Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

For NOT-AI-15-045, areas of possible comment include but are not limited to:

1. **Best practices in maintaining public data sharing repositories.***
2. Innovative bioinformatics or data analysis tools or methods for research data visualization that are currently missing from or need to be improved upon in ImmPort.
3. Metadata analysis tools and methodology for extracting new information and knowledge from studies in public data repositories that are currently missing from or need to be improved upon in ImmPort.
4. Existing barriers that prevent maximum utilization of ImmPort including specific obstacles related to accessibility, readability, or usability of data from ImmPort or to the data submission process.
5. Outcomes from utilizing the ImmPort dataset and tools including, but not limited to: new collaborations, manuscripts, grant proposals, research proposals, research funding, and consultations.
6. Ability to use ImmPort in conjunction with other databases and analytical tools.
7. **Other emerging technologies or research initiatives that may impact the future development of ImmPort.***
8. **Data model and data repository infrastructure that support efficient data collection, curation, annotation, integration, and public sharing.***
9. **Data standards and transformation methods for integrating disparate datasets.***
10. Suggestions for improving ImmPort.

*****_Responses below are provided for the **BOLDED** areas above_

Elsevier is appreciative for the opportunity to provide a response to NOT-AI-15-045, a Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services. Our response is split into two parts (this is Part I) and were submitted by Holly Falk-Krzesinski, PhD, Vice President, Strategic Alliances, Global Academic Relations, on behalf of Elsevier, July 30, 2015

1. Best Practices in Maintaining Public Data Sharing Repositories

Regarding research data repositories, we think it is most useful to think in terms of data management plans and preferably discipline-specific data repositories. Elsevier is supportive of mandates for data management plans where researchers/authors have the flexibility to choose where to deposit their data and that data sharing routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). We also recognize that deposit into repositories is not an end in itself, the goal of depositing data should be on enabling reuse, thus it is essential to focus on making repositories and the data therein readily discoverable, e.g., through linking. Importantly, as efforts on research data repositories advance, it will be essential for the NIH to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience are leveraged, a duplication of effort and resources are minimized, quality and trustworthy data is separated from other types of data, data discoverability across multiple repositories is guaranteed, and cost savings and administrative efficiency are maximized.

The new NIH's Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (the Plan) indicates that, "the NIH will expect funded researchers to deposit data in 'appropriate, existing, publicly accessible repositories before considering other means of making data available,' but where needed, NIH will take steps to support the development of 'selected community-based data repositories and standards.' To help researchers find an appropriate repository to deposit their data, NIH will expand its database of existing repositories and plans to develop guidance and criteria to aid researchers in identifying 'acceptable repositories' not funded by NIH." While we are assuredly in favor of establishing authentication methods for data repositories we contend that researchers/authors must have the flexibility to choose where to deposit their research data into repositories as they are most knowledgeable about determining the repository best suited to their data and research. This principle should be at the center of any criteria NIH seeks to develop, and the NIH criteria should not inadvertently limit data publication routes, such as linking data, data journals, interactive data plots, etc.

Rigid repository-prescribing funder-specific mandates might lead to direct depositing of research data to a limited number of more generic repositories, running the risk of losing discipline- and domain-specific repositories that add significant value for data reuse and reproducibility. Similarly, mandates that require depositing to a single funder's repository will lead to fragmentation on the basis of country, which is counterproductive to the ever-expanding global nature of (biomedical) science and creation and use of (biomedical) research data by international teams of researchers working across sectors. Research data should be created in formats that allow deposition in a multitude of repositories, and published or deposited in any repository that best suits the research and the discipline. It is also important for the NIH not to put a policy in place that requires undue burden on researchers. It should take special care to ensure that NIH-supported investigators working in international collaborations don't find that they are required to meet multiple—and especially not disparate—funder data posting mandates.

The NIH needs to be a strong partner in defining data repository quality requirements and ensuring that repositories are validated. This would offer the NIH the opportunity for a more flexible policy that allows research data to be stored at repositories that meet specific the quality levels; more flexibility will facilitate compliance on the part of researchers and their institutions. Moreover, quality of repositories must also relate to unfettered access and linking abilities by multiple stakeholders. Recognizing that quality of data repositories is critical, Elsevier encourages the development of data repository certification standards building on initiatives like the Data Seal of Approval, an effort by several data repositories (working in partnership with other research data community stakeholder groups) to ensure sustainable and trusted data repositories. Data validation and data publishing are areas in which Elsevier has deep expertise that we can lend to this effort. Elsevier's data articles and microarticles (see below) are part of the continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing research data systems and repositories.

