Workshop ‘FAIRification of Nordic+Baltic data repositories’

April 22, 2020

Authors: Andreas O Jaunsen, Tuomas Altera, Josefine Nordling, Mari-Elisa Kuusniemi, Bert Meerman, Erik Schultes

Goal

We wish to provide here a relatively short and concise *background* for the recommendations on how communities / digital repositories can become FAIRer. The recommendations will be based on the results (specifically on the tests that were *not* passed) from the ‘FAIR data’ team’s FAIR maturity evaluations, which were performed using Mark Wilkinson’s FAIR Maturity Evaluator service.

Introduction

The European Open Science Cloud (EOSC) is an European Commission initiative that started in 2015 and the EOSC Portal was launched in November 2017. It is an infrastructure consisting of open science promoting services which thus enables access and reuse of research data. EOSC aims to serve three objectives: (1) to increase the value of scientific data assets by making them easily available to a greater number of researchers, across disciplines (interdisciplinarity) and borders (EU added value) and (2) to reduce the costs of scientific data management, while (3) ensuring adequate protection of information/personal data according to applicable EU rules.

EOSC currently supports the development of FAIR in various ways and through various approaches. There are many projects and working groups within the EOSC project ecosystem that work with landscaping tasks and directly FAIR promoting activities. EOSC-Nordic is one of the five EOSC regional projects (‘5b’), all with the aim of connecting national initiatives, policies, infrastructure services and people to ESOC. The other four regional projects are EOSC Pillar,

EOSC Synergy, ExPaNDS, and NI4OS-Europe. Note that each of the regional implementation projects have independent strategies and plans for how they intend to implement the EOSC in their respective regions.

The FAIRsFAIR project - Fostering FAIR Data Practices in Europe, is also an EOSC funded project ('5c'), which aims to embed FAIR data practices in the research data life cycle. The ESFRI clusters projects are serving the purpose of domain specific disciplines (ENVRI-FAIR, PaNOSC, ESCAPE, SSHOC and EOSC-Life) and are similarly supporting and enforcing the EOSC community with its offering and expertise.

The EOSC Executive Board has established five EOSC Working Groups: FAIR, Landscape, Rules of Participation, Architecture and Sustainability. The Working Group on FAIR is charged with implementing FAIR data principles by defining the associated requirements for the development of EOSC services in order to foster cross-disciplinary interoperability. The EOSC Governance includes the Governing Board, Executive Board and Stakeholder Forum. The EOSC Governance is assisted by a Synchronisation Force developed from within the FAIRsFAIR project, which seeks to map all relevant FAIR related activities within the EOSC in order to avoid duplicated work and to foster synergies. The Synchronisation Force seeks input from the Expert Group of FAIR Champions (ECFG), the ESFRI research clusters, and the thematic and regional '5B' projects.

Background

The EOSC-Nordic has pledged to 'implement FAIR' in the region. This implementation will happen primarily by;

> i) disseminating the benefits of going FAIR to a broad science community,
> ii) providing an evaluation-based recommendation on how to FAIRify a data repository
> iii) supporting communities by hosting a handful of hackathons and/or metadata-4-machines events.

The EOSC-Nordic FAIR work package activities started off by conducting landscaping activities with the aim to get an overview of Nordic and Baltic data repositories. Repositories were collected from re3data.org and suggestions from the participating partners in the WP, provided they were linked to one of the countries of the Nordic+Baltic region. The aim was then to evaluate the level of FAIR maturity for each of the selected data repositories at the start of the project. Towards the end of the project we will remeasure the FAIR maturity in order to detect (positive) changes to the FAIRness of the data repositories in the sample.

Here, repositories may be any data deposit, archive or registry that provides access to research data. By research data, we mean any data that is suitable for research and it may consist of raw or processed data, structured or unstructured data. Data repositories may be broad

cross-discipline repositories, archives for infrastructures or community specific repositories that are limited to a narrow scientific field. The only exception to this is repositories that primarily (or exclusively) host research publications (documents) – these have been kept out of the sample intentionally. The reason for this is that publications archived in a repository are not considered research data; they are largely institutional collections of work that is published elsewhere and due to the sheer number of such repositories, they would bias the results of our study.

