Archive II

19. desember 2013

UNINETT

What is an archive?

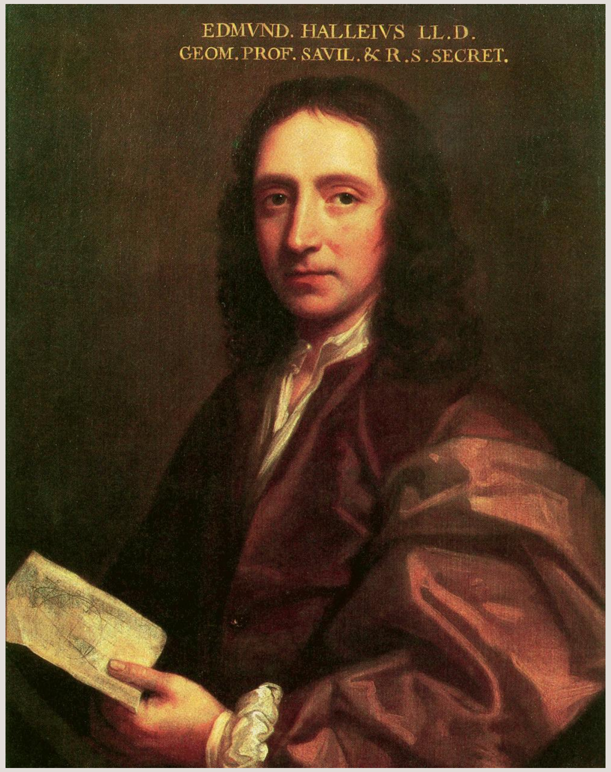
- Is a service that provides long-term storage and access of data.
- Long-term usually means ~5years or more.
- Archive is strictly not the same as a backup.
  - Backup is a snapshot of data that may change over time (eg Tuesday’s backup of file X != Wednesday’s backup).
  - Once the data reaches a mature state (ie doesn’t change) then we talk about archiving the data.

What data should I put in an archive?

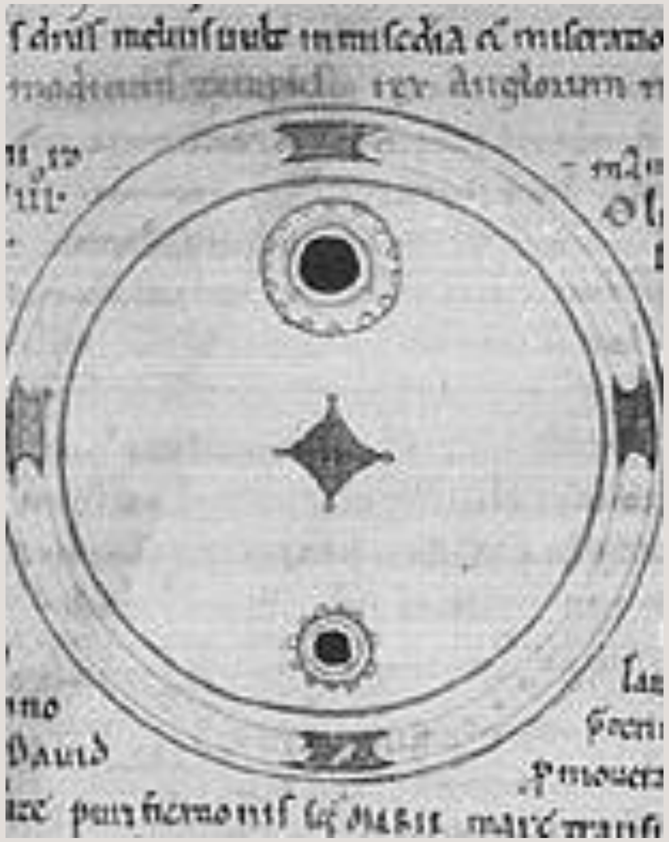
- Data that has matured can be a candidate for an archive.
  - Means data will not be modified (ie is considered ‘closed’ or ‘complete’).
- Data on which a research work has been published either directly or indirectly.
- Data that is considered to be valuable to the community.
- Data that cannot easily be reproduced (either because of resources required, or environment being unique).

Why archive?

