SHared Access Research Ecosystem (SHARE)

June 7, 2013

DRAFT

Association of American Universities (AAU)
Association of Public and Land-grant Universities (APLU)
Association of Research Libraries (ARL)

This draft proposal is under review for adoption and implementation by AAU, APLU, and ARL.

SHared Access Research Ecosystem—SHARE

Research universities are long-lived and are mission-driven to generate, make accessible, and preserve over time new knowledge and understanding. Research universities collectively have the assets needed for a national solution for enhanced public access to federally funded research output. As the principal producers of the resources that are to be made publicly available under the new White House Office of Science and Technology Policy (OSTP)[1] memorandum, and that are critical to the continuing success of higher education in the United States, universities have invested in the infrastructure, tools, and services necessary to provide effective and efficient access to their research and scholarship. The new White House directive provides a compelling reason to integrate higher education's investments to date into a system of cross-institutional digital repositories that will be known as SHared Access Research Ecosystem (SHARE).

SHARE envisions that universities will collaborate with the Federal Government and others to host cross-institutional digital repositories of public access research publications that meet federal requirements for public availability and preservation. Universities already own and operate key pieces of the infrastructure, including digital institutional repositories, Internet2, Digital Preservation Network (DPN)[2], and more. These current capacities and capabilities will naturally be extended over time. Universities have also invested in recent years in working with Principal Investigators and other campus partners on developing digital data management plans to comply with agency requirements.

There are also compelling business interests for higher education investments in this system. The current publishing structure for research and scholarly literature effectively manages peer review and editing, but limits its usefulness by restricting access to the breadth and depth of the literature. Limited access, particularly to the research that results from federal funding, constrains new academic programs, such as those that seek to engage in computational analysis of the research corpus and more. In this business context, an important goal of both SHARE and the Federal Government is to ensure that we maximize the value of research funding. SHARE fully embodies the spirit of the OSTP directive of February 22, 2013, and will efficiently and effectively provide all functionalities that the directive envisions to the public, commercial, and scientific communities.

How SHARE Works

University-based digital repositories will become a public access and long-term preservation system for the results of federally funded research. SHARE achieves the mission of higher education by providing access to and preserving the intellectual assets produced by the academy, in particular those that are made openly available. Adopting a common, brief set of metadata requirements and exposing that metadata to search engines and other discovery tools will federate existing university-based digital

[1] White House Memorandum on "Increasing Access to the Results of Federally Funded Research," February 22, 2013.
[2] http://www.dpn.org

repositories, obviating the need for a central digital repository and leveraging the considerable investments already made by universities and their libraries over the last decade.

Agencies that choose to develop their own digital repositories, or work with an existing repository such as PubMed Central, could simply adopt the same metadata fields and practices to become a linked node in this federated, consensus-based system. Discipline-based repositories, some of which are housed at universities, will be included. Minimum standard SHARE metadata fields will include author, article title, journal title, abstract, award number,[3] Principal Investigator ID (e.g., ORCID or other similar tools), and designated repository number. These metadata fields could change over time, with additional fields added. A prudent, predictable level of content and infrastructure redundancy among repositories is anticipated and consistent with best practices for digital preservation.

University digital repositories exist widely and together with others have the capacity to house the corpus of articles arising from federally funded research, but not all universities have digital repositories. Within the SHARE framework, every university or research institute that accepts federal research funding will have the opportunity to designate an existing university digital repository (its own repository or another as outlined in the following paragraph) as the site where its articles will be deposited for public access and long-term preservation. Institutions will build the designation and identification of repositories into their local grants management function at the time a grant is awarded.

Every state in the U.S. has one or more state-funded universities; most of those institutions already have repositories that can fulfill the public deposit requirements for any Principal Investigator (PI) at an institution in a state that does not already have a digital repository or has not otherwise designated one. If in some state, no state-funded university has and none can build a digital repository to fill this function, an institution there will partner with another state-funded university. Such partnerships are consistent with existing culture and practice of sharing within the Association of Public and Land-grant Universities (APLU) and other inter-institutional relationships, including a number of private-public university collaborations.

SHARE will be functional for all Principal Investigators when the federal agency policies go into effect. It is envisioned that SHARE will have a policy advisory board that will include representatives of all stakeholders including representatives of federal agencies. This will ensure interoperability of policies and a single point of contact for federal agencies.

For the White House policy to succeed, as a condition of awarding grants, federal agencies will need to require the following of universities, and universities will then need to require of their Principal Investigators:

[3] Use of the term "award" encompasses grant, contract, and award.

