Building a research data archive - what flavour should it be?

Adil Hasan, Sigma2 AS Norway
Philipp Conzett, UiT The Arctic University of Norway

Datataverse Community Meeting 2022

Outline

- Sigma2, NIRD, and the Archive2021 project (Adil)
- Key findings from the concept phase of Archive2021 (Adil)
- Useful insights for the Dataverse community? (Philipp)
- Q&A and discussion (all)

Sigma2, NIRD, and the Archive2021 project (Adil)

- Current archive been running since 2014
  - More than 600TB of data spread over 12M files. Minimal metadata schema based on Dublin Core.
- Challenges have motivated the need for a new archive:
  - Current software derived from solution used by Comp Bio community who have now moved away from s/w. No local expertise in the s/w. Metadata doesn’t support all use-cases (collections of datasets)
  - Support for Open Science and FAIR data support essential
- Archive2021 project aims to learn from other archives and experts to provide a solution that takes into account best practices, supports open science and is adaptable to change.
- More info on the project: https://www.sigma2.no/project-new-research-data-archive

Key findings from the concept phase of Archive2021 (Adil)

- Researchers familiar with data management
    - Would like to have archiving part of their data management process.
    - Would like to archive detailed metadata to make data discovery and reuse easier.
    - Would prefer not to have to enter metadata manually, or enter same metadata in multiple locations.
    - Would like to have domain-specific applications interfaced to the archived data to make reuse easier.
- =>
    - would very much appreciate rich APIs to archive data.

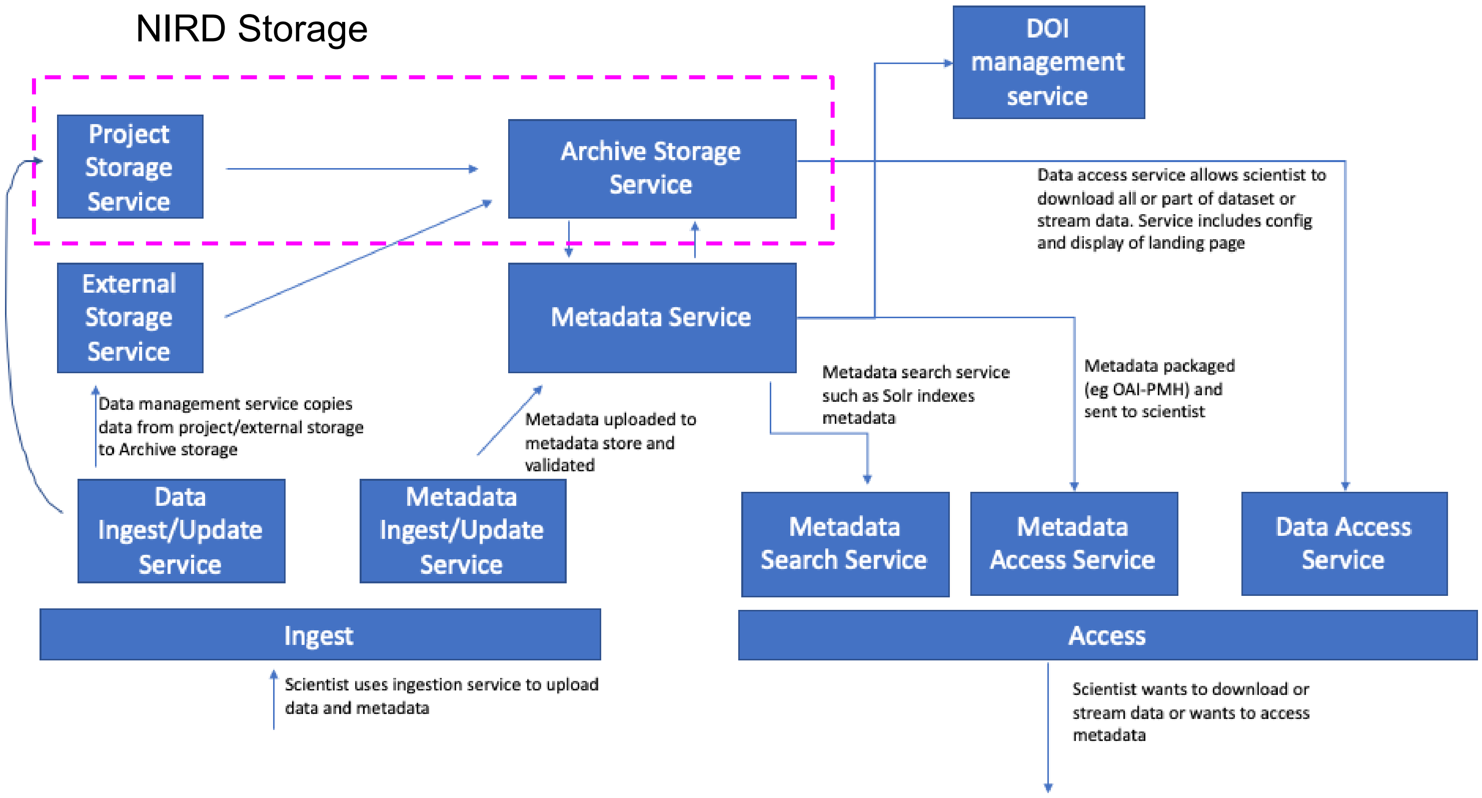
Key findings from the concept phase of Archive2021 (Adil)

