Publication database.

EUCALL Workshop: User Access at Advanced Laser Light Source

Alexander Wagner

ELI-Beamlines

Dolní Břežany, 22.09.2017

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY

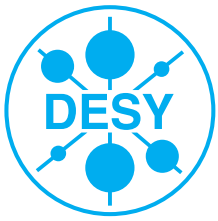
Overview

- General considerations
- Content
- Data

DESY

Publication database.

Part I: General considerations



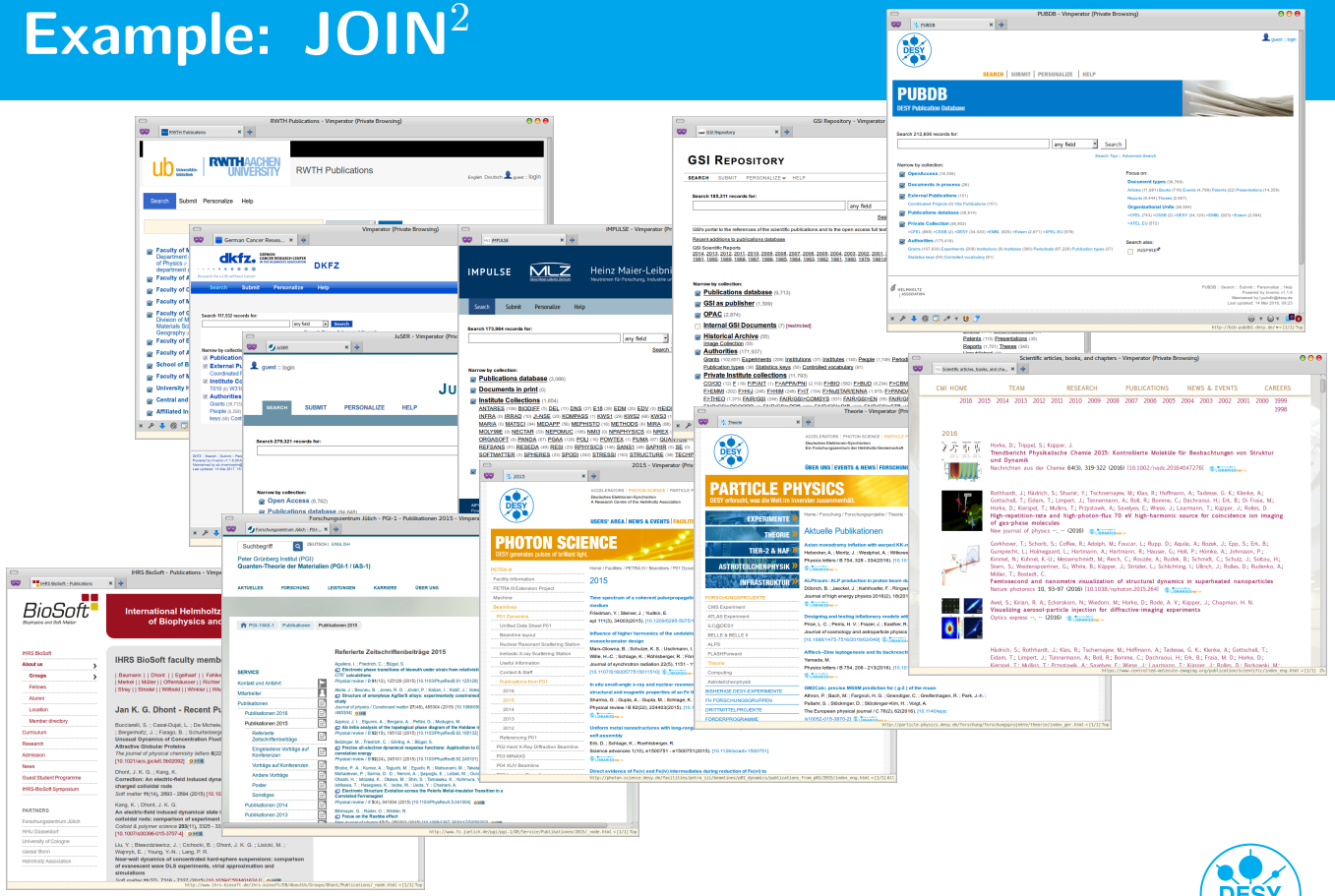
Stakeholders

Who is using your pubdb? How are they using it?

