SHARE Proposal - DSpace Notes

> (i) **DSpace Notes on SHARE Proposal from ARL/AAU/APLU**
>
> This page contains DSpace-software specific notes in relation to the SHARE Proposal. More general questions/comments on the SHARE Proposal are also available on the page: SHARE Proposal
>
> Some of these notes are also borrowed from the OR13 DSpace Developers Meeting, where the SHARE proposal was discussed briefly:
>
> - SHARE Discussion notes start at: https://docs.google.com/document/d/1QzrmHDvpJlpsVo6e8tJvWk2rXZm1p_F3TaUcEKIeODU/edit#heading=h.salbnzj9k0li
>
> DSpace-specific comments/questions are made throughout the below summary (in italics) and are marked as such.

High Level Summary

This High Level Summary is copied from the SHARE Proposal parent page.

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:

General DSpace Note: DSpace lets you add new metadata schemas or fields (must be of a general format: [schema].[element].[qualifier]). So, it would be possible to add necessary new metadata fields related to SHARE. More details below

1. author
    - *DSpace stores in 'dc.contributor.author' or 'dc.creator'*
2. article title
    - *DSpace tends to capture article titles as the 'dc.title'*
3. journal title
    - *Currently, DSpace only captures a Journal title as a part of the "dc.identifier.citation" field (which is a human-readable citation)*
    - *This likely would need to be a new metadata field added to DSpace*
4. abstract
    - *DSpace captures in 'dc.description.abstract'*
5. award number
    - *Would need to be a new metadata field in DSpace. Unclear whether this is a single field, or if we'd also need to store the Agency information (the agency who assigned this number)*
6. Principal Investigator ID (ORCID or ISNI)
    - http://www.isni.org/isni_and_orcid
    - *Would need to be a new metadata field in DSpace.*
        - *Possible issue: currently, it's not possible to associate related metadata fields in DSpace. So, simply storing an ORCID as metadata may be problematic, as it may not be associated with the proper author entry.*
    - *Some additional notes on ORCID + DSpace are at: ORCID Integration*
    - ***Unclear if this is just storage of an ORCID, or something beyond that?***
7. designated repository number
    - *Perhaps this is just the "site" handle for a DSpace repository? e.g. [handle-prefix]/0 is the Handle / Presistent Identifier for a site (although currently it does not "resolve" via hdl.handle.net)*

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"
    - *DSpace supports/recommends deposit licenses. The sample license provided with DSpace (based on MIT deposit license) seems like it may cover all these use cases.*

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"
   - ***Unclear what this means for DSpace. May need more clarification around use cases.***
   - *DSpace does support SWORD (v1 & v2) which could be used for this. But, it just treats XML documents like a digital document (and cannot do anything special with them by default)*
2. Make article available to search engines
   - Google, Google Scholar, Yahoo, Bing, etc
   - *DSpace tries to keep up with SEO (Search Engine Optimization). We've worked directly with Google Scholar folks to make ourselves more easily indexable.*
3. Must be able to link to publisher's website
   - ***Unclear how we obtain this link. Would it be actually "stored" in DSpace, or would it be a "lookup" against an external service?*** *Either way there is likely some work to be done in DSpace to support this.*
4. Support embargo
   - link to publisher's website until embargo period expires
   - make full-text of article available post-embargo
   - *DSpace has embargo functionality that should meet these needs, provided we can determine the link to the publisher's website.*
5. Certify compliance with agencies
   - Automatically notify "both the funding agency and the PI's institutional research office that a deposit has occurred"
   - ***Unclear how this would work.***
     - *If this is a "pull" (agency can query repositories), then DSpace could already support this via OAI-PMH.*
     - *If it's a "push" (repository needs to notify agency), then it might be possible to add a new email notification feature (though we'd need to know who to notify via email).*

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)
   - *See notes above under "Minimum SHARE metadata fields"*
2. Award Identification Number - assigned by Federal agencies
   - *See notes above under "Minimum SHARE metadata fields"*
3. Copyright License Terms - "requires a standardized and coded expression ... for machine processing"
   - *How would this be "coded"? We'd need a centrally defined "standard" representation that all repositories can attempt to implement.*
   - *DSpace currently only stores licenses as plain text.*
   - *DSpace stores embargo information in the database, so that part is "machine actionable"*
   - *Present, but may need cleanup*
     - *Creative commons is possible, on the item level*
     - *Embargos are possible, even on the level of individual bitstreams*
     - *There are collection & community license text*
     - *There's the item license text that is accepted at the end of submission.*
4. Repository Designation ID Number - "to identify the repository access location"
   - *See notes above under "Minimum SHARE metadata fields"*
5. Preservation Rights - "required to be coded into the metadata residing with the record"
   - *How would this be "coded" (PREMIS?)? We'd need a centrally defined "standard" representation that all repositories can attempt to implement.*
   - *DSpace doesn't make it entirely clear what the difference is between copyright license and preservation rights. Depends on how the institution fills out the different license texts.*

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 notes above under "Minimum SHARE metadata fields"*
2. Award Number (*Also mentioned in "Requisite Conditions"*)
   - *See notes above under "Minimum SHARE metadata fields"*
3. Publication ID - "unique, persistent identifier to reference the journal article of the publication"
   - *For DSpace, this could be the item handle assigned by the repository*
4. Data Set ID - "resolvable, persistent identifier to location of stored data or data sets that are linked to the published article"
   - *For DSpace, this could be the item handle assigned by the repository.*
   - *However, it might be tricky to link a data set to the associated article.*
   - *If the data set resides outside of the repository, this could be captured by a metadata field on the journal article which stores the location (URL) of the data set*
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"
   - *Could just be a new metadata field in DSpace*
   - *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"
   - *For DSpace, this could be supported via OAI-PMH, if the agencies regularly harvest this information from repositories.*
   - *But, it's unclear if the repository is expected to push this information to the agencies (currently not supported by DSpace)*

8. Core Usage Statistics - "Reports to authors (and agencies, if desired) include statistical data on usage activity and downloads of their publications."
    - *DSpace currently captures usage statistics (views/downloads) on all items in the repository*
    - *However, statistics are just displayed in the User Interface. There are not any statistical reports (e.g. emailed reports) generated at this time*
9. Metadata Exposed to Search Engines
    - *DSpace exposes all its metadata (for public items) to search engines and tries to keep up with latest SEO best practices.*
10. SWORD
    - *DSpace already supports both SWORD v1 and v2 (servers). It also has a SWORD v1 client which can submit content to another system via SWORD.*
11. OpenURL
    - *May already be present in DSpace, depending what the exact requirements are.*
    - https://wiki.duraspace.org/display/DSDOC3x/Configuration#Configuration-JSPUI/XMLUISFXServer
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