One element that Elsevier is interested in working with the NIH on is defining the difference between data posting and data publishing. When researchers *post* a description of their research on the web, it is not validated by peers. When the text describing research is *published*, then others know that the associated research is peer-reviewed and validated, and thus can be trusted. It is important to make a similar distinction between *data posting* and *data publishing*: validating and quality stamping the data is becoming an ever more important element of a data-driven research community. Elsevier has developed a hierarchy of trust levels of data, where all of these issues are being addressed in a step-wise manner (see Figure 1 below). We also developed best-practice solutions for pushing data up in this hierarchy (like data journals, data profiles, data citations. and data linking), and are continuing to develop others (data repositories, data management, and data search). We are furthermore interested in collaborating with NIH and others to increase data trust through development of methods to identify data fabrication and data falsification.

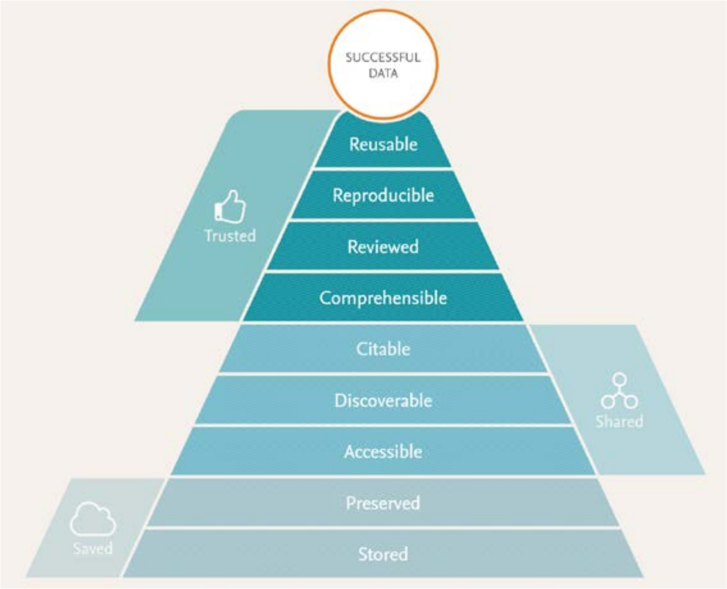

Figure 1: A hierarchy of research data needs. First, research data need to be stored and preserved, so that the data is saved for future use. Second, it needs to be accessible, discoverable and citable, so that other researchers can find and retrieve the data. Last, it needs to be comprehensible, reviewed, reproducible and reusable, so that it can be trusted and built upon.

Data fraud detection tools will need to be an important focal point for NIH as well. In recent scientific fraud causes, fraud was detected as data that was statistically, "too good to be true." Similarly, image manipulation for scientific articles has been observed and is being addressed by a number of publishers at high cost due to the manual labor involved. To avoid future problems and resulting distrust in our data-drive scientific approaches, NLM and publishers will need to work together to find efficient and effective ways to detect data fraud before data sharing and publication.

Elsevier's research data policy (http://www.elsevier.com/about/research-data) commits us to encouraging and supporting researchers to making their research data freely available with minimal reuse restrictions wherever possible. Alongside our policy, we have developed a range of best-practice tools and services to support researchers to store, share, access, and preserve research data. These include our Open Data and Data Profile pilots, our DataLink search tool and database linking program, and our data journals, such as *Genomics Data and Data in Brief*.

Collectively, the NIH should work with other stakeholders in thinking about the big picture goal of enabling researchers to properly collect and annotate their research data initially in ways that lead to archiving, auditing, reproducibility, and interoperability. This might include making vocabularies and other data models available in the researchers' workflow (e.g., controlled vocabularies and drop-downs in Electronic Lab Notebooks; preferred use of DOI's for data sets). This is especially for vocabularies, databases, and other data models that identify entities that define research data (anatomy, diseases, organisms, etc.). Making this available in formats that foster interoperability is a big part of this. This way, unique identifiers and codes are captured early on and can stay with the research data through its entire lifecycle (whether or not research ends up getting published).