- Sufficient copyright licenses to enable permanent archiving, access, and reuse of publications.[4]

- That scholarly manuscripts arising from grants and submitted for publication to scholarly journals include the award identifier, PI number, and the digital institutional repository in which it will reside post-publication.

The SHARE workflow is straightforward, and using existing protocols can be fully automated.

1. PI or author submits manuscript to journal as currently occurs.

2. Journal publisher coordinates peer review, accepts, and edits manuscripts as currently occurs.

3. Journal submits XML version of the final peer reviewed manuscript (including the abstract) to the PI's designated repository, or the author submits the final peer-reviewed and edited manuscript accepted for publication (including the abstract) to the PI's designated digital repository.

Upon ingest of the article, designated SHARE repositories will make abstracts and metadata available to commercial search engines (e.g., Google, Google Scholar, Yahoo, Bing, etc.) and other discovery tools.[5] If there is an embargo period, the repository will link to the publisher's website for the duration of that period and will make the full text of the article available upon its expiration. The repository will continue to link to the publisher's website post-embargo.

Designated SHARE repositories will use automated methods of certifying compliance with the agency requirements by notifying both the funding agency and the PI's institutional research office that deposit has occurred. Existing protocols such as SWORD (http://swordapp.org/about/) will enable the repositories to designate any additional recipients of either notice of deposit or a copy of the article itself, subject to embargo restrictions.[6]

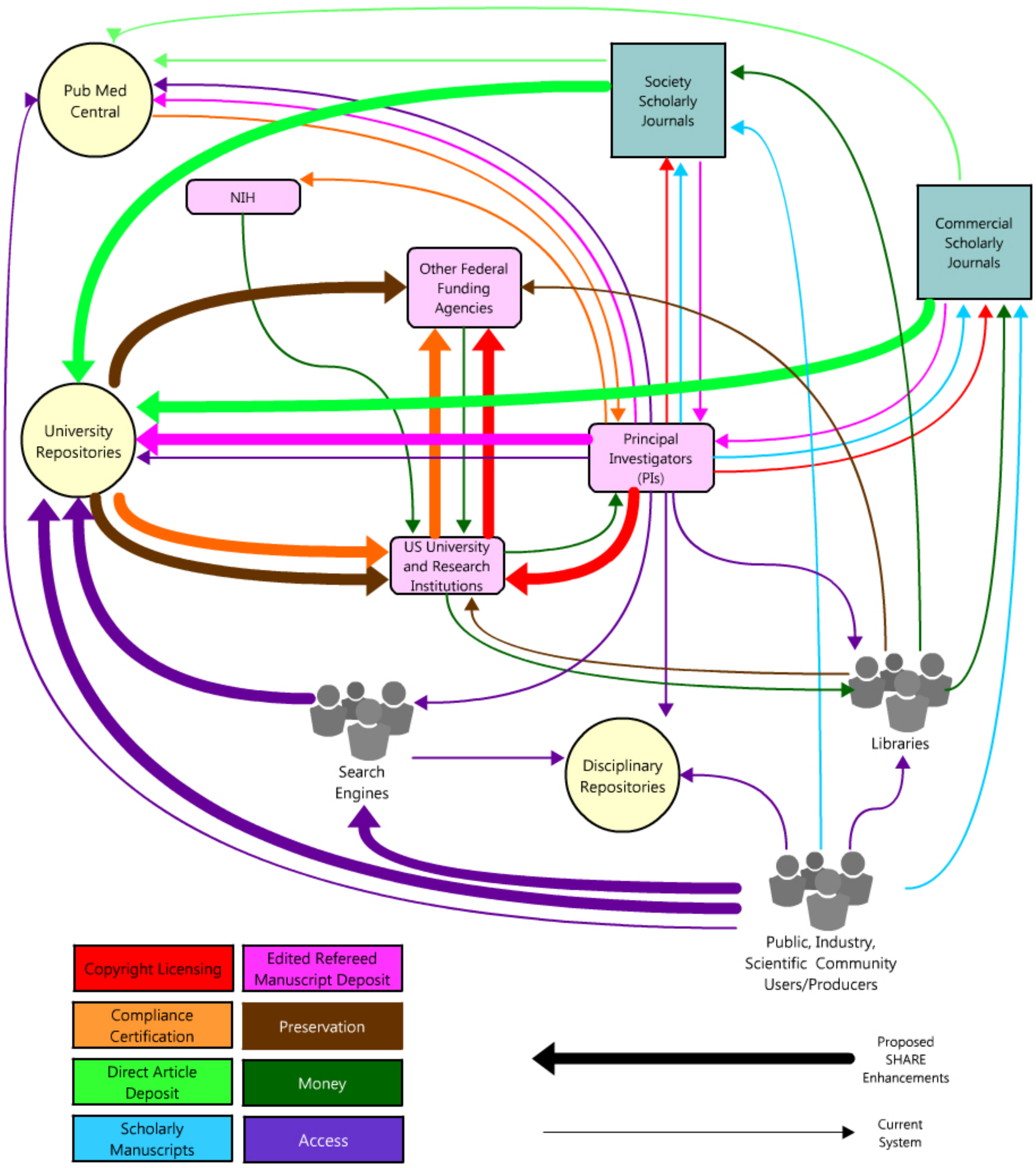
In sum, in collaboration with others in the public and private sectors, SHARE will make federally funded research resources publicly available. Technical descriptions of