So for the repositories that do enter our sample, we measure them all based on the same premises using generic machine actionable metadata. Machine-actionable, means that a bot or harvester ("the machine") can understand what kind of data repository it is interacting with, including what kind of data is hosted (research domain, formats etc), what the usage licenses is, potential provenance information, FAIR vocabularies that are supported – to name some of the tested aspects. In short, the term machine-actionable boils down to "the machine knows what I mean". The term 'generic' refers to the fact that we are testing metadata terms that are established concepts for datasets such as license, metadata identifier, data identifier, searchable, persistence, outward references and more.

The goals of the evaluation are:

- to demonstrate the benefits of machine actionable metadata
- to raise awareness of FAIR related to data repositories
- to enable monitoring of repository FAIRness so that FAIRification efforts are detected and quantified

One of the primary goals of our effort is to monitor the evolution of FAIR metrics for individual data repositories in order to trace the (positive) development of FAIRer research data and, if possible, to gauge its effect on data reuse and scientific quality.

FAIR Digital Objects

A (FAIR) digital object is an element that is identified by a persistent identifier (PID) and contains metadata and data elements (see Figure 1).

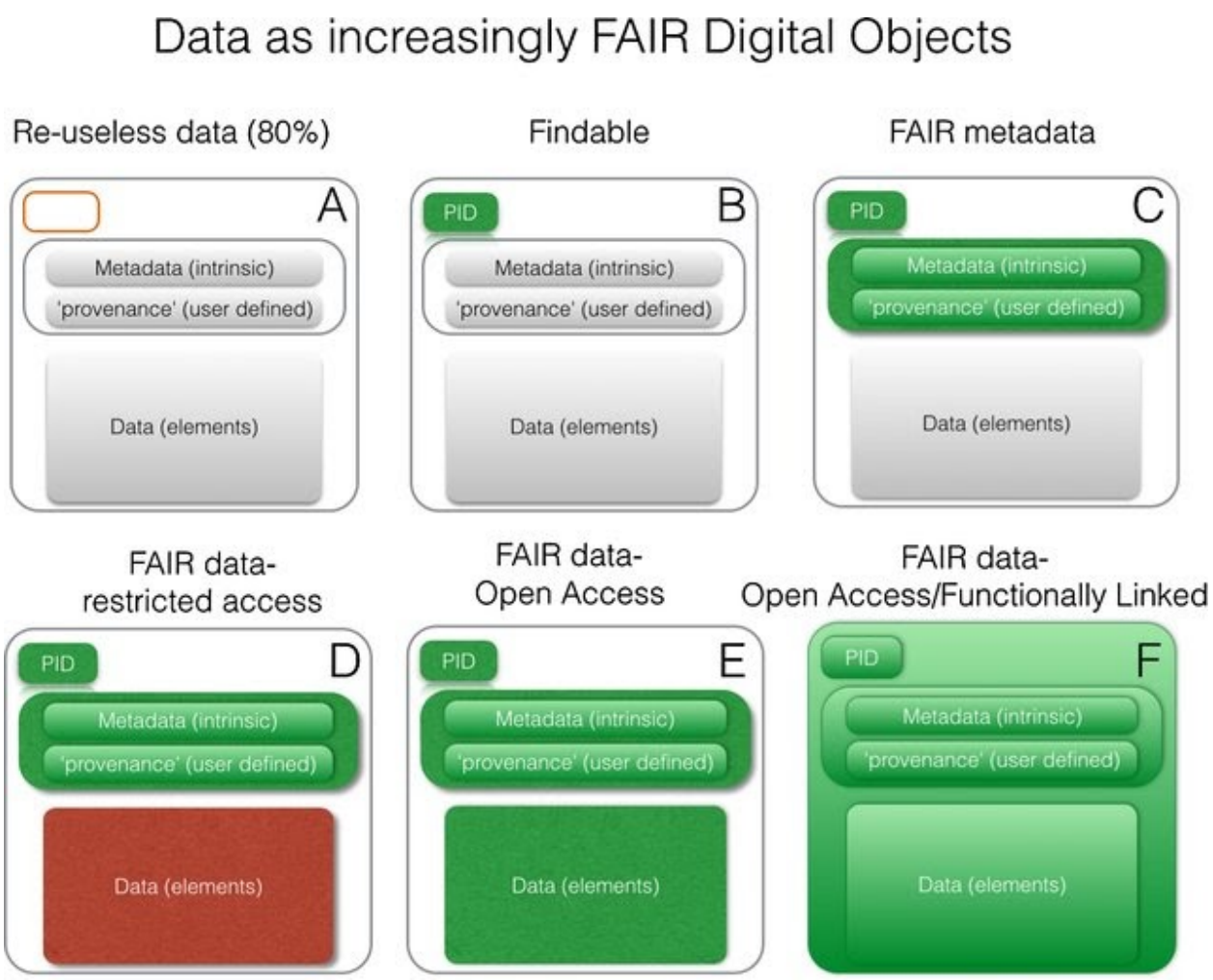

Figure 1: This schematic (from Mons et al. 2017) shows levels of FAIR digital objects (here also referred to as *datasets*). Panel A (upper-left) represents the majority (80%) of research data – such datasets lack PIDs, metadata and data are unlinked. Panel B illustrates datasets that can be uniquely addressed using a PID. In panel C the dataset has both an identifier and a set of FAIR metadata. In some cases data is restricted and unavailable to the general public (Panel D), while in panel E the dataset is a digital object that consists of a PID that contains unique identifiers for the metadata and the data elements. In the final panel F all elements of the FAIR digital object is additionally functionally linked with explicit outward references.

Evaluation methodology

The FAIR Evaluation Services is based on the FAIR principles (Wilkinson et al 2016), and aims to measure aspects of the FAIR principles using metrics (Wilkinson et al 2018). The approach and design of the FAIR Maturity evaluator is detailed in Wilkinson et al. 2019. In short, it is a fully automated and formally objective evaluator that we have chosen to employ on selected datasets from data repositories in the sample in order to avoid biases introduced by e.g. manual

/ human evaluations. Clearly the objectiveness of the evaluator only goes as far as that it only tests what it has been designed to check (the metrics or maturity indicators).