Part I: Elsevier's Response to Request for Information NOT-AI-15-045: Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services

Innovation is central area in promoting use of research data and maintaining an open ecosystem while allowing for the creation of services that provide added value. Innovations can range from search services to aggregators and analytical tools. For example, the Open PHACTS project in Europe provides a developer friendly API that enables applications to build across public domain pharmacology data. Their service is supported by pharmaceutical companies through a foundation. Importantly, this service allows proprietary commercial data to sit alongside public data. Three lessons for the NIH arise from this example:

1) Innovation developments should ensure that it is possible to develop a range of services with different business models that store, access, and query various forms of research data. In providing an open model, both in funding and with respect to technological solutions, the NIH can create a flexible framework that allows academic and industry parties to develop components that optimally mesh together and enable systems that can change over time and are tailored to the needs of specific medical and scientific communities;
2) The NIH should seek to develop reporting mechanisms such that downstream aggregators and users can ensure that upstream, publicly funded data providers can receive credit; and,
3) While standardization is helpful for downstream data users, it is important to note that a flexible and open ecosystem can help manage complexity. Therefore, it is preferable to recommend vs. mandate data standards, and any mandates must have the flexibility to allow for change in capabilities and community practice over time.

Elsevier is very interested in supporting a system that evaluates the performance of various components of the biomedical Research Data Management cycle. We are currently actively engaged in a number of conversations with academic and industry partners to enable components to such a shared set of metrics, and systems to support them. We are interested in working in partnership with the NIH and other stakeholders on a workbench that enables quantitative evaluation of the usefulness and usability of different tools pertaining to research data storage, sharing, and search. Questions that one can ask of such a system could include:

- Which data standards, metadata systems, and curation efforts optimally improve outcome of a particular use case, such as data search, or data reuse?
- What metrics can be used for successful data storage or curation: reuse, amount of queries/downloads, or other—possibly social—metrics?
- What systems can act across the spectrum of biomedical repositories, publications, and other research outcomes to track and combine these metrics?

Finally, the NIH should seek opportunities to collaborate effectively with publishers to avoid duplication of effort and costs associated with research data sharing and to minimize administrative costs to research institutions and burden to researchers. By way of example, in conjunction with the Professional and Scholarly Publishing Division (PSP) of the Association of American Publishers (AAP), Elsevier has been involved with the CHORUS service; which leverages existing infrastructure, tools, and services across publishers that have committed to collaboration with federal funding agencies around the public access of research articles.

7. Other Emerging Technologies or Research Initiatives That May Impact the Future Development of ImmPort

Understanding that a recognition economy is the dominant environment in which academic and government researchers operate, it is essential to consider the drivers of research data sharing at the individual researcher level to maximize rapid and efficacious sharing. The NIH needs to address data sharing incentives and rewards for researchers in development of its policies and procedures. Relying only on the "stick" of mandated policy compliance, the full potential to stimulate and motivate broad sharing of research data will go unmet and will face challenges similar to those related to posting to PubMed Central and ClinicalTrials.gov. Elsevier encourages the NIH to review and operationalize the literature that provides an evidence base for understanding what drives researchers to be participatory data donors and we encourage the NIH to develop *new* research funding programs to extend empirical knowledge about this area of science policy. One approach might be for the NIH to partner with the NSF's Science of Science Innovation and Policy (SciSIP) program to develop a research data stream and funding resources to support new research grants in this area.

The free, public Mendeley Research Data Sharing group contains a rich library of such research data sharing resources. Contained therein, references describe the need to develop a reward and recognition system that affords researchers ongoing attribution, recognition, and professional reward for their sharing efforts. The literature also calls on policy makers, funders, and research organizations to consider the resources necessary for researchers and their institutions to comply with policy mandates, such as necessary skills, time & effort, and ongoing finances. Furthermore, the literature demonstrates the need for stakeholders to take into account the impact of sharing and potential for misuse on individual competitiveness, an essential consideration given the current hypercompetitive funding landscape.

8. Data Model and Data Repository Infrastructure That Support Efficient Data Collection, Curation, Annotation, Integration, and Public Sharing

Much of what was presented in Section 1 above is relevant here as well. For example, Elsevier's data articles and data linking program are proven parts of an effective larger data infrastructure.

