is less straightforward, and these communities would probably still benefit from a questionnaire.
- One can also observe that experts who filled in the questionnaire and do not know details do not take the step to invite their knowledgeable colleagues to answer the questions. A possible solution to this is similar to the solution for the fact that people do not like filling in questionnaires: link it to activities where they can do it together with colleagues.
- There are currently no clear job descriptions/profiles, i.e. people with different skills and backgrounds, different organisational embeddings in the organisations and different terminologies participated in answering the questions. A possible solution to this is automation of some of the questions, plus again to make sure that people do not have to fill it in in isolation.
- We can distinguish two groups of experts: a) those that come from "big data labs" who understood most of the questions and b) those who have more of a library background. Most accurate answers came from those who take responsibility for repositories in "big data initiatives". There is a great learning opportunity here for the library world, which can be

seen as a challenge but also a possibility. Common events with support from colleagues might again work here.

- The answers are already much better than 5 years ago in RDA's first survey where it was finally necessary to make many interviews instead of using a formal questionnaire. When we first sent around a questionnaire, we could hardly find people who had a deep enough knowledge about concepts behind terms such as "repository, access technology, aggregation of metadata via common protocols, persistent identifiers, etc" and thus who were willing to become active. However, during the interaction we were very quickly able to map terminologies and thus to understand each other. The most positive aspect, however, was that a questionnaire is "only taking" while an interaction also includes "giving". People were eager to speak with us since we had already deep experience in many fields, i.e. people expected insights from us about new developments. At that moment where a formal questionnaire can be turned into a source of information that will help people, they may have a positive attitude.
- There are many different types of repositories (data, metadata, ontologies, books, etc.) and the engaged experts believe that they are different. Only the notion of Digital Objects will clarify this confusion, but it is not known to all experts yet.
- The questionnaire has a few weak points:
    - The questionnaire is addressing people with different profiles when asking about access technologies and semantic vocabularies which does not fit. There are only few experts who have that broad knowledge.
    - Some questions are too detailed[1].
    - The answers on AAI, access methods and protocols are the most diverse and messy indicating also the divergence in solutions. These questions might benefit the most from being automized from one side or both, so as to exclude human involvement. A clarifying diagram would help people to understand in detail where we are talking about.
    - The term provenance is ill-defined, since some people speak about "provenance" when categories such as "author" and "creation data" are included which are basically part of any metadata set. Other people mean distinguished vocabularies such defined by "PROV" when talking about provenance.
    - The DMP questions are obsolete in this context, since mostly repository managers or data scientists responsible for data stores were responding. These experts do not create DMPs.

Summarizing our general impressions, we can expect that

- Reality is far from being FAIR compliant, even though responders reply that they are FAIR-compliant. There is a mismatch in understanding what FAIR means.
- Reality does not offer possibilities for automatic FAIR assessments with a few exceptions.
- We need to discuss whether having at least two different questionnaires, one addressing the "big data folks" and one addressing those that deal with metadata, publications, etc., would improve the willingness to participate and the quality of the results.

Repositories

- The following approaches are being used: 7 use relational databases, 4 own software stacks, 3 iRODS, 2 File Systems, 2 Dataverse, 2 Triple Stores, DSpace, Fedora, Kite, CMS (people did not make a difference between metadata and others, for big data most seem to use file systems)
- 6 repositories are certified following the CTS rule set

[1] The questions whether "ORCID is used for authentication" or whether "headers are used to include metadata" are two such examples which are too specific.

- 8 claim to have some policies described at one of their web-pages
- 4 repositories are being registered via re3data, 2 with datacite, 2 with FAIRSharing, WDS, GEOSS, etc.
- 7 repositories gave specifications for persistency, 5 for 10 years, 1 for 20 years, 1 for 50 years

PIDs

- 9 use DOIs for final collections, 7 from Datacite, 2 from Crossref; 6 of these add DOIs manually which indicates small amounts of PIDs
- 5 use Handles, all being “big data people” assigning Handles automatically, 3 use ePIC Handles, 2 have own services
- DOIs are being resolved to resolved to HTML landing pages, the others resolve to structured records with different attribute sets or are planning to do so

Access

- There is much variation in the answers and solutions. Terminology is not at all clear.
- 13 say that metadata is open.
- 6 specify that they use some kind of CC license for data and metadata. One says that they use a code of conduct, since there is no clear legal basis.
- ORCID IDs are used only in one case for authentication.
- 13 are using an own user database for authentication, sometimes in addition to distributed AAI.
- Local LDAP is used in most cases. In addition, Shibboleth, SAML, OpenID, eduGain and CPAuth were mentioned for distributed solutions.
- For authorisation local ACLs are used. One is using a higher-level Authorisation DB (no one has a distributed solution).

Data

- There is a large variation in types and formats. 1 is just focusing on ontologies, 2 just on metadata and 1 on statistical survey data.
- People using “big data” are often using netCDF and HDF5 formats well-known in natural sciences.
- 8 claim that they are using registered schemas/formats for their data.
- It is not clear whether search on data is being offered.

Metadata

- 8 are using custom metadata schemas, 4 are using generic schemas (DC, Datacite, ISO, etc.)
- 7 claim to use registered metadata categories.
- The term “provenance” is not used systematically, since it is not well defined. Three refer to standard terminology such as PROV.
- 7 support OAI-PMH as exchange protocol, 3 offer Json/XML/CSV downloads.
- 10 claim to enter their PIDs in the metadata.
- 10 offer search on metadata, in some cases the answers were not so clear.
- Only 3 state that they have long-term plans for metadata.
- The notion of “machine actionability” is not well understood yet.

Semantic Vocabularies

- Mostly the answers are not so clear and people refer mostly to home-grown vocabularies. There is an opportunity here to link with a variety of services already being used by the different communities that could help.

DMP

- 3 are using DMP tools, but the questions do not fit with the people who answered the questionnaire.
- The question about an “explicit publication step” was not well-understood. Only two speak about explicit validation and quality control before publication.

Processing

- 7 state that they are using some form of local workflows operating on the data.
- The question about distributed workflows is not well-understood.
- 6 state to offer some analytic support on data, 3 are working on it, 3 state that they are offering certain data products.

FAIRness

- 11 claim to be compliant with F
- 12 claim to be compliant with A
- 12 claim to offer F compliance to a certain extent
- 12 claim to be compliant with R

Comments

- 3 explicitly state that the questionnaire is good – these are all “big data” people.
- 1 states that the questionnaire is comprehensive.
- 1 states that many questions are difficult to interpret.
- 1 has much critique on the questionnaire and doubts whether the authors are serious. This colleague only deals with metadata.