- Scientists (e. g. publication lists)
- non-scientific Staff (e. g. administration)
- General public
- Facility (e. g. web pages)
- Funders (e. g. EU: OpenAIRE)

What services can you provide to them?
Are people using the database or only the data?

Example: JOIN^2



Goals

What do you want to achieve?

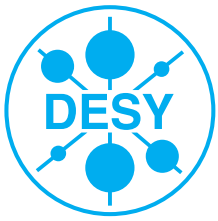
- Visibility (think web, OpenAccess)
- OpenAccess (requires much more than a database; e. g. policies, copyright, work flows, interfaces)
- Repository and archive (e. g. What was our impact after *n* years of operation?)
- Reporting (also: application for new funding, new proposals)

Always think OpenAccess.
Never forget Green OpenAccess: go for 100%

Remember: "All that is gold does not glitter..." (J.R.R. Tolkien, *The Riddle of Strider* in "The Fellowship of the Rings")

Publication database.

Part II: Content



Items

Which items to *whom* and *how* to access?

- Bibliographic data (OpenAccess mandates: will no longer be sufficient)
- Full text documents (easy future access, build an archive)
- Access rights (restrictions, embargoes)

Document types:

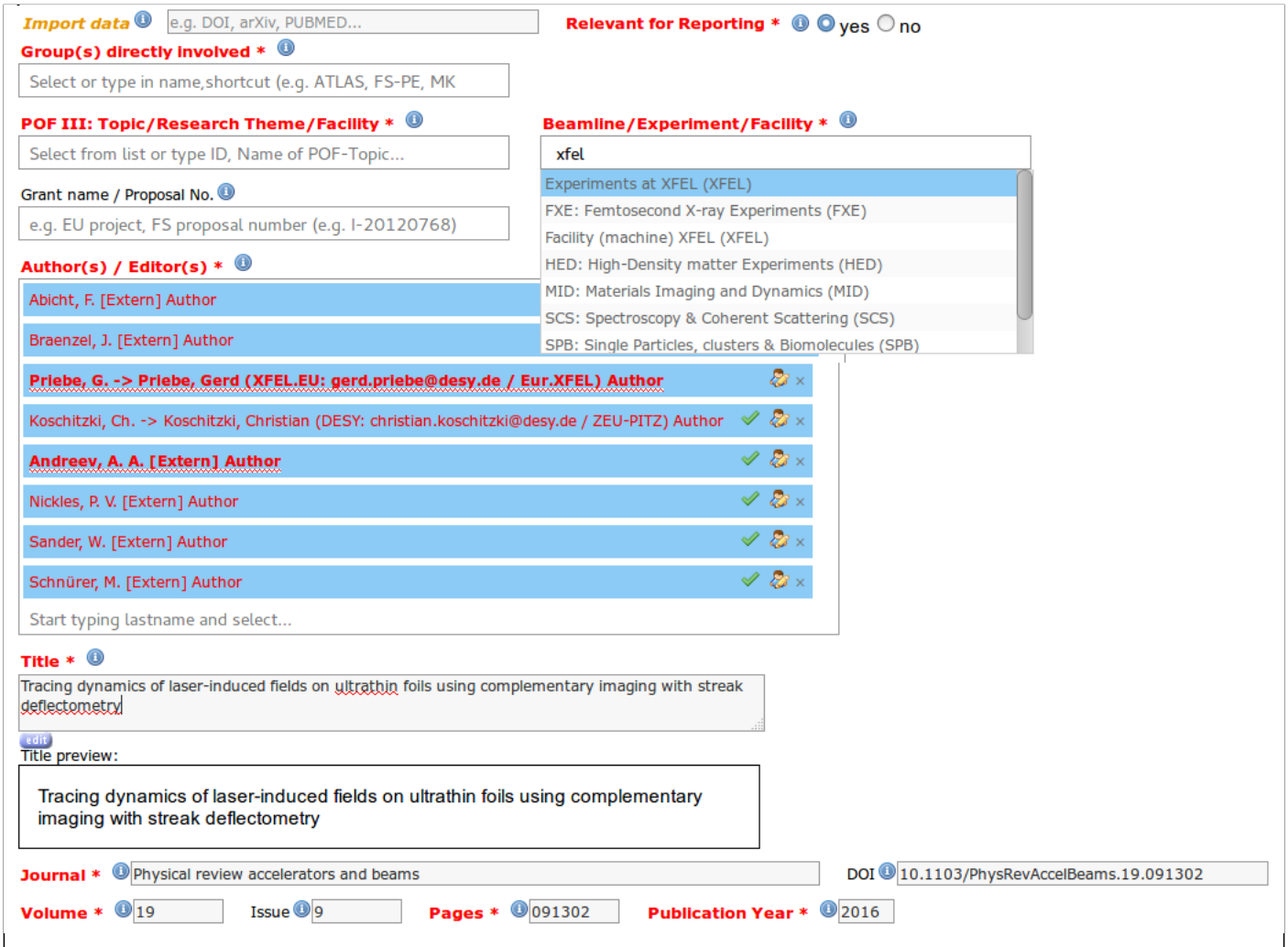
- Publications (really only "own" results?)
- Data(?) (will pose different problems)
- Other documents (theses, instructions, guidelines, documentations)