In its new Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (the Plan), it is very good to see that the, "NIH recognizes the benefit of collaborating with other federal agencies and public and private stakeholders to adopt consistent practices for citation of data sets across scientific communities and other data set attribution systems and will work toward this goal." And a broader context for this can also be found in the HSS Guiding Principles document, which talks about developing healthdata.gov as the basis for a "data commons approach across agencies," specifically the development of an internal HHS Enterprise Data Inventory that will serve as the internal catalog for all HHS data assets and be linked to healthdata.gov, the external-facing platform through which the public will be able locate and access federally funded research data. Next to Elsevier being co-creator of the Force11 Data Citation Principles, it has best-practice linking services that could add to this initiative by expanding the reach of healthdata.gov datasets.

The NIH's recent Plan also explains that "As part of the data discovery index, a system for unique identifiers for datasets generated by NIH-funded research will be developed, analogous to the PubMed Central identification number (PMCID) that is assigned to all submitted publications resulting from NIH-funded research. The identifier would also provide a means of linking the data with the biomedical literature via associated PubMed records." We would like to take this opportunity to share our thoughts around the NIH participating in development of an open, international standard identifier system built on DOIs.

Data DOI's are becoming a globally recognized standard for biomedical and other types of research data identification. Worthy of noting, a number of big data repositories, including the NIH Protein Data Bank (PDB), have assigned DOIs for all its accession numbers. DataCite, for example, has a valuable set of services connected with it offered at no cost and that make it easier to connect with other systems and DataCite has plans to expand its services to accommodate use cases that it currently cannot support (e.g., unpublished data that is early on in the lifecycle, and which is still subject to change). DataCite could be positioned to become a resolver for all other data accession numbers, which simplifies the entire research data infrastructure. The mapping of the Data DOI to an accession number is in the DataCite metadata, and so the DataCite API can be used to map accession numbers and then benefit from metadata for that record in DataCite. Other organizations are also focused on collaborative digital data standards development, including: APARSEN; Opportunities for Data Exchange (ODE); CoData; and, NISO/NFAIS Supplemental Journal Article Materials Project.

Elsevier recommends that NIH focus on the use of Data DOIs as the primary open, international identifier option for data that is published in any formal sense, rather than developing a identifier schema. And if the NIH is to develop a new accession number schema, then it must include assigned DOIs as well.

Elsevier further encourages the NIH to leverage the significant amount of work that has gone into developing common ways to *expose and cite* data. For example, the community effort of the FORCE11 Joint Data Citation Implementation Group has led to the creation of a standard for citing data within article publishing (the NISO JATS 1.1d2 XML schema). The Joint Data Citation Principles has been endorsed by over 90 institutions. The paper, "Achieving human and machine accessibility of cited data in scholarly publications," describes how to operationalize those principles. As described in the Partnership section above, this effort further exemplifies the benefits of collaboration between major stakeholders in the scholarly communication ecosystem, focused on biomedical research and other types of research and data more broadly. By leveraging these community-driven efforts, a common basis for new models of sustainability will emerge.

Elsevier is an active partner with the Research Data Alliance (RDA) and ICSU World Data System (ICSU WDS). With such a wide range of stakeholders across for-profit and nonprofit sectors around the world, and an understanding that biomedical research data is a subset of research data more broadly, it is crucial for the NIH to be partner with these collaborative efforts so as not to duplicate work nor move in a direction specific only to research funded by the NIH.

The basis for Elsevier's involvement in partnerships is that we recognize that creating a research data infrastructure (including the technical infrastructure but also policies, best practices, standards, etc.) has to be a collaborative, cross-stakeholder and international effort where all the different players work together. Elsevier is proud to contribute our deep expertise and perspective from our position as a world leader in research information and appreciate having a voice in development of a synergistic and interoperable emerging research data infrastructure.

The RDA is a great forum for such an approach, as it brings together thought leaders in research data from various stakeholder groups (data centers, research institutes, libraries, publishers, funders, interest group, etc.) and individuals working in the research data field with different expertise and focus, all the way from deep technical expertise to policy-making. The primary value of the RDA is that it has become the forum where stakeholder groups come together to interact and work on issues and focus on making realistic progress on a swift timescale (e.g., 18 mos is the typical lifespan of an RDA working group).