- Researchers unfamiliar with data management:
  - Would like an easy-to-use interface to archive their data.
  - Either support for metadata standards for their domain, or a generally approved standard.
  - Ideally would like not to have to fill in lots of metadata.
  - Definitely need guidance on what information to supply, license to choose, etc.
- =>
  - Very much would prefer a web-interface wizard for archiving their data.

Key findings from the concept phase of Archive2021

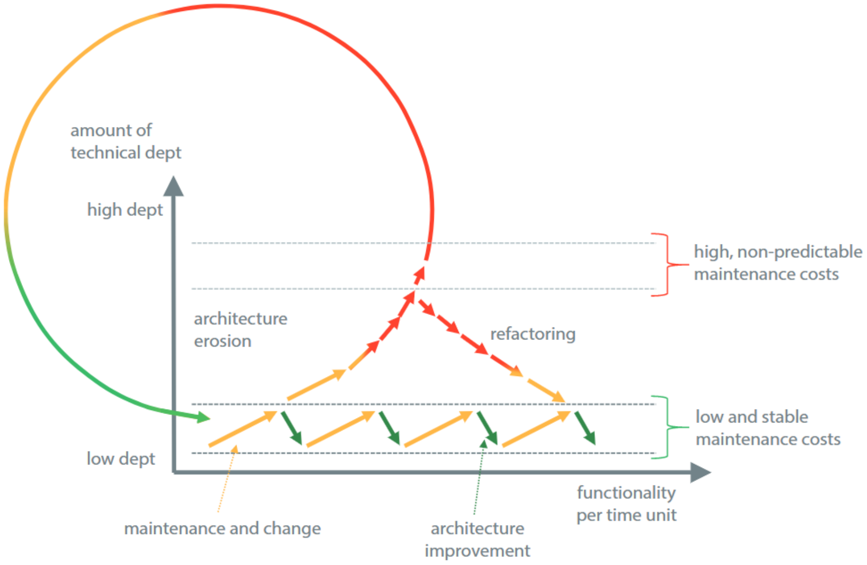
- Current archive main users have large volumes of data and established data management procedures.
- New archive must have rich API and be able to integrate with existing NIRD storage.
- No existing solution currently offers support for large data volumes and easy integration with existing infrastructure.
  - Looked at Dataverse, CKAN, Invenio, Domain-specific services
- Have decided on a component based approach:
  - Each component is good at one activity and can be replaced by better ones when they come along.
  - Everything changes (current archive is on 3rd iteration of storage), so system needs to be able to adapt to change.

Key findings from the concept phase of Archive2021

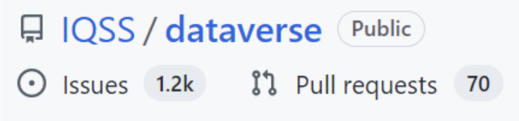

NIRD Storage
DOI management service
Project Storage Service
Archive Storage Service
External Storage Service
Metadata Service
Data access service allows scientist to download all or part of dataset or stream data. Service includes config and display of landing page
Metadata search service such as Solr indexes metadata
Metadata packaged (eg OAI-PMH) and sent to scientist
Data management service copies data from project/external storage to Archive storage
Metadata uploaded to metadata store and validated
Data Ingest/Update Service
Metadata Ingest/Update Service
Metadata Search Service
Metadata Access Service
Data Access Service
Ingest
Access
Scientist uses ingestion service to upload data and metadata
Scientist wants to download or stream data or wants to access metadata

Useful insights for the Dataverse community? (Philipp)

- Continuous **growth** of Dataverse
  - installations and community
  - needs
  - feature requests



