Towards a Model for FAIR Data Information Infrastructures

Caspar Terheggen, Ed Simons, Radboud University
SCOPE OF THE PRESENTATION

• About Research Data *Information* Infrastructures, dealing with aspects captured by the term “FAIR”:
  - *Findability*
  - *Accessibility*
  - *Interoperability*
  - *Reusability*
  of datasets

Where I would like to add the term “*Interpretability*” for the “I” of FAIR.

• So it is NOT about data storage infrastructures.

• The aspect of “FAIR” has to do with (optimal supply and availability) of METADATA about datasets.
POINTS OF DEPARTURE

• *Silo-ed registration of metadata on datasets*, meaning in a separate system, only dealing with dataset metadata out of their broader context is not optimal.

• A model concerning *FAIR data infrastructures should include the local (institutional), national, international and discipline level.*

• An optimal FAIR data infrastructure consists of two complementary parts:
  - A technological or “systems” part.
  - A service and support part.

• *FAIR data information infrastructures and services* (built upon these infrastructures), *can in principle be developed and managed by different organisations/entities than the ones that deal with storage infrastructures.*
FAIR Model – Technological Part: THE ROLE OF CRIS’s

• Silo-ed registration of metadata on datasets, meaning in a separate system, only dealing with dataset metadata out of their broader context is not optimal and should be avoided.

• Why is this so?
  - *Combining dataset metadata with additional metadata* on (the related) publications the researchers involved, the project the dataset resulted from, the institute(s), the cooperations, the funders, etc… provides a much richer information source for “FAIR-ness” of datasets and – not a minor point – for finding interesting related datasets.

• Given this CRIS’s, and registering metadata on datasets in the CRIS, can (and in my view should) take a central position in FAIR data infrastructures.
THE ROLE OF CRIS’s: a bit of history

• CRIS’s started (1990’s) as administrative, reporting systems to the government, and were mostly conceived by researchers as “an administrative overhead and nuisance”.

• In the course of the years 2000: CRIS’s became instruments for research policy and management and also as a basic resource for institution’s OA Repositories.

• The last 5-10 years: CRIS’s being used as instruments for on line profiling of research(ers) and research institutes. As a result: CRIS’s more and more accepted by the researchers.

To summarize: CRIS’s have evolved from systems for administrators and managers to (also) interesting tools for the researchers themselves.
THE ROLE OF CRIS’s: a bit of history

• The researcher’s “traditional” view on CRIS’s.

Research activities

“My World”

Research information activities (CRIS)

“Theyir (administration) World”
THE ROLE OF CRIS’s: a bit of history

• Is changing....

Research activities

Research information activities (CRIS)

“My World”
FAIR Model – Technological Part: THE ROLE OF CRIS’s

Central Position of CRIS in the RI System Landscape

**INPUT FROM**
- External Publication / Data Resources: (WoS, Scopus, Google Scholar...)
- Researchers
- Administrative Resources: (HRM, Finance, Project Man.)
- Inst. Secretariat
- Internet

**OUTPUT TO**
- Publication Repositories
- Dataset Repositories / Archives
- Profiling & Management Applications
- ((inter)national) Research Portals
- Other RIS-systems / formats
- ORCiD
- VIVO

Institutional CRIS

CERIF
FAIR Model – Technological Part: a look at the DATA LIFE CYCLE

- **Data Collection**: Selection
- **Analysis**: Temporary storage
- **CRIS**: Registration of (metadata) for dataset
- **CRIS**: Linking publications to dataset + archiving of data.
- **Publication of Results**: Final (shared) dataset
- **Data Life Cycle Management**
- **Articles**
FAIR Model – Technological Part: integrating CRIS-functionality in the Virtual Research Environment of the Researcher (VRE)

A “traditional” VRE may look like this

A “CRIS-integrated” VRE may look like this.
Metadata Layer Structure: A CONCRETE EXAMPLE

• Suppose: Marine Biology research into the water quality in a certain region of the Indian Ocean and its effects the Coral Reefs in that region.
• Water samples are taken regularly in various parts of this region.
• In layer 2, the central generic layer (CRIS), metadata are stored such as:
  • Unique identifier of the dataset (e.g. DOI)
  • Name/title of the dataset
  • Language
  • Conditions for access and re-use
  • Possible restrictions for public access
  • Institute which conducted the research
  • (names, titles, roles, etc... of) Researchers involved
  • Responsible person and contact person for the dataset
  • Project as part of which the research was carried out
  • Publications based upon/linked to the dataset
  • Geographical coordinates of the of the Indian Ocean region the research applied to
  • Pointer (URI) to the data-specific metadata in layer 3, etc...
From layer 2 the discovery metadata (e.g. DC) for layer 1 are automatically generated and “pushed” into the data repository (or either directly exposed on the Internet).

In layer 3 the discipline- or data-specific metadata are stored, e.g. the geographical coordinates of the specific parts of the region, specific classification schemes of the coral flora and fauna, chemical analysis procedures applied, equipment used etc...

