# Open Public Response to Request for Information: Public Access to Digital Data Resulting from FFSR

**1/12/2012**

This is a public and open document intended to draft a collective response to the request of information posted by the Science and Technology Policy Office (OSTP), on whether digital data resulting from federally funded research should be required to be made publicly available.

**Dear Office of Science and Technology Policy,**

Kitware applauds the initiative of the OSTP on seeking public feedback on these matters of high relevance to the scientific community and to the American public. However, please note that this is not an official Kitware response.

In order to contribute to this process, we reached out to our many collaborators and invited them to join us in writing a collective and thoughtful response to the insightful questions of the RFI. The result is the document attached to this submission letter. The names of the contributors and those in favor of this response are found at the end of the document.

Please find below our response to the RFI on “Public Access to Peer-Reviewed Publications from FFSR”. NOTE: In the responses below we use the following acronyms:

**FFSR**: Federally Funded Scientific Research

**Preservation, Discoverability, and Access**

Question 1: What specific federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

**Response:**
In summary our response advocates:

- Immediate release of acquired data
- Disclosure of  a broad estimation of acquisition cost
- Proper open licensing
- Adoption of open standards for data files
- Adoption of extensible standards for metadata

**Immediate Release**
Federal agencies funding scientific research must establish policies by which the data acquired in federally funded scientific research (FFSR) must be made immediately and fully available in public data repositories. These policies should include provisions for protecting private information in the case of human subjects participating in medical research.

The policies should follow the model of:

- Bermuda Principles
- Pantom Principles (http://pantonprinciples.org/)
- http://sciencecommons.org/projects/publishing/open-access-data-protocol/

In particular on:

- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute, and reorganize the data.

**Disclosure of Acquisition Cost**
When reviewing proposals for funding opportunities, federal agencies should require that the sections requesting public funds for data acquisition activities provide a clear estimation of the cost of acquiring the data. If funded, researchers should be required to make data available in public repositories within 24 hours after acquisition, and in the metadata used to describe a dataset; researchers should also be required to include a disclosure of the cost of acquisition.

The goal will be to develop a sense of the economic cost of not releasing data. For example, not releasing a dataset that cost $1M to be acquired is a loss for the federal government from the $1M funds provided by taxpayers. This is the direct value lost from the overall economy; the actual value lost is much larger since it includes the missed opportunities that could have resulted from the exploitation of the data.

The European Commission, for example, recently adopted a policy of open data dissemination http://www.kitware.com/blog/home/post/212. The principle, rooted in the arguments that Yochai Benkler makes in his book “The Wealth of Networks” is that data is more valuable when shared; in economic terms, data is an “anti-rival good”. It is a good that becomes more valuable when more people have access to it and use it.

**Proper Open Licensing**

Current copyright legislation has been strongly focused on protecting the creators of artistic works, and in the process have created an inhospitable environment for the daily sharing of scientific information. The litigious tendencies of many institutions regarding copyrighted materials also results in over-cautious behaviors from potential data users and documents resulting from scientific research activities.

To dispel this environment of uncertainty, it is fundamental to clarify the rights of the public to make use of data acquired as a result of FFSR. The most effective way of achieving this goal is by affixing a clear statement of licensing, indicating what the recipients of the data are legally allowed to do with the data, to every released dataset. Licensing issues are expanded on in both the Science Commons Protocol for Implementing Open Access Data and the Panton Principles for Open Data in Science.

Some of the best examples of proper licenses are:

- The Creative Commons Zero Waiver: http://creativecommons.org/publicdomain/zero/1.0/
- The Open Data Commons licenses: http://opendatacommons.org/licenses/

Federal agencies should identify a set of licenses that ensure the rights of the general public to deal with the data, in particular to copy, distribute, and create derivative works, and in this way ensure that the data get to reach their maximum economic potential to foster the growth of the U.S. economy. It should then require federally funded researchers to make their data publicly-available under those selected licenses. The pool of licenses must be small, two at the most, to prevent confusion and to maximize the ease by which data can be integrated into subsequent research activities.

**Adoption of Open Standards**

Federal agencies must ensure that data are released in a usable form. The first step in that direction is to require the adoption of open standards for file formats, and forbid the use of proprietary formats that could prevent the general public from having access to the data.

Standard file formats used for digital storage of scientific data are abundant and vary greatly from one domain to the next. Therefore, the scientific community will have to be engaged with the federal agencies in identifying the proper open standard to be used on each discipline, and to create new standards in the cases where no suitable standard file format exists yet.

For standards to reach their full potential, it is fundamental to have an open source reference implementation of the standard, and to encourage the development of an ecosystem in which commercial applications implement the standard as well. In this way, it becomes possible to maximize the use of the data acquired as a result of FFSR. The standards themselves must be unencumbered by patents and copyrights.

**Open Standards for Metadata**

In order to make use of FFSR data, the public must first be able to find it. This is typically done by implementing search engines that rely on publicly available metadata that is affixed to the actual FFSR datasets. The effectiveness of the search engines can be improved by the adoption of open standards that define the form and content of these metadata entries.