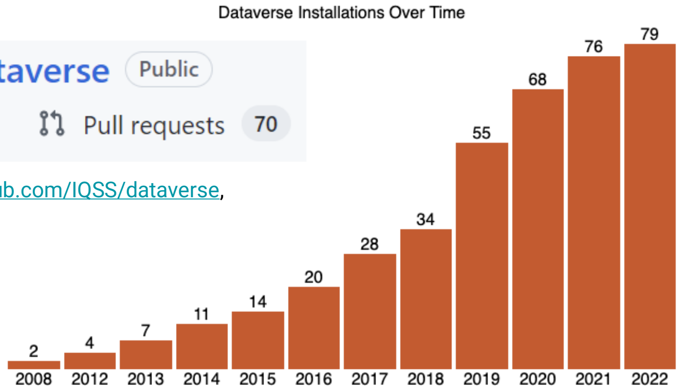
(Source: Lilienthal 2019, p. 5. All rights reserved dpunkt.verlag)



IQSS / dataverse Public

Issues 1.2k Pull requests 70

(Source: https://github.com/IQSS/dataverse, June 9, 2022)



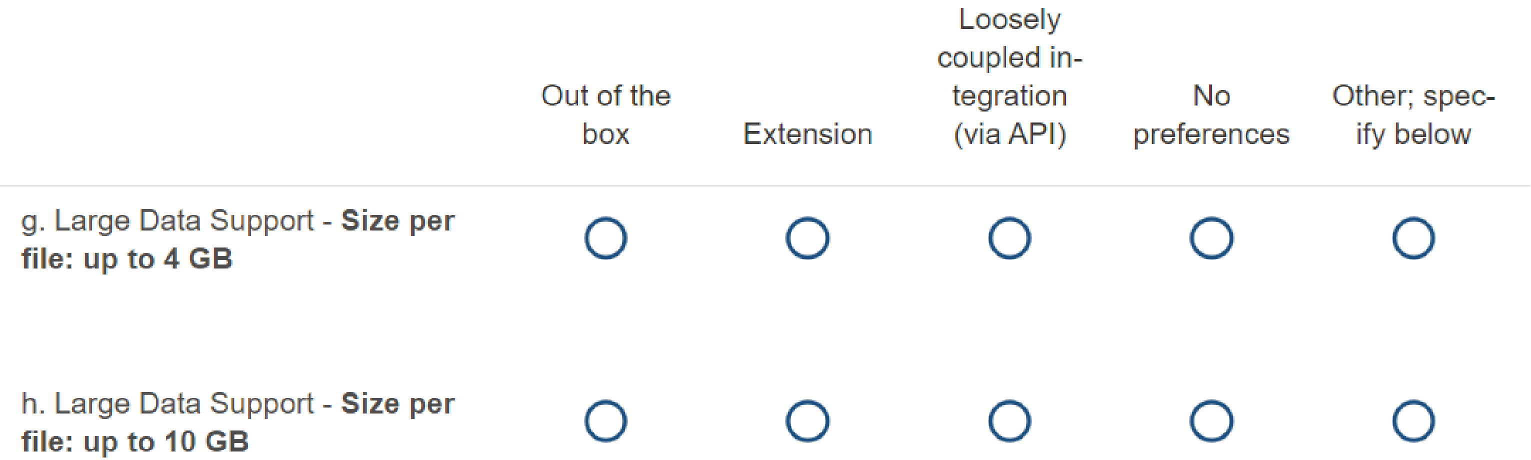
(Source: Phil Durbin on Dataverse Slack, June 6, 2022)

- Need for strengthening the **sustainable development and maintenance** of Dataverse software and ecosystem of associated tools and services