For each layer specific services may be implemented, e.g.: harvesting services on layer 1, dataset profile creation or linked data services on layer 2, integrity control services on layer 3, etc...
FAIR Model – The Service Part: LEVELS and PLAYERS

- There are four levels or “players” included in the FAIR service model:
  - The *local or institutional* level (institutional repository/CRIS)
  - The *national* level (national registry/repository)
  - The *discipline* level (disciplinary registry/repository)
  - The *international* level (international repository).
FAIR Model – Service Part: LEVELS and PLAYERS
THE SERVICE MODEL: LAYERS and PLAYERS

• The *national / international registry* is the obvious party to:
  - register metadata to the global search engines;
  - offer a national repository service (ideally via cloud storage);
  - implement and provide the request-approve-receive workflow (see further);
  - assist research institutions that choose to host (some of) their datasets on premise or in the cloud.

• For whatever reason some research institutions may choose to host (some of) their datasets on premise or in the cloud themselves, others will store their datasets in the national repository.
Discipline Specific Repositories

• Researchers working in a discipline with a discipline specific repository (such as astrophysics and genomics) will in general access this repository directly, both for registering their own datasets, and searching for other researcher’s datasets.

• Such repositories offer elaborate, discipline specific search capabilities, access request and granting workflows, and download functionality.

• For the purpose of completeness it is desirable that national registries hold the (generic) metadata of datasets in discipline specific repositories. To this end the research institution should assist their researchers (through some automated process) to upload the generic metadata to the national registry without requiring double work from the researcher.
IMPLEMENTING THE “FAIR DATA INFORMATION MODEL”

• To implement FAIR infrastructures in a scalable way, we need to:

  - Implement the following functions: *Find, Request, Approve, Receive.*

  - Ensure that datasets are stored safely, and remain accessible.

  - Ensure that the metadata are correct, standardized and detailed and complete enough.
IMPLEMENTING THE FAIR DATA MODEL: integrating the technological and service part
Example

• A researcher searches for a dataset based on certain meta data.
• The search engine produces a reference to a dataset registered in a national registry. The dataset itself is hosted in some research institution.
• The researcher issues a request for the dataset via the national registry services interface.
• The national registry forwards the request to the research institution hosting the dataset.
• If the request is granted, the national registry communicates the request specific download link to the researcher.
• The researcher accesses the research institution repository to download the dataset. (In case the dataset resides on the national repository, the registry provides the appropriate link as well.)

• Discipline specific repositories generally provide all four steps, and the searching capabilities are much more discipline specific.
ADDING USER SUPPORT: The Front Office-Back Office Model

• To make the model work, an optimal user support structure is necessary. An example of such a structure is the so-called “Front Office-Back Office” model.
ADDING USER SUPPORT: The Front Office-Back Office Model

- The Front Office is located on the local (institutional) level, *with the library being an obvious location for it*, close to the researcher and takes care of the 1st line support.
  
  This includes a.o.t.:
  - Help desk for user support of the dataset registration, e.g.
    - how to use the registration interface,
    - how to structure the data file set according to the requirements regulations of the data hosting organisation,
  - Working out multimedia demo’s as well giving hands-on training to institutes researchers.
  - Check on completeness and correct structure and format of the dataset files
  - Help in writing data management plans.
  - To institutes: support in developing / formulating a data management policy.
  - Development and management of an information web site.
  - etc...
ADDING USER SUPPORT TO THE MODEL: The Front Office-Back Office Model

- The “Back-Office” functionality is located at the (inter)national or disciplinary level (data hosting services level) and concerns 2nd level support to and through the Front Office.
ADDING USER SUPPORT: The Front Office-Back Office Model

- (with an example from The Netherlands currently being implemented)
SUMMARY: “FAIR DATA INFORMATION INFRASTRUCTURES”

• The model concerns the “FAIR” aspect of data, meaning it deals with information (metadata) promoting and supporting the findability, interoperability, interpretability and reusability of datasets.

• The model consists of two parts:
  - A technological solution.
  - A service and support solution

• **CRIS systems are core elements of the technological solution** since they provide rich additional metadata on datasets and put the datasets and their metadata into their proper context, and so significantly enhance the FAIR-ness of datasets.

• **Another crucial element** underlying the technological solution is the 3-layer metadata model developed by euroCRIS.
SUMMARY: “FAIR DATA INFORMATION INFRASTRUCTURES”

• Exchange of information between parts and levels of the model should be based on a standard. CERIF is an obvious choice for this.

• To make the infrastructure and the (FAIR) services built on it work, organisations on the local, national, international and disciplinary level are involved.

• An optimal support organisation, for which the “Front Office – Back Office” is an appropriate model is of the utmost importance. Especially the Front Office function is vital to make the model work for researchers and institutes.

• Last but not least: the “FAIR Data Information Infrastructure” should be an integrated part of national and international Research Data Infrastructures (such as the EU Open Science Cloud), in order for these infrastructures to work optimally.