PAUL SCHERRER INSTITUT

PSI

Stephan Egli :: Paul Scherrer Institut :: Photon Science Department

Publication Requirements from Data Policy and other considerations

Publication Workflows for Data Curation, July 3rd 2018

PSI Data Policy

- Has become public recently, see https://www.psi.ch/science/psi-data-policy, based on work done within PanData EU Projects
- Main points
  - The policy defines the data ownership, curation, archiving and open access
  - Data custodian (Head of research division) and data owner (Principal investigator)
  - Data retention for minimum five years, striving for ten years
  - Default embargo period of three years (with option for five years)
  - Implementation bases on Metadata Catalogue, PetaByte Archive and remote access functionality

FAIR principles

- Required by SNSF (Swiss National Science Foundation),see http://www.snf.ch/en/theSNSF/research-policies/open_research_data
- **Findable**: Data and Metadata should be easy to find by both humans and computer
- **Accessible**: longterm storage and access via standard communication protocols
- **Interoperable**: Data should be ready to be exchanged, interpreted and combined in a (semi-) automatic way
- **Reusable**: Data and Metadata are sufficiently well described to allow data to be reused in future research

Relevant points of Data Policy for Publication Workflow

- 2.6 Open Access means belonging to the public at large, unprotected by most copyrights or patents and subject to appropriation by anyone. Those data will be made available under CC-BY-SA (https://creativecommons.org/licenses/by-sa/4.0/).
- 3.1.1 All Raw Data and Metadata obtained as a Result of Public Research will be Open Access after an initial embargo period during which access is restricted to the Experimental Team, represented by the PI.
- 3.2.5 Each experiment and data set will have a unique persistent identifier. Anybody publishing Results based on open access data must quote the same identifier (and related publications if available & required).
- 3.2.6 High level Metadata such as title, authors, abstract, specific Research infrastructure will be made public as soon as the experiment has been carried out. This information will be available via the persistent identifier landing page on the web.
- 3.3.2 Access to the On-line Catalogue of PSI will be restricted to registered users of the On-line Catalogue.

Relevant points of Data Policy for Publication Workflow

- 3.3.3. Access to Raw Data and Metadata obtained from an experiment is restricted to the Experimental Team for an embargo period of three (3) years after the end of the experiment. Thereafter, the data will become openly accessible. Any PI that wishes data to retain restricted access for a period longer than three (3) years will have this possibility on a yearly basis on a maximum prolongation of two (2) years... Data can always be made openly accessible earlier on simple request of the PI…
- 3.3.4. Raw Data and Metadata explicitly used for peer-reviewed publication will become Open Access at the time of such publication.
- 3.3.7. The On-line Catalogue will enable linking experimental data to experimental proposals. Access to the full proposal text will only be provided to the experimental Team and appropriate facility staff, unless otherwise authorized by the PI.
- 6. Publications related to data from experiments carried out at PSI must cite the persistent identifier of the experiment and data in their publication.

