Elsevier Responses to NIH Research Data-Related Requests for Information (RFIs)

- NOT-OD-15-067, Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM) → Refer to Comment 5 of the response only
- NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services → Part I and Part II
- NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories
- NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories
- NOT-OD-17-015, Strategies for NIH Data Management, Sharing, and Citation

Request for Information (RFI): Request for Information (RFI): Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM)

Thank You - Your Comments Have Been Received. You may want to print this page with your comments for your records.

03/09/2015 at 06:04:49:363 PM

Name:

Holly J Falk-Krzesinski, PhD

Email Address:

h.falk-krzesinski@elsevier.com

Name of Organization:

Elsevier

City and State:

New York, NY, USA

Comment 1:

Current NLM elements that are of the most, or least, value to the research community (including biomedical, clinical, behavioral, health services, public health, and historical researchers) and future capabilities that will be needed to support evolving scientific and technological activities and needs.

Responses in Comment 4 and Comment 5

Comment 2:

Current NLM elements that are of the most, or least, value to health professionals (e.g., those working in health care, emergency response, toxicology, environmental health, and public health) and future capabilities that will be needed to enable health professionals to integrate data and knowledge from biomedical research into effective practice.

Responses in Comment 4 and Comment 5

Comment 3:

Current NLM elements that are of most, or least, value to patients and the public (including students, teachers, and the media) and future capabilities that will be needed to ensure a trusted source for rapid dissemination of health knowledge into the public domain.

Responses in Comment 4 and Comment 5

Comment 4:

Current NLM elements that are of most, or least, value to other libraries, publishers, organizations, companies, and individuals who use NLM data, software tools, and systems in developing and providing value-added or complementary services and products and future capabilities that would facilitate the development of products and services that make use of NLM resources.

Elsevier values its multi-faceted and synergistic relationship with the National Library of Medicine (NLM) and is appreciative for the opportunity to provide a response to NOT-OD-15-067, a Request for Information (RFI) Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the NLM.
TAXONOMIES/THESAURI/DATABASES: 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 as 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. Our Natural Language Processing group is a user of both the MeSH thesaurus and its supplementals (mostly drugs and chemical compounds) and of UMLS. We follow the annual update cycle and are quite satisfied in doing so. MeSH is the de facto standard for general Life Sciences /Medical concept annotation. There are domain specific ontologies / thesauri but none beats MeSH in the 'general' area. Recently NLM took the initiative to put out MeSH in RDF format, to connect it to the world as 'linked data', starting from concepts are unique URI identifiers. This is an important initiative that the Elsevier Labs group is glad to be part of, particularly if this project continues to evolve so that the data is truly linked to other resources (PubMed at least) and easily accessible. Our Corporate R&D business unit utilizes and incorporates several NLM elements such as data streams of raw data, bibliographic information, and taxonomies, into several different products. The NLM resources are invaluable as they contain high quality standardized information which we leveraged together with Elsevier content to advance biomedical and health related science. The availability of this high quality and standardized information for researchers is extremely important, and should continue to be an important part of NLM's efforts. As technologies and platforms evolve, the demand for high throughput data retrieval and analysis workflows continues to increase, so it will be beneficial for researchers/corporations if NLM continues to develop its access mechanisms for NLM elements to meet this demand. Specifically, our products Text Mining and Pathway Studio leverage bibliographic, text, and taxonomic/vocabulary data, among other NLM elements. As we merge many data elements together for comprehensive solutions for our customers, we have identified some areas we hope NLM will consider for future development: 1) Convene stakeholder groups in standardizing structures of other biomedical research and health data elements. Similar to the development of the NISO JATS XML standard, NLM could work with stakeholder partners towards either extending this standard or developing new standards such that other data types/formats could also be captured and delivered in a standardized way e.g. electronic health records; 2) While NLM has been involved with ORCID and other unique author identifiers, which are gaining wider use, it would be good if the public could have a better understanding of how these elements are intended to be disseminated, i.e. as part of which data fields in particular record types; and, 3) Further map *and* provide mappings between taxonomies, e.g., UMLS–RxNorm-SNOMED are all very integrated but mapping files between them are complex and somewhat difficult to discern. A potential solution could be API for mapping translation. NLM seems to be keen to improve their services to the community, which we applaud. We'd be interested in a number of developments in this regard: 1) Linking MeSH to other resources that are in the linked-data sphere; provide equivalences (exactMatch, partMatch, etc.) between MeSH concepts and concepts in other taxonomies that are linked-data-enabled, such as NAL, DBPedia etc. 2) Make all NLM vocabularies available by API on a day-to-day basis. Getting access to MeSH is currently non-trivial and cumbersome. Elsevier would appreciate having a query API that allows us to receive updates on at least a weekly basis.
PUBMED/MEDLINE: From the traffic we receive from PubMed to our Health & Life Sciences Content on ScienceDirect we can see how important it is as a discovery tool for researchers in these disciplines. We appreciate how our content is indexed for MEDLINE, especially the assignment of MeSH terms and making these terms available to other search and discovery services. This has a great contribution to the discoverability and dissemination of the content Elsevier publishes. Our general analytical services reporting (commercial and extensive pro bono activities) also benefits from PubMed/MEDLINE through Elsevier's Scopus because of the well-defined/assigned PMIDs. The PMID-DOI converter and API are especially useful.
PUBMED CENTRAL/PUBLIC ACCESS POLICY: Elsevier welcomes the opportunity to enhance delivery of public access through collaboration and interoperability with NLM to avoid duplication and wasted resources. There are opportunities for NLM to collaborate more effectively with publishers in the context of PubMed Central (PMC) to avoid duplication of effort and cost and to minimize administrative costs to research institutions and burden to researchers. One of the significant collaboration opportunities in facilitating public access is via the CHORUS service (http://www.chorusaccess.org/). At Elsevier, we are concerned that the NIH is the only US federal funding agency that has not met directly with representatives of the CHORUS service, and has not considered how this new approach presents opportunities for cost-savings within the NIH budget and for institutions receiving NIH research support. NLM should actively seek opportunities to work with publishers, including integration with CHORUS, to develop and implement open access publication options that leverage existing

infrastructure, tools, and services that support sharing, access, discoverability, reporting, and preservation. It is also important for NLM to recognize that its public access policy's one-size-fits all 12-month embargo period is not suitable for all journals nor for all publishers, and to introduce a petition mechanism, as outlined in the OSTP memo, so publishers can signal these exceptional cases and provide supporting evidence. We would welcome greater sensitivity from NLM colleagues to more clearly distinguish approaches that are effective in the life and biomedical sciences from other disciplinary domains. Finally, while the NLM claims PMC to be a public-private partnership, in practice, the opportunities for collaboration with Elsevier and other publishers have been minimal. Collaboration is a recursive process that relies on continuous lines of open communication; with partners working together to develop and meet shared goals and involves shared governance and review procedures. Elsevier urges NLM to focus on engaging in more genuine collaboration around public access policy and policy implementation. Elsevier requests that NLM share COUNTER-compliant distributed usage statistics for manuscripts in PMC so that publishers can continue to report on impact and usage to authors and to their institutions that subscribe to these publications and pay their publication costs. It is also critical that NLM cease reformatting and enhancing manuscripts to make them appear more like, and substitute for, the final version-of-record of articles. Moreover, it is essential that PMC ensure readers are presented with the best version of the article available, which means that entitled users are transparently linked to the final published version. Finally we believe NLM needs to commit to taking concrete steps to prevent commercial re-use of manuscripts archived in PMC that is not authorized by the copyright holders of these works. PUBLISHING: As a health, medical, life, and biomedical sciences publisher and our involvement with the International Committee of Medical Journal Editors, Elsevier deeply values its collaboration with NLM in setting standards for journal articles and for developing and strengthening policies and practices in the field of publication ethics. NLM's leadership in publication standards makes it a unique participant in the national library space. In particular, we value NLM's commitment to PubMed and ClinicalTrials.gov. PubMed is the medical community standard reference point for article search and ClinicalTrials.gov is a vital mechanism for ensuring accountability, helping to deliver accurate published randomized trial reports and holding authors accountable not only for reporting standards but also for the timely release of their findings. There are areas we believe NLM can make further strides. We feel strongly that NLM should adopt a more global role in fulfilling its mission and responsibilities, with these specific recommendations: 1) Invest in advocacy and infrastructure to advance sustainable platforms for information access in low and middle income country settings to support the health dimensions of the Sustainable Development Goals, e.g., in library services, human resources, national leadership, in partnership with country health sectors; 2) Work to assist countries in developing their capacities for research information generation, publication, and implementation; 3) Partner with journals and publishers to advance these global goals; and 4) Make global equity in information access core to NLM's mission. Elsevier is a proud partner and promoter of the NLM's Emergency Access Initiative (EAI), through which we provide free access to our primary online clinical information and reference tool, ClinicalKey, and to a corpus of relevant literature on our ScienceDirect platform. As a member of the NLM-Publisher Panel, Elsevier is pleased to have a forum to discuss issues of common interest. Topics discussed at recent meetings include the 'Article of the Future' initiative; the MEDLINE submission and review process, including the Literature Selection Technical Review Committee; Emergency Access Initiative; improving access to publisher full-text content; and reproducibility and rigor of research findings. The Panel has provided essential collaboration on these and other initiatives. The Panel can continue to increase its usefulness by addressing additional matters of common interest, for example: 1) Increasing the acceptance rates of evaluated journals and book serials, which would lead to additional high-quality content being available via PubMed; 2) Indexing book content beyond serials as books offer a unique view into biomedical and health related information that is not mirrored in journals, providing an integration of research across time and subject areas, consolidating disparate literatures into one source, and synthesizing research advances and applications; 3) Linking of all information relating to clinical trials, including all articles published as a result of a trial; and, 4) Sharing of clinical trial data, including protocols for how to cite data, where to store data, and how to share data. Elsevier looks forward to our continued participation in the Panel and collaboration with the NLM and other publisher representatives.

Comment 5:

How NLM could be better positioned to help address the broader and growing challenges associated with:

- Biomedical informatics, "big data", and data science;
- Electronic health records;
- Digital publications; or
- Other emerging challenges/elements warranting special consideration.

RESEARCH DATA: Elsevier would like to see the NLM allow mining of all database content inside the suite of databases managed and curated by the NLM and provide actionable copyright metadata elements on all NLM content so we understand what we can mine/use for commercial and non-commercial purposes. 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 tools and services to support researchers to store, share, access, and preserve research data. These include our open data pilot, our database linking program, and our data journals, such as Genomics Data and Data in Brief. Collectively, Elsevier as partners with NLM, we should to be thinking about the big picture goal of enabling researchers to properly collect and annotate their research data 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). 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). 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 NLM 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. Data fraud detection tools will need to be an important focal point for NLM. 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. Regarding research data repositories, we think it is most useful to think in terms of data management plans and data archives. Elsevier is supportive of mandates for data management plans where researchers have the flexibility to choose where to deposit their data and that data publication routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). Importantly, as efforts on research data repositories advance, it will be essential for the NLM to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience as leveraged, a duplication of effort and resources are minimized, and cost savings and administrative efficiency are maximized. 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 NLM, 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 NLM 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. NOTE: Elsevier is also developing a separate and detailed response to the NIH RFI NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories, which we will submit by the deadline of March 18, 2015. ELECTRONIC HEALTH/MEDICAL RECORDS: Since HIPAA issues make it near impossible to obtain actual health records, a test/gold set of anonymized Electronic Health Records would be a great resource to Elsevier to develop and test point of care applications we are currently developing. Also, a test bed EHR/EMR system would be incredibly valuable, where different content providers could plug in applications to show added value of relevant data at the point of care. Elsevier's Health Analytics group is especially interested in developments with regards to EHR/EMR. We are supportive of: 1) Central, anonymized linked patient databases (including detailed clinical encounters in primary and secondary care, medication, genetic data, etc.) for research; and, 2) Central patient records, or at least interoperability standards (including federated search or HIEs) as a method of improving care delivery to individual patients. We encourage the NLM to continue working in coordination with the Office of the National Coordinator for Health Information Technology to drive both of these initiatives. We also want to make sure that NLM is aware of our high-performance computing (HPC) capabilities to analyze data for patterns. Reed Elsevier, Elsevier's parent company, is one of the very few companies in the world that has analytical HPC capabilities and is expert and experienced in dealing with highly confidential and very private data. Regarding the linked patient databases, more (diverse) and bigger (simply more) is better. Broad accessibility (under appropriate safeguards) to the anonymized, longitudinally linked for-research data, including by industry, is desirable for Elsevier's Health Analytics. Industry finances applied research and product development that brings universities' basic research to the point of care and to actually benefit patients. Broad accessibility will also drive innovation from big data, which is currently hindered by selective access. Heath Analytics currently conducts substantial research projects granting us securely anonymized patient data access together with healthcare systems in Europe. Regarding central patient records, comprehensive (all individual patient encounters) and timely is better. As an example, Denmark has introduced a shared medication record. Physicians there can see their colleagues' prescriptions. This transparency among providers is dramatically transforming the Danish healthcare system, already one of the best in the world. Physicians now feel responsible for the full array of prescriptions, even those of their colleagues. Also patients can access and review their complete personal health record, which makes them a responsible partner in their health management. The networking of all players improves patient outcomes substantially.