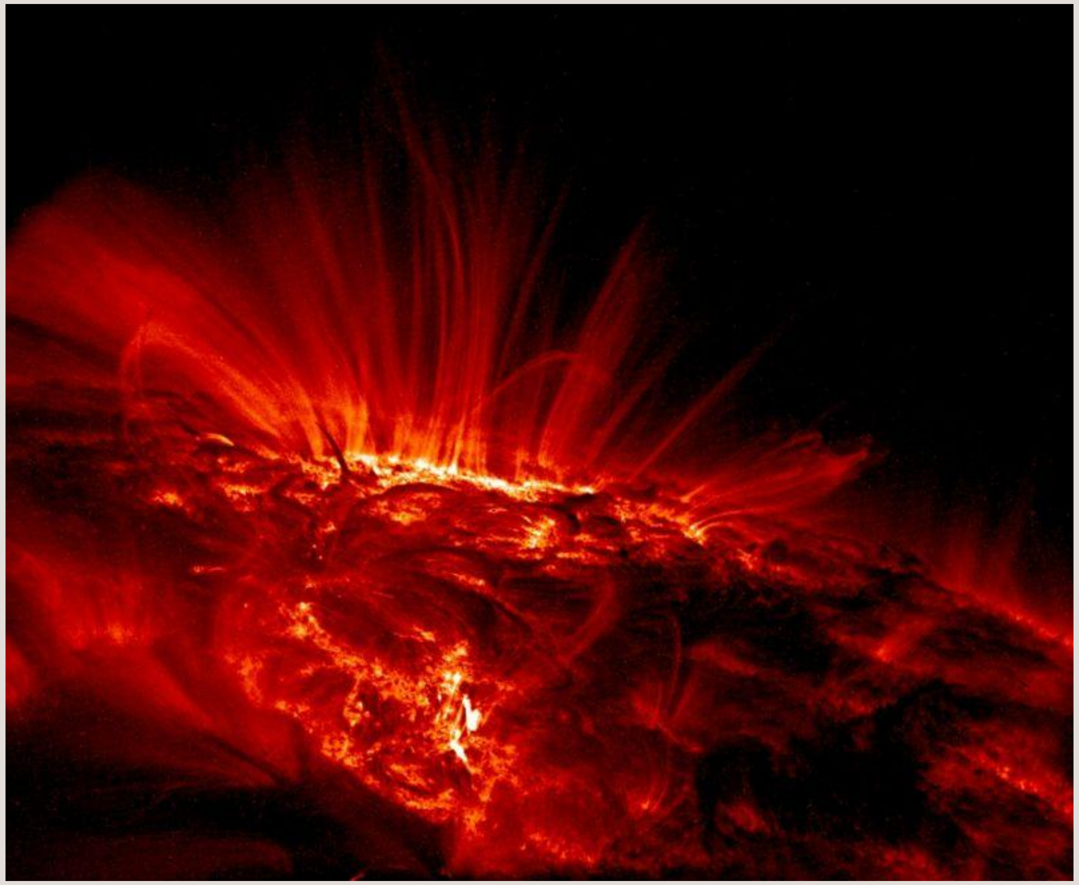
- Edmond Halley (C18) used historical data to determine trajectory of comet and provide validation of Newton's theory of gravitation.
- Due to period of comet being ~70 years historical data essential.
- A nice example of historic data being used for different purpose than originally intended.



Why archive?



- Sunspots guide models describing the nature of the Sun.
- Long history of sunspot observations (initially had religious significance).
- Utilizing historical data helps development of geophysical models.
- Another example of data collected for one purpose being used for a different purpose



Why archive?

- Many examples of data that was collected with one purpose in mind, but later reused for a totally different purpose.
  - One person’s background is another person’s signal.
- Very difficult to anticipate just what purpose the data you collect today will be used for in the future.
- How can we accommodate these unknown purposes?

Roles

- The Norstore archive recognises 5 different types of user: **Creator**, **Contributor**, **Data Manager**, **Rights Holder**, **Access User**.
- All types can be a person or an organisation (although in the case of an organisation a contact person is needed).
- You are not required to define the **Access Users** (unless you want to restrict access to the data).
- It is possible that the different types can resolve to the same person or organisation.
- It's important to assign these roles to the dataset in case of questions.