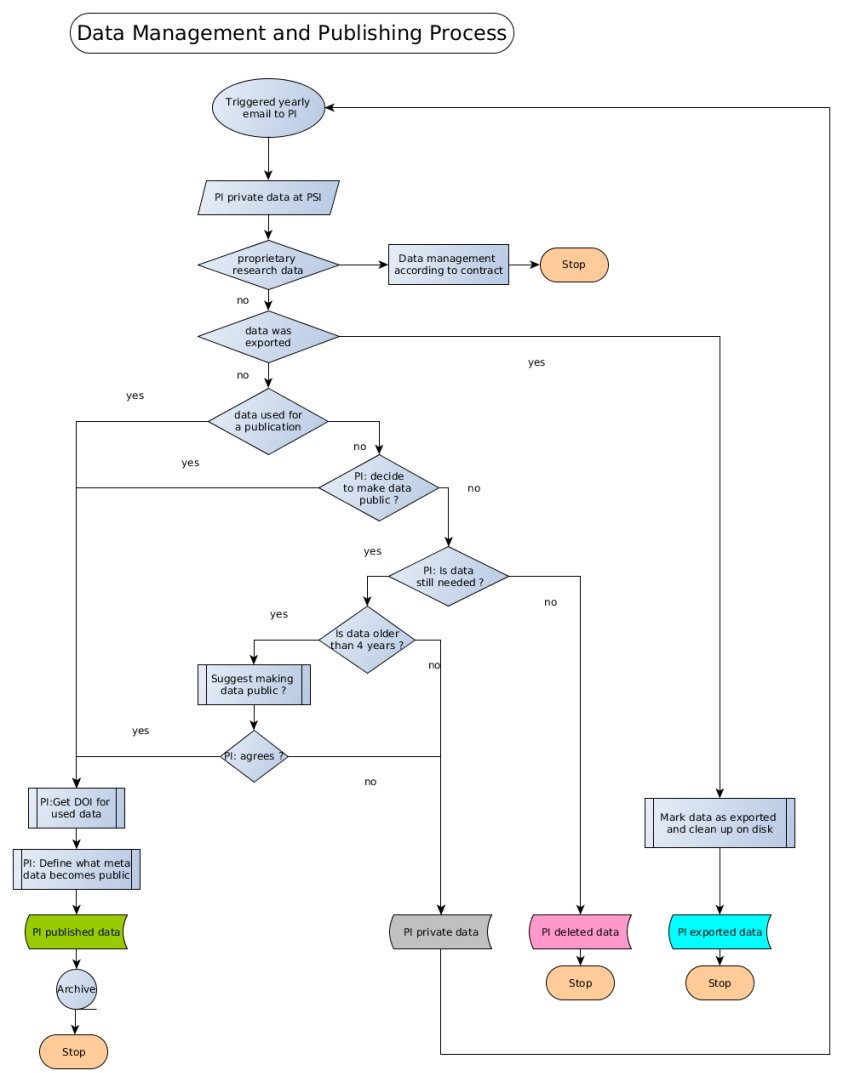
Relevant points of FAIR principles for Publication Workflow

- The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set's PID in the metadata.
- Metadata are used to build easily searchable indexes of data sets.
- If one knows a data set's identifier and the location where it is archived, one can access at least the metadata. Furthermore, the user knows how to proceed to get access to the data.
- It often makes sense to request users to create a user account on a repository. This allows to authenticate the owner (or contributor) of each dataset, and to potentially set user specific rights.
- Metadata are available even when the data are no longer available
- (Meta)data archived on the repository is accessible using a standardized protocol

Some Thoughts about Publication workflow

- As part of a publication workflow datasets must be linked to DOIs in order to become citable.
- DOIs are globally unique way identifiers that allow to link from the DOI to the associated data and metadata via so called landing page servers (LPS)
- This assignment needs to be done as part of a publication process usually triggered by the author of the publication
- The DOIs can link to both raw and/or derived datasets.
- However there is an additional possible workflow that makes datasets public, even if no publication is linked to it so far, triggered by the timeout of the embargo-period
- - see flowdiagram