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.

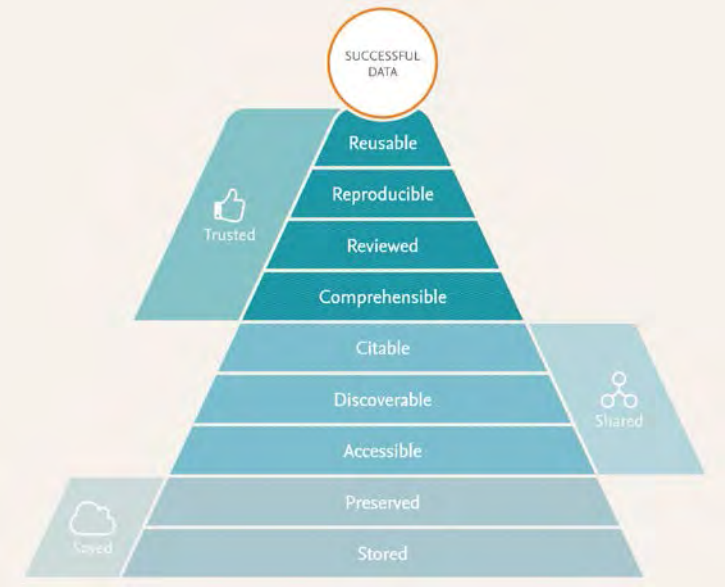


Figure 1: A hierarchy of research data needs. First, research data need to be stored and preserved, so that the data is saved for future use. Second, it needs to be accessible, discoverable and citable, so that other researchers can find and retrieve the data. Last, it needs to be comprehensible, reviewed,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

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

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.

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

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?

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

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.

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

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.

Part II: 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 II) and were submitted by Holly Falk-Krzesinski, PhD, Vice President, Strategic Alliances, Global Academic Relations, on behalf of Elsevier, July 30, 2015

3. METADATA ANALYSIS TOOLS AND METHODOLOGY FOR EXTRACTING NEW INFORMATION AND KNOWLEDGE FROM STUDIES IN PUBLIC DATA REPOSITORIES

Elsevier has a long track record of data and metadata standards, dating back to the 1990s when we led the TULIP project. The Elsevier XML specifications for journal articles and book chapters are widely known and in use for 3000+ propriety and society journals and the metadata for 20,000+ journals. Content, including 12M journal articles, resides in a content repository that is accessible through restful APIs. Its metadata model is described using RDF serialized as JSON-LD. The API payloads and responses in JSON-LD are treated in the same way as our main content standards.

Our content is stored in multiple content-type-specific "warehouses." Through a metadata repository, this is made in to a virtual whole, called our Virtual Total Warehouse. Our content model and metadata standards are especially focused on content versioning. "Generations" of content assets keep various files together that together constitute a version. This Virtual Total Warehouse (VTW) plays a role in acquisition, editing and curating content (in our case, journal articles, book chapters, drug monographs, patents, patient education, and much more) and a Content Enrichment Framework takes this content and can, in principle, run any semantic process on the content, depositing the results back in VTW.

Elsevier also has a linked data repository adhering to the standards of linked data and linked open data.

Part II: 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

Elsevier's approach to unstructured information: The vast majority of information exists as an unstructured text which makes it unsuitable for efficient analysis by humans. The area of computational assistance to analysis of large volumes of textual information is traditionally split into two (somewhat overlapping) approaches - information retrieval and information extraction.

Information Retrieval (IR) systems concentrate on finding documents containing information deemed relevant to a particular topic of interest. Usually this is done by analyzing the word content of the documents using statistical methods based on keywords or word co-occurrence. IR methods are by their nature generic and to a large degree language-independent; the output of IR systems is *intended for human readers*.

Unlike IR, Information Extraction (IE) focuses on extracting information contained within the documents in a form *suitable for automatic processing*. IE systems use an *ontology* (or knowledge representation schema) as a model of a particular domain, and thus are domain-specific. The simplest form of an ontology is a list (or, even better, a hierarchical tree) of concepts relevant to the domain. More advanced forms of ontology also specify possible semantic types of relationships between the concepts. Extracting information with high precision involves deep understanding of the actual meaning of the text; as a result, IE systems are language-specific.

In developing solutions for vertical markets, Elsevier takes the IE road. Instead of building one generic, language- and domain-independent system that deals with large number of topics but provides little depth when it comes to the subject matter, we focus on extracting structured information specific for a particular domain from English text.

Elsevier's Information Extraction (IE) technology: Elsevier Text Mining

Within its Elsevier Text Mining portfolio, Elsevier has developed a proprietary natural language processing (NLP)-based technology called MedScan for extraction of structured information from unstructured text. It is a good fit for automatic indexing of NIH's content as the MedScan Thesaurus/Taxonomy was built mostly based on NIH thesauri and has all the NIH identifiers integrated (MeSH Headings, NCI Metathesaurus IDs, Entrez Gene IDs, Organism Tax IDs, etc.). The technology works by first recognizing domain-specific named entities (concepts) in the input text, and then uses natural language processing techniques to extract *attributed, directional semantic relationships* between them. The relationships can be of any complexity from simplest binary (X affects Y) to n-ary (X protects Y from Z) and complex multi-level nested ones (effect of X on Y depends on Z).

Elsevier IE technology has modular architecture. Each module performs its specific function and has well-defined and documented input/output format. Modules with compatible interfaces can be combined into different text processing pipelines, as required by the application. All modules are written from scratch to achieve our flexibility/precision/performance goals. The modules are portable C/C++ applications interacting via files and pipes.

Part II: 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

MedScan Technical Description: MedScan is a proprietary natural language processing (NLP)-based technology for extraction of structured information from unstructured text. Structured information is captured and formally represented using a conceptual model (ontology) of the domain. The ontology consists of a set of conceptual named entities (e.g. Proteins, Small molecules, Cellular processes, Diseases, etc) and a set of categorized relationships (Binding, Protein Modification, Expression regulation, Molecular Transport, etc) between them.

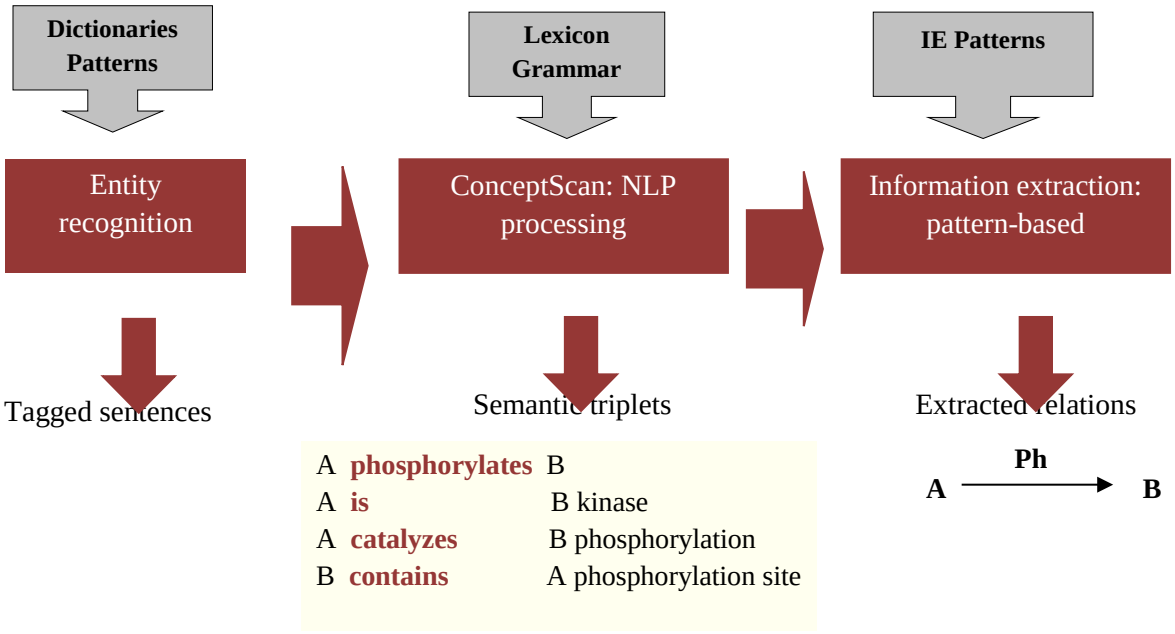


Figure 2. An overview of MedScan Architecture

MedScan first recognizes different domain-specific named entities (gene/protein names, cellular processes, cellular components, diseases, tissues, organs, etc.) in the input text, and then extracts functional relations (binding, regulation, association, molecular transport, etc.) between them. Figure 2 shows an overview of MedScan architecture.

The Entity Recognizer module utilizes hand-crafted dictionaries of domain-specific entities in combination with an advanced matching algorithm to detect them in input text.

To extract entity relationships from the text, MedScan utilizes two modules. The natural language processing module, ConceptScan, analyzes the sentence structure and decomposes each sentence into a deterministic set of Subject-Verb-Object triplets, each representing a single semantic relationship between two singular noun phrases. Next, Pattern Matcher matches carefully designed linguistic patterns over the triplets to extract and encode the entity relationships.

MedScan has been field-tested and is proven to be fast, efficient, and accurate information extraction technology. It is currently used to process the content of the entire Medline database along with more than 40 freely available full-text journals in order to extract more than 3.5 million individual facts (relations) about functions of proteins

with an overall accuracy of 90% and recall of 70%. The entire processing cycle can be completed in less than 24 hours on a regular PC.

Dictionaries and Named Entity Recognition:

Entity type	Number	Main sources
Proteins	136,000	Entrez Gene
Prot. Classes	7,500	GO, Enzymes, PubMed
Cell components	740	GO, PubMed
Cell processes	5,200	GO, PubMed
Diseases	6,300	MESH, PubMed
Small Molecules	270,000	MESH, PubChem, PubMed
Tissues	100	MESH, UMLS, NCI, EVoc
Cell types	360	MESH, UMLS, NCI, EVoc
Organs	2,875	MESH, UMLS, NCI, EVoc
Clinical parameters	1,786	Pubmed, ClinicalTrials.gov
Cell lines	2,500	PubMed

Table 1. MedScan Dictionaries

The Entity recognition module of MedScan utilizes hand-curated dictionaries of biomedical entities to detect them in the input text. Dictionaries are manually compiled and curated from the number of various public-domain resources (EntrezGene and SwissProt for protein names, PubChem and MESH for small molecules, GO for cell processes and components, MESH for diseases, NCI thesaurus for organs, tissues and cells, etc). Whenever possible the entities are hyperlinked to those outside resources for reference. Many additional aliases and terms are also added directly from the literature resources, e.g. PubMed. Table 1 shows the content of MedScan dictionaries. MedScan uses number of different algorithms to achieve accurate detection of entities in text. It can also use rule- and regular expression- based approaches to detect specific types of entities (abbreviations, numbers, dates, etc). The dictionaries are in a simple tab-delimited format so they can be easily extended or modified.

