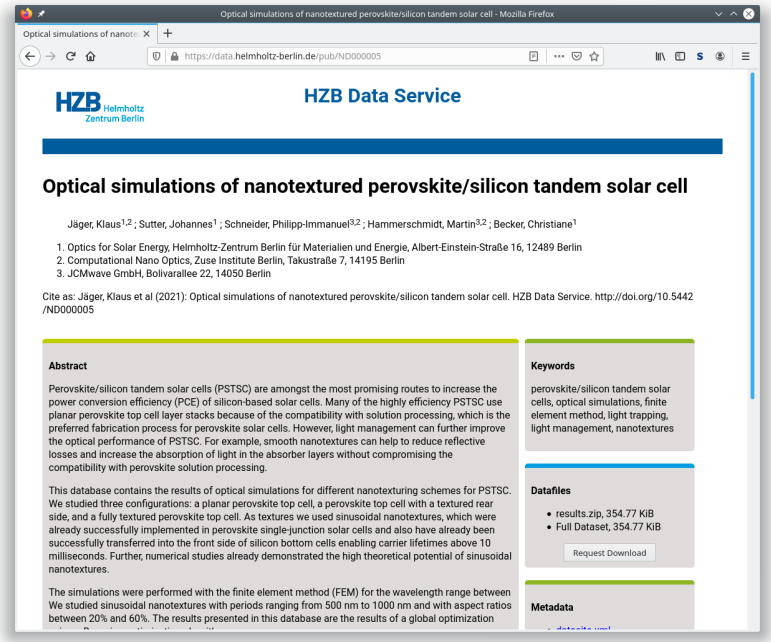
Data Publication Workflows

Rolf Krahl

ICAT Schema Discussion, September 2021

Rationale: Manage Curated Data Publications

- Need to manage curated data publications.
- Generate the publication landing page out of ICAT.
- Need to be able to store the full set of bibliographic metadata in ICAT.
- Take DataCite 4 as the standard for the definition of metadata.
- At present, there is no place in ICAT to put that metadata.



DOIs for Public Raw Data versus Curated Data Publications

Need to distinguish:

- Open access raw data:
  - Indiscriminate: all raw data from our experiments become openly accessible, regardless of relevance or scientific value.
  - Poor bibliographic metadata, only what can be collected automatically. No manual curation practicable.
- Data publications:
  - Selected: only carefully selected data get published, most often the data supporting a journal article.
  - Often a dedicated dataset is created specifically for the publication.
  - Manually curated and rich bibliographic metadata.

Minimum Viable Version

- Add classes:

  DataPublication
  - attributes: doi, title, publicationDate, subject, description
  - relation: DataCollection

  DataPublicationUser (see next slide)

  DataPublicationDate
  - attributes: dateType, date

  RelatedIdentifier
  - attributes: identifier, relationType, fullReference

  FundingReference
  - attributes: funderName, funderIdentifier, awardNumber, awardTitle

- Advantage: no interference with any other ICAT content.
- See: #200

Minimum Viable Version: DataPublicationUser

- Add classes:
  DataPublicationUser
  - attributes: contributorType, orderKey, fullName, familyName, givenName, email, affiliation, affiliationIdentifier
  - relations: DataPublication, User
- Note duplication of some attributes (fullName, familyName, givenName, email, affiliation) with User.
- Rationale for this duplication: these attributes may change over time for a given person. We need to keep a record (e.g. a "snapshot in time") of the values at the time of the publication.

Extension: Add Affiliation

- Add a separate class Affiliation.
  - attributes: name, pid
  - Relation: DataPublicationUser
- Rationale:
  - Users may have more than one affiliation, even in one single publication.
  - Simplifies handling of PID, e.g. ROR.

Extension: Allow FundingReference also for Investigation

- Instead of a many-to-one relation from FundingReference to DataPublication, add two many-to-many relations to DataPublication and Investigation respectively.
- Pro: allow funding references also to be attached to investigations and thus to be taken from the proposal.
- Con: slightly more complicated.
- See: #249

Extension: Add class UserInfo

- Rationale: the issue that we may need to keep a record of volatile user attributes does not only occur in curated data publications.
- Make the "snapshot facility" for these attributes from DataPublicationUser generally available.
- Idea: separate the identity of the user from potientially volatile attributes.
- Add a new class UserInfo and move all potentially volatile attribues from User to UserInfo.
- Change the relations from
  Investigation $\leftarrow$ InvestigationUser $\rightarrow$ User
  to
  Investigation $\leftarrow$ InvestigationUser $\rightarrow$ UserInfo $\rightarrow$ User
- Issues:
  - disruptive change, many potential compatibility issues.
  - may have an impact on performance.
- See: #248

Publication Metadata

Based on DataCite 4, the following metadata would be needed:

- DOI.
- User (creator and contributor, include given name, family name, and affiliation).
- Title.
- Publisher (facility).
- Dates: date of creation, date of submission, date of publication.
- Resource type ("Dataset").
- Subject, keyword, classification code, or key phrase.
- Related identifiers. Relevant relation types may include: "IsCitedBy", "Cites", "IsSupplementTo", "IsPartOf", "IsCompiledBy", "IsDerivedFrom", "IsSourceOf", "IsReferencedBy", "References".
- Description, e.g. abstract.
- Funding information.