Publication Workflow Meeting Minutes

Date:
July 3-4. 2018

Location:
MAX IV Laboratory Fotongatan 2 224 84 Lund Sweden

Attendees
Fredrik Bolmsten
Stephan Egli
Luke Gorman
Maria Johnsson
Monica Lassi
Gareth Murphy
Hannes Petri
Tobias Richter
Darren Spruce

Agenda:
https://indico.esss.lu.se/event/1066/

What is the relation between PIDs and DOIs for our data?

- Many to Many
- DOI can aggregate many PIDs
- PIDs could be part of many publications/DOIs

Workflow for Publication driven publishing of datasets vs embargo-period driven **publishing**

- Embargo-driven DOI, grouped by proposal (1 DOI for each expired proposal)
- PIs can make subsets of data for DOI
- External scientists cannot make subsets but can/must cite full set
- PIs can combine data from multiple proposals, only their own data or public

"Merge" process of metadata from PIDs to DOIs

- Just have common metadata
- Need proposal info, short abstract
- Can have array of proposals
- 6 mandatory properties from DataCite https://schema.datacite.org/meta/kernel-4.0/ (see this document for mandatory, recommended, optional fields)

    1. Creator
    2. Affiliation
    3. Title
    4. Publisher
    5. Year
    6. Resource type

- Automerge mandatory fields from metadata plus user written comment field
- Use proposal info for automatic merging
- (OpenAIRE schema makes it more findable)

Do we allow only anonymous access or authenticated access or both to LPS ?

Any DOI must resolve to anonymous page with basic metadata, which can then link to pages for full download which redirect to login page if needed

If the search functionality should implemented in LPS or simply as part of global DOI systems ?

- Search field for convenience of user, can redirect to SciCat?
- No search in landing page, but links to related information
- Can search through SciCat
- Need to have branding, be pretty

If the landing page server should only serve published data or also data still under embargo period

- ESS not allowed to publish metadata before embargo expires
- PSI needs to publish metadata before expiry date - create DOI for proposal once its accepted

Which fields of the meta data become public when ?

- at creation time (e.g Title, authors, abstract, used infrastructure PI
- after embargo period ends
- never (e.g. full Proposal text for PSI)
- No site-specific rules, to begin with

How raw/derived data can be fetched?

- via HTTPS, HDF5 server, dedicated export server ?
- Small enough data, hdf5, HDF5
- 100s of TB use gridftp
- (iRODS)
- globus-online

Who is authorised to publish data? PI, Co-PI etc

PI, with delegation

Mockup of a GUI for the scientists to select data for a DOI.

Shopping cart for PI

*Do we need an additional PID local handle server to make PIDs themselves public ?

• Do we require a dataset to be available before it can be part of a publication process ?

• If authentication used: via federated ID systems (e.g. umbrellaId/Eduid) or local accounts ? Or via DUO accounts ?

======

PSI scicat is not externally accessible

Need static html produced by script

LPS can't access scicat outside PSI.

Move to github

Site-specific functionality

Protocol for site-specific

Day 2 -

Publication Workflow - tasks including LPS

Separate server

Client app

Gets info from catamel

Server side rendering
Angular 6
Start from scratch

BAckend

New model

PublishedData

{
Doi:
Mandatory datacite fields:
Creator
Affiliation
Title
Publisher
Year
Resource type
PIDs Array:
Full URL: route https://www.esss.dk/doi/10.1799.XEHJFH
DOI_posted_successfully_date:
Last_modification_date:
}

PI and delegatee can add PIDs of their own data plus calibration, open data

Email generation/notification

Embargo runs out - here is notification

Publishing is automatic
Per proposal - emails PI or everyone in group with notice
On a defined date, end of beamline run/cycle + 3 years

APE: Automatic Publication Engine

PSI use atom feed to provide

Task List

- Task 1: Server side rendering in Landanie - PSI
- * Task 1b: independent landing page server - Gareth & Luke & Hannes
- * Task 2a: Backend: Catamel extra model - Gareth & Luke
- Task 2b: Catamel: Policy for who can manage, create PublishedData - Stephan & Luke
- Task 3: Catanie shopping cart - creates PublishedData - Hannes
- Task 3b: CronJob - separate instance Embargo runs out + automatically proposal data published under one DOI - PSI

- Task 4: Catanie: Overview/table/GUI of published data for admin purposes and also PIs - Hannes
- Task 5: Catamel: Talk to DataCite and mint/create DOIs and upload to DataCite - Gareth PublishedData.DOIminter() Check if it worked? Is it published? Check connection. CheckDOI hasbeen submitted
- Task 6: Check with ETH/DataCite if they can accept xml vs atom feed - PSI
- Task 7: Think about email notification logic and dependencies

---

Generic manager interface

See his pgroups

Set flags for publications

Manager Interface

My Proposals
My PublishedData
My Archive settings/Data policy

My Delegates

1. PI actions
2. User cannot publish
3. PI can publish

Gitlab move to Github
on friday - Gareth does the tickets and sends out an announcement

Site specific functionality
How do we decide a site specific solution is needed and how it is “provided/enabled/configured”?

Should be avoided where possible
Failing that it should be contained to UI code when at all possible

UI customizations move from CI folder to sites

Branding when logged in

Footer or header for ESS/MAXIV/PSI

Help number

creationLocation => instrument or beamline at different sites

Site localisation

Lang file for strings
Env file for features

Sites folder for branding

Header/footer

Placeholder

Variable for {beamline, instrument, location etc}

Angular internationalisation pipes

- Task 8: Branding and localisation
- Task: 8b: move CI to sites folder

It would be nice to have documentation in the source code repo
So the same merge request merges documentation

- Task 9: Move documentation out of documentation repo into catanie andf catamel
- Task 10: Have the github pages autobuild with Jekyll from the catanie/catamel repo

---

Ingestion Stories

STM complexity

Central storage

Data collection software

3 beamlines try to connect

Biomax has MXcube which is react, easy to connect to and get parameters

NanoMax which is using Sardana

Signal that scan has started
Noone knows the scan has finished
“Fire and forget” difficult for metadata catalogue

Windows Scienta software

/data/visitor/<proposal_id>/<date_of_visit>/<raw| derived>

This folder structure is created

- For windows machine, write to e.g. /tmp login with linux and make a symbolic link
- Use 2 operating systems
- Similar CSAX

- Difficult to monitor distirbuted file systems for file write completions
- Scrape hdf5 data

Data Curation Workshop

Present FAIR data
Success story from TomCat, BioMax, NanoMax
Encourage adaptation of FAIR principles

IT perspective user
First meeting to spark interest
Esko
ISPyB (http://www.esrf.eu/ispyb)

Adding images
Edna

Get image working more nicely

Add images to table for November

Profile image in corner upper right

Move the big search bar into the filter box

No slidey menu

User can hide the navigation part (filtering part)

- More eye candy (header footer)
-
- More user focus (
-
- Raw icon
- Derived icon
-
- Users prefer text
-
- Sample name
- Experiment title
- Source folder too long
- PID is not interesting/ID
- Shorten the PID
- Change to run number
- Instrument

- Mockup of what to show in november
- Data Reduction, Analysis and Modelling Group
- Different column structure for derived data

- Live update - experiment in progress - very nice
- We don’t have web sockets
- Need loopback support
- Displaying archive and retrieve jobs would be nice to have sockets

- Sample preparation is foreseen in database
- Model is present

Federation and External Metadata

- Umbrella ID
- Public PIDs
- Federated IDs

September PSI vacation
October better to visit PSI

https://indico.esss.lu.se/event/1066/

https://indico.esss.lu.se/event/1056/