Do and Don't

- Build tools scientists want to use and non-scientists can use
- Do not build a lighthouse
- Minimize necessary manual work (importer, but don't fall for batch ingestion!)
- Maximize re-use (structured exports, web integration)
- Integrate with your library (catalogue, document delivery, repository, publishing house…)
- Imagine workflows:
  - Who knows what at when?
  - What can be derived, what is only internal knowledge?
  - How can available data be reused? (What derives from the proposal e. g.?)

…give me an universal, unique identifier and …

Archimedes?



Input

Model data, not answers to todays questions

- Think big (e. g. **JOIN**^{2}: 7 institutions, serving 25.000 people, about 355.000 records)
- Define publicly available values (e. g. bibliographic values)
- Define specific values (data nobody else has/knows, e. g. proposal id)

The more details, the more work for **Scientists**

Input

- Author disambiguation: ORCiD (login, proposals, submissions)
- Completeness and correctness (review work flow, **web output**)
- Derive from proposals e. g.
  - funding information
  - scientific subject
  - coupling to submission

Rethink "Call for Submission"

- Not fast enough (e. g. EU OpenAccess Mandate)
- Asymmetric work load (usually everything in a few weeks)
- Incomplete (something happend a year ago, even worse 12 months...)

Submission should happen right after publication, better before.

(approval forms?)

Storage

- Use established standards ("established" may be old)
- "Authority controlled " data (normalize semantically)
- Persistent, common, globally unique IDs (e. g. ORCiD, DOI, arXiv etc.)
- Run archival safe (OAIS)

Involve your Library

...but also believe them if something is impossible.

Output

- Think in structured outputs (BibTeX, EndNote, RIS...)
- Your web pages are *life online reports* (will require caching)
- Enhance visibility for *machines* (schema.org, microformats, semantic markup...)
- Facilitate item interchange (e. g. OAI-PMH, especially beyond Dublin Core)
- Report only based on your publication database
- Minimize manual efforts for common reports
  - Structured exports (e. g. BibTEX)
  - Batch processing (e. g. LATEX or pandoc)
  - Common formats (e. g. BibLATEX, CSL)

Publication database.