Specifically, Elsevier is involved in a number of working groups under the "Data Publication" umbrella Interest Group (IG) of the Research Data Alliance (RDA) and encourages NIH to join in the partnership. All of these working groups began as ICSU WSD working groups and now have dual ICSU WDS/RDA mandate:

- Data Publication Bibliometrics
- Publishing Data Cost Recovery for Data Centres (for more details, see previous paragraph)
- Data Publication Services

The joint RDA/ ICSU World Data System Publishing Data Cost Recovery for Data Centres scope aligns with this RFI. Co-chair Anita de Waard of Elsevier and her colleagues recently interviewed 22 data centers about their ideas around cost recovery methods, now and in the future. In summary, Elsevier supports the collaborative efforts of the joint RDA/ICSU WDS Interest Group (IG) to elucidate the full cost of data management throughout its lifecycle–from inception through publication to storage and curation—by engaging funders, researchers, repositories, and other stakeholders in the research data management lifecycle. Specifically, the IG finds that data repositories are looking for new funding mechanisms – including charging deposit fees, access fees, and working through public-private partnerships—but are having trouble finding the time and resources to actively explore these new models. Elsevier is very interested in supporting further work regarding these questions, whether within the scope of the RDA or in direct collaboration with the repositories and/or the NIH.

The NIH should strive to work in partnership with other stakeholder groups to develop consistent preservation criteria. To do so, it will be important to address some key questions, such as: Should all versions of data be preserved? Should research data be overwritten with newer data? For how long should data be preserved? Is indefinite preservation sustainable?

9. DATA STANDARDS AND TRANSFORMATION METHODS FOR INTEGRATING DISPARATE DATASETS

Research data adds huge value to the users of published research articles. An important focus is twofold: 1) Attach and make available publicly the methods and data underlying published research; and, 2) Develop standard markups (XML) to allow machine interpretation of the data (this is an area that Elsevier’s Mendeley team is currently working on). It will be important for NIH to work in close partnership with a broad stakeholder group to consider the most effective approach to enforcing data transparency and developing a set of markup standards.

There is a need for data standards, but it should also be recognized that such standards do develop continuously. So any standardization proposal should include a proposal for continuous maintenance and further development of the standard. It should also be noted that data standards have to be discipline, perhaps even subdiscipline, specific, and will always have some element of least common denominator as science, by definition, goes beyond what has been standardized.

Tools for automatic mapping of data would indeed be extremely useful as they can provide the input for data search engines. Furthermore, such tools can help scientists to better comply with funder requirements to share data in a meaningful way, especially when such tools are combined with proper (provenance) annotation capabilities.

Elsevier would be very interested in working with the NIH, other publishers, and data archive managers on mechanisms to connect articles and related datasets. It would be valuable for publishers to link plug-ins into their systems, such that authors could submit the data to the archive of their choice and simultaneously link this to an article.

We also feel that it is important that the NIH work with stakeholders on developing capabilities (at a variety of levels) to validate data and mark it as “OK” following a certain hierarchy of quality, from data has been well-described to data that has been fully reproduced in a different environment by a different team. Elsevier’s data articles and microarticles do provide one of the steps in this continuum of quality/integrity validation, but there are additional levels beyond peer-review that need to be considered and built into developing systems.

With regards to the quality criteria and quality stamps for data archives, there has been considerable discussion in this space, especially in the EU, but it is essential that there be commonly shared view on what a data repositories should adhere to, e.g., the National Digital Stewardship Alliance (NDSA) levels of preservation do make a step in one dimension of data repositories (archives), but there are many more dimensions to consider.

UMLS provides a wide range of medical vocabularies. These by themselves are valuable for determining names of medical concepts and alternative names for the same concepts. More importantly, UMLS maps equivalent notions from different vocabularies. Those notions are classified into a reasonable number of semantic groups, which is helpful for us at Elsevier processes our content and looks for relations between things such as classes of drugs and types of diseases. The UMLS browser is helpful for quick lookups of vocabulary and relation data. NLM also provides tagging tools like MetaMap, useful in work on recognizing medial entity mentions. Elsevier’s EMMeT Taxonomy uses UMLS as the primary source for the taxonomy. ClinicalKey licenses the PubMed taxonomy and proposes its content in the ClinicalKey suite of products. GoldStandard sends its drug data to RxNorm to get it coded. These three resources are very important contributors to our product offerings.

In terms of vocabularies representation and alignment, MeSH and MedDRA are critical resources for our projects. What would be useful in the future would be a “graph of biomedical data” linking biomedical data across MeSH and MedDRA (and ideally all of UMLS) using Linked Data formats. The current work on representing MeSH in RDF is a very exciting step, but a SKOS/SKOS-XL representation would also have a lot of value and would make the integration with our own datasets easier. Elsevier is also interested in the multi-lingual aspect of some UMLS vocabularies, for building cross-language bridges; here again, MeSH and MedDRA are key.