SHARE Proposal

General Notes on SHARE Proposal from ARL/AAU/APLU

This page includes general notes on the SHARE proposal from ARL/AAU/APLU, in response to the White House OSTP memoradum entitled "Increasing Access to the Results of Federally Funded Scientific Research".

It is not a formal statement from DuraSpace, but just a way to gather feedback (which will be passed along to ARL). As such, members of the community are welcome to comment on these notes or help to enhance them.

General comments/questions are made throughout the below summary (in italics) and are marked as such.

For additional software-specific notes, see the child pages:

The Proposal

The full text (PDF) of the proposal is available at: http://www.arl.org/publications-resources/2772-shared-access-research-ecosystem-share-proposal

High Level Summary

The SHARE proposal suggests a number of functions and metadata fields that would need to be captured by repositories. We've attempted to briefly summarize them below. But, the full text of the Proposal has additional details.

Minimum SHARE metadata fields

These are the listed minimum SHARE metadata fields as noted near the beginning of the "How SHARE Works" section of the proposal:

1. author
2. article title
3. journal title
4. abstract
5. award number
6. Principal Investigator ID (ORCID or ISNI)
    - http://www.isni.org/isni_and_orcid
7. designated repository number

In Support of Principal Investigators

As described in the paragraph about requirements of Principal Investigators (PIs), repositories may need to be able to "capture" or log the following:

1. "Sufficient copyright licenses to enable permanent archiving, access, and reuse of publications"
    - ***General Comments***
        - *Many repository platforms do have an option to require a "deposit license" which often covers these scenarios. However, the text of the "deposit license" is decided by the institution. There may need to be "recommended copyright license language" provided by a central entity, to help ensure locally created licenses are "SHARE-compliant".*
        - *Does this need to be machine actionable / verifiable?*

General Repository Functions

As described in the "SHARE workflow" paragraphs, a repository would need to support the following functions:

1. Be able to accept XML versions of manuscripts from Journal publishers
    - "Journal submits XML version of final peer reviewed manuscript to the PI's designated repository
    - ***General Comments***
        - *Who defines this XML format? It would need to be defined by a central entity.*
        - *Is there a reason why XML is chosen as the transmission format instead of a protocol like SWORD (with a common packaging format)?*

- As XML is not human-readable, this implies we'd need a more human-readable format as well (PDF or similar), which is why SWORD may be useful here.

2. Make article available to search engines
   - Google, Google Scholar, Yahoo, Bing, etc
3. Must be able to link to publisher's website
   - **General Comments**
     - *How is the publisher's website link obtained by the repository? Is there a way to "look it up" via a central service, or would it be a required metadata field? If it is the latter, what happens if the publisher's website changes it's URL?*
     - *Is a link to the DOI acceptable?*
4. Support embargo
   - link to publisher's website until embargo period expires
     - *See comments above about how do we obtain the publisher's website link.*
   - make full-text of article available post-embargo
5. Certify compliance with agencies
   - Automatically notify "both the funding agency and the PI's institutional research office that a deposit has occurred"
   - **General Comments**
     - *How would repositories know where to send notifications to? What type of notification?*
     - *Is this a "push" notification (e.g. automated email to agency), or is it more of a "pull" notification (where an agency could query repositories for recent deposits)?*
       - *If the agency just needs to query the repository for recent deposits, perhaps would could use OAI-PMH. But, at the same time, the funding agency couldn't be expected to query 100's of repositories for this data. It'd need to be a centralized location that could be queried*

Requisite Conditions

As noted in the proposal, the "following precursors are required immediately to implement SHARE as a solution to the OSTP memorandum.":

1. Principal Investigator (PI) Identifier (recommended to use either ORCID or ISNI)
   - **General Comments:**
     - *Is capturing this identifier as a simple metadata field "good enough"?*
     - *Are researchers expected to just enter their own ORCID? Or do we need some sort of more complex "lookup" for each author entered?*
2. Award Identification Number - assigned by Federal agencies
   - **General Comments:**
     - *Unclear if this is a single field or multiple fields. It seems as though we'd need to also store information about the agency who assigned the number (for uniqueness), unless the "identification number" includes a code which identifies the agency.*
3. Copyright License Terms - "requires a standardized and coded expression ... for machine processing"
   - **General Comments:**
     - *How would this be "coded"? We'd need a centrally defined "standard" representation that all repositories can attempt to implement.*
4. Repository Designation ID Number - "to identify the repository access location"
   - **General Comments:**
     - *Who defines this "number"? Could this simply be the repository URL, or a persistent identifier which resolves to the repository URL?*
5. Preservation Rights - "required to be coded into the metadata residing with the record"
   - **General Comments:**
     - *How would this be "coded"? We'd need a centrally defined "standard" representation that all repositories can attempt to implement.*

Phase ONE (12-18 months)

Additional requirements for Phase One, after which "the SHARE system will be available for both deposit and access".