The input text can be in various formats (plain text, Microsoft Office, HTML, reasonable forms of PDF, zip/tar/gzip archives of the above, etc.) The output of the entity recognition step consists of individual sentences labeled to preserve their origin with identified named entities marked up with entity IDs, using **ID{***number***=….}** format (shown in red):

15986412:5 Enzyme assay, Western blot and ID{4000000,4106278=reverse-transcription} polymerase chain reaction (RT-PCR) results demonstrated that protein and mRNA expressions of human simple **ID{445329=phenol sulfotransferase}** (**ID{6799=P-PST}**), human ID{6818=monoamine sulfotransferase} (ID{6818=M-PST}), human ID{6822=dehydroepiandrosterone sulfotransferase} (ID{6822=DHEA-ST}) and human ID{6783=estrogen sulfotransferase} (ID{6783=EST}) were induced in ID{10000000,11012376=Hep G2

cells}; ID{6818=M-PST} and ID{6822=DHEA-ST} were induced in ID{10000000,11010382=Caco-2 cells}. The type of entity is encoded in its numerical range.

Natural Language Processing: The central idea of Elsevier's NLP algorithm (called ConceptScan) is decomposing natural language sentences into semantic relationships (which we will also call semantic triplets). Each triplet is designed to represent a single semantic relationship between two singular noun phrases (NPs). An example below illustrates this paradigm using a complex artificially constructed sentence.

11940574:7 Because **Axin2** has been shown to associate with and inhibit **beta-catenin** abundance and function, we hypothesized that **Axin2**, which is affecting proliferation of MEF cells can work in a negative feedback pathway, regulating **Wnt** signaling and thus controlling apoptotic process.

Triplets:

Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	associate	beta-catenin abundance
Axin2	inhibit	beta-catenin function
Axin2	affect	MEF cell line proliferation
Axin2	work	negative feedback pathway
Axin2	regulate	Wnt signaling
Axin2	control	apoptotic process

The extracted triplets capture the main facts expressed in a sentence. The ConceptScan is used in conjunction with named entity detection algorithm to index relationships between biomedical entities and to extract entity relationships.

ConceptScan parses sentences in several sequential algorithmic steps (See figure below)

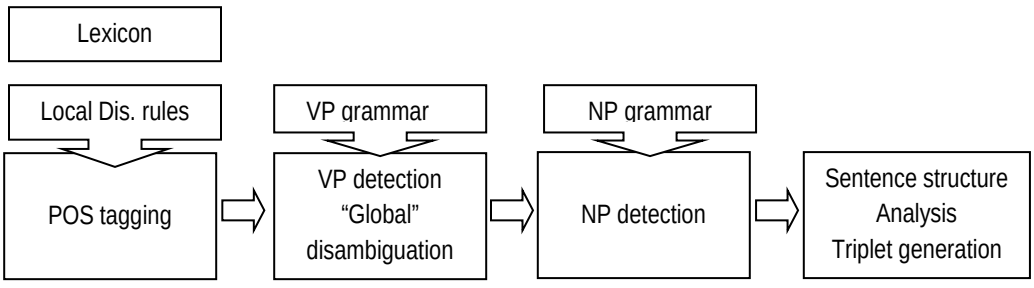


Figure 3. ConceptScan algorithm

The first step of NLP is part-of-speech tagging and local disambiguation. During this step, the words in a sentence are reduced to all possible uninflected forms, looked up in the lexicon and annotated with the respective syntactic categories. After initial POS tagging, the local disambiguation algorithm, encoded by a set of contextual regular expression-like rules, is applied. Notably, not all ambiguities can be resolved locally. The unresolved ambiguities are preserved for subsequent processing steps. The next step is identification of verbal phrases. Verbal phrase (VP) grammar is encoded in a single but complex deterministic finite-state automaton (DFA), with more than 25,000 states. It is matched over the sequence of syntactic categories assigned to sentence words at the POS-

tagging step. NP grammar is matched after detection of verbal phrases is complete. Similarly to VP grammar, it is encoded by a DFA. The structure of NP grammar covers prepositional attachment, conjunctions, relational constructs, appositions and exemplifications. Once VPs and NPs have been identified, ConceptScan analyzes the structure of the entire sentence.

Information extraction: The specific relationships between entities are extracted using separate module - Pattern Matcher. It utilizes a formalism closely resembling regular expressions to detect specific linguistic constructs expressing entity relations and to capture the expressed relations. It is specifically tailored to deal with linguistic input; it operates on the level of individual words rather than symbols and supports advanced linguistic features like matching all word forms and multi-word lexemes. Pattern matching also supports all regular expression features: wildcards, sets, negation, etc. The figure below shows a sample information extraction pattern.

```
CONTROL
{
  ControlType = "ProtModification"
  in  = %Protein1(Protein)
  out = %Protein2(Protein)
}
:
%Protein1 $MODAL? $ADV* phosphorylate~ %Protein2        |
%Protein2 $MODAL? $BE $ADV* phosphorylated by %Protein1 |
Phosphorylation of %Protein2 by %Protein1               |
;
```

Figure 4. An example of the information extraction pattern. The head template encodes the name of the output frame and templates for the values of its slots, which can be literals or other frames. Named entity variables (%Protein1 and %Protein2) are distinguished by the leading '%'. The head template can restrict the named entity variables to take values of specific semantic type(s) by providing the list of types in parentheses. Named word sets are distinguished by the leading '$'. They can be defined anywhere in the pattern file and can be used in multiple patterns. In the above example $MODAL is the set of modal verbs (can, may, might, etc). The '~' postfix indicates that the preceding word can be matched in any grammatical form. Multiple patterns extracting identical information are separated by the '|' separator.

MedScan output: The output of MedScan is in an XML-based format describing entities and relation between them (see an example below):

```
<resnet mref="16377759:4" msrc="The catalytic domain of ID{820019=S6K1} could be
phosphorylated by Arabidopsis ID{841259=3-phosphoinositide-dependent protein kinase-1}
(ID{830330=PDK1}), indicating the involvement of ID{830330=PDK1} in the regulation of
ID{820019=S6K}.">
  <nodes>
   <node local_id="N1" urn="urn:agi-llid:841259">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="at1g48390" />
   </node>
   <node local_id="N2" urn="urn:agi-llid:820019">
    <attr name="NodeType" value="Protein" />
    <attr name="Name" value="AT3G08720" />
   </node>
  </nodes>
  <controls>
   <control local_id="L1">
    <link type="in" ref="N1" />
    <link type="out" ref="N2" />
    <attr name="ControlType" value="ProtModification" />
    <attr name="ModificationType" value="phosphorylation" />
   </control>
  </controls>
 </resnet>
```

Figure 5. An example of a MedScan output

MedScan Ontology of Relationships: Elsevier has developed ontology of different types of relations between biological entities. Each type of relation has a very specific semantic definition and is typically attributed with additional information, e.g. sign of relations (e.g. positive, negative or unknown) or mechanism (e.g. phosphorylation, methylation, etc). There are three set of patters currently used by MedScan to extract biological relations – patterns focused on extraction of different aspects of protein functions, small molecule functions and disease biomarkers. The Table 2 below shows the scope of biological relationships currently extracted by MedScan.

The current scope of the information extracted by MedScan can be extended by developing new dictionaries covering other aspects of biomedical domain (e.g. focused more on medical or clinical entities) and/or by developing novel information extraction patterns to capture other types of entity relationships.

The Pattern Matcher is extremely fast: it runs through more than 16,000,000 entity-tagged sentences from the entirety of Medline in less than 20 minutes.

- Protein -> Protein
  - Binding
  - Protein modification
  - Expression (positive/negative/unknown)
  - Promoter regulation/Binding
  - Regulation (positive/negative/unknown)
- Protein -> Small Molecules
  - Synthesis/Degradation
  - Mol. Transport
- Protein -> Cell processes
- Protein -> Disease
  - Positive/negative regulation
- Disease -> Protein/Small molecules
  - Changed concentration/expression (positive/negative/unknown)
  - Mutations
  - Activity (positive/negative/unknown)
- Small molecules -> Protein
  - Binding
  - Direct regulation
  - Expression
  - Indirect regulation (positive/negative)
- Small molecules -> Disease/Cell processes (positive/negative/unknown)

Table 2. Relationships currently extracted by MedScan

MedScan Customizations: MedScan is flexible platform open for two types of end-user modifications. First, MedScan taxonomy and dictionaries can be extended to include new concepts and even new concept classes. Dictionaries are provided in a simple text-based tabular format and new concepts and concept aliases can be added to the files. Second, the scope of extracted information can be extended to include new relationships by modifying information extraction rules. The rules are recorded in a well-documented textual format and new rules can be created and added to MedScan.

MedScan Features and competitive advantages: Elsevier's IE engine has been designed and implemented from scratch to address flexibility, precision/recall and performance problems of the off-the-shelf NLP tools. Our design efforts focused on issues specific for texts in vertical application domains characterized by complex sentence and relationship structure, highly specialized entity notation, proliferation of abbreviations and synonyms. As a result of this focus, we have surpassed the 90% precision / 60% coverage mark on technical texts in our current application domains (biology and medicine). Our engine has an unmatched performance – it can process up to 1000 sentences per second on a regular PC, which is 2-3 orders of magnitude faster than prevailing NLP technologies. High performance allowed us to achieve clean separation between modules where traditional approaches intertwine distinct functions like parsing and ontology-based information extraction to cut down on the amount of information exchanged between modules. Also, much attention has been paid to keep domain-specific information in dictionaries and rule files, to simplify maintenance and extending the coverage to other domains.

The engine achieved production quality in 2003 and since then has been installed on many sites, including both individual and corporate-wide licenses.

Elsevier's Information Extraction (IE) technology: Fingerprint Engine

A back-end software system, the Elsevier Fingerprint Engine mines the text of scientific documents – publication abstracts, funding announcements and awards, project summaries, patents, proposals/applications, and other sources – to create an index of weighted terms which defines the text, known as a Fingerprint™ visualization.

By aggregating and comparing Fingerprints, the Elsevier Fingerprint Engine enables institutions to look even beyond metadata and expose valuable connections among people, publications, funding opportunities and ideas.

The Elsevier Fingerprint Engine powers many solutions including Pure, comprehensive information management system, and Reviewer Finder, Elsevier's tool for finding reviewers.

The Elsevier Fingerprint Engine uses a variety of thesauri to support applications pertaining to different subject areas. By applying a wide range of thesauri, Elsevier can develop solutions in but not limited to: the life sciences, engineering, earth and environmental sciences, arts and humanities, social sciences, mathematics and agriculture. Thesauri provided by an institution or specific research domain can also be incorporated.