[4] "Copyright licenses" refers to the copyright rights of the Principal Investigator (PI) that must be licensed to federal agencies.
[5] Researchers, scholars, members of the public, and commercial interests all rely on commercial search engines for discovery. As SHARE evolves, there will be a continued examination of privacy practices of these commercial search engines and their suitability for SHARE searches.
[6] SWORD (Simple Web-service Offering Repository Deposit) is an interoperability standard that allows digital repositories to accept the deposit of content from multiple sources in different formats (such as XML documents) via a standardized protocol. Wide adoption will lower the administrative overhead on PI's trying to manage and track deposits in more than one repository or between publisher and repositories.

SHARE, phases of deployment and functional properties of a public access system follow.[7]

The following flowchart depicts the current research funding and dissemination system and proposed enhancements to it by SHARE.

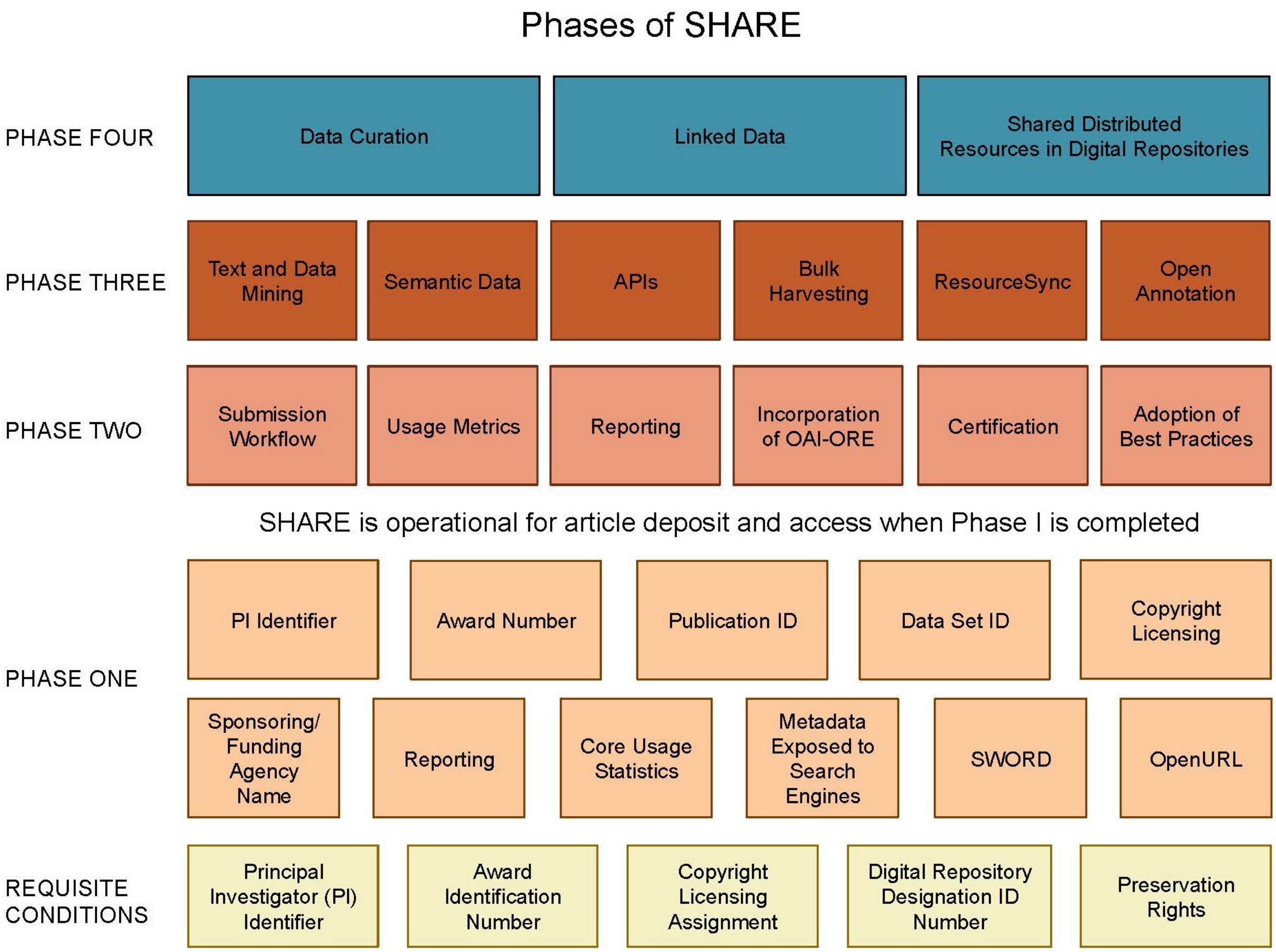
SHared Access Research Ecosystem


Pub Med Central
Society Scholarly Journals
NIH
Other Federal Funding Agencies
Commercial Scholarly Journals
University Repositories
Principal Investigators (PIs)
US University and Research Institutions
Search Engines
Disciplinary Repositories
Libraries
Public, Industry, Scientific Community Users/Producers
Copyright Licensing
Edited Refereed Manuscript Deposit
Compliance Certification
Preservation
Direct Article Deposit
Money
Scholarly Manuscripts
Access
Proposed SHARE Enhancements
Current System

[7] There may be some movement within the phases of SHARE as certain functionalities may be available within a shorter time span.

Phases of SHARE (SHared Access Research Ecosystem)

SHARE will have a phased rollout with increasing functionality that will be developed within the community, both during initial implementation and over time.



Phases of SHARE

PHASE FOUR	Data Curation	Linked Data	Shared Distributed Resources in Digital Repositories			
PHASE THREE	Text and Data Mining	Semantic Data	APIs	Bulk Harvesting	ResourceSync	Open Annotation
PHASE TWO	Submission Workflow	Usage Metrics	Reporting	Incorporation of OAI-ORE	Certification	Adoption of Best Practices

SHARE is operational for article deposit and access when Phase I is completed

PHASE ONE	PI Identifier	Award Number	Publication ID	Data Set ID	Copyright Licensing	
	Sponsoring/ Funding Agency Name	Reporting	Core Usage Statistics	Metadata Exposed to Search Engines	SWORD	OpenURL
REQUISITE CONDITIONS	Principal Investigator (PI) Identifier	Award Identification Number	Copyright Licensing Assignment	Digital Repository Designation ID Number	Preservation Rights	

REQUISITE CONDITIONS: The following precursors are required immediately to implement SHARE as a solution to the OSTP memorandum.

- *Principal Investigator (PI) Identifier*—Used to disambiguate author names, this identifier would be required to resolve the problem of consistency of referring or referencing author names across their publication history. Adoption of either ORCID or ISNI (http://www.isni.org/isni_and_orcid) to provide a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. ORCID and ISNI are interoperable.

- *Award Identification Number*—Assigned by Federal agencies, the award number is used to tie together awards with research output, including publications and other outcomes such as data management plans and data.

- *Copyright License Terms*—Statement of what copyright provisions apply to the publication. Requires a standardized and coded expression that is embedded in the associated metadata for machine processing.

- *Repository Designation ID Number*—Data field required to identify the repository access location.

- *Preservation Rights*—Assignment by author to the hosting repository and any subsequent preservation archive; the preservation rights to the final published version (e.g., in the event the publisher ceases business). Required to be coded into the metadata residing with the record.

PHASE ONE: The target for completion of Phase One requirements and capabilities is within 12–18 months. Existing standards and protocols will be utilized. When Phase One is complete, the SHARE system will be available for both deposit and access and sufficient capacity will be available to include all peer-reviewed scholarly publications produced by all PIs that are to be made publicly available under the White House Office of Science and Technology Policy (OSTP)[8] memorandum.

- *PI Identifier*—Used to disambiguate author names, this identifier would be required to resolve the problem of consistency of referring or referencing author names across their publication history. Identifiers such as ORCID or ISNI could be employed. Mandatory use of adopted convention to provide a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers.

- *Award Number*—Assigned by federal agency, the award number follows the publication as part of the award's metadata.

- *Publication ID*—Unique, persistent identifier to reference the journal article of the publication.

- *Data Set ID*—Resolvable, persistent identifier to location of stored data or data sets that are linked to the published article.

- *Copyright License Conditions*—Resolve questions regarding embargoes upfront, allows repositories to know when to make accessible the deposited publication.

[8] White House Memorandum on "Increasing Access to the Results of Federally Funded Research," February 22, 2013.

- *Sponsoring/Funding Agency Name*—Link to agency providing funding so that reports can be automatically returned, leveraging other existing investments, when possible.

- *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.

- *Core Usage Statistics*—Reports to authors (and agencies, if desired) include statistical data on usage activity and downloads of their publications.

- *Metadata Exposed to Search Engines*—SHARE will expose its content to public and commercial search engines to ensure the widest possible public access to federally funded research.

- *SWORD*—Incorporation and use of the SWORD protocol to lower the barriers to deposit.

- *OpenURL*—Incorporation of a standardized format of Uniform Resource Locator (URL) intended to enable Internet users to more easily find a copy of a resource that they are allowed to access. The OpenURL is tagged with the metadata record to enhance connectivity between the repository and the article on the publisher's website for final published version. The National Information Standards Organization (NISO) has developed OpenURL and its data container (the ContextObject) as American National Standards Institute (ANSI) standard Z39.88.

PHASE TWO: Development of software required to support Phase Two should begin concurrently with Phase One activities. The target for completion of Phase Two requirements and capabilities is 24 months from initiation of implementation, i.e., 6–12 months after completion of Phase One.

- *Submission Workflow*—Development of software to automate and optimize article submission from author through repository and to publisher. Envisions automatically assigning and populating fields specified in Phase One, and requires publishers to comply with a single standardized submission mechanism that will be deployed.

- *Usage Metrics*—Development of more sophisticated metrics which move beyond individual article usage and place the article in the context of the wider ecosystem and reporting aggregate network effects.

- *Reporting*—Generation of reports to funding agencies as requirements are determined, as part of an interactive feedback loop.

- *Incorporation of OAI-ORE*—Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of web resources (http://www.openarchives.org/). These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation.

- *Certification*—Evidence that the PI has complied with agency mandated compliance requirements. Due to scale considerations, requires a standardized way to encode metadata so exceptions can be identified and reported by machine.

- *Adoption of Best Practices*—Because the nature of the decentralized and federated repository system will contain a variety of different repository systems, best practices will be shared within the community to optimize the system.

PHASE THREE: Envisions more complex interactions within SHARE, where real added value from federating the content of federally funded research investment can be demonstrated.

- *Text and Data Mining*—Full text of a small number of articles in XML format will enable advanced search and discovery as well as reuse. As mentioned previously, copyright licenses need to be granted on a non-exclusive basis to the funding agency and, in turn, to universities.

- *Bulk Harvesting*—Supports any requestor, under a defined set of circumstances, to request a bulk feed or download the corpus of records and publicly available associated articles from target repository server.

- *Semantic Data*—Incorporating the ability to extract meaning from the relationships among publications.

- *Application Programming Interface Specifications (APIs)*—Development of support for standardized APIs to improve interaction with repositories.

- *ResourceSync* (http://www.niso.org/workrooms/resourcesync/)—Leverage the research and development of new open standards on the real-time synchronization of web resources. Increasingly, large-scale digital collections are available from multiple hosting locations, are cached at multiple servers, and leveraged by several services. The proliferation of replicated copies of works or data on the Internet has created an increasingly challenging problem of keeping the repositories' holdings and the services that leverage them up-to-date and accurate. As we move from a web of documents to a web of data, synchronization becomes even more important: decisions made based on unsynchronized or incoherent scientific or economic data can have serious deleterious impacts. Incorporation or adoption of the tool will save a tremendous amount of time, effort, and resources for repository managers through the automation of the replication and updating process.

- *Open Annotation*—Aimed at specifying a web-centric annotation framework useable for scholarly applications to annotate articles in the ecosystem, i.e., that supports, linking, relating, comparing, referencing, illustrating, teaching, and other activities that are integral to scholarship.

Note: academic research programs are rapidly developing strategies centered on the challenges of big data and correspondingly the development of data science or data analytics. The corpus of digital repository content, both full text articles as well as the associated data sets, will provide a rich resource for these research programs to experiment with, test and develop new methods to extract meaning and relationships from the repositories.

PHASE FOUR: Development of infrastructure relationships to support data requirements of federal agencies. Although this functionality will exist across SHARE, this phase will likely entail a smaller subset of participating share repositories.

- *Data Curation and Associated Software*—Potential development of repository mechanisms to meet federal requirements for data submission.

- *Linked Data*—Build on semantic web concepts to provide access to and rich associations between open data.

- *Shared Distributed Resources in Repositories*—Explore use of output of the Shared Canvas project (Stanford University and Herbert Van de Sompel; http://www.shared-canvas.org/) to establish a standard mechanism for annotating and rendering potentially distributed resources that exist in physically and digitally separate repositories.

Note: All phases connect with and take advantage of the Digital Preservation Network (DPN) now under development and funded predominantly by research libraries.

Functional Properties of Public Access Repositories

The OSTP memorandum wisely provides agencies with latitude in developing and maintaining the public access repositories called for, stipulating that such repositories could be maintained by the agency funding the research, another federal agency, or through a public/private partnership with external parties. AAU, APLU, and ARL believe that the following functional attributes are needed for any system to achieve the public access goals of the OSTP policy directive. The SHARE proposal achieves these functionalities:

1. Copyright licenses to allow public access uses of publications resulting from federal awards need to be awarded on a non-exclusive basis to the funding agency responsible for deposit in order for that system of public deposit to work. Agencies should therefore add several new conditions to research awards in order to ensure the successful implementation of the White House public access policy, including enabling universities and other entities to better manage compliance with agency regulations. To the extent possible, requirements should be comparable across agencies to minimize the burden on universities of mandated compliance requirements.

    - Federal funding agencies need to receive sufficient copyright licenses to peer-reviewed scholarly publications (either final accepted manuscripts or preferably final published articles) resulting from their grants to enable them to carry out their roles in the national public access scheme. Such licenses would enable the placement of peer-reviewed content in publicly accessible repositories capable of preservation, discovery, sharing, and machine-based services such as text mining, once an embargo has expired.

    - Federal agencies should require the use of persistent, unique identifiers for awards, publications, data, and authors to foster reuse of content and enable better grant tracking and the development of new services by individuals and machines.

2. Federal agency policies should stimulate the development of new tools and services (human and machine), and licensing arrangements should ensure that no single entity or group secures exclusive rights to publications resulting from federally funded research that would inhibit or prohibit access and use of such services and tools.

3. Publications resulting from federal funding and subject to public deposit requirements include the final published version of the peer-reviewed article or the final, peer-reviewed manuscript accepted by a journal for publication.

4. Federal agency compliance requirements should be transparent, and deposit requirements should be easy for the researcher—or institution or publisher depositing on behalf of the researcher—to accomplish. Articles shall either be deposited in a public access digital repository by the journal that publishes the article or by the author directly, and in a manner that “ensures full public access to publications” and includes metadata without charge in a data format that ensures

interoperability with current and future search technologies. The metadata should be publicly available once a peer-reviewed manuscript is published and should provide a link to the location where the full text and associated supplemental materials will be made available after the embargo period.

5. Once the embargo period for public access to an article in the digital repository has expired, restrictions on the manner in which the article is accessed, utilized, or downloaded from the designated depository will be those established by the copyright licenses granted to the federal funding agency by the author and awardee at the time the agency funded the research from which the publication arose.

6. Following an embargo period, full-text search of items in repositories shall be permitted, as well as keyword and metadata-based searches. Open standards are necessary to ensure interoperability in the digital repository system design for search and discovery, and the metadata describing publications should be based on open standards to ensure that the public can read, download, and perform text mining on the publications.

7. Final peer-reviewed scholarly publications in repositories should be linked openly to their source data to the extent possible to allow for reuse and replication of results, and such links should be established in a generalizable, sustainable manner.

8. Metrics and identifiers should be supported to provide information on access, use, and impact of final peer-reviewed scholarly publications.

9. Access to the repositories by the scientific community, industry, and members of the public, should be permitted and encouraged without login, credentialing, or individual tracking, consistent with requirements for maintaining the integrity of those repositories.

10. Final peer-reviewed scholarly publications resulting from publicly funded research should be accessible to persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973. XML is the optimal accessible format.

11. Bulk downloads of the federally funded corpus of scholarly publications for research purposes should be allowed under terms and conditions fostered by the agencies' copyright license terms intended to encourage research while protecting the integrity of the scientific record.

6/7:2:00