1. PI Identifier (*Also mentioned in "Requisite Conditions"*)
   - *See comments under "Requisite Conditions" above*
2. Award Number (*Also mentioned in "Requisite Conditions"*)
3. Publication ID - "unique, persistent identifier to reference the journal article of the publication"
   - **General Comments:**
     - *Is this ID assigned by the repository? It's unclear if this is something the repository needs to "lookup" or just assign.*
4. Data Set ID - "resolvable, persistent identifier to location of stored data or data sets that are linked to the published article"
   - **General Comments:**
     - *Where are these data sets expected to reside? Is the repository capturing the dataset and assigning the identifier, or is it assigned by an external system?*
5. Copyright License Conditions (*Also mentioned in "Requisite Conditions"*)
   - includes embargo information
   - *See comments under "Requisite Conditions" above*
6. Sponsoring/Funding Agency Name - "Link to agency providing funding so that reports can be automatically returned"
   - **General Comments:**
     - *If this is primarily used for reporting, it's likely we also need to capture an email address or a URL / identifier. It depends on the decisions around reporting.*
7. Reporting - "Creates a feedback loop to the federal agency and the PI's research office providing tracking of publications resulting from awards funded by the agency"
   - **General Comments:**
     - *What type(s) of reports are expected? How would these be made available to the agency / research office?*
     - *Is this a "pull" (agency/research office can visit the repository and view/request necessary reports), or a "push" (reports are automatically sent from the repository to the agency / research office by some means)?*
       - *As far as repositories are concerned, obviously a "pull" is easier. A "push" would require the repository to know where to send such reports (up-to-date email addresses or similar)*

8. Core Usage Statistics - "Reports to authors (and agencies, if desired) include statistical data on usage activity and downloads of their publications."
    - ***General Comments:***
        - *What type(s) of statistical reports are expected? Would there need to be some "minimal required statistics" to capture/report? How would the reports be made available to the authors and agencies?*
        - *Is this a "pull" (authors/agencies can visit the repository and view/request necessary reports), or a "push" (reports are automatically sent from the repository to the author / agency by some means)?*
            - *As far as repositories are concerned, obviously a "pull" is easier. A "push" would require the repository to know where to send such reports (up-to-date email addresses or similar)*
9. Metadata Exposed to Search Engines
10. SWORD
    - ***General Comments:***
        - *We would need to standardize on a SWORD submission profile / packaging format. As a protocol, SWORD just transmits content and doesn't require a specific format.*
11. OpenURL
12. Some connections to Digital Preservation Network (DPN)? - "All phases connect with and take advantage of the Digital Preservation Network (DPN)"

Phase TWO (6-12 months after phase one)

We have not added any comments on Phase TWO yet, as its vision is still vague. Much of the Phase TWO listed features refer to requirements that are yet to be determined. Others refer to possible enhancements to Phase ONE features, based on usage needs.

Required in support of phase two. Begun "concurrently with Phase One activities".

1. Submission Workflow - "Development of software to automate and optimize article submission from author through repository and to publisher"
    - Requires publishers to comply with single, standardized submission mechanism
2. Usage Metrics
3. Reporting
4. Incorporate OAI-ORE
5. Certification
6. Adoption of Best Practices

Phase THREE

We have not added any comments on Phase THREE yet, as its vision is still vague. Phase THREE features don't have very specific use cases defined, and seem to be almost "brainstorms" of possible future interactions with SHARE.

Phase Three envisions "more complex interactions with SHARE", and includes:

1. Text and Data Mining
2. Bulk Harvesting
3. Semantic Data
    - Relationships among publications
4. API Specifications
    - In support of interation with repositories
5. ResourceSync
    - http://www.niso.org/workrooms/resourcesync/
6. Open Annotation
    - Web-centric annotation framework

Phase FOUR

We have not added any comments on Phase FOUR yet, as is vision is still vague. Phase FOUR features refer to the yet-to-be defined "data requirements of federal agencies". They seem to almost be "brainstorms" of possible options based on those unknown requirements.

Phase Four involves "development of infrastructure relationships to support data requirements of federal agencies"

1. Data Curation and Associated Software
2. Linked Data
3. Shared Distributed Resources in Repositories