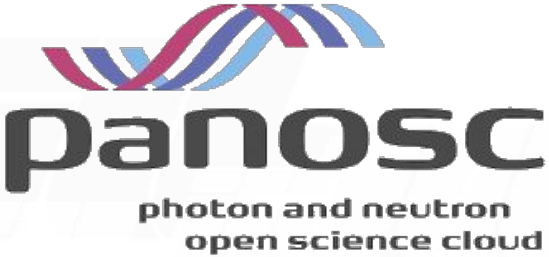


Data Ingestion @ ILL

PaNOSC WP3
Catalogue Integration Best Practices Workshop

20th of May 2021

Outline

- ILL overview
- User Portal and Proposal System
- Data acquisition and file storage
- Datafile scanning
- DOI minting
- Data Portal
- OAI-PMH and Search API
- Data ingestion workflow

Institut Laue-Langevin

- Located in Grenoble, France
- High-flux reactor neutron source
- 40 instruments + 10 test instruments
- First experiments in 1972
- 500 employees
- Usually 3 reactor cycles (50 days each) per year
- Approx 1200 experiments per year
- Approx 2500 scientific visitors per year

NEUTRONS FOR SOCIETY

User Portal and Proposal System

- Proposal submissions via UserClub
- Proposal includes rich metadata
  - Title, abstract
  - Team, user IDs (OrcID)
  - Beam parameters (eg wavelength)
  - Sample parameters
    - Formulae, environment, form
  - Instruments
- Review and validation by Scientific Colleges
  - Beam time allocated for following cycle
- All proposal and schedule data stored in Oracle DB

Data Acquisition and File Storage

- Experiments performed using Nomad instrument control software (primarily)
- User enters proposal identifier before starting experiment
  - Ensures link between proposal data and experimental data (proposal ID stored in datafile)
- Experimental data acquired and stored on local hard disk
  - Remove any potential network problems from instrument control
- Datafiles transferred to permanent storage on NFS filesystem
  - File sizes can range from a few KBs to tens of GBs
  - Files folders organised by cycle, instrument, proposal
    - Eg data/192/d11/2-05-132/rawdata
  - Each datafile has a unique, consecutive number

Datafile scanning

- Scheduled process to regularly read all datafile folders
- If new datafiles exist they are scanned for metadata:
  - Title & Sample formula
  - *Dataset* defined by these two elements (called *data range* at ILL)
- Most instruments use Nexus format
  - Instruments have different file structures or custom entries
  - Depending on acquisition type the file format also changes (for a particular instrument)
- Very difficult (time consuming) to obtain more metadata
- Data is not very clean either (free text)

DOI Minting

- Minting process runs at the end of a cycle
- Only generates DOIs for proposals where raw experimental data exists
  - Found from Datafile scanning
  - Exclude industrial proposals
- Request DOI creation from DataCite
  - REST API to create and update DOIs
  - Include public metadata
    - Title and authors
- Landing page at ILL
  - Eg https://doi.ill.fr/10.5291/ILL-DATA.4-01-1266
- Minted 5,111 DOIs since 2012 (start of the ILL Data Policy)
- Looking into generating DOIs for instruments, samples, software etc.

