N web servi

EUROPEAN CLINICAL RESEARCH
INFRASTRUCTURE NETWORK

eXtreme DataCloud

XDC WP2 ECRIN

- Status of data acquisition and data transfer from ECRIN to to Onedata/INFN –

Authors: Goryanin (ECRIN), Ohmann (ECRIN)
Version: 1, final
Date: 24 April 2019

1. Acquisition of metadata from data sources

Responsible: ECRIN

Depending on different data sources, the collection of data may vary. For the demonstrator, original metadata were successfully extracted by ECRIN from the following data sources (status 23.04.2019):

Data source	Type of data source	Metadata schema	Metadata down-loadable	No of records	Time period covered	Size	Last extraction date
ClinicalTrials.gov	Trial registry		Download the DB	296657	2009-2019	4.4 Gb	18.2.2019
WHO ICTPR	Trial registry	None	crawling web-pages content	113547	2009-2018	44.3 MB	23.01.2019
Pubmed	Bibliographic DB	Dublin Core elements	.XLM data files and API-services OAI-MPH	474523	2011-2019	1.6 Gb	23.2.2019
ZENODO	Generic repository	DataCite with enrichments	RESTful API-services OAI-PMH	117392	2010-2018	396 MB	23.2.2019
Bio-LINCC	NIH-Repository	none	No, metadata	390	2008-2018	5.4 MB	21.1.2019

			provided by data source (.CSV data files)				
Data Dryad	Repository of datasets	Dublin Core Metadata Initiative Abstract Model (DCAM)	RESTful API-services OAI-MPH	4675	2012-2018	27.4 MB	16.12.2018
WWARN	Repository of datasets	None	crawling web-pages content	72	2010-2018	45 KB	21.1.2019
Edinburgh DataShare	Institutional data repository	Modified Dublin Core Metadata Schema	OAI-PMH	2198	2010-2018	34.6 MB	23.2.2019

Retrieved metadata are stored as JSON objects on OneData (to make them easy to use by ElasticSearch component) with the original structure (MySQL, PostgresSQL) (1 data source = 1 database) on the following servers – Testbed server, provided by ReCaS Bari.
The original metadata will be real-time updated once a week.

2. Mapping metadata to ECRIN metadata schema

Responsible: ECRIN
The metadata acquired from the different data sources were mapped to the ECRIN metadata schema (published on Zenodo). For this process the actual version of the ECRIN metadata schema was used (Canham, personal communication, 11 February 2019). Two JSON templates originating from this metadata schema have been created: one for studies and one for data objects related to a study. The structure of the JSON objects is displayed in the appendix.

Additional task: 'Metadata mapper' tool
To make the process of metadata converting easier and automatic, ECRIN developed a converting tool – 'Metadata mapper'. To date 8 data sources have been acquired (original metadata from 8 data sources were imported) and for 5 data sources metadata were manually mapped, but in the future, assuming to work with tens of data sources, manual mapping process will take a lot of time. Therefore the mapper tool seems to be the best solution, in order to have a simple user interface which automatically converts metadata from different repositories to the ECRIN schema.

3. Pumping metadata into OneData

Responsible: ECRIN
By requesting OneData RESTful API (link: link: https://onedata.org/#/home/api/latest/oneprovider?anchor=section/Overview/API-structure) it is possible

to upload metadata in JSON format with a single structure to OneData platform, where the metadata will be available for ElasticSearch (Preparator: INFN) and users via web interface (Preparator: OneData).

In OneData data files and the metadata belonging to a data file are stored. For the ECRIN use case the data files are kept empty and only the metadata are uploaded.

To date, OneData and INFN have data from the following data sources: CT.gov, WWARN and BioLinCC. Total amount of records: 296657 studies and 298927 data objects.
It is planned to send a second package with more data sources latest in June 2019 to Onedata/INFN for integration into the demonstrator.

4. User interface

Responsible: OneData & INFN

OneData, together with INFN is developing the user interface and the platform according to the requirements and the mock-ups. The demonstrator will cover all data sources imported into Onedata/INFN. ElasticSearch contains the algorithms to search metadata

Appendix:

JSON object for study
JSON object for data object belonging to a study