Part III: Data

Data

There is no *best practice* for "Data" in 2017

However, note:

- Only published Data is (re-)usable
- Publishing implies availability for ever ("citation safe")
- The publisher is responsible for eternal storage and access
- There is no fall back (no National Libraries, Portico or the like)
- Publishing is not (only) an IT problem:
  - Metadata (findability)
  - Consistency (reliability)
  - Formats (availability)
  - Long term preservation and migration (a bit stream is usually not enough)

Data

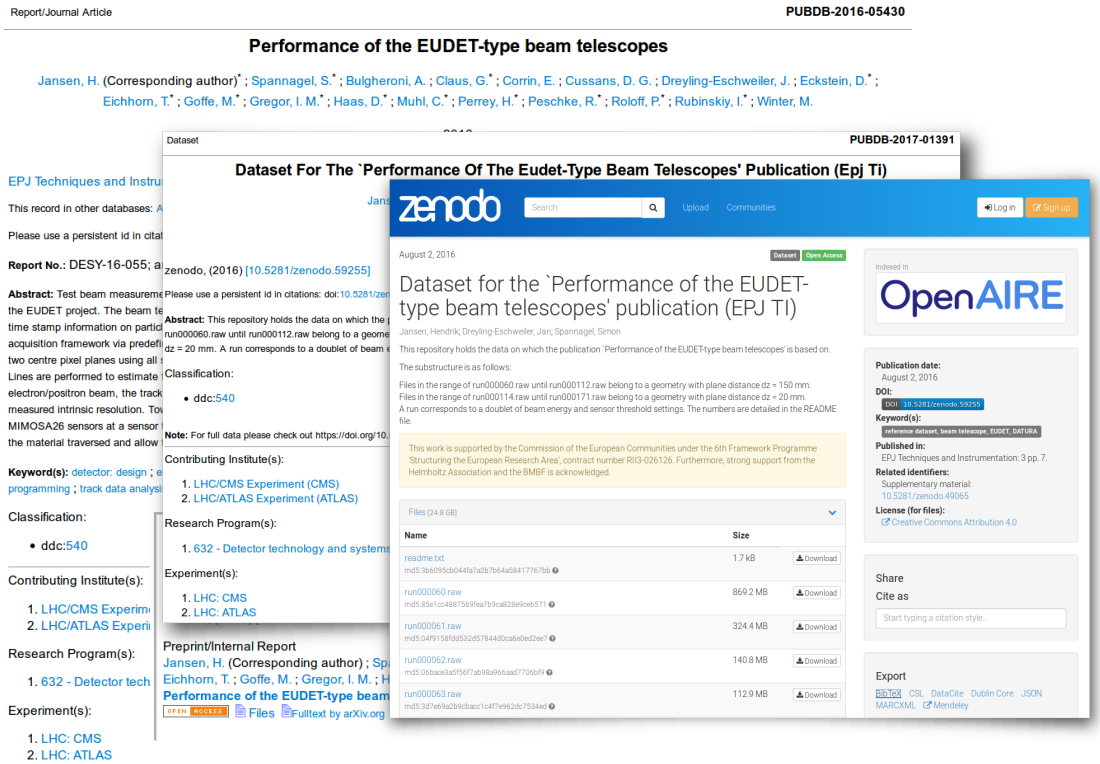
- In your Repository: Small records (non-IT-issues still apply)
- Global data centres: Large records

Major headache

Intermediary sized data

- Samples: HEPData, Pangaea, DRYAD ...
- Registry: re3data.org (if there is something suitable, use it)
- General & Software: zenodo (up to 50GB per dataset, more on request)

Data @ PubDB



Thank you!

Alexander Wagner
Deutsches Elektronen-Synchrotron
Central Library

Tel.: +49–40–8998–1758
alexander.wagner@desy.de
0000-0001-9846-5516

http://library.desy.de

This document is available as https://doi.org/10.3204/PUBDB-2017-10906

Typeset by lualaTeX