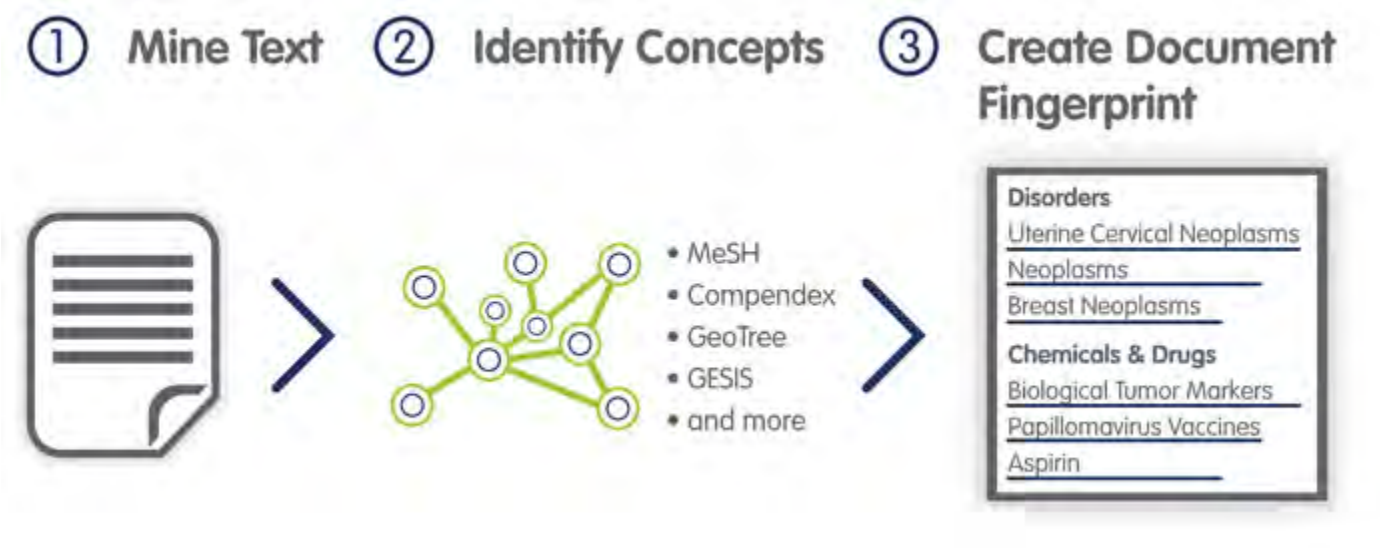


Figure 6: The Elsevier Fingerprint Engine creates Fingerprints via a three-step process

1. The Elsevier Fingerprint Engine applies a variety of Natural Language Processing (NLP) techniques to mine the text of scientific documents including publication abstracts, funding announcements and awards, project summaries, patents, proposals, applications and other sources

2. Key concepts that define the text are identified in thesauri spanning all the major disciplines

3. The Elsevier Fingerprint Engine creates an index of weighted terms that defines the text, known as a Fingerprint.

Applying Fingerprints to inform decision making: By aggregating and comparing Fingerprints of people, publications, funding opportunities and ideas, the Elsevier Fingerprint Engine can reveal insightful connections with practical applications. Here are some examples of how Fingerprints are currently used to bring scholarly business intelligence to institutional data.

- Pure aggregates the Fingerprints of individual documents to create unique Fingerprints that reveal your researchers' distinctive expertise. Pure also matches the Fingerprints of funding opportunities in SciVal® Funding to researchers' Fingerprints, recommending appropriate funding opportunities and suggested collaborators.

- Reviewer Finder compares document Fingerprints with researcher Fingerprints, making it easier to identify reviewers and raise awareness about potential conflicts of interest.

- Elsevier Journal Finder helps researchers find journals that could be best suited for publishing their articles. Journal Finder matches abstracts to Elsevier journals, scanning Elsevier's 2,200+ titles in the Health Sciences, Life Sciences, Physical Sciences and Social Sciences.

NIH Request for Information (RFI): Input on Sustaining Biomedical Data Repositories

For NOT-ES-15-011, the NIH is seeking information that addresses, but is not limited to, the following areas:

- Financial Models – New business models for sustaining digital repositories, including but not limited to examples cited in http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf and http://www.sr.ithaka.org/research-publications/guide-best-revenue-models-and-funding-sources-your-digital-resources.
- Innovation – Sustaining data repositories while enabling new innovations in finding, accessing, integrating and reusing their contents by a wide variety of stakeholders.
- Evaluation - Criteria to determine which data repositories require sustained funding models or no longer need to be sustained, including, but not limited to metrics for measuring the value of given repositories and data within those repositories.
- Best Practices - Current, new, and emerging means or practices to sustain data repositories for the long-term.
- Partnerships - The type, form, and governance of partnerships to ensure long-term access to essential data repositories including, but not limited to, private-sector organizations, non-profit foundations, universities, national and international government agencies, and combinations thereof.
- Technical – Technological developments needed to sustain data repositories in a more cost-effective way while furthering accessibility and usability to a broad set of stakeholders.
- Human Capital – Models to enhance efficiency in the application of human capital associated with data repositories.
- Life Cycle – Consideration of the evolution of value, cost, and scale as data repositories emerge, reach maturity, and either gain or lose relevance in the long term.

Response submitted by Holly Falk-Krzesinski, PhD on behalf of Elsevier, March 18, 2015

Elsevier values the NIH focus on research data and research data repositories and is appreciative for the opportunity to provide a response to NOT-ES-15-011, a Request for Information (RFI) on **Input on Sustaining Biomedical Data Repositories**.

Financial Models

Elsevier is involved in a number of working groups under the "Data Publication" umbrella Interest Group (IG) of the Research Data Alliance, notably the joint RDA/ ICSU World Data System Publishing Data Cost Recovery for Data Centres. The scope of this IG is greatly overlapping 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 RDA/ICSU WDS IG is submitting a separate, detailed response to this RFI.

Innovation

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.

Evaluation

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 the data is *published*, then others know that the associated research data 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. We need to develop a hierarchy of trust levels of data where at some moment reproducibility levels and algorithms to detect data become a part of that as well. Data validation and data publishing are areas in which Elsevier has deep expertise that we can lend to this.

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?

Best Practices/Policy

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. Elsevier has linking services that could add to this initiative by expanding the reach of healthdata.gov datasets.

The Plan also 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 need the flexibility to choose where to deposit their research data into repositories and are the 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 its criteria should not inadvertently limit data publication routes, such as linking data, data journals, interactive data plots, etc.

Rigid funder-specific mandates lead to directing 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.

That said, the NIH should 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.

Partnerships

As stated above, 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) 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

Technical

The NIH's recent Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research 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.

Finally, Elsevier is very interested for the NIH to develop open architectures to which other parties (including commercial) can contribute.

Human Capital

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.

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.

Life Cycle

With regards to life cycle, 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?

Previous RFI Responses

Elsevier recently submitted a response that included information about research data and data repositories to NOT-OD-15-067, a Request for Information (RFI) on Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the NLM (NLM Elements RFI). The following is excerpted verbatim from that NLM Elements RFI response. In addition, we wish to call your attention to the NLM Elements RFI response that was submitted by the Professional & Scholarly Publishing Division (PSP) of the Association of American Publishers (AAP; refer to 'Research data' in Comment 5). In addition, the PSP/AAP will be submitting a response to this RFI as well.

Submitted by Holly Falk-Krzesinski, PhD on behalf of Elsevier on March 13, 2015:
Research Data: Elsevier would like to see the NLM allow mining of all database content inside the suite of databases managed and curated by the NLM and provide actionable copyright metadata elements on all NLM content so we understand what we can mine/use for commercial and non-commercial purposes.

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 tools and services to support researchers to store, share, access, and preserve research data. These include our open data pilot, our database linking program, and our data journals, such as *Genomics Data and Data in Brief*. Collectively, Elsevier as partners with NLM, we should to be thinking about the big picture goal of enabling researchers to properly collect and annotate their research data 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). 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).

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 NLM 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.

Data fraud detection tools will need to be an important focal point for NLM. 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.

Regarding research data repositories, we think it is most useful to think in terms of data management plans and data archives. Elsevier is supportive of mandates for data management plans where researchers have the flexibility to choose where to deposit their data and that data publication routes are not limited (e.g., linking data, data journals, interactive data plots, etc.). Importantly, as efforts on research data repositories advance, it will be essential for the NLM to seek out collaboration opportunities with a broad and diverse range of stakeholders across sectors to ensure that collective expertise and experience as leveraged, a duplication of effort and resources are minimized, and cost savings and administrative efficiency are maximized.

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 NLM, 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 NLM 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.

Elsevier values its multi-faceted and synergistic relationship with the NIH and appreciates the opportunity to provide a response to NOT-OD-16-133, Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories. Submitted on behalf of Elsevier by:

Holly J Falk-Krzesinski, PhD
Vice President, Strategic Alliances
Global Academic Relations
Elsevier
h.falk-krzesinski@elsevier.com
New York, NY, USA

Response Contents

Part 1: Research Data Definition and Research Data Metrics

Definition and Disciplinarity

Research Data Definition

Elsevier's working definition is, "research data refers to the results of observations or experimentation that validate research findings." Research data can also be defined as, "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings."[1] Research data covers a broad range of information types[2], and digital data can be structured and stored in a variety of file formats.

The main goal of research data sharing is to enable other researchers to reuse data. Thus, reusability should always be taken into account when designing systems that create and store research data. We believe that data reuse could be optimized by aligning the 10 aspects of data listed below, Figure 1. This pyramid[3] – loosely modeled on Maslow's hierarchy of human

[1] OMB Circular 110, https://www.whitehouse.gov/omb/fedreg_a110-finalnotice

[2] From 'Defining Research Data' by the University of Oregon Libraries: Documents (text, Word), spreadsheets; Laboratory notebooks, field notebooks, diaries; Questionnaires, transcripts, codebooks; Audiotapes, videotapes Photographs, films; Protein or genetic sequences; Spectra; Test responses; Slides, artifacts, specimens, samples; Collection of digital objects acquired and generated during the process of research; Database contents (video, audio, text, images); Models, algorithms, scripts; Contents of an application (input, output, logfiles for analysis software, simulation software, schemas); Methodologies and workflows; and, Standard operating procedures and protocols.

[3] See figure in '10 aspects of highly effective research data' at https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data.

needs – can be seen as an extension of the FAIR Data Principles[4] (data should be Findable, Accessible, Interoperable and Reusable) and can function as a roadmap for the development of better data management processes and systems throughout the data lifecycle.

Figure 1: This pyramid can function as a roadmap for the development of better data management processes and systems.

Disciplinarity of Data

While this RFI specifically indicates *biomedical* repositories, it is important to recognize the increasingly interdisciplinary nature of biomedical, life sciences, and health sciences research and the overlaps of research data types from other disciplines.

In a parallel effort, the NSF has been focused on open data and research data through the Open Data Workshop Series[5], the first of which was held in November, 2015. While the workshop's initial focus was on generating discipline-specific responses from the Mathematical and Physical Sciences research communities to the federal policy requiring open data and the recently-released NSF policy statement on open data, there is considerable alignment with the NIH biomedical domain as it relates to research data: decide how and what to preserve in terms of research data for public consumption; the manner by which research data will be stored and accessed; and, the level of burden implied by conservation that is placed on the individual investigator.

[4] Force11 The FAIR Data Principles, https://www.force11.org/group/fairgroup/fairprinciples
[5] NSF MPS Open Data Workshop Series, https://mpsopendata.crc.nd.edu/index.php

International standards organizations, such as the National Data Service (NDS)[6], Research Data Alliance (RDA),[7] and ICSU-World Data System (WDS)[8], have been leading the charge to develop consensus and standards related to research data across disciplines. Elsevier, along with other publishers and research information providers, and additional research ecosystem stakeholders have been working in close partnership with these organizations, and have been engaged with the NSF initiative, as well as working with NIST[9]. These joint efforts have already begun to make significant strides in defining how to publish, find, and reuse research data. We thus recommend that the NIH also participate in this collaborative approach to:

1. Adopt flexible, broad standards and principles related to research data so that all disciplines have the maximum opportunity to interpret research data metrics and demonstrate research impact according to their field and across domains;

2. Consider how to combine quantitative with qualitative inputs; this to ensure that all disciplines, and all agencies and institutions regardless of their disciplinary focus, can share and interpret outcomes and research impact in a similar way;

3. Highlight the full range of types of research data deposit and reuse relevant to many research disciplines, so researchers have the widest opportunity to demonstrate maximum research impact of their work.

Research Metrics

This response focuses on research data, which constitutes an important part of the comprehensive ecosystem of research recognition. We would like to note the following types of research impact that should be considered across the research workflow (Figure 2, below):

1. Research activity – production of outputs leading to enhanced knowledge and understanding, such as original research in journal publications and books, research data, reports, designs, software, etc.; securing income to support ongoing research activities.