Creator & Contributor Roles

- A person uploading data into the archive takes the role of the **Contributor**.
    - There can be more than one contributor for a dataset.
- The **Contributor** uploads the data and fills-in the metadata for the dataset.
- The **Contributor** shares the responsibility of ensuring the dataset is complete, abides by the Terms and Conditions and the metadata is accurate.
- The **Creator** is the person or group that created the data.

Data Manager Role

- To address the problem of datasets being used in different situations than originally anticipated need to have an ‘expert’ or ‘contact person’ for the dataset.
  - The Contributor does not need to maintain a connection with the dataset (eg contributor could be a PostDoc or PhD student).
- **Data Manager** responsible for fielding questions or comments regarding the dataset during its lifetime.
  - Doesn’t have to be an expert on the dataset, but should know whom to contact.
- Similar to what happens with publications (contact person or corresponding author is mentioned).

Rights Holder Role

- The **Rights Holder** is the person or group that controls or owns the rights to the dataset.
  - This includes intellectual property.
- There may be more than one **Rights Holder** for a dataset.
- They control the copyright. If the access restrictions exist on the use of the dataset the **Rights Holder** will need to be contacted for permission to use the dataset.
- In most cases (those abiding by the NLOD or CCv4 license) the role of the **Rights Holder** is less important (but it still relevant).
- It is **IMPORTANT** that you check with your Institution, funding agency as to who owns the rights for your dataset.

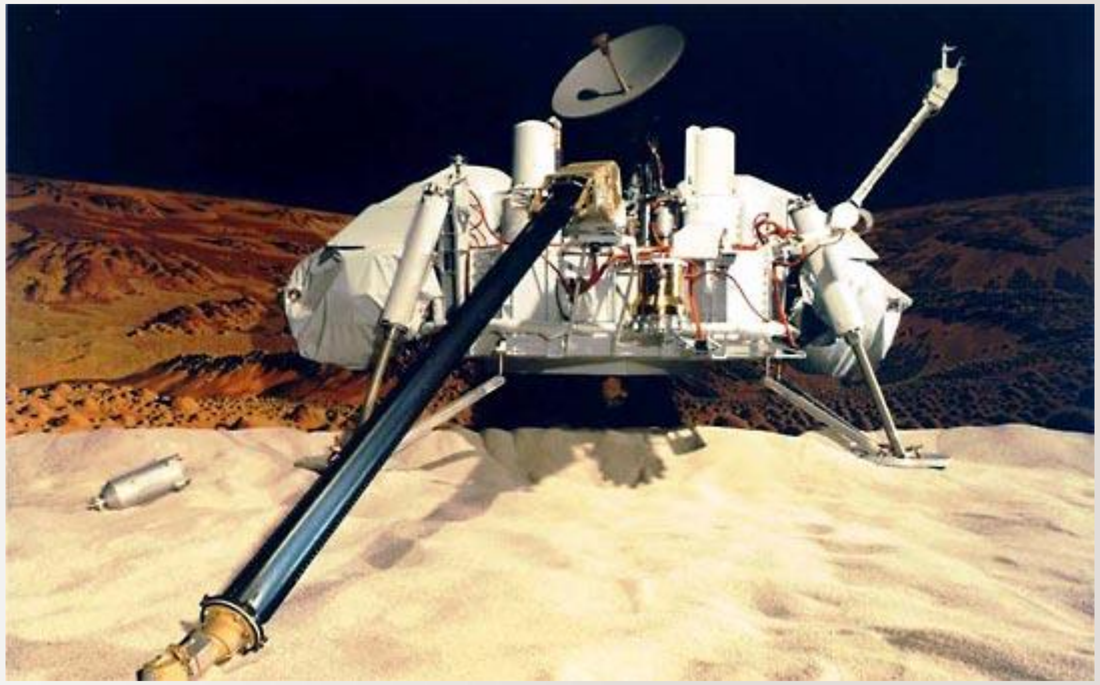
Access User Role

- Any person querying the archive or using the data in the archive assumes the role of an Access User.
- Metadata for all published datasets is accessible by all Access Users.
- Datasets <5GB in size are accessible only requiring an email address
  - The download link needs to be sent to the user.
- Datasets > 5GB in size are accessible upon request
  - Mainly due to the issue of managing large datasets.
- Using datasets assumes you abide by the access licence.

