FAIRSFAIR *in Action*

Fostering Fair Data Practices in Europe

Assessing the FAIRness of the ARCHIVER long-term data preservation services with F-UJI

As part of its R&D validation process, ARCHIVER needed to assess the FAIRness of the resulting ARCHIVER repository services. The F-UJI tool (https://github.com/pangaea-data-publisher/fuji) developed in the context of FAIRsFAIR responded to this need as it provides programmatic assessment of FAIRness of research data objects based on metrics developed by the FAIRsFAIR project, breaking it down in concrete tests that could be included in the ARCHIVER.

ARCHIVER
ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

www.archiver-project.eu

The Challenge

Digital preservation has emerged in recent years as a fast-moving and growing community of practice that is of ubiquitous relevance, but in which capability is unevenly distributed. Digital preservation in the research community has a close alignment to the FAIR principles and is delivered, albeit unevenly, through a complex specialist infrastructure comprising not simply technology but also the capacity of staff and 'know why' of policy[1].

The European Open Science Cloud (EOSC) initiative has extensively worked to promote and enable access to Open Science data with the stated aim of ensuring that researchers can maximize the value of their research processes, sharing large-scale Research Infrastructures (RIs). The importance of advanced long-term preservation to allow reproducibility of research results is emphasized by the EOSC Strategic Research and Innovation Agenda (SRIA)[2] and different reports of relevant bodies such as the Digital Preservation Coalition.

ARCHIVER, funded from the European Union's Horizon 2020 research and innovation programme, is providing a substantial contribution to this vision. Started in January 2019, ARCHIVER is a unique initiative currently running in the EOSC framework that is competitively procuring R&D services for archiving and digital preservation. The ARCHIVER tenderers were selected through an open and competitive procurement process. Between December 2020 and August 2021 three consortia worked on innovative, prototype solutions for long-term data preservation, in close collaboration with CERN, EMBL-EBI, DESY and PIC.

ARCHIVER procured R&D services that address the long-term preservation needs across the entire research data management cycle. The resulting services are sustainable and provide the needed functionality at scale that can implement FAIR Data Management Plans, using Trustworthy Digital Repositories (TDRs) certified according to best practices (e.g. ISO 16363 and CoreTrustSeal).

As part of its R&D validation process, ARCHIVER needs to assess the FAIRness of the resulting ARCHIVER repository services. The F-UJI tool (https://github.com/pangaea-data-publisher/fuji) developed in the context of FAIRsFAIR might respond to this need as it provides programatic assessment of FAIRness of research data objects based on metrics developed by the FAIRsFAIR project, breaking it down in concrete tests that could be included in the ARCHIVER .

The F-UJI team is willing to collaborate with ARCHIVER. In the broader context, FAIRsFAIR and ARCHIVER are both making significant contributions for FAIR in the scope of the EOSC.

1 Currie, Amy, & Kilbride, William. (2021). FAIR Forever? Long Term Data Preservation Roles and Responsibilities, Final Report (Version 7). Zenodo. https://doi.org/10.5281/zenodo.4574234

2 https://www.eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant Agreement No. 831558.

The implementation

The initial assessment started by gathering some basic information about the current repositories from the organisations involved in the ARCIHVER project, namely EMBL-EBI, DESY, PIC and CERN to get familiarised with the tool.

The following information was shared:

- Data domains (scientific discipline, community)
- Assessment Target (e.g. subset of data holdings)
- Data access level (e.g. if restrictions in place)
- Meta(data) dissemination (OAI-PMH, REST, Content Negotiation, Schema.org)
- Metadata standards (e.g. DDI Dublin Core, schema.org etc.)
- Semantics (SPARQL endpoint, Vocabularies)
- Data formats (e.g. discipline specific formats)

The following data sets were used for a preliminary test of the FAIR assessment tool:

EMBL-EBI

EMBL-EBI. The '1000 genomes' dataset contains 1000 human genomes, all publicly available with no restriction. The data is available from an AWS S3 bucket, s3://1000genomes.org/data. The total data volume is about 64 TB, split evenly across ~2.7K subdirectories all to be considered a good sample to practice with.

Link: https://registry.opendata.aws/1000-genomes/

DESY

DESY. An example of serial femtosecond crystallography data was taken to perform the assessment. The web page contains some metadata and links to CrystFEL Beam File, CrystFEL Geometry File, Processing Scripts and diffraction patterns, e.g. https://www.cxidb.org/data/21/cxidb-21-run0190.tar

Link: https://www.cxidb.org/id-21.html

CERN

CERN/1. The audiovisual recordings of each talk of a conference (WWW94) was subject to the test. The metadata was available in the 2/3 MARC21 format (https://cds.cern.ch/record/423168/export/xm) and contained descriptive and preservation information as well as the locations of the mp4, mov and mkv files.

Link: https://cds.cern.ch/record/423168

CERN/2. Example of a CMS collision dataset in AOD format. The dataset consists of 2916 files and is 2.7 TB large. The metadata is represented in the JSON format, following a custom JSON schema. The metadata is machine-accessible by appending ".../export/json" to the above URL. The individual files are accessible via "file indexes" that are listed in the metadata and that can be downloaded in either JSON or TXT formats from the ".../export/json" page. Each index file contains a list of individual dataset files in the ROOT format that belong to a given dataset volume. An individual ROOT data file can be downloaded over HTTP and XRootD protocols.

Link: http://opendata.cern.ch/record/1395

CERN/3. An example of a CMS simulated dataset in AODSIM format. The dataset consists of 15410 files and is 6.2 TB large. The metadata and data can be accessed as described above for the CMS collision dataset. Note that there are more than 15,000 individual ROOT data files belonging to this dataset.

Link: http://opendata.cern.ch/record/1386

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant Agreement No. 831558.

CERN/4. A simple example of an OPERA neutrino event dataset. The dataset consists of 15 files and is 8 KB large. The metadata is again accessible by appending ".../export/json" to the above URL. The location of individual data files in the CSV format is now directly listed in the metadata JSON snippet, i.e. there is no need for intermediate "index" files as was the case for large CMS collision and simulated datasets. Note that the dataset semantics expresses the meaning of CSV columns.

PIC. Fake dataset mimicking one night of raw data from the MAGIC Telescopes. The dataset contains fake data mimicking one night of data taking of the MAGIC Telescopes. The file names are realistic and named according to the MAGIC File Naming Convention. The content of the files is random binary data, except that realistic metadata is embedded at offset 64. Initially, the dataset was static. Once Demo Platform testing was ramped up, the data set were overwritten daily mimicking the arrival of the next night of observations.

Link: https://archiver01.pic.es/TSTRAW/

The F-UJI tool was installed on-prem at CERN, DESY, PIC, EMBL-EBI to be part of the testing activity of the pilot phase.

Next steps

Further testing is planned during the Pilot Phase of the project with an assessment the ARCHIVER repositories after scientific data ingestion, as part of the R&D sprint process of ARCHIVER.

The results of the assessment will be part of the degree of FAIRness evaluation of the 2 trustworthy data repositories developed by ARCHIVER.

Adopting Organisation/Body

ARCHIVER project

Which stakeholder category do you best represent?

- Research infrastructures & e-infrastructures (including ESFRI clusters)
- Repositories
- Service providers

Contact person
João Fernandes (CERN)
joao.fernandes@cern.ch
ARCHIVER project coordinator

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant Agreement No. 831558.