Technical Integration of Data Repositories: status and challenges

28/01/2020

Isabel Bernal (DIGITAL.CSIC)
Slava Tykhonov (DANS-KNAW)
Fernando Aguilar (CSIC)

Outline

- Repositories
- Previous Integration initiatives
- Technical barriers and challenges
- Potential integration

Repositories: DIGITAL.CSIC enabling Open Science

- CSIC Green Open Access/Open Data Mandate since April 2019
- Repository's Data Policy since 2014: nearly 12,000 datasets, training to CSIC researchers and librarians, support to comply with funders and journals data sharing policies
- Involvement in several EOSC / FAIR Data initiatives
- DOI assignation to datasets, software, notebooks, preprints and other outputs via DataCite
- Member of DataCite Metadata Working Group
- Current aggregation with OpenAIRE, CORE, SHARE, DataCite Search, BASE…
- Registered in Re3Data and Repository Finder

DSpace-CRIS state of art

- Publications and datasets (organized in communities and collections) supported in standard DSpace
- DSpace-CRIS extends this functionality and involves the other entities that are part of the research landscape:
  - Researchers
  - Projects
  - Organization Units (Groups, Departments)
  - Second Level Dynamic Objects
- DSpace 5.x has useful software integrations like CKAN and ORCID and data integrations (Archivematica) and can be deployed from Docker images

SSHOC Dataverse

Makes use of Dataverse software developed by Harvard IQSS

4 ERICs: DARIAH, CLARIN, EHRIS and CESSDA

Building mature infrastructure based on requirements of involved EOSC communities (Docker and Kubernetes)

Investigating sustainable governance models

Training Service Providers and institutes how to use Dataverse as a service

Different levels of repositories integration

Metadata integration

- aggregation by OAI-PMH and ResourceSync

Software integration

- data repository provides functionality delivered by another services (DSPACE-CRIS with CKAN and ORCID support)

Data integration

- controlled vocabularies support (COAR, OpenAire, LOC, FundRef)

Data archiving

- Datasets with files can be moved from one repository to another by SWORD protocol (for example, from DSpace-CRIS to the long term archive like Archivematica)

Previous Integration initiatives

- OAI-PMH, Resource sync (THIS ONE IS EMERGING, TRYING TO REPLACE OAI)
- Aggregators
- OpenAIRE
- …

Previous Integration initiatives - Metadata

OAI-PMH

Open Archives Initiative Protocol for Metadata Harvesting

6 Actions. Any metadata schema (pre-defined). Simple but enough.

Used by aggregators. Information about the resource, but not the resource itself. (there is a further development which is OAI-ORE)

Previous Integration initiatives - Metadata

ResourceSync

ResourceSync supports synchronization of both Resources and Metadata about Resources with the relationships clearly indicated.

Tracks the status of a resource (updated, deleted).

Access to the resource.

Aggregators - Integration of repos

OpenAIRE explorer

CORE

Google Dataset Search

DataCite Search

SHARE

OpenAIRE

OpenAIRE has grown through a series of project phases funded by the European Commission: from the DRIVER projects to link Europe's repository infrastructure, to the first OpenAIRE project aimed to assist the EC in implementing its initial pilot for Open Access (OA) to publications.

Services:

- Explorer (aggregator)
- Monitor (compliance with EC Open Access/Open Data Mandate)
- Scholix - Scholexplorer (links papers-datasets)
- AMNESIA
- DMPs

WP3.3 FAIR data integration

- Thematic services usually keeping data inside of its own database, not storing in some data repository with persistence identifiers. Data derivatives often located in the temporary storage, aren’t sustainable and can be cleaned (Findable)
- Metadata schema varies in the different scientific communities, should be supported by all data repositories, however citation block should be common (Accessible)
- controlled vocabularies (CV)/ontologies are different, should be standardized and preferably provided by common CV services maintained by EOSC (Interoperable)
- most of metadata contain descriptions in the native language of researcher, should be translated to English and linked to common CV (Interoperable)
- most of Thematic services don’t have provenance information exposed as PROV-O or other standard required to be stored in datasets together with metadata and data files (Reusable)
- different types of licenses for data access should be supported by all EOSC data repositories, sensitive data should be handled differently (Reusable)

WP3 Software integration challenges

- The maintenance of the distributed applications with external services is very difficult and expensive
- requires the highest level of service maturity
- increasing the **code coverage** does not necessarily lead to more **functionality coverage**
- writing integration tests even more important than adding more unit tests
- it’s almost not possible to run distributed services without help from community