Tips on Structuring Data

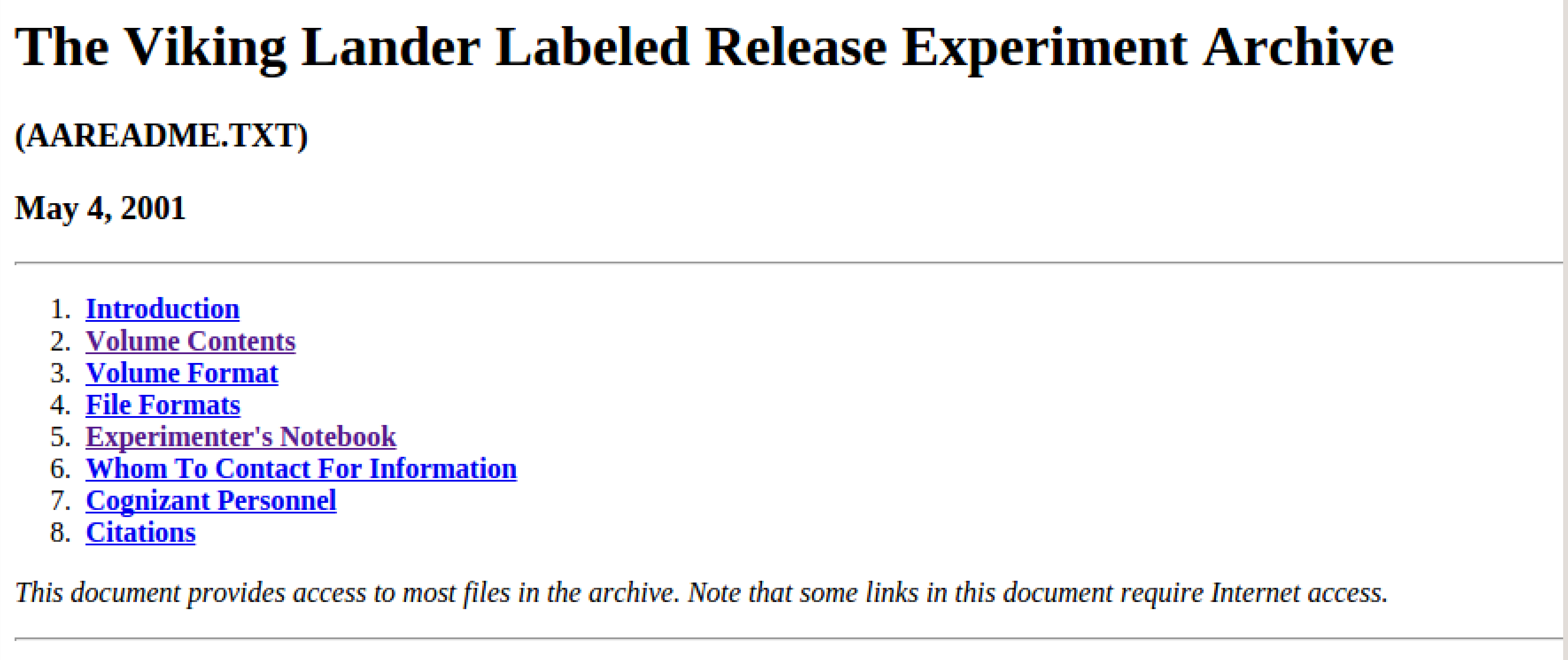
- When archiving data you need to bear in mind that data needs to be accessible for 10 years (or more).
  - Try to use open standards are used for data format (if at all possible).
  - Try to ensure the internal structure of the data is well documented.
  - Try to see that the important features of the data are documented. E.g. in case of images perhaps the number of pixels and colour depth etc.
    - This information is important in case of a need to migrate data to a new format.
  - Try to ensure integrity of data is documented (checksums).
  - Try to make sure documentation exists on how to use the data (what applications, what workflows, etc)
  - Try to make sure all the auxiliary data is included.