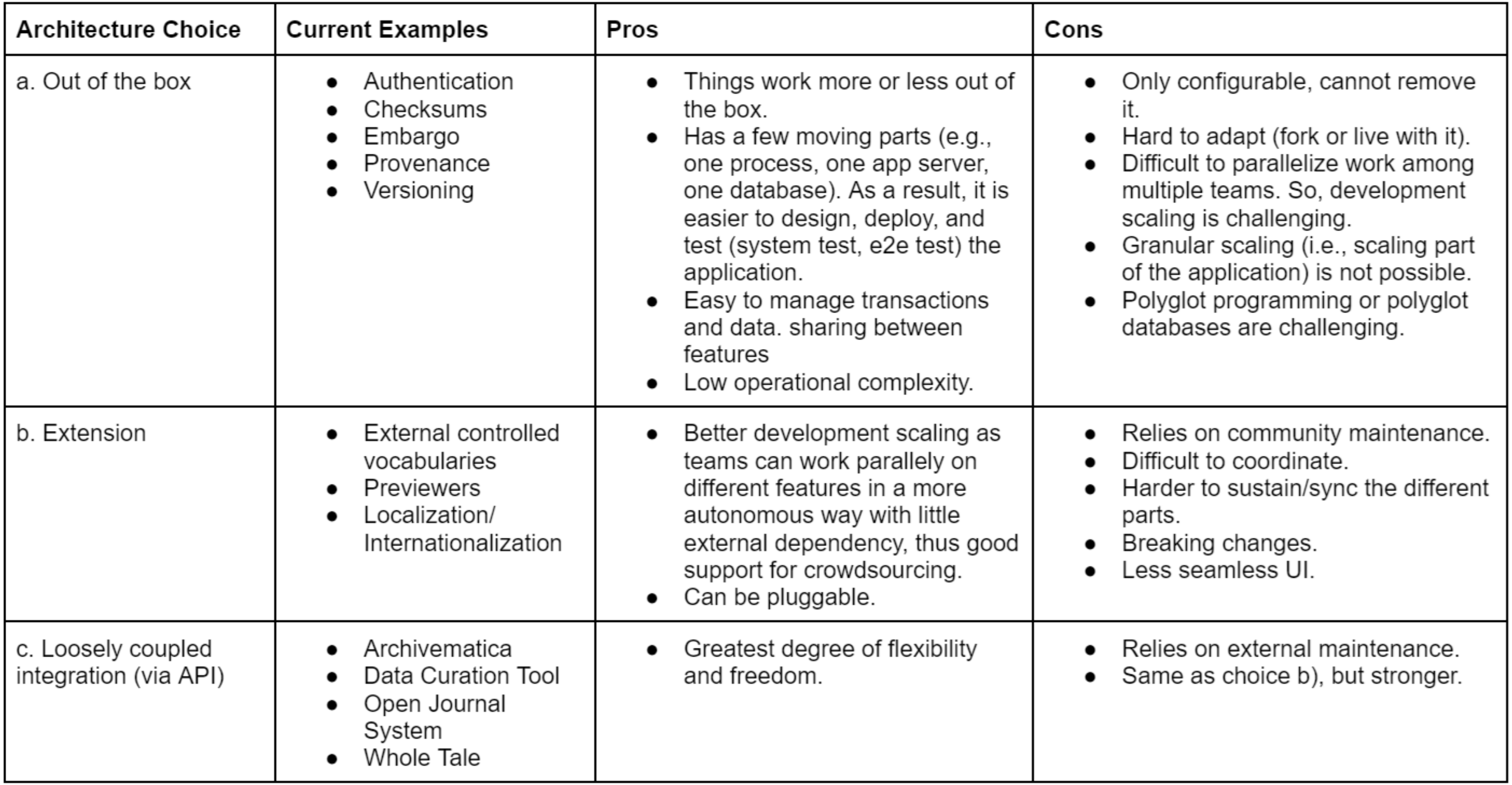
- Goal: avoiding/mitigating **technical debt**

Useful insights for the Dataverse community?

- First step: **Dataverse Community Survey 2022**
- Aim: Mapping **roadmaps and priorities** of installations
- Section about repository features asks about preference of **software architecture choices**:



	Out of the box	Extension	Loosely coupled integration (via API)	No preferences	Other; specify below
g. Large Data Support - **Size per file: up to 4 GB**	○	○	○	○	○
h. Large Data Support - **Size per file: up to 10 GB**	○	○	○	○	○



Architecture Choice	Current Examples	Pros	Cons
a. Out of the box	● Authentication ● Checksums ● Embargo ● Provenance ● Versioning	● Things work more or less out of the box. ● Has a few moving parts (e.g., one process, one app server, one database). As a result, it is easier to design, deploy, and test (system test, e2e test) the application. ● Easy to manage transactions and data. sharing between features ● Low operational complexity.	● Only configurable, cannot remove it. ● Hard to adapt (fork or live with it). ● Difficult to parallelize work among multiple teams. So, development scaling is challenging. ● Granular scaling (i.e., scaling part of the application) is not possible. ● Polyglot programming or polyglot databases are challenging.
b. Extension	● External controlled vocabularies ● Previewers ● Localization/ Internationalization	● Better development scaling as teams can work parallely on different features in a more autonomous way with little external dependency, thus good support for crowdsourcing. ● Can be pluggable.	● Relies on community maintenance. ● Difficult to coordinate. ● Harder to sustain/sync the different parts. ● Breaking changes. ● Less seamless UI.
c. Loosely coupled integration (via API)	● Archivematica ● Data Curation Tool ● Open Journal System ● Whole Tale	● Greatest degree of flexibility and freedom.	● Relies on external maintenance. ● Same as choice b), but stronger.

Useful insights for the Dataverse community?

- Some of the key findings from the Sigma2 Archive2021 project, together with some of the results from the Dataverse Community Survey will be useful input for community discussions about how to strengthen the sustainable development and maintenance of Dataverse software and ecosystem of associated tools and services.

Q&A and discussion (all)

References

Lilienthal, C. (2019). *Sustainable software architecture: Analyze and reduce technical debt* (1st edition.). dpunkt.verlag.