We need Docker to increase the maturity of the infrastructure

Docker advantages

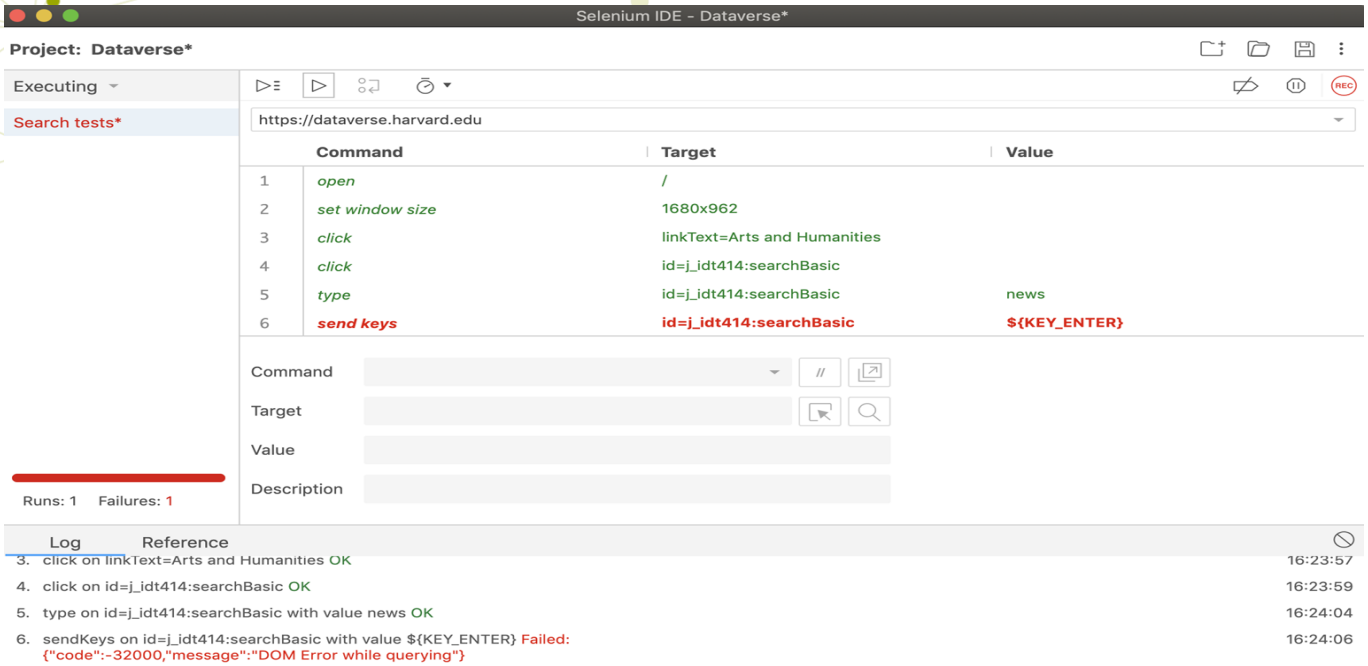
- Faster development and deployments
- Isolation of running containers allows to scale up apps
- Portability saves time to run the same image on the local computer or in the cloud
- Snapshotting allows to archive Docker images state
- Resource limitation can be adjusted
- Increasing reproducibility

SQA Service maturity requires Docker on Kubernetes!

Still, there are some technical barriers...

- The maturity of software and services is different, common baseline in the development state (WP3)
- Maintenance of the ecosystem consisting of variety of tools and services is complicated and consumes a lot of human resources (upgrade of servers, bug fixing, security updates)
- fast technological development requires continuous training and knowledge transfer

Service for Quality Assurance



Selenium IDE allows to create and replay all UI tests in your browser

Shared tests can be reused by community to increase reproducibility

SQA for the service maturity = unit tests + integration tests

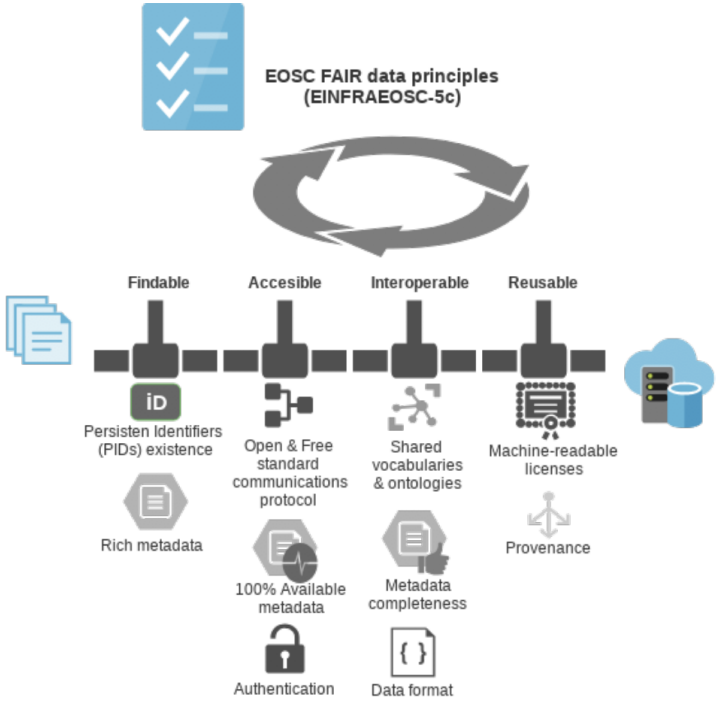
WP3.3 FAIR baseline on the technical integration

Integration of data repositories in EOSC means:

- Findable: data have to be findable in a standard way by the users and other core and thematic services
- Accessible: data accessible via standard interfaces supported in EOSC
- Interoperable: data combinable with other data in EOSC repositories
- Reusable: data exploitable by the EOSC services enabling data analysis to be performed using resources and services from the EOSC infrastructure providers, keeping datasets versions and provenance.

(from EOSC-Synergy proposal)

WP3.3 FAIR adoption



Evaluation of FAIRness of data:

- check if PID exists (F)
- metadata accessible via standard protocols (A)
- shared vocabularies used (I)
- machine readable data contain provenance information (R)

Outcome: “EOSC-ready” badge should be released by SQA service for all other services suitable to be delivered under EOSC

WP3.1 Maturity of thematic services

Thematic services should:

- follow common SQA baseline for Software and Services
- increase the maturity by adding unit and integration tests, preferably by community
- improving Cloud infrastructure both horizontally and vertically requires a good testing strategy

Potential Integration

~~Integration of repositories metadata - data~~

Enrich repositories services ✓

Focused on:

- FAIRness
- Metadata and data formats flexibility
- Connection to software
- Connection to computing

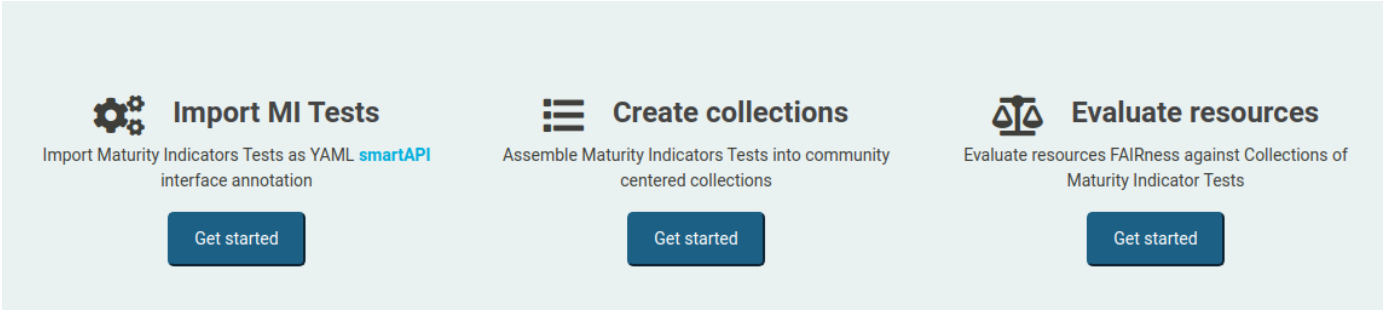
Examples - FAIRness evaluation

WP3 - WP4

Automatic or semi-automatic FAIRness evaluation

Based on recommendations (e.g. FAIRsFAIR)

Some solutions available: #FAIREvaluator



Example - Jupyter integration

- Search and retrieve data/software from repositories.
- Ensure reproducibility.
- Workflows
- Publication of results.
- Etc.