Example – Viking Lander experiment



http://pds-geosciences.wustl.edu/viking/vl1_vl2-m-lr-2-edr-v1/vl_9010/aareadme.htm

Viking Lander Example



The Viking Lander Labeled Release Experiment Archive

(AAREADME.TXT)

May 4, 2001

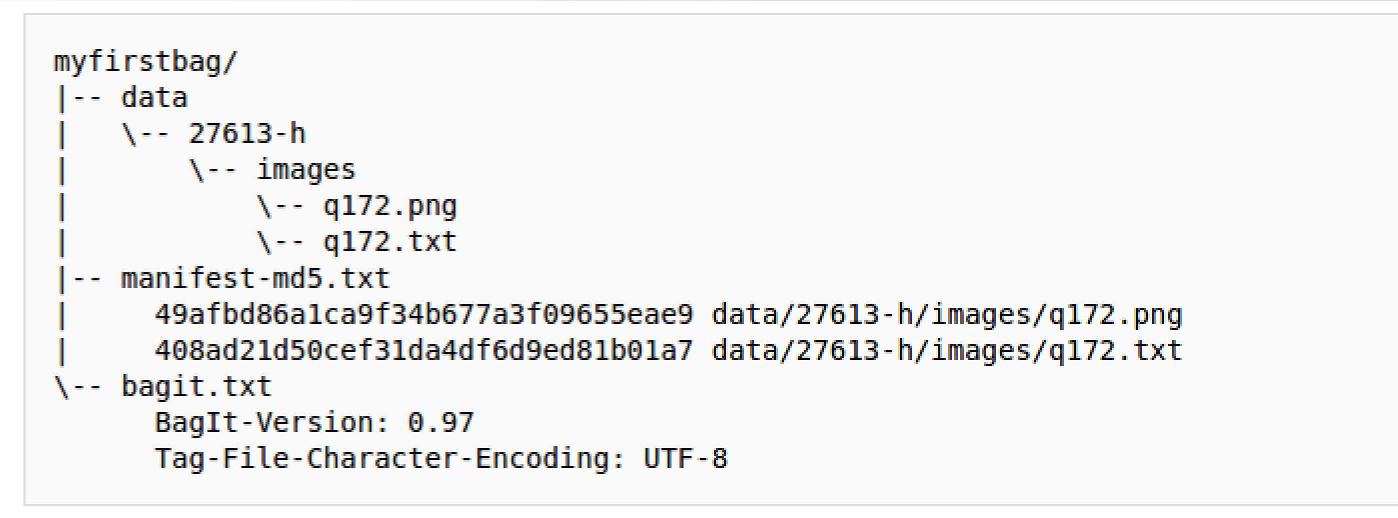
1. Introduction
2. Volume Contents
3. Volume Format
4. File Formats
5. Experimenter's Notebook
6. Whom To Contact For Information
7. Cognizant Personnel
8. Citations

This document provides access to most files in the archive. Note that some links in this document require Internet access.

- Top-level document (README) describes the content of the dataset.
- Experimenter’s Notebook essentially contains data resulting from analyses plus workflow

Tips for Structuring Data

- Internet Engineering Task Force proposal for structuring related data – **BagIt** (http://tools.ietf.org/html/draft-kunze-bagit-10)
- Used by a variety of institutions (eg Library of Congress)
- Essentially:



```
myfirstbag/
|-- data
|   \-- 27613-h
|       \-- images
|           \-- q172.png
|           \-- q172.txt
|-- manifest-md5.txt
|     49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
|     408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
\-- bagit.txt
      BagIt-Version: 0.97
      Tag-File-Character-Encoding: UTF-8
```