2. Research impact – recognition of the influence of research activity on subsequent research through viewing activity, and the receipt of citations from that subsequent research.

[6] NDS, http://www.nationaldataservice.org/
[7] Research Data Alliance, https://rd-alliance.org/
[8] ICSU-WDS, https://www.icsu-wds.org/
[9] Public Access to NIST Research, https://www.nist.gov/open

3. Scholarly impact – the wider recognition of research, beyond citing previous work, within the scholarly community, such as the receipt of prizes, requests to edit a journal and to peer review funding applications, and so on.

4. Economic impact – the production of commercializable outputs such as registered and granted patents and spin-out companies, and income generated from these outputs.

5. Social impact – the achievement of societally relevant outcomes, the enhancement of well-being to society as a result of research outputs and/or activities.

A well-rounded, inclusive recognition system can be assessed on all of the facets mentioned above, including research data, by the responsible use of research metrics as good approximations (proxies) of the actual level of performance. The research metrics that are selected should be complemented by the occasional use of narrative inputs such as case studies, firstly as a sanity check that the research metrics are indeed reflective of performance, and secondly in cases where research metrics cannot capture the full value of the research output or outcome.

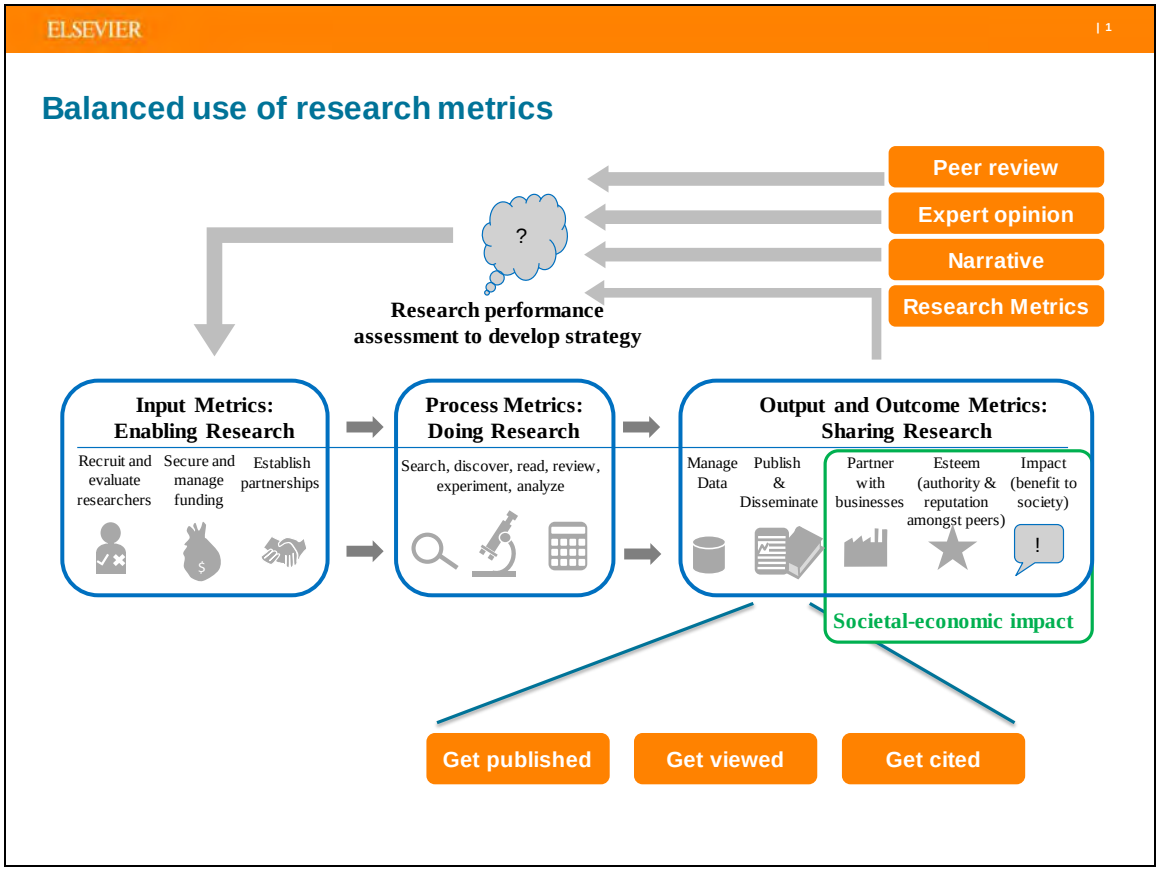


Figure 2: Balanced use of research metrics across the full research workflow.

Golden Rules

Elsevier's work with the research community has led us to recommend two "Golden Rules" for working with research metrics:

1. All decisions and participants benefit from a combination of both quantitative indicators and qualitative (e.g., case studies) input;

2. Quantitative input should always be based on at least two metrics (refer to Table 1 and Table 2 below for examples).

These Golden Rules are a practical reflection of the fact that the highest confidence in decision making is achieved when based on the most complete picture possible, which in turn depends on diverse inputs. Indicators reflect a version of the complete truth that is represented in research data repositories, and as such are an effective proxy for performance. The combination of these indicators can create a good impression of a comprehensive picture, as when a jigsaw has enough pieces in place to gain a good impression of the image, but the indicator jigsaw retains gaps, even when the underlying data sources are comprehensive and a broad set of indicators are used. Consequently, we recommend always complementing quantitative input from indicators with qualitative input from narratives to bring the view into sharper focus, and equally, we recommend that qualitative inputs are always used in combination with indicators.

Basket of Metrics

In close partnership with the research community, we have developed a 'basket of metrics' approach to using research metrics representing all types of research activity across the research workflow (Figure 2); research data metrics are no exception. In the next section, we list research data metrics that would be useful to help measure research impact, but would like to make some general comments about the advantages of an approach that builds on a multiplicity of research metrics here. The advantages of a 'basket of metrics' are:

1. Research excellence, even in one area such as research data, covers a broad range of concepts, and this diversity is best captured by considering a broad range of research metrics.

2. Funders and institutions need flexibility to determine the most appropriate research metrics to demonstrate research impact.

3. The set of research metrics offered can be read out in different ways, which accommodates the expectation by the research community for both simple research

metrics and more sophisticated, but complex, ones. Our research[10] shows that both types are needed and appreciated by users, and both types are important in offering the most complete picture of performance.

a. Simple research metrics such as total counts of activity, and counts normalised by university or faculty size (expressing the indicator as a proportion (%) of total, or by dividing the total count by number of researchers or outputs), are useful for offering transparency and clarity on the underlying data, and for showing the magnitude of activity in absolute terms.

b. More complex research metrics, such as field-normalised algorithms, take into account different behavior between fields and so enable the fair comparison of relative performance in physics with that in biology, for instance.

4. Our work with the community has led us to recommend Two Golden Rules of using research metrics. We discussed the first, always using quantitative measurements together with qualitative inputs, in question 4. The second Golden Rule is to always use at least two quantitative indicators as input into any decision. We recommend that any instance of research impact is demonstrated by using at least two research metrics, because:

a. Every single indicator has its weaknesses as well as its strengths, and these weaknesses can be complemented, or balanced, by the strengths of other indicators.

b. It reduces the likelihood of game playing. There is not, and will never be, one single research metrics that encompasses all aspects of excellent performance. If we try to reduce excellent performance to any single research metric, we will almost certainly drive unbalanced, undesirable behaviour; the researchers could work out how to optimise their performance according to that one research metric. It is much more difficult to see how researchers could adjust their behaviour when the outcomes of that behaviour are measured by using two, or three, or five different research metrics, except by doing genuinely better research across a range of outcomes – which is a result that the NIH is aiming to encourage.

[10] Extensive user research is represented in L. Colledge and C. James, 2015, A "basket of metrics"—the best support for understanding journal merit, European Science Editing 41(3), p61-65; http://europeanscienceediting.eu/articles/a-basket-of-metrics-the-best-support-for-understanding-journal-merit/

Metrics for Research Data

According to the Digital Curation Centre (DCC)[11], "a key measure of the worth of research is the impact it has or, to put it another way, the difference it is making both within the academic community and beyond." It is therefore in the interests of researchers, institutions, and funders to track the impact of research, starting with the impact of research outputs. Historically, research output used to evaluate impact was primarily peer-reviewed research articles. In recent years, other forms of research output are being recognized. The NIH now identifies research data as a legitimate type of 'research product' that can be listed in the "Contributions to Science" section of biosketches submitted as part of a grant application, carrying equal weight with publications.

Elsevier, through Scopus, is leading the way in displaying and collecting journal, article, and author level metrics around scientific literature[12], and intends to do the same for research data (see more below in the "Citation in Practice – The Scopus Model" section). Elsevier's Metrics team, with input from members of the NIH Big Data to Knowledge (BD2K) team, has developed an initial set of quantitative research data metrics (Table 1 and Table 2). All of the research data metrics presented in both tables can be calculated at multiple levels of aggregation (e.g., institution or discipline).

Table 1: Types of Research Data Metrics

Category	Research Data Metric	Description
Collaboration	Collaboration	Proportion of research data outputs with international, or national, or institutional, or no co-authors
Posting	Research Data Outputs	Total count of research data outputs
Get Viewed	Search Count	Total count of times research data outputs have been returned in a search
Get Viewed	Views Count	Total count of views
Get Viewed	Views Percentile measurement	For an individual piece of research data, this would be its percentile according to views received, compared to similar research data outputs For an aggregate entity like an institution, this will be proportion of research data outputs that fall into the top 1%, 5%, 10% or 25% of the world of research data outputs
Get Cited	Citation Count	Total count of citations

[11] Why measure the impact of research data?, http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics#why-measure-the-impact-of-research-data

[12] Scopus metrics, https://www.elsevier.com/solutions/scopus/features/metrics

Get Cited	Cited Research Data Outputs	Proportion of Research Data Outputs that have been cited at least once
Get Cited	Citations Percentile measurement	As Views Percentile Measurement
Economic Impact	Academic-Corporate Collaboration	Proportion of research data outputs with both academic and corporate co-authors
Scholarly Impact	Scholarly Activity	This is the total of Mendeley deposits, CiteULike deposits, and similar kind of activity. You can then slice and dice by each individually
Scholarly Impact	Scholarly Commentary	Total mentions in e.g. F1000. You can then slice and dice by each individually
Social Impact	Social Activity	This is the total of Tweets, Facebook likes, and similar kind of activity. You can then slice and dice by each individually
Social Impact	Mass Media	Total mentions in mass media. There are a few variants of this metric we have worked on for publications and which could be applied

Table 2: Research Data Repository Metrics

Category	Research Data Metric	Description
Data Reuse	Data Linkage	Proportion of papers with research data associated with them
Data Reuse	Data Depositing	Proportion of researchers that deposit research data within a certain time frame

Part 2: Research Data Repositories

Defining Trustworthiness

Elsevier has been actively working in robust and deep partnership with numerous national and international research data organizations developing standards for research data repositories. These organizations have made significant strides in defining the criteria that should be used to develop and certify trusted research data repositories.

The most advanced existing data repository certification schemes are:

- Data Seal of Approval (DSA)[13]
- World Data Scheme (WDS) Certification[14]
- Trusted Repositories Audit & Certification (TRAC)[15]

[13] DSA, http://www.datasealofapproval.org/en/
[14] WDS Certification, https://www.icsu-wds.org/services/certification

- Digital Curation Centre (DCC)'s Nestor Catalogue of Criteria for Trusted Digital Repositories[16]

DSA and WDS, whose schemas both rely on self-assessment, are combining their efforts through the Research Data Alliance (RDA)'s Repository Audit and Certification DSA–WDS Partnership Working Group[17] for "realizing efficiencies, simplifying assessment options, stimulating more certifications, and increasing impact on the community. The output from this WG is envisioned as a possible first step towards developing a common framework for certification and a service of trusted data repositories."