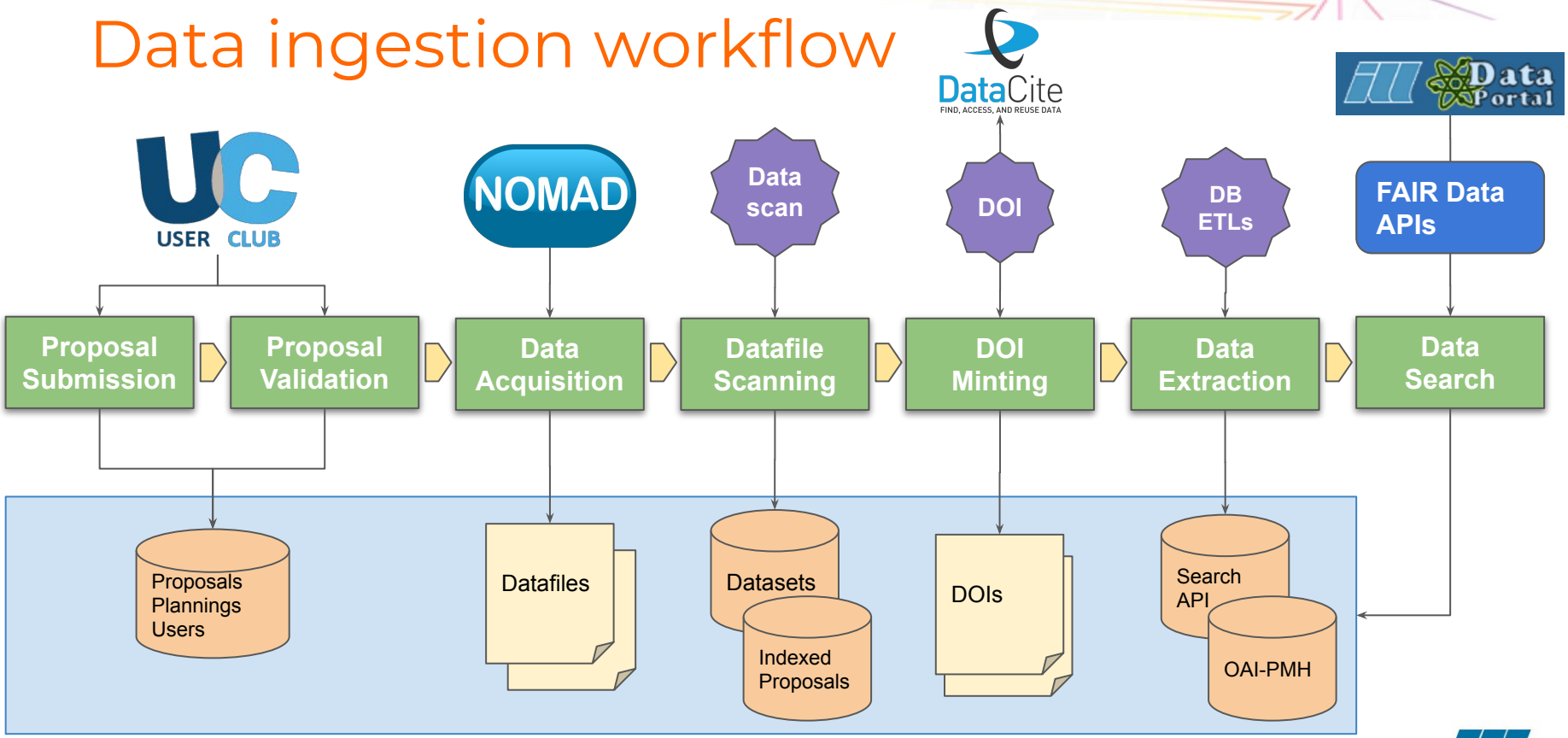
Data Portal

- The ILL Data Portal regroups data from both proposals and datafiles
  - https://data.ill.fr
- Used principally to obtain private/embargoed data
- Proposals are indexed using Solr for efficient searching
- Users can search by
  - Date (or reactor cycle)
  - Proposal identifier
  - Instrument
  - Formula
  - Full text search
- Allows users to download data (individual files or full datasets)
- Allows proposal members to give access to the data to other scientists (ACLs)

OAI-PMH and Search-API

- Open Data is harvested by OpenAire using the OAI-PMH endpoint at the ILL
  - https://fairdata.ill.fr/openaire/oai
  - ~1800 open proposals since 2012
- Search API provides access to ILL Proposals and Datasets using WP3 API
  - https://fairdata.ill.fr/fairdata/api
  - Currently offering Open data
  - Also supports authentication for private data
- Data extracted from the Data Portal and formatted to suit these two applications (ETLs)
  - Nightly cron job

Data ingestion workflow



Questions

INSTITUT LAUE LANGEVIN

NEUTRONS FOR SOCIETY