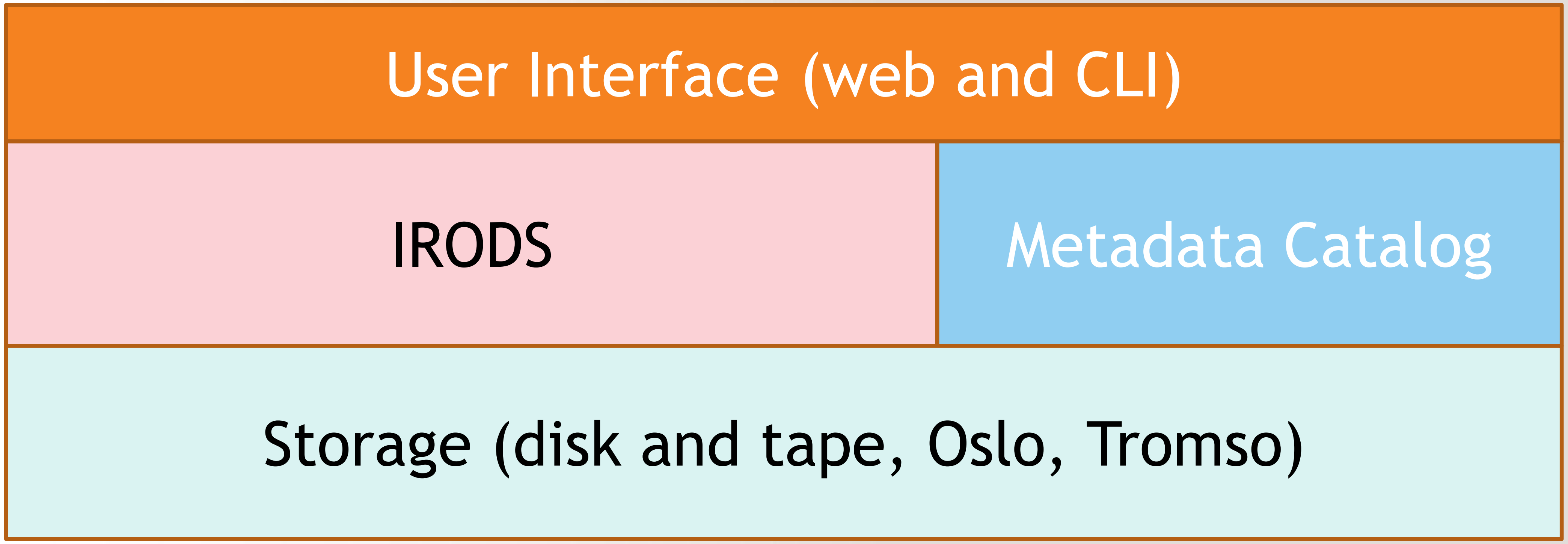
Source: wikipedia

Tips for Structuring Data

- BagIt data directory contains sub-structure. Suggest dividing into:
  - doc – for documentation (including table of contents of layout)
  - src – for any source code needed to read the data (and possibly that generated the data)
  - aux – auxiliary data file
  - <data type> – for data files of that data type
- Or any other layout. But, try to provide a ‘doc’ directory containing documentation and a ‘src’ containing source code.
- Can then zip or tar the BagIt hierarchy and upload to the archive.

The Archive Details

- Designed the archive so it’s possible to replace any component with minimal impact.



User Interface (web and CLI)

IRODS

Metadata Catalog

Storage (disk and tape, Oslo, Tromso)

