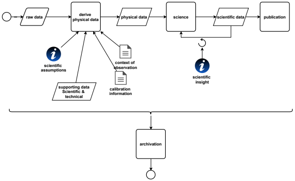
RIAF — a Repository Infrastructure that Accommodates Files

Daniel Mohr, Björn Brötz

abstract

RIAF (dlr-pa.gitlab.io/riaf) is a repository infrastructure to accommodate files. It enables to hold the data with the FAIR principles (oceanrep.geomar.de/id/eprint/55269, see also fair-principles www.go-fair.org/fair-principles).
RIAF is designed to enable provenance and reproducibility of the research data in the early part of the data life cycle, i. e. prior to publication. It further is designed to enable checks on metadata relevant to research data management as defined e.g. in a machine actionable data management plan (maDMP).
This new concept of using CI pipelines for research data allows interesting features. The server could create cryptographic timestamps to inhibit silent changes of the history. Research data management can define relevant checks on metadata. From given metadata a public accessible landing page can be created.
In our concept most data is stored in a repository and can be easily distributed. This allows the data genesis in a private environment (e. g. aircraft, campaigns, ...) without network access and later share the data using a central server instance. Also already during data genesis (e. g. raw data, physical data, scientific data) the possibility to share data and track changes is given. And in the end after preparing a publication the data can be transported to a public data repository.
The primary focus is to work as an in-house solution to handle digital assets. It should be possible to use the data without downloading a complete digital asset.

idea



- create data (e. g. measurement)
- store data in a good way
- analyze data, share data (in-house)
- create derivative work
- preparation of publication of derivative work (e. g. paper or data publication)
- create publication of derivative work
- archive: data, analyzed data and publication
- reuse of data and/or analyzed data

open source software

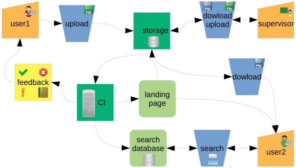
For this purpose we use open source software in a composability design (Unix philosophy):

- gitolite
- sskm
- git
- WebDAV
- git-annex
- OpenSSH
- apache
- GitLab

Further we implemented some additional open source software:

- fuse_git_bare_fs
- gitolite_web_interface
- pydabu

example use case



usage of CI

Using CI pipelines is a typical concept in software engineering to run automatic tasks (e. g. tests, deployment). This is triggered by a developer with new code. Here we want to trigger automatic tasks by new or adapted (research) data:

- cryptographic timestamp on server inhibits silent changes of the history
- checks on metadata relevant to research data management as defined e. g. in a machine actionable data management plan (maDMP) (interoperable, reusable)
- landing page (accessible)
- put data in search engine data base (findable)

roadmap

- requirements analysis, 2021 (done)
- create testing infrastructure, 2021 (done)
- create testing environment inside DLR, 2021 (done)
- develop possible solution(s), 2021 (done)
- design RIAF, 2021-2023 (work in progress)
- create RIAF documentation, 2021-2023 (work in progress)
- test gitolite infrastructure, 2022 (done)
- test GitLab infrastructure, 2022 (work in progress)
- gitolite-trigger, git hook, GitLab file hook, 2021-2022 (work in progress)
- search engine for metadata, 2022 (on going)
- maDMP check, 2022-2023 (done)
- PA user testing phase, 2022 (on going)
- PA user beta phase, 2022-2023 (on going)
- PA production, 2023-2024

further readings

- technical manual 'RIAF': dlr-pa.gitlab.io/riaf
- technical report 'An interpretation of the FAIR principles to guide implementations in the HMC digital ecosystem': https://doi.org/10.3289/HMC_publ_01
- software 'fuse_git_bare_fs': dlr-pa.github.io/fuse_git_bare_fs
- software 'gitolite_web_interface': github.com/dlr-pa/gitolite_web_interface
- software 'pydabu': dlr-pa.gitlab.io/pydabu

outlook

During the project progression manifold ideas arise that can be the basis for future work.

- Storing research data in self-contained/impartial/decentralized and federated systems like GitLab instances (maybe GitHub) using extensions for large data.
- Using CI pipelines to check suitability.
- Using CI pipelines to push landing page including metadata to centralized platform.

DLR
Deutsches Zentrum
für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft