arXiv Feedback on the Guidance on the Implementation of Plan, January 31, 2019

We applaud the cOAlition S's efforts towards a system of scholarly publishing that is more accessible, efficient, fair, and transparent. We appreciate this opportunity to provide feedback on the Guidance on the Implementation of Plan S. We full heartedly support a vision of a scholarly publishing system that provides immediate, free, and largely unrestricted use and re-use of scholarly publications. We would like to comment on the compliance requirements that apply to institutional and subject repositories such as arXiv. First, we will share some background information about arXiv operations to highlight the importance of factoring in the current capabilities and business models of the existing services to the Plan S implementation requirements.

About arXiv

Twenty-seven years ago, Paul Ginsparg began a project on his desktop to allow fellow physicists to share unpublished academic manuscripts efficiently without photocopying and paper mail. Today, arXiv boasts 1.5 million papers with 600 papers submitted each day and 7 paper downloads every second from users from all around the world. There are no fees associated with accessing or depositing content to arXiv. The repository has grown significantly and transformed the scholarly communication infrastructure of multiple fields of physics, and it plays an increasingly prominent role in mathematics, computer science, quantitative biology, quantitative finance, and statistics. Last year, it expanded its scope to include electrical engineering and systems science and economics.

From the users' perspective, arXiv continues to be a successful, prominent service, meeting the needs of many scientists around the world. It has an international scope, with submissions and readership from around the world, and collaborations with U.S. and foreign professional societies and other international organizations. It is an essential component of scientific communication, necessary to rapidly and widely disseminate their findings, establish priority of their discoveries, and seek feedback to help improve their work. arXiv is a repository for scholarly materials, including preprints, Version of Record, Author Accepted Manuscript, and postprints. We keep a permanent record of every submission and version posted, in order to provide perpetual access to the scholarly record. DOI and journal reference fields are provided so that the authors can indicate when their papers are formally published. Additionally, arXiv collaborates with some publishers and service providers (such as Inspire) to automatically update arXiv metadata with the DOI and journal references of published versions. Currently, approximately 60% of arXiv articles include DOIs or journal reference numbers.

Cornell University holds the overall administrative and financial responsibility for arXiv's operation and development, with guidance from its Member Advisory Board (MAB) and its Scientific Advisory Board (SAB). Since 2010, arXiv's funding and governance has been based on a membership program engaging libraries and research laboratories worldwide that represent the repository's heaviest institutional users.[1] arXiv operates based on a shoestring budget as the current sustainability model supports the baseline

[1] The financial model entails three sources of annual revenues: $170,000 in-kind contribution from Cornell, $400,000 from the Simons Foundation, and $550,000 from about 230 libraries and research labs from 26 countries that represent the heaviest users of arXiv.

operation with a focus on maintenance and daily operation—in other words, what we call "keeping the lights on." In 2016, based on grants and gifts, the arXiv team embarked on the next-generation arXiv (arXiv-NG) initiative and has been following a strategy of incremental and modular renewal of the existing arXiv system. arXiv will need to identity additional resources to implement some of the changes required to make the service Plan S compliant, taking into consideration both one-time development and ongoing maintenance costs.

arXiv Feedback on the Guidance on the Implementation of Plan S

Below, please find our comments and questions that apply to specifications provided in Section 10 of the implementation document, *10. Deposition of Scholarly Content in Open Access Repositories.* These are our preliminary comments based on our interpretation of the requirements as we try to understand the role of arXiv as a subject repository in ensuring compliance, especially vis-à-vis the publishing (version-of-record) functionality of journals. Also, we aim to explore the compliance requirements of the overlay journals that are built on the arXiv platform.

10.2 Requirements for Plan S compliant Open Access repositories:

Automated manuscript ingest facility

- Our interpretation of this requirement is that compliant repositories should provide mechanisms for programmatic accession of Plan S compliant publications from publishers or others. We recommend making it clear that providing an API or an FTP-based protocol for deposit to publishers or other authorized entities satisfies this requirement.

Full text stored in XML in JATS standard (or equivalent)

- Text & data mining is a big tent, and different user groups have differing expectations about delivery formats. We recommend striking the reference to JATS/XML, as it gives the misleading impression that this specific delivery format is significant for compliance.
- Our interpretation of this requirement is that the ultimate goal is to make content readily available for text and data mining by whatever means practicable. We recommend clarifying the phrase "or equivalent" by enumerating salient requirements, specifically:
    - That plain text content suitable for TDM be made freely available, e.g., via an API.
    - That metadata and content be delivered in a format suitable for computational/programmatic consumption (e.g., as JSON, XML, or another serialization format).
- As articulated in the Confederation of Open Access Repositories' response, we recommend that rather than requiring support for TDM to be a part of a compliant repository system itself, TDM may be facilitated through external services that aggregate and convert resources into text-minable format. There already exist numerous such services that specialize in TDM

use-cases, such as Semantic Scholar and Open MinTeD, and that are in a position to address TDM requirements most effectively.

Quality assured metadata…including information on the DOI of the original publication, on the version deposited (AAM/VoR), on the open access status and the license of the deposited version. The metadata must fulfill the same quality criteria as Open Access journals and platforms (see above). In particular, metadata must include complete and reliable information on funding provided by cOAlition S funders.

- We recommend that this phrase be clarified to emphasize the role of the author and/or publisher in assuring the quality of metadata for Plan S compliant publications, specifically that Plan S compliant repositories are responsible for accepting, storing, and making available those metadata when provided by the author and/or publisher.

...in standard interoperable format…

- We recommend that this phrase be clarified to emphasize the wide range of formal schema adopted by scholarly repositories around the world. Our interpretation of this requirement is that metadata be made available in a common serialization format such as JSON or XML, that the metadata conform to a formal schema, and that the schema be made publicly available.
- With regard to metadata encoding for machine readability, Plan S should consider recommending the Signposting protocol as an acceptable convention for delivering standardized metadata along with publication content.

Open API to allow others (including machines) to access the content

- We recommend that this statement be clarified to emphasize that a compliant API must be free to access, allowing for the best-practice of requiring authentication information from clients making API requests.

QA process to integrate full text with core abstract and indexing services (for example PubMed)

- We recommend that this point be clarified by emphasizing the role of disciplinary information services in addition to general-purpose indexing platforms like PubMed. We recommend that repositories be required to support abstract and indexing services relevant to their stakeholder communities. This may be fulfilled by delivering standardized metadata (quality assured by the author and/or publisher) and plain text content of Plan S compliant publications in a timely manner, either by providing freely accessible APIs or by pushing content to those services.

Submitted by:

Dr. Erick Peirson, arXiv Lead System Architect, Cornell Computing and Information Science
Dr. Oya Y. Rieger, arXiv Program Director, Cornell Computing and Information Science
Professor Steinn Sigurdsson, arXiv Scientific Director, Pennsylvania State University, Department of Astronomy & Astrophysics
Professor Licia Verde, arXiv Scientific Advisory Board, ICREA & Institute of Cosmological Sciences, University of Barcelona