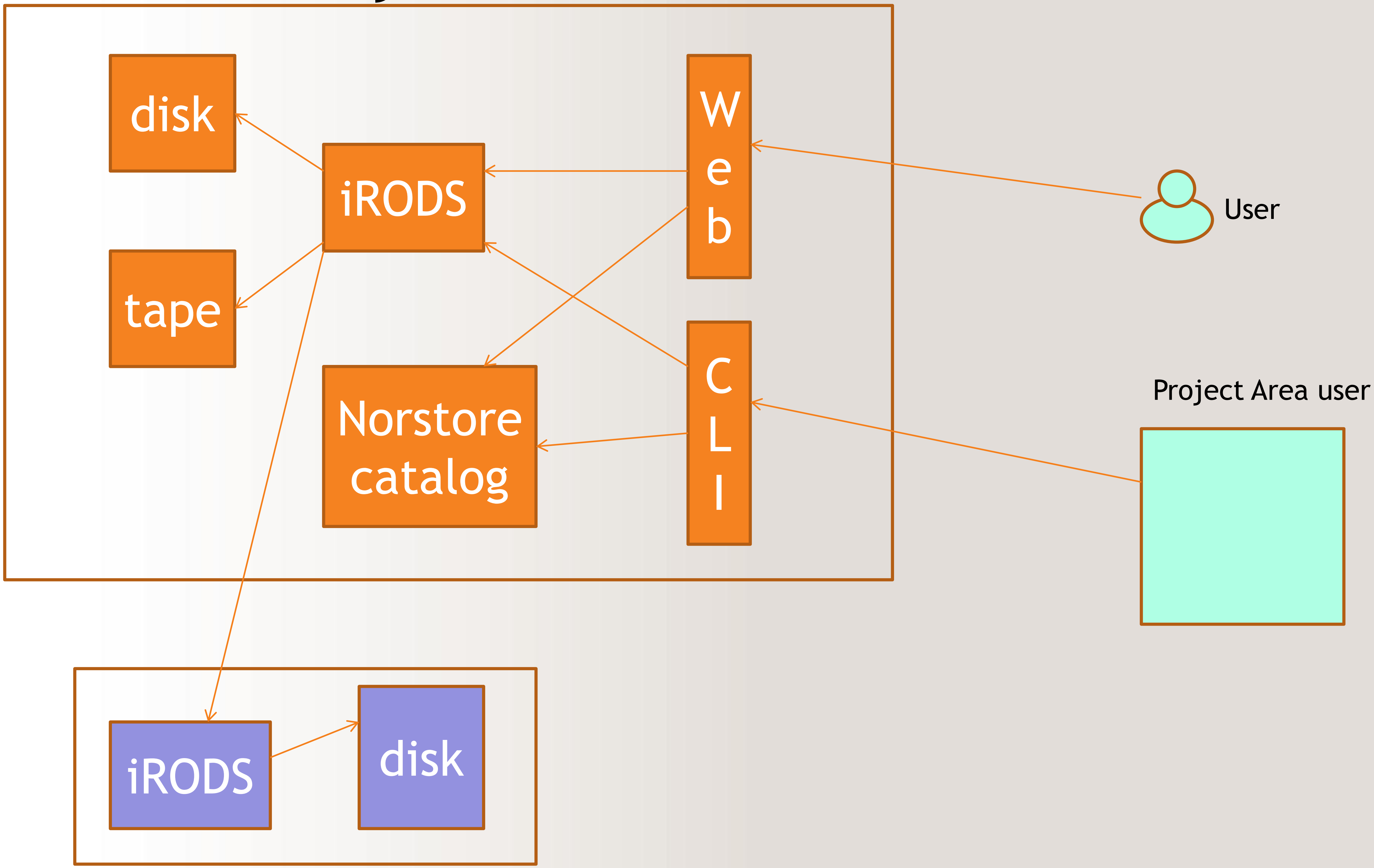
Archive Details: User Interface

- The primary user interface is web-based.
- Command line interface used for large dataset interaction with the project area.
- Interfaces to norstore metadata catalogue.
  - PostgreSQL database. All metadata and state information held there.
- Also interfaces to the iRODS system.

iRODS

- Rule oriented data management system
- Abstracts details of distributed storage by providing logical-layer
- Logical-physical mapping held in iRODS metadata catalogue
  - PostgreSQL database.
- Provides access control and interfaces to authentication such as GSI and Kerberos
  - Norstore makes use of just one archive user to manage the data
  - Users don’t interact directly with iRODS, but through the web interface or command line tools.

