Virtual definition of data sets according to RDA recommendations

EOSC-Pillar
Coordination and Harmonisation of National &Thematic Initiatives to support EOSC

USE CASE

EMAIL
javier@gfz-potsdam.de

COMMUNITIES
Geosciences

June 2022

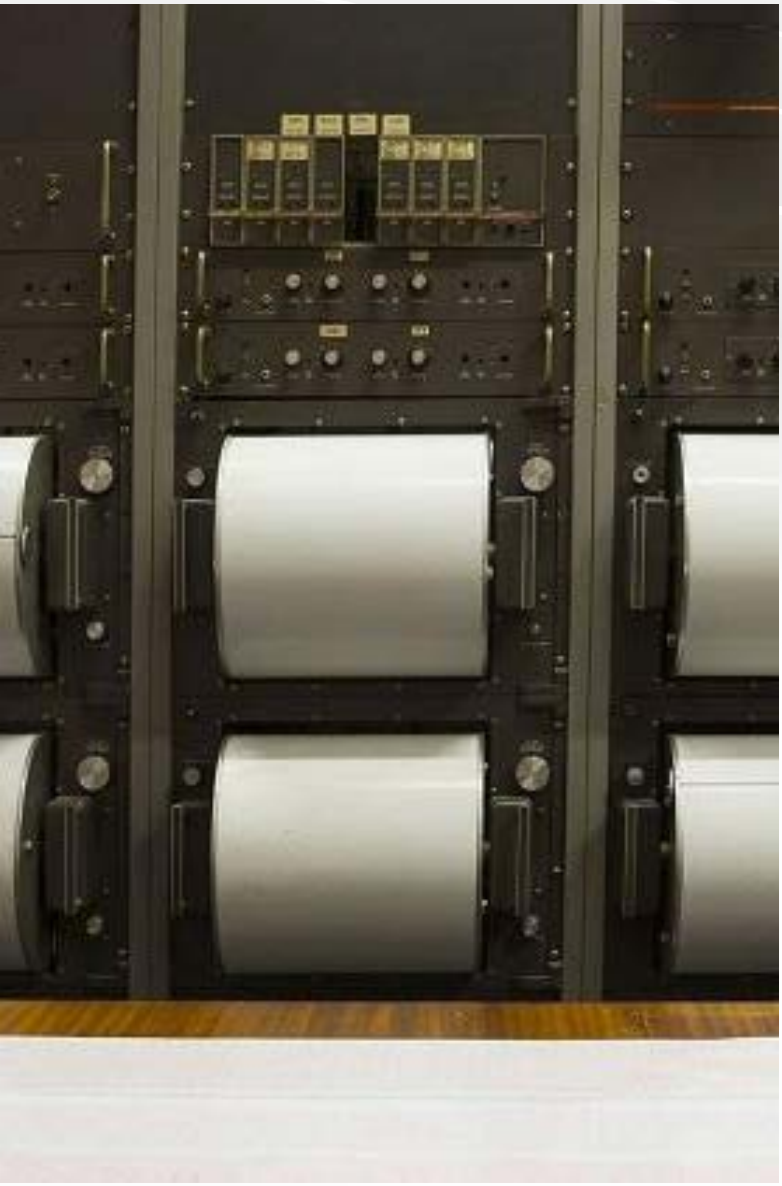
PARTNERS

GFZ Helmholtz Centre Potsdam
KIT Karlsruhe Institute of Technology
DKRZ Deutsches Klimarechenzentrum
Fraunhofer

Introduction

At GEOFON data centre is very difficult to offer big pre-assembled datasets to be downloaded, due to the resources needed for their storage. In this context, the idea of using a Data Collections System (DCS) in order to define and save this type of dataset is very appealing, because we can define collections containing only "pointers" (e.g. PIDs, URLs) to the files which are included. This implies almost no extra storage, as only the pointers are saved.

Therefore, we implemented this DCS based on an extended version of RDA WG specificiation on research data collection and make it generic enough and ready to be adopted by different communities within EOSC. Currently, 6000+ collections and 1.5+ million members were defined in the internal service.



Challenges addressed

This use case aims at making the Data Collections System, designed within the activities of RDA, generic enough and ready to be adopted by different communities within EOSC.

The system aims at :

- Building aggregations or data collections of different data objects as an essential element in the Research Data Management practice.
- Describing the research data collections in a standardized way to make them actionable by automated processes in order to be able to cope with the increasing amounts and volumes of data.

Benefits through EOSC-Pillar

This use-case relies on – and demonstrates the benefits of – EOSC-Pillar services as follows:

- Improve the current system by revisiting the RDA specifications based on the feedback we received from our partners in EOSC-Pillar.
- Deploy the service in production and foster its usage within communities partaking the project,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857650.

Virtual definition of data sets according to RDA recommendations

EOSC-Pillar
Coordination and Harmonisation of National &Thematic Initiatives to support EOSC

USE CASE

Highlights

Achievements as of May 2021:

- MS28 Code for the improved Data Collection System.
- We revised the requirements from the seismological community for such a system.
- An implementation following this specification had been in use for some time at GEOFON, where more than 6000 collections and 1.5 million members for datasets had been pre-defined by the data centre operators. However, this system was only of internal use.
- After revising the requirements, we modified the system in order to make completely generic and ready to be tested by other communities.
- One of the best aspects, regarding the resources needed to put the service in production, is that the members of a collection are identified by PIDs (DOIs, ePIC), what means that almost no space is needed to define them. Within the context of this project, we added the capability to identify resources by URL, making the system independent from a Handle server in case that some resources (or the collection itself) needs to be identified.
- The identification of improvements to the current system by revisiting the RDA specifications is already advanced.
- A first set of requirements were collected from communities and other Use cases of the project. In particular, from Climatology, taking DKRZ as an example, and the Federated FAIR Data Space (F2DS).

Next steps:

- Deploy Data Collection System as a service to be evaluated how feasible is to open it to more communities.
- Contact other communities to foster usage of the system.
- Implement extensions and improvements to the RDA Recommendations. For instance, the automatic export to the Federated FAIR Data Space developed.
- Demonstrate the service in production.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857650.