CEPLAS-ARC-Finder – a simple, locally-deployed tool to find your peer's research data

Konzept für das Projektmodul (Modul 9) des Zertifikatskurs FDM (15.07) 2021 / 2020
Teilnehmer: Dominik Brilhaus
ORCID: https://orcid.org/0000-0001-9021-3197
Email: brilhaus@hhu.de

Motivation

Research is a highly collaborative endeavour that builds on synergistic interaction between different stakeholders enabled by efficient knowledge exchange. Gaining a prompt overview of the ongoing research efforts – both pre- and post-publication – is oftentimes hindered (for social, legal or technical reasons) even between parties of spatially closest and well trusted surroundings of a collaborative consortium such as the Cluster of Excellence on Plant Sciences (CEPLAS[1]). The key to enable discussion on and exchange of research data is *findability*, the first layer of the FAIR principles of data stewardship. The project presented here aims to address this layer, by making CEPLAS research easily findable and visible amongst CEPLAS researchers and showcase the beauty and ease of data sharing to spike fruitful collaborations with peers.

State of the art

Research data management within CEPLAS is closely aligned with DataPLANT[2], the NFDI consortium for plant sciences. DataPLANT has developed the Annotated Research Context (ARC[3]), a directory structure for research objects. Annotation of research data in the ARC is based on the metadata schema ISA[4] (for investigation – study – assay). Serialized in spread sheet format as *ISA-tab* this allows intuitive, flexible and yet structured and conclusive metadata annotation of the versatile data types produced in plant sciences. ARCs are git[5] repositories that can be shared via DataPLANT's DataHUB[6], a customized GitLab[7] instance with a federated authentication interface to allow controlled access across institute borders. Although the ARC environment is continuously being developed, the choice of these key technical pillars are set: (a) ARC as the structure, (b) ISA as the metadata language, (c) git as version control logic and (d) gitlab for ARC collaboration and user management. This allows to leverage the ARC and develop at least intermediate solutions for data findability, knowing that time and efforts are well-invested, since both (meta)data inputs in as well as secondary outputs dependent on the ARC will be adoptable and migratable in the future.

Approach

This project focuses on metadata at the highest project and least sensitive (i.e. ISA's "investigation") level to minimize user input or possible discomfort with data sharing and will be achieved in four concerted, but independent modules of metadata

1. collection,
2. retrieval,
3. restructure, and
4. representation.

First, metadata is collected – manually or supported by automation – in the ISA investigation spread sheet, packaged in ARCs and submitted to the DataHUB by individual volunteers. Here, access to the ARCs can be controlled to share them publicly or with invited collaborators. The CEPLAS-ARC-Finder selectively retrieves, downloads and dumps the metadata locally on the user's machine. The CEPLAS-ARC-Finder then restructures the metadata into a simple spreadsheet-based database. From the database the investigation data is finally read and represented by a user interface that enables finding the data available to the individual user.

[1] https://ceplas.eu "CEPLAS"
[2] https://nfdi4plants.de "DataPLANT"
[3] https://github.com/nfdi4plants/ARC "ARC specifications"
[4] https://isa-tools.org/ "ISA Metadata Schema"
[5] https://git-scm.com/ "Git"
[6] https://git.nfdi4plants.org "ARC DataHUB"
[7] https://gitlab.com "GitLab"