Archive Layout



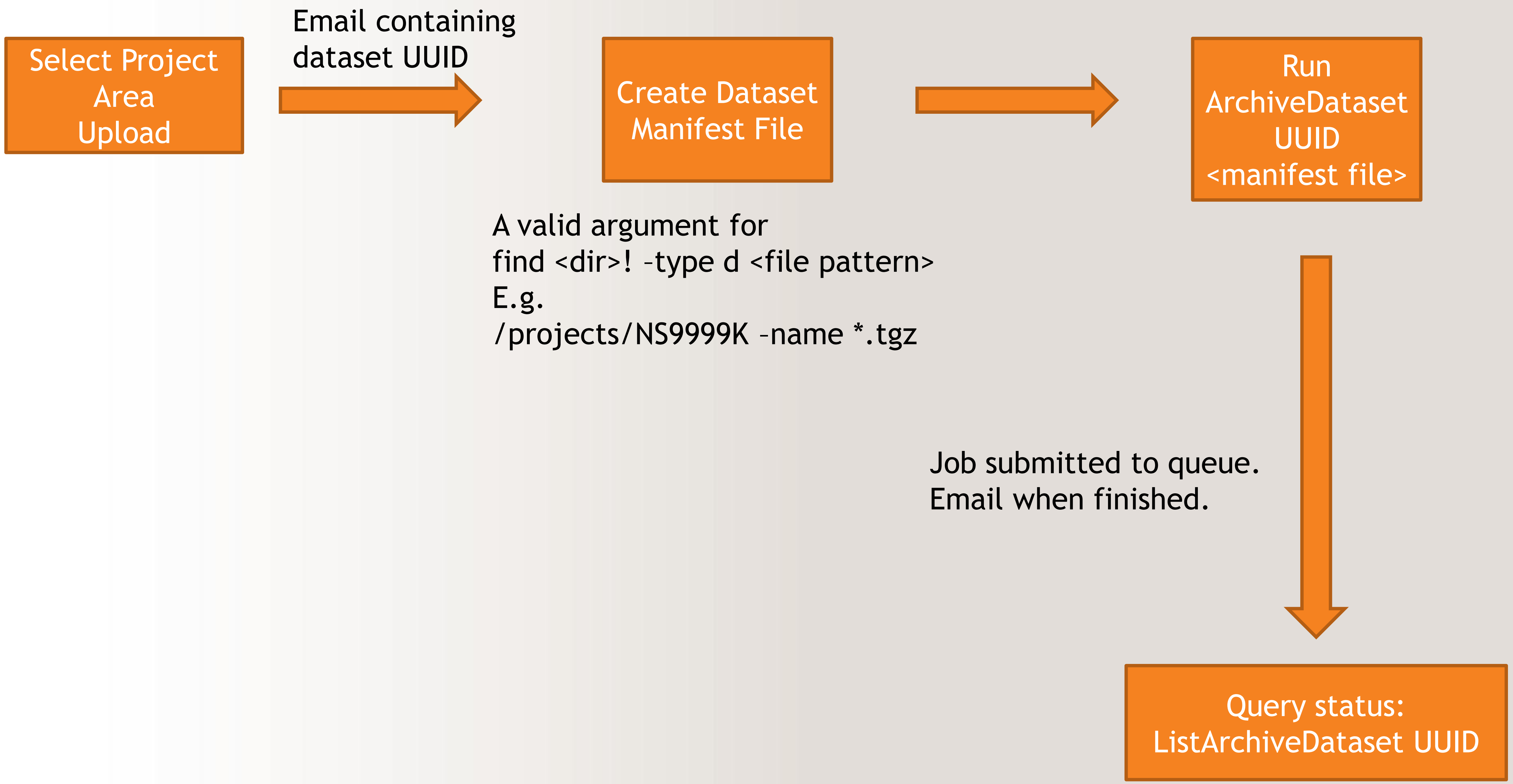
iRODS

- Allows policies to be placed on the data
- Norstore policy to replicate data to 3 resources
  - Also have a policy to remove data from one resource and replicate to a new resource
- Also policy to regularly checksum data

Archiving a Dataset

- Demo illustrates archiving a dataset from the users computer
  - Currently allowed for datasets < 5GB in size (but will remove the limit soon)
- For datasets larger than 5GB upload is currently allowed via the norstore project area:
  - Requires user is registered with a valid project
  - Once dataset is uploaded metadata needs to be filled in via the web interface.
  - More details on uploading datasets from the project area in: https://www.norstore.no/services/archive/cmds-to-archive

Project Area Upload


Select Project Area Upload
Create Dataset Manifest File
Run ArchiveDataset UUID <manifest file>
Query status: ListArchiveDataset UUID

Email containing dataset UUID

A valid argument for
find <dir>! –type d <file pattern>
E.g.
/projects/NS9999K –name *.tgz

Job submitted to queue.
Email when finished.

UNINETT

Publishing Data

- Necessary in order to be able to cite datasets.
- Currently using DataCite node in Denmark to issue Digital Object Identifiers.
  - DOI are standard, unique identifier that can be used to identify a resource.
  - Originally developed for documents, but now being used for data.
  - Each DOI must point to metadata about the object and may contain a link to the dataset itself.
  - Resolver services are used to resolve the DOI to a URI.
- Structure of DOI meaningful doi:10.1000/182
  - 10 refers to the DOI registry, 1000 refers to the entity that registered the data, 182 refers to the actual object.
- Once a dataset is published it cannot be modified
  - Some metadata may be updated