The maturity evaluations are in principle reproducible. A caveat, however, is that repositories develop continuously and there is no version control of how they are presented via the repository interface. Therefore, we archive the evaluator output. The harvester and the 22 indicators tests generate result files that we store for future reference and comparison.
In order for such results to be comparable over time, we will operate with a single version of the harvester and indicators at all times during the project.

Based on the successfully passed tests we calculate average 'FAIR scores', both a total score and for each of the F, A, I and R letters. The standard deviation for each score is also calculated using the dataset evaluations. A non-zero deviation indicates that scores vary over different datasets and one should be cautious about drawing conclusions about the repository based on the small number of tested datasets. The recommendations and technical details for each of the indicator tests are detailed in the supplemental document '*FAIRification recommendations*'.

Principles R1.2/R1.3: domain specific and provenance metadata (not part of the current indicators)

It is worth noting that the FAIR principles R1.2 ((Meta)data to be associated with detailed provenance) and R1.3 ((Meta)data to meet domain relevant community standards) are currently not being evaluated by the FAIR Maturity Evaluator. These two principles are not included because they rely on the respective science communities to define the group of predicates to indicate relevant provenance and community standards. If communities were to publish the domain specific metadata for a given community on e.g. FAIRsharing.org, we may soon see these supported test requirements being used by the FAIR Maturity Evaluator.

How to access the 'public' FAIR Maturity evaluator

We have deployed a dedicated server for executing a controlled version of the evaluator in order to maintain version control and to enable parallelisation of evaluations. We have also worked in close dialogue with the developer (Mark Wilkinson) to improve the harvester and indicator tests.

A public version of the evaluator is available online at:
https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/collections/new/evaluate

Note, however, that this may not give the very results as those run by our study due to our version being locked to Hvst-1.1.1 and mostly Tst-0.2.1 for the indicators (indicator versions vary, depending on how many times they were altered to account for improvements/fixes).

Repository and dataset selection criteria

Data repositories have been selected based on their being listed in re3data.org and associated with a Nordic+Baltic member state. Additionally we have made an internal survey among the 14 partners, listing repositories that host research data of any kind. We have excluded repositories that only host publications (typically pre-prints or reprints) and those that have no obvious potential use in science. This yielded 136 repositories associated with one or more of the Nordic + Baltic countries.