DSA includes 16 guidelines[18] covering data producers, data repositories, and data consumers. DSA already has a process in place for the full range of research data repositories to obtain certification, and it maintains a directory of repositories that have successfully acquired certification. The developing DSA-WDS Common Requirements[19] creates a harmonized set of criteria for certification of repositories at the core level addressing research data repository sustainability issues in the areas of organizational infrastructure, digital object management, technology, financial, and legal, etc. Furthermore, the DSA-WDS joint initiative plans to collaborate on a global framework for repository certification that moves from the core to the extended (NESTOR-Seal[20]), to the formal (ISO 16363[21]) level.

Rather than constructing schemas anew specific to biomedical repositories, the current DSA and WDS guidelines and developing Common Requirements must be applied to biomedical repositories to ensure the greatest potential for discoverability and reuse of research data that results from NIH-funded studies and other biomedical research.

Obtaining Certification

From its inception, Elsevier has incorporated the guidance developed by the aforementioned organizations into the development of our multidisciplinary data repository, **Mendeley Data**[22].

---

[15] TRAC, http://www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories

[16] DCC Nestor Catalogue of Criteria for Trusted Digital Repositories, http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor

[17] Repository Audit and Certification DSA–WDS Partnership WG, https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html

[18] DSA Guidelines, http://www.datasealofapproval.org/en/information/guidelines/

[19] DSA-WDS Common Requirements, https://rd-alliance.org/system/files/DSA%E2%80%93WDS%20Catalogue%20of%20Common%20Requirements%20V2.2.pdf

[20] NESTOR Seal for Trustworthy Digital Archives, http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

[21] ISO 16363 Trusted Digital Repositories Management Systems, http://anab.org/programs/isoiec-17021/ms-accreditation-programs/digital-repositories-iso-16363/

[22] Mendeley Data, https://data.mendeley.com/

A critical and absolute criterion of a trusted repository, but one often overlooked by many data repositories, is a mechanism for *long-term* preservation of digital assets. Elsevier has long been a leader in the area of permanent e-journal preservation and an advocate of publisher and research information provider responsibility for digital archiving. Just as Elsevier has done for content published in our journals, we teamed up with DANS (Data Archiving and Networking Services)[23] to ensure that all research datasets within Mendeley Data will be sent offsite to DANS, where they will ensure that the research data is safely archived.

Elsevier is also in the process of obtaining the Data Seal of Approval for Mendeley Data.

Part 3: Data Discoverability

Data Indexing

Elsevier's DataSearch[24] is a prototype research data search engine developed by Elsevier's Research Data Management team that allows users to search for research data across domains and types, from domain-specific, cross-domain, and institutional data repositories. The tool is an exploration of what a search engine for research data needs to look like (versus a web search engine or a document search engine). DataSearch currently indexes images, tables and supplementary data from content sources[25], considered 'research data components.' DataSearch also indexes a series of domain-specific repositories, as well as non-domain specific ones[26]. We are exploring how we might integrate DataSearch with our other offerings, such as Mendeley Data, Scopus, and Pure, to provide robust research data management solutions across the research workflow. And for both, we are working with BD2K on the inclusion of Mendeley Data and DataSearch into the NIH Data Commons.

DataSearch harvests data through APIs (application program interfaces) from various repositories or, in some cases, through database dump files provided to the project. We then normalize the data to our data model, index the data to make it searchable, and generate previews of data where possible. Users can go directly to the source repository from the preview page.

[23] DANS, https://dans.knaw.nl/en
[24] Elsevier DataSearch, https://datasearch.elsevier.com/
[25] Other than from Elsevier's ScienceDirect, DataSearch only indexes open data from open access repositories
[26] As of June 2016, DataSearch is indexing the following content sources: Tables, figures and supplementary data associated with papers in ScienceDirect, arXiv and PubMed Central; Mendeley Data; NeuroElectro; Dryad; PetDB; ICPSR; Harvard Dataverse; and ThemoML at NIST Thermodynamic Research Center (TRC). We are currently investigating DataSearch being able to index all of the NIH-supported data repositories (see https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html for list). We will continue to add more content sources in the future.

Elsevier uses a pilot set of criteria to select repositories to index in DataSearch, including the number of users, the ease of our ability to index the repository data, and relationships we have with data repository managers. We are committed to indexing all 63 NIH-supported repositories[27] in DataSearch; we cannot do them all at once, however, so we will seek input from the NIH on ranking/prioritization.

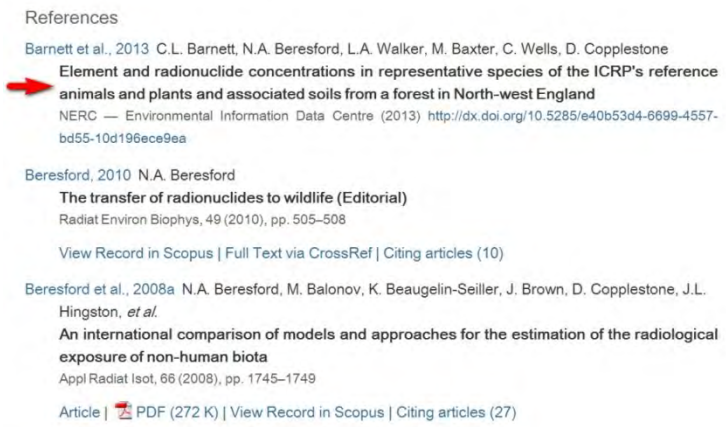
We are also engaging with data repositories to investigate how we can most effectively combine efforts regarding data discovery options, including having DataSearch power search on the repositories themselves. The DataSearch team is working with the NIH-funded bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem)[28] team, which has been developing a data discovery index prototype[29] that indexes data that are stored elsewhere, and Elsevier is exploring how we can better collaborate through shared interfaces and API's.

Data Citation

For data to be discovered and acknowledged it must be widely accessible and cited in a consistent and clear manner in the scientific literature. Elsevier endorses the Joint Declaration of Data Citation Principles[30], which will render research data an integral part of the scholarly record, properly preserved and easily accessible, ensuring that researchers get proper credit for their work. The citation principles focus on Importance, Credit and Attribution, Evidence, Unique Identification, Access, Persistence, Specificity and Verifiability, and Interoperability and Flexibility. A data citation is included in the standard References list of an article, and treated on equal footing with article citations.

In Elsevier's ScienceDirect platform, this means readers will enjoy the same benefits with data as they do with article citations, including one-click deep links to the referenced material and the ability to quickly jump to the point in the article where the work was first cited (see Figure 3 below).

[27] NIH Data Sharing Repositories, https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
[28] bioCADDIE, https://biocaddie.org/about
[29] DataMed, https://datamed.org/
[30] Joint Declaration of Data Citation Principles, https://www.force11.org/group/joint-declaration-data-citation-principles-final


References

Barnett et al., 2013 C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone
Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England
NERC — Environmental Information Data Centre (2013) http://dx.doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

Beresford, 2010 N.A. Beresford
The transfer of radionuclides to wildlife (Editorial)
Radiat Environ Biophys, 49 (2010), pp. 505–508

View Record in Scopus | Full Text via CrossRef | Citing articles (10)

Beresford et al., 2008a N.A. Beresford, M. Balonov, K. Beaugelin-Seiller, J. Brown, D. Copplestone, J.L. Hingston, *et al.*
An international comparison of models and approaches for the estimation of the radiological exposure of non-human biota
Appl Radiat Isot, 66 (2008), pp. 1745–1749

Article | PDF (272 K) | View Record in Scopus | Citing articles (27)

Figure 3: The image shows a reference list from the article "A new approach to predicting environmental transfer of radionuclides to wildlife: A demonstration for freshwater fish and caesium," published in *Science of the Total Environment* 2013.

Citation in Practice – The Scopus Model

Elsevier's Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books/book chapters, and conference proceedings. Delivering a comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences, and arts and humanities, Scopus features smart tools to track, analyze, and visualize research and its impact. Scopus' vision of research data aligns with the Force11 Joint Declaration of Data Citation Principles[31] which state that research data is as integral to recognizing and assessing the research output of modern researchers as are articles, reviews, books and all other "traditional" forms of research output (refer to Figure 2). Thus, research data must be:

- Discoverable
- Trustworthy
- Included in the author profile
- Creditable

DataSearch and Scopus are taking a complementary approach. Whereas DataSearch indexes a number of data sources and allows researchers to discover, access, and preview relevant data sets in multiple formats, the goal for Scopus is to integrate and curate DataSearch results to ensure that the research data discoverable via Scopus.com is trustworthy, in a manner consistent with the approach we take toward traditional content inclusion by way our independent Content Selection & Advisory Board (CSAB)[32].

[31] https://www.force11.org/group/joint-declaration-data-citation-principles-final
[32] Scopus CSAB, https://www.elsevier.com/solutions/scopus/content/scopus-content-selection-and-advisory-board

Presently the Scopus CSAB vets all journals indexed in Scopus to ensure high quality standards. We believe that a similar methodology should be applied to data repositories, to ensure transparent, consistent, high quality content

Integrating a research data search engine such as DataSearch in Scopus as a prototype will require a combination of human and algorithmic curation techniques to ensure that Scopus users can trust and rely on the results. In order to achieve this, we intend to apply rigorous selection criteria to both data repositories and data types (refer to the sections on "Research Data Definition" and "Defining Trustworthiness" above for criteria that we will consider).

After ensuring research data is discoverable, the next step will be for Scopus to integrate research data citations in Scopus Author Profiles, to appropriately link and assign credit to the author. Metrics can be applied to research data citations in Scopus just as they are now for articles (refer to the section above, "Metrics for Research Data").

Scopus is leading the way in displaying and collecting journal, article, and author level metrics around scientific literature[33], and intends to do the same for research data. Several parameters will be developed to attribute metrics to data. Scopus will collect and display these metrics in a way that is clear and imparts meaning and value to each metric. Through these efforts, Elsevier can enhance recognition across the research workflow (Figure 2) through enhancement of data search and credit for research data output.

Part 4: Recognition and Reward

While this RFI doesn't specifically identify the topic recognition and reward of research data to support widespread research data sharing, we think that the issue is inextricably linked to the sustainability of research data repositories.

At the SciDataCon 2016 conference in September, 2016, there was a session entitled, *Getting the incentives right: Removing social, institutional and economic barriers to data sharing*[34]. The session description indicates that while "much work has been done relating to the technical aspects of scientific data sharing…[progress toward research data sharing]…has been particularly hampered by a lack of awareness that the barriers and risks to be addressed are socio-technical concerns, with the non-technical concerns –the social, institutional and economic aspects of data sharing, often overlooked."

[33] Scopus metrics, https://www.elsevier.com/solutions/scopus/features/metrics
[34] http://www.scidatacon.org/2016/sessions/37/

Elsevier has been working with the research data community to compile a body of literature addressing the socio-technical aspects of research data sharing rewards and incentives, as well as relevant references on knowledge sharing incentive systems (Table 3)[35]. We recommend that this literature be comprehensively evaluated with the goal of developing recommendations for effective policies and practices that the NIH (and other funders), research institutions, and faculty promotion & tenure committees can employ to promote research data sharing.