Just as with the data formats themselves, open metadata standards require an open source reference implementation, combined with an ecosystem where commercial applications implement the same open standard, to be effective. This wide adoption leads to interoperability, ease of communications, and data exchange.

These standards may have to be defined by different groups in different disciplines. For example, the genomics community will have different needs and interests than the astronomy community, the nano-sciences community, etc.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

**Response:**

In addition to the stakeholders listed in this question, it is critical to note that the general public (the American taxpayer) is the primary stakeholder to be considered here. Given that in the context of FFSR, the public’s tax dollars are paying for the scientific research being undertaken, and thus the public’s interest is the first one that should be considered when making trade-offs between available options.

In order to have a productive discussion on intellectual property, it is important to first deconstruct the term “intellectual property” and clarify its meaning in the context of current U.S. laws. We do this in **Appendix A** and conclude that **copyright** is the only concept of intellectual property that is relevant for the purpose of this RFI.

Under U.S. copyright laws, the only aspect of scientific data that is subject to copyright protection is the creation of organized collections of data. Beyond that unique exceptional case, scientific data are not copyrightable, given that scientific data are factual and must never contain material resulting from the creative labor of artistic work. No scientific endeavor should include data that are the result of the “creative work” of the researcher. Such a practice would be unethical in the context of scientific research. Scientific data must be the result of systematic measurement of real world parameters, or the outcome of computational models that operate on such real world measurements as inputs. In either case, such data do not fit the nature of “creative work” for which U.S. copyright laws provide protection.

Researchers may have applied creative works in the process of designing the experiments and methodologies that lead to the data acquisition. However, the actual data acquired must be factual, and therefore free of creative content, if it is to be considered worthy of the scientific process.

Regarding the copyright for organized collections of data, it is required by U.S. copyright laws that the data organization be non-trivial. For example, the simple ordering of temperature data acquired through time is not worthy of copyright protection. A novel and non-obvious approach to organizing data in such a way that it can be exploited for analysis, or that it reveals patterns and trends never seen before, is more aligned with the kind of creative work that copyright is intended to protect.

That being said, U.S. copyright laws are rooted in the economic bargain by which the government grant creators provide the exclusive right of exploitation of their creations for a limited time, as a way to provide an incentive for the production of such creative works.

In the context of FFSR, such an economic copyright incentive is not needed at all, because the federal government has already provided the funding for researchers to engage in the gathering and organization of the data in the first place. Therefore, the economic incentive has already been provided in the very concrete form of public funds awarded to federally funded researchers. Hence, the economic problem of provisioning “public goods” has already been solved proactively by paying up front for the scientific research using the monetary contributions of American taxpayers. Therefore, attention should turn to making sure that the American taxpayers get unfettered access to the data resulting from FFSR, which they have already paid for.

Scientists who gather data in FFSR do so as part of their job duties, and therefore under U.S. copyright laws they are performing “work for hire”. This means that their employers are the copyright holders of any creative aspect of that data gathering (as pointed above, that only includes the organization of data collections). Given that the scientists’ employers received funds from the federal government, it should be expected that they will be subject to the same demands of the Federal Acquisition Regulations (FAR) as other contractors of the federal government. In particular, with respect to the licensing of data acquired as part of federal contracts.

In the past, it has been a common practice for publishers to demand from researchers the transfer of copyrights related to the materials encompassed in a published scientific article, as a requirement for the publication of such article. No monetary compensation is given by publishers to researchers in exchange for that transfer. The policies of federal agencies should establish that the copyright of FFSR data collections should no longer be transferred to publishers, given that publishers do not provide researchers, their employing institutions, or the federal government with any monetary compensation for such transfer of value.

To maximize the value of data to the public, federal agencies should require researchers to make FFSR data publicly available immediately upon acquisition by using open licenses that clearly state the rights of the general public when dealing with the data.

(3) How could federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

**Response:**
Working groups should be established for different disciplines, involving representatives of leading research institutions for each discipline.

Working groups should define differences on how the data are represented, indexed, stored and exchanged, but should **not** have the latitude to restrict the free dissemination of information in any way. All the policies should consistently have a common requirement for immediate and full release of data, unconstrained by any embargo periods or licensing restrictions. Credit for the acquisition of data could be ensured by data publications (eg http://datacite.org) that can be cited by further works.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

**Response:**
The working groups in the different disciplines (from Question 3) should establish guidelines on practices for dissemination and storage of different types of data. For example, in genomics, it may be reasonable to store the secondary sequence information but not the primary sequence (given their great difference in data size). Analogously, the guidelines may require primary sequences to be stored only for 2 years, while the secondary sequences should be stored for 10 years.

In astronomy, it may be required that certain types of images be stored for different periods of time. Some images may be required to be stored with different compression ratios, and therefore correlate their storage cost with the potential expected benefit for future studies. In this cost-benefit evaluation, the original cost of acquiring the data should be taken into account. For example, a project that invested $50M in acquiring data should not attempt to make savings of a few hundred dollars in storage.