The FAIR Maturity Evaluator takes as input a global unique identifier (GUID) and explores the machine-actionable metadata provided within the dataset landing page (whatever the GUID directs to). The basic requirement for a successful FAIR machine-actionable evaluation is that the data repository provides uniquely identifiable, persistent and resolvable identifiers for datasets/data records called GUIDs (Global Unique Identifiers).

Datasets could only be identified and referenced when they were equipped with globally unique identifiers (GUIDs)[1] that point to a landing page that is dedicated to the metadata/data of the dataset. Repositories that do not provide GUIDs/PIDs for individual datasets could not be evaluated further. The minimum requirement, therefore, was a GUID[2] pointing to the landing page containing links to metadata and the data itself.

Among the 136 repositories, 36 repositories do not satisfy the selection criteria and a further 28 are discarded on account of not providing a GUID to each of the datasets. The remaining 72 repositories were possible to evaluate using the FAIR Evaluator Service and for this we randomly picked out ten datasets for each data repository (spreading the selection across disciplines when appropriate). The Evaluator was then executed on all collected datasets, approximately 710 datasets. An evaluation of a dataset consists of a metadata harvester that explores the provided GUID and all its outward references to search for machine-actionable metadata. An evaluation will take between a few minutes upto more than an hour to complete. To speed up the process we have implemented parallelisation for the execution of the evaluations with upto ten workers running simultaneously. The total execution time for the sample was a little over 4 days.

Standardised communication protocols

FAIR principle A1 states “(Meta)data are retrievable by their identifier using a standardised communication protocol”. This means that FAIR data retrieval should be mediated without specialised or proprietary tools or communication methods. This principle focuses on how data and metadata can be retrieved from their identifiers and is fundamental to the way we test and explore the FAIRness of data repositories.

---

[1] also known as persistent identifiers (PIDs)

[2] currently the evaluator supports URI, DOI, Handle and INCHI keys

Some repositories employ specific APIs to expose their metadata and data (datasets). The evaluator uses the REST interface to retrieve information from a GUID. This is based on HTTP to resolve and harvest metadata about resources. The evaluator does not support alternative protocols such as FTP, SOAP, OAI-PMH, OpenDAP etc.

Data repositories that expose metadata through other protocols than the REST interface (HTTP) will not be possible to test, as the Evaluator only operates using the REST interface.

How to interpret the results

Communities that have had their data repository evaluated by EOSC-Nordic WP4 should take the results as constructive input that measures FAIR aspects on the basis of its machine-actionability (recall that machine actionable metadata is core to the FAIR principles). The evaluation results and accompanying FAIRification recommendations can be used to guide any efforts to make a repository FAIRer.

Maturity level 0

A null score means that the evaluator could not be run on the repository due to the lack of uniquely identified datasets. Repositories that do not provide a form of GUID for each dataset can not be tested using the evaluator. Recognised GUIDs (or persistent identifiers) can be of various types, but typically URI or handle/DOIs.

Maturity level 1 (low)

These three indicators will pass by simply providing a URI and represents a unique identifier, open free protocol for metadata retrieval and metadata authentication & authorisation.
The majority of repositories fall in the category with 3 out of 22 passed tests. Although the datasets are equipped with GUIDs, there is not much trace of machine-actionable metadata compatible with the FAIR principles.

Maturity level 2 (medium)

For this medium maturity level, scores fall between 8-12 passed tests, obtaining between 35-60% on the FAIR score. Repositories at this level are employing some machine-actionable metadata and the datasets are consequently more easily discoverable and FAIR. Often their FAIRness can be improved by adopting more extensive use of FAIR vocabularies, a license predicate, identifying the digital object explicitly in the metadata to name a few.

Maturity level 3 (high)

Very few repositories fall into the high maturity level, with 13-18 passed tests, these repositories obtain 60-80% on the FAIR score, depending on how stable those results are over multiple

datasets. At the moment there are not many additional tests in the current generation of FAIR evaluation indicators that can explore the FAIRness aspects further for these repositories. The most natural advancement would likely be related to principles R1.2 and R1.3 – provenance metadata and domain specific standards.