Table 3: References on Rewards and Incentives for Research Data Sharing

1. Anderson MS, Ronning E a., De Vries R, Martinson BC. The perverse effects of competition on scientists' work and relationships. Sci Eng Ethics. 2007;13:437–61.
2. Arzi S, Rabanifard N, Nassajtarshizi S, Omran N. Relationship among Reward System, Knowledge Sharing and Innovation Performance. Interdiscip J Contemp Res Bus. 2013;5(6):115–41.
3. Bartol K. Encouraging Knowledge Sharing: The Role of Organizational Reward Systems. J Leadersh & Organ Stud. 2002;9(1):64–76.
4. Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, et al. Prepublication data sharing. Nature [Internet]. 2009;461(7261):168–70. Available from: http://dx.doi.org/10.1038/461168a
5. Borgman CL. The Conundrum of Sharing Research Data. SSRN Electron J [Internet]. 2011;63(6):1–40. Available from: http://www.ssrn.com/abstract=1869155
6. Boudreau KJ, Lakhani KR. "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. Res Policy. 2015;44(1):4–19.
7. Bourne PE, Lorsch JR, Green ED. OUTLOOK BIG DATA IN BIOMEDICINE Sustaining the big-data ecosystem. 2015;
8. Carrara W, Fischer S, Steenbergen E van. Open Data Maturity in Europe 2015: Insights into the European state of play. European Data Portal Open. 2015.
9. Chia-Shen C, Shih-Feng C, Chih-Hsing L. Understanding Knowledge-Sharing Motivation, Incentive Mechanisms, and Satisfaction in Virtual Communities. Soc Behav Personal An Int J [Internet]. 2012;40(4):639–47. Available from: http://proxy.indianatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=75245509&site=ehost-live&scope=site
10. Costello MJ. Motivating Online Publication of Data. Bioscience [Internet]. 2009;59(5):418–27. Available from: http://www.jstor.org/stable/10.1525/bio.2009.59.5.9
11. Cress U, Barquero B, Schwan S, Hesse FW. Improving quality and quantity of contributions: Two models for promoting knowledge exchange with shared databases. Comput Educ [Internet]. 2007 [cited 2016 Sep 15];49(2):423–40. Available from: http://www.sciencedirect.com/science/article/pii/S0360131505001375
12. Denis J, Goëta S. Exploration, Extraction and "Rawification". The Shaping of Transparency in the Back Rooms of Open Data. Soc Sci Res Netw [Internet]. 2014; Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2403069&download=yes
13. Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. Science friction: Data, metadata, and collaboration. Soc Stud Sci [Internet]. 2011 Aug 15 [cited 2012 Mar 22];41(5):667–90. Available from: http://sss.sagepub.com/cgi/content/abstract/0306312711413314v1
14. Ember C, Hanisch R, Alter G, Berman H, Hedstrom M, Vardiagn M. Sustaining Domain Repositories for Digital Data: A White Paper. 2013;(February):1–16.
15. Enke N, Thessen A, Bach K, Bendix J, Seeger B, Gemeinholzer B. The user's view on biodiversity data sharing - Investigating facts of acceptance and requirements to realize a sustainable use of research data -. Ecol Inform. 2012;11:25–33.
16. Fecher B, Friesike S, Hebing M. What Drives Academic Data Sharing? SSRN Electron J [Internet]. Berlin, Germany; 2014;10(2). Available from: http://papers.ssrn.com/abstract=2439645
17. Fecher B, Friesike S, Hebing M, Linek S, Sauermann A. A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing [Internet]. Berlin, Germany; 2015. Report No.: 1454. Available from: http://d.repec.org/n?u=RePEc:diw:diwwpp:dp1454&r=sog
18. Fitch P, Craglia M, Pollock R, Cox S, Fowler D. Getting the incentives right: removing social, institutional and economic barriers to data sharing [Internet]. International Data Week Conference Session. Denver, CO, USA; 2016. Available

[35] Holly Falk-Krzesinski, PhD, at Elsevier can be contacted directly to be added to the growing reference group, h.falk-krzesinski@elsevier.com

from: http://www.scidatacon.org/2016/sessions/37/
19. Friesike S, Fecher B, Hebing M, Linek S. Reputation Instead of Obligation : Why We Need to Forge New Policies to Motivate Academic Data Sharing [Internet]. Blog Post. Alexander von Humboldt Institute for Internet and Society; 2015. p. 5–8. Available from: http://www.hiig.de/en/23202/
20. Gardner D, Toga AW, Ascoli G a, Beatty JT, Brinkley JF, Dale AM, et al. Towards effective and rewarding data sharing. Neuroinformatics. 2003;1(3):289–95.
21. Gaulé P, Maystre N. Getting cited: Does open access help? Res Policy. 2011;40(10):1332–8.
22. Goëta S. Instaurer des données, instaurer des publics- Une enquête sociologique dans les coulisses de l'open data: Instantiate data, instantiate publics- a sociological inquiry in the backrooms of open data. [Paris, France, France]: Télécom Paris Tech; 2016.
23. Gorgolewski KJ, Margulies DS, Milham MP. Making data sharing count: A publication-based solution. Front Neurosci. 2013;(7 FEB).
24. Hung SY, Durcikova A, Lai HM, Lin WM. The influence of intrinsic and extrinsic motivation on individuals knowledge sharing behavior. Int J Hum Comput Stud. 2011;69(6):415–27.
25. Ingram C. How and why you should manage your research data: a guide for researchers [Internet]. JISC; 2016. Available from: https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data?mkt_tok=3RkMMJWWfF9wsRonuqjMZKXonjHpfsX56%2B4pW6S%2BlMI%2F0ER3fOvrPUfGjI4ATMRmI%2BSLDwEYGJlv6SgFTrLHMa1izLgNUhA%3D
26. Jahani. Is Reward System and Leadership Important in Knowledge Sharing Among Academics? American Journal of Economics and Business Administration. 2011. p. 87–94.
27. Kim Y, Adler M. Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. Int J Inf Manage. 2015;35(4):408–18.
28. Koers H. How do we make it easy and rewarding for researchers to share their data? – a publisher's perspective. J Clin Epidemiol [Internet]. 2015 Jul [cited 2015 Jul 14]; Available from: http://www.sciencedirect.com/science/article/pii/S089543561500325X
29. Li Y-M, Jhang-Li J-H. Knowledge sharing in communities of practice: A game theoretic analysis. Eur J Oper Res [Internet]. 2010;207(2):1052–64. Available from: http://www.sciencedirect.com/science/article/pii/S0377221710003899
30. Lin S-W, Lo LY-S. Mechanisms to motivate knowledge sharing: integrating the reward systems and social network perspectives. J Knowl Manag [Internet]. 2015;19(2):212–35. Available from: http://www.emeraldinsight.com.ezp.lib.unimelb.edu.au/doi/full/10.1108/JKM-05-2014-0209
31. Longo DL, Drazen JM. Data Sharing. N Engl J Med [Internet]. 2016;374(3):276–7. Available from: http://dx.doi.org/10.1056/NEJMe1516564\nhttp://www.nejm.org/doi/full/10.1056/NEJMe1516564
32. Mayernik MS, Callaghan S, Leigh R, Tedds J, Worley S. Peer Review of Datasets: When, Why, How. Bull Am Meteorol Soc. 2014;1–32.
33. Mueller-Langer F, Andreoli-Versbach P. Open Access to Research Data: Strategic Delay and the Ambiguous Welfare Effects of Mandatory Data Disclosure. Berlin, Germany; 2014. Report No.: 239.
34. Muller R, Spiliopoulou M, Lenz H. The Influence of Incentives and Culture on Knowledge Sharing. System Sciences, 2005 HICSS '05 Proceedings of the 38th Annual Hawaii International Conference on [Internet]. 2005. p. 247b--247b. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385745
35. Nelson B. Data sharing: Empty archives. Nature [Internet]. 2009;461:106–63. Available from: http://www.nature.com/news/2009/090909/full/461160a.html
36. Niu JJ. Reward and Punishment Mechanism for Research Data Sharing. IASSIST Q. 2006 May;(Winter).
37. Pham-Kanter G, Zinner DE, Campbell EG. Codifying collegiality: Recent developments in data sharing policy in the life sciences. PLoS One [Internet]. 2014;9(9). Available from: http://journals.plos.org/plosone/article/asset?id=10.1371/journal.pone.0108451.PDF
38. Pisani E, AbouZahr C. Sharing health data: good intentions are not enough. Bull World Health Organ [Internet]. 2010;88(6):462–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2878150&tool=pmcentrez&rendertype=abstract
39. Piwowar HA, Chapman WW. A review of journal policies for sharing research data. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 20 - Proceedings of the 12th International Conference on Electronic Publishing, ELPUB 2008 [Internet]. 2008. p. 1–14. Available from: http://elpub.scix.net/cgi-bin/works/Show?001_elpub2008
40. Poline J-B, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. Front Neuroinform [Internet]. 2012;6:9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319918&tool=pmcentrez&rendertype=abstract
41. Ranganathan K, Ripeanu M, Sarin a., Foster I. Incentive mechanisms for large collaborative resource sharing. IEEE Int Symp Clust Comput Grid, 2004 CCGrid 2004 [Internet]. 2004;1–8. Available from:

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1336542
42. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol [Internet]. 2015;13(11):e1002295. Available from: http://dx.plos.org/10.1371/journal.pbio.1002295
43. Rolland B. Data Sharing and Reuse: Expanding Our Concept of Collaboration [Internet]. Team Science Toolkit Blog. 2016 [cited 2016 Jan 1]. Available from: https://www.teamsciencetoolkit.cancer.gov/Public/ExpertBlog.aspx?tid=4
44. Rood RB, Edwards PN. Climate Informatics: Human Experts and the End-to-end System. Earthzine [Internet]. 2014;1–14. Available from: http://earthzine.org/2014/05/22/climate-informatics-human-experts-and-the-end-to-end-system/
45. Šajeva S. Encouraging Knowledge Sharing among Employees: How Reward Matters. Procedia - Soc Behav Sci [Internet]. 2014;156(April):130–4. Available from: http://www.sciencedirect.com/science/article/pii/S1877042814059540
46. Sarathy R, Muralidhar K. Secure and useful data sharing. Decis Support Syst. 2006;42(1):204–20.
47. Sayogo DS, Pardo TA. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. Gov Inf Q. 2013;30(SUPPL. 1).
48. Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, et al. Post-publication sharing of data and tools. Nature [Internet]. 2009;461(7261):171–3. Available from: http://dx.doi.org/10.1038/461171a\nhttp://www.nature.com/doifinder/10.1038/461171a
49. Seonghee K, Boryung J. An analysis of faculty perceptions: Attitudes toward knowledge sharing and collaboration in an academic institution. Libr Inf Sci Res. 2008;30(4):282–90.
50. Shuman LJ. Data Sharing in Engineering Education. Adv Eng Educ [Internet]. 2016;5(2). Available from: http://advances.asee.org/summer-2016-volume-5-issue-2/
51. Stanley B, Stanley M. Data sharing: The primary researcher's perspective. Law Hum Behav. 1988;12(2):173–80.
52. Sterling TD, Weinkam JJ. Sharing scientific data. Commun ACM [Internet]. 1990;33(8):112–9. Available from: http://portal.acm.org/citation.cfm?id=79182\nhttp://delivery.acm.org.ezp2.bath.ac.uk/10.1145/80000/79182/p112-sterling.pdf?key1=79182&key2=8630174821&coll=GUIDE&dl=GUIDE&CFID=104901268&CFTOKEN=61271004
53. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data Sharing by Scientists: Practices and Perceptions. Neylon C, editor. PLoS One [Internet]. 2011 Jun 29 [cited 2011 Jun 30];6(6):e21101. Available from: http://dx.plos.org/10.1371/journal.pone.0021101
54. Thorisson G a. Accreditation and attribution in data sharing. Nat Biotechnol [Internet]. 2009;27(11):984–5. Available from: http://dx.doi.org/10.1038/nbt1109-984b
55. Toronto International Data Release Workshop Authors. Prepublication data sharing. Nature. 2009;461(September):168–70.
56. Van Noorden R. Data-sharing: Everything on display. Nature. 2013;500:243–5.
57. Wallis JC, Rolando E, Borgman CL. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS One [Internet]. 2013;8(7):e67332. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332
58. Weber NM, Baker KS, Thomer AK, Chao TC, Palmer CL. Value and context in data use: Domain analysis revisited. Proc Assoc Inf Sci Technol [Internet]. 2012;49(1):1–10. Available from: http://onlinelibrary.wiley.com/doi/10.1002/meet.14504901168/abstract
59. Yang HL, Wu TCT. Knowledge sharing in an organization. Technol Forecast Soc Change. 2008;75(8):1128–56.
60. Zimmerman A. Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists [Internet] [phdthesis]. 2003. Available from: http://deepblue.lib.umich.edu/dspace/handle/2027.42/39373
61. Zinner D, Pham-Kanter G, Campbell E. The Changing Nature of Scientific Sharing and Withholding in Academic Life Sciences Research: Trends From National Surveys in 2000 and 2013. Acad Med [Internet]. 2016;91(3):433–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26675188

Elsevier's Response to NOT-OD-17-015,
Request for Information (RFI): Strategies for NIH Data Management, Sharing, and Citation

Submitter Name
Holly Falk-Krzesinski

Name of Organization
Elsevier

Type of Organization
Other Type of Organization
Research Information & Technology

Role
Institutional Official

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology)
All areas

Type of Data That You Primarily Plan to Generate and Share
Non-Human
Other Type
Research data of all types

Repositories You or Your Organization Primarily Utilize (Maximum: 250 words)
There are over 75 repositories Elsevier currently supports: for a full listing, refer to https://www.elsevier.com/books-and-journals/enrichments/data-base-linking/supported-data-repositories.