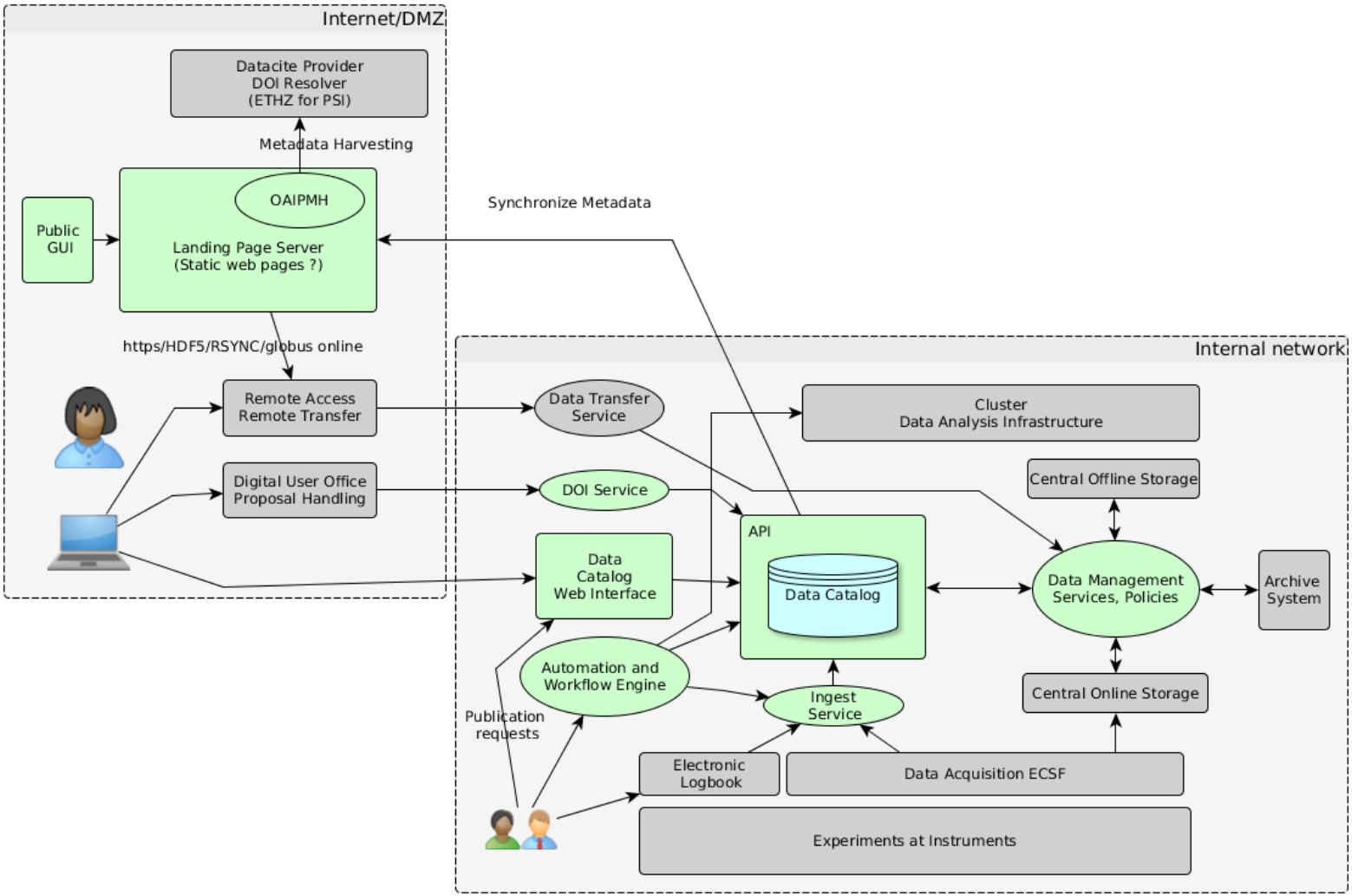
Data Management and Publishing Process (Early Draft from 2016)



Relation of PIDs and DOIs

- There is a **one-to-one** relation between **datasets and PIDs** (Persistent identifiers)
- PIDS have the form PID-Prefix/Suffix, e.g. for PSI
  - 20.500.11935/0dde3c6a-a82b-4dd4-af21-ba07618afe9c
  - (Prefix given from Handle http://www.handle.net/prefix.html)
  - Suffix auto-generated from Database ID. Guarantees globally unique ID
- FOR DOIs: Same structure as PIDs: Prefix/Suffix doi:/10.16907/PSI-Suffix
- The relation between DOIs and PIDs (=Datasets) can in principle be
  - one-to-one : DOIs and PIDs would be equivalent, (DOIs could still be a sub-set)
  - many-to-one : Would allow to collect many PIDs/datasets under one DOI
  - many-to-many : would in addition allow to use the same dataset in different publications (Probably we have to go for many-to-many relation)
- Possible implementation DOIs get their own model ("table") with list to links of PIDs. This would allow to define metadata at the level of the DOI as "merged" metadata from the linked PIDs and to extend metadata by further comments relevant for the publication
- Note: unfortunately most documents do only talk about persistent identifiers, and make no distinction between PIDs and DOIs

Architecture Update for Publication Workflow (Input for Discussion)



Things we need to define

- Relation between PIDs and DOIs
- Workflow for Publication driven publishing of datasets vs embargo-period driven publishing
- Mockup of a GUI for the scientists to select data for a DOI.
- "Merge" process of metadata from PIDs to DOIs
- If we allow only anonymous access or authenticated access or both to LPS ?
- If the search functionality should implemented in LPS or simply as part of global DOI systems ?
- If the landing page server should only serve published data or also data still under embargo period
- 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)
- How raw/derived data can be fetched: via HTTPS, HDF5 server, dedicated export server ?

Some more questions

- Do we need an additional PID local handle server to make PIDs themselves public ?
- Do we require a dataset to be archived 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 ?

Additional Material
Overall Data Catalog Architecture