Moreover, Elsevier hosts its own open research data repository, Mendeley Data https://data.mendeley.com/, which enables researchers to store and share any type of research data.

SECTION I. Data Sharing Strategy Development

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)
The definition of research data differs from field to field, but broadly speaking it refers to the result of observations or experimentations that validate research findings and can include, but are not limited to: raw data, processed data, software, algorithms, protocols, methods, materials and methods descriptions as well as lab notebook entries. Research data do not include text in manuscript or final published article form, nor do they include other/supplementary materials submitted and published as part of a journal article.

In principle, all data should be stored, with an emphasis on two points:
1) When it is harder, more costly, or even impossible to reproduce a dataset, data storage is essential, for example:
- Human-subject studies, where patients or other participants have spent time and effort to make themselves available and samples have been gathered;
- Data where animals have been sacrificed to enable research;
- Non-replicable data such as environmental observation studies, where the conditions of study cannot physically be reproduced because they present a view on a moment in time.

2) Where possible, the rawest form of data should be stored, as well as any software, scripts and methods to interpret or reformat this data. For example, in the case of questionnaires the original answers as well as any software to process these must be preserved. This approach enables a replication of the analysis work, which can support rigor; secondly, it allows other researchers to reuse the raw data, which supports data reuse.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications (Maximum: 250 words)

Obviously, the longer data can be stored, the better, but also obviously, there are costs involved. In general it seems important that raw data is stored well beyond the period in which the data might be reexamined: in most domains this will mean a period of 10 years or more, though in the case of human study or observations of natural phenomena (e.g. in ecology, epidemiology, etc) it would be worth looking at much longer time spans in the order of 10 – 50 years.

Where it is not possible to preserve data in perpetuity, an indexing and abstracting service such a Scopus could preserve metadata associated to a specific dataset, and allow for citations and permanent reference.

In Elsevier's Mendeley Data repository, for example, data is preserved in perpetuity via an agreement with DANS (Data Archiving and Networked Services), whereby DANS archives every dataset posted to Mendeley Data which passes the internal review process : refer to https://www.knaw.nl/en/news/news/collaboration-dans-and-mendeley-on-archiving-datasets. Mendeley Data can therefore guarantee that any data deposited will always be available at the DOI provided.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)

A barrier to data stewardship is uncertainty of long-term funding of data repositories: the fact that many repositories are judged and funded in competition with research leads to great uncertainty: refer to https://www.rd-alliance.org/sites/default/files/case_statement/RDA_WDS_IG_Publishing_Costs.pdf. The fact that many repositories are funded via many different routes further enhances the burdens to seek funding sources by the repository directors, who should be focusing their efforts on providing the best possible data curation support. In other cases, repositories are informed that their grants will not be renewed and are encouraged to seek alternate funding models, without being given the time or resources to procure those funding sources. It's critical the NIH work cooperatively with other funders globally to develop models for long-term data infrastructure support and develop clear guidelines within and between agencies and divisions as to the type of work that will be supported.

Additionally, career opportunities and the acknowledgement and promotion trajectories of data stewards/curators are currently under-supported.

Lastly, we point to the barriers mentioned in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118053#pone.0118053.ref059, specifically, a lack of clarity regarding rights and privacy issues concerning human data. A clear legal understanding of the rights of use of research data are needed, especially in medicine and in the social sciences. Funding agencies could play an important role here, and educate researchers on the copyright and need for anonymization of human subjects data they collect.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)

Elsevier strongly supports the NIH's efforts to make all research data as well as software, methods, and protocols openly available where possible. We urge the NIH to move beyond requiring the creation of Data Management Plans, and are interested in helping to define and support a set of requirements for data storage, sharing and preservation. Of all the components of a data sharing process, obviously storing the data in a long-term preservable format is of the highest priority. Once the data is preserved, appending any metadata concerning the methods which were involved in creating the data and maintaining a description of the provenance of any — raw or derived — dataset is key. Ultimately, the preservation of data itself and a clear and easily interpretable set of metadata describing how the data was created will lead to greater rigor and reproducibility, and a lack of

duplication of efforts on behalf of the researchers: see https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data for a further elaboration of this concept.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

1. The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing (Maximum words: 250)

Reporting the software and data researchers' create can provide additional evidence of usefulness of the products of funded research and may help enforce data sharing mandates. Elsevier, the NIH, and other stakeholders can work together to create a coherent ecosystem that allows many different paths (dependent on domain, role, and personal preference) for scientists/scholars to identify, report, and track data sharing and reuse practices.

Funding agencies' data sharing policies are named as a key factor to encourage academic data sharing (e.g. http://dx.doi.org/10.1371/journal.pone.0118053 and http://onlinelibrary.wiley.com/doi/10.1111/j.1755-263X.2012.00259.x/pdf). However, funding policies still show varying degrees of enforcement: achieving clarity and correspondence between funding programs (within/between funding agencies) is a key factor to encourage compliance with data sharing mandates: http://journals.sagepub.com/doi/abs/10.1177/1745691613491579.

Elsevier is a leader in highlighting the association between articles and data, and including data as an output associated with a specific author/institution. Using quality filters similar to that Scopus uses to index articles, we enable data and software citations for evaluating the scientific output of a single researcher/institution.

2. Important features of technical guidance for data and software citation in reports to NIH, which may include:

2a. Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) * (Maximum words: 250)

Regarding citation of software and data, Elsevier is an active supporter of the Force11 Data Citation group, https://www.force11.org/group/dcip, as shown by the recent implementation of these standards in our over 1,800 journals, https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data. Within the Force11 DCIP Publisher Early Adopters group, https://www.force11.org/group/dcip/eg3publisherearlyadopters, we are co-leading efforts to develop a joint Force11 DCIP Data Citation Roadmap for science publishers, due for publication shortly.

Regarding the specific use of persistent identifiers, we fully support the recommendations provided by Force11 DCIP Repositories Early adopters group, https://www.force11.org/group/dcip/eg4repository, who pre-published their Roadmap, https://doi.org/10.1101/097196, which explicitly states:

- All datasets intended for citation must have a globally unique persistent identifier that can be expressed as unambiguous URL.
- Persistent identifiers for datasets must support multiple levels of granularity, where appropriate.
- This persistent identifier expressed as URL must resolve to a landing page specific for that dataset.

Within Elsevier's data repository, Mendeley Data, we enable unique identification of data versions, as well. When a published dataset is edited, the last digits of the data DOI will change to reflect a new version of the dataset, https://data.mendeley.com/faq.

Regarding software citations, we support the principles published by the Force11 Software Citation Working Group, https://www.force11.org/software-citation-principles, on Unique Identification, which states: "A software citation should include a method for identification that is machine actionable, globally unique, interoperable, and recognized by at least a community of the corresponding domain experts, and preferably by general public researchers."

To enable the accurate reporting of funded research, we strongly encourage the NIH and other funding agencies to identify grants by Unique Identifiers, and make these available through a portable or externally accessible database, such as CrossRef's 'Funding Data,' http://www.crossref.org/fundingdata/.

2b. Inclusion of a link to the data/software resource with the citation in the report (Maximum: 250 words)

With regards to links between publications and data, within the aegis of the RDA Data Publishing Group, https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html, we have helped lead the development of a Linked data demonstrator and set of guidelines, the Scholix Initiative, http://www.scholix.org/. Scholix and the accompanying DLI aggregation service offers a high level interoperability framework for exchanging information about the links between scholarly literature and data, https://www.icsu-wds.org/news/news-archive/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature.

With regards to software access, we again concur with the Software Citation principles on accessibility: "Software citations should facilitate access to the software itself and to its associated metadata, documentation, data, and other materials necessary for both humans and machines to make informed use of the referenced software"

2c. Identification of the authors of the Data/Software products (Maximum: 250 words)

To unambiguously assign credit it is highly recommended that authors use a Unique Identifier, such as their ORCID/Scopus/Mendeley profile ID. Next to this, it is advisable that authors include a unique identifier of the grant number which was used to collect and analyze the data included. This assumes such a grant ID is unique and readily accessible.

Specifically, we are interested in matching up our unique author IDs with those in the funder's information systems, so we can support institutions and individual researchers in developing reporting systems that correctly identify individuals. To that end, it would be useful to be able to have access to the NIH's systems of identification of individuals, institutions and departments.

For software citations, we again concur with the Software Citation Principles:
Credit and Attribution: Software citations should facilitate giving scholarly credit and normative, legal attribution to all contributors to the software, recognizing that a single style or mechanism of attribution may not be applicable to all software.

2d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately (Maximum words: 250)
On Mendeley Data, citations currently point to either datasets or to data files. In future citations will be possible to collections of datasets.

For articles, editors and reviewers are rejecting articles that don't contain sufficient novelty; maybe there should be some sort of responsibility for repositories on how to aggregate data. NIH is doing it in one of the most important dataset ever collected: www.cdc.gov/nchs/nhanes/nhanes_citation.htm.

For software citations, we again concur with the Software Citation Principles on Specificity: "Software citations should facilitate identification of, and access to, the specific version of software that was used. Software identification should be as specific as necessary, such as using version numbers, revision numbers, or variants such as platforms."

2e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed (Maximum words: 250)
In accordance with the DCIP report mentioned earlier, we support the unambiguous identification and creation of a Landing Page containing a PID for each dataset. We have contributed to and are in support of the example set by the Force11 Resource Identifier Initiative, https://www.force11.org/group/resource-identification-initiative, to provide an unambiguous identifier to any electronic resource utilized in a research report.

The Scholix project, mentioned above, also supports the creation of Linked Data Systems to enable unambiguous data citation and identification.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications (Maximum: 250 words)
We would like to suggest that that researchers should not only be encouraged to document and report on how they share data and software but also how they use and contribute to existing data sets and software outputs. This would encourage community cooperation around common data sets and software, and reward the creators of the original data and software.

Proper authorship for data and software allows attribution and credit to both the author and their institution. Tools such as Scopus can provide metrics and analytics around high quality scientific output of any form, software and data as well as articles and books. Using metrics and quality assessments based on Scopus data, especially if it would include data, can provide a valuable tool to give credit to researchers and evaluate the impact of their output.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum: 250 words)
Elsevier is eager and enthusiastic to remain involved in the next stages of discussion on this important topic, and we are looking forward to continuing and expanding our current engagement and collaboration related to research data with the NIH.

In addition to the current RFI response, Elsevier has submitted responses to all research data-related NIH RFIs in the last two years, including:

- NOT-OD-15-067, Soliciting Input into the Deliberations of the Advisory Committee to the NIH Director (ACD) Working Group on the National Library of Medicine (NLM) --> Refer to Comment 5 only
- NOT-AI-15-045, Input on NIAID Data Sharing Repository, Immunology Database and Analysis Portal (ImmPort), and Services
- NOT-ES-15-011, Input on Sustaining Biomedical Data Repositories
- NOT-OD-16-133, Metrics to Assess Value of Biomedical Digital Repositories

Copies of these other related RFI responses are appended here as an attachment for reference.