Economists must be involved in the working groups with the mission of providing guidelines for storage and dissemination, as that this is a problem in which the trade-off for the benefit of society-at-large must be continually evaluated.

The recommendations of these working groups should be reviewed and updated regularly in order to keep up with the constant advances in storage technology and the rapid decrease in the cost of storage. The federal government should stimulate the development of storage technology, either by creating large storage decentralized facilities, creating consortia to manage data storage services, involving the public in facilitating distributed (and redundant) storage systems based on peer-to-peer network technology that has already proven to handle large amounts of data.

All these guidelines should be prepared following open and transparent procedures in order to prevent proprietary standards and vendor lock-in situations that would prevent the policies from maximizing the utility of FFSR to the general public.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

**Response**:
They can join the working groups established in their respective disciplines of interest that will define practices for data management, including consortia combining universities, commercial companies, and government agencies.

As standards and agreements are developed, working groups can help implement and test such plans in pilot projects. It will be of great help if federal agencies provide seed funding for these pilot projects.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

**Response**:
**A.** Specific funding streams should be created for researchers and institutions that dedicate themselves to hosting and distributing data. Today, there are very few (if any) funding opportunities for institutions that provide data storage services to their scientific communities, despite the fact that such services are of immense value for fostering the progress of their fields.

**B.** The rewards and merit systems of federal funding agencies must be adjusted to give proper incentives to researchers (and their institutions) who dedicate themselves to facilitate the storage and free dissemination of scientific data. These activities must be valued when researchers and their institutions pursue further funding. Today, only peer-reviewed publications are counted as part of the merit system of researchers when they apply for further funding opportunities. Therefore, researchers have no incentive to engage in public data sharing, and instead have self-interest in retaining data with the hope that it can help them produce more peer-reviewed publications that will contribute to fostering their careers.

**C.** Standard funding streams (such as R01 grants) must include provisions to fund the initial storage and dissemination of data acquired during a research project. This should be enough to cover the period of performance and two years after the end of the project. After that period, data should be moved to dedicated storage services. This practice will replace the current approach of having data storage and processing as an “afterthought,” which leads to inadequate data management, therefore data loses and underutilization of data. See a blog on“Software Forethought” by Kitware CEO Will Schroeder: http://www.kitware.com/blog/home/post/196.

**D.** Federal agencies should track researchers’ compliance with releasing data resulting from previous funding when considering new proposals from those same researchers.

**E.** The Data Sharing plans in grant proposals should be evaluated based on specific provisions for storage and dissemination of the data to be acquired. Review panels should include reviewers with expertise on data storage and web-based distribution services.

(7) What approaches could agencies take to measure, verify, and improve compliance with federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

**Response**:
**A.** Define standard annotations that include information about the funding stream (e.g. grant number, researcher identification, dates of funding) that supported the acquisition of the FFSR datasets.

**B.** Require awardees to tag their data releases with the type of annotations defined in (A) when they post the FFSR datasets to public repositories.

**C.** Fund the creation of a distributed indexing system that allows many institutions to consistently index the annotations (A), and helps the public search those indexes to efficiently locate and gain access to the data. These systems must be decentralized and be open for other organizations and individuals to introduce innovative searching and indexing mechanisms.

**D.** Provide a public Dashboard where the record of data releases for every funded researcher will be displayed publicly. The information should be provided in such a way that it can easily be harvested and data-mined by any other institution for the purpose of generating statistics and comparative studies. Public, open and transparent reporting of compliance with data release policies is the most effective way to ensure that researchers adopt data dissemination practices as a regular and standard activity.

**E.** Award institutions and researchers who excel at data dissemination. For example, a federal agency could provide honorary awards to the researchers each year who excel at sharing data.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

**Response**:
**A. Clear Licensing**
Identify licensing practices that provide clarity on the types of activities that users can perform with the data. It is particularly important for companies and start-ups to invest in the utilization

of the data for fostering businesses and creating jobs. Business must be able to trust that they will not end up in litigation for having used data that was generated as a result of FFSR.

Work by multiple scholars have covered this topic, see for example: Stodden, Victoria, “Enabling Reproducible Research: Open Licensing for Scientific Innovation” (March 3, 2009). International Journal of Communications Law and Policy, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1362040 .

### A.1 Creative Commons Public Domain

One of the best licenses to be considered for scientific data is the **CC0** license: http://creativecommons.org/publicdomain/zero/1.0/

This license has been defined as the closest we can get to put resources in the public domain. The **CC0** license lowers the bar of requirements and controls what the potential rights holders can impose on the recipients (the downloaders and users) of the data.

Our recommendation is to adopt the **CC0** license as the default standard of data sharing in order to ensure that American taxpayers get the maximum return on investment on the resources that they have put in the scientific research enterprise. The **CC0** license removes the majority of obstacles that can be imposed to the free dissemination of scientific information.

For a licensing discussion, see this podcast: http://inscight.org/2012/01/08/episode-22-public-access-to-federally-funded-research/

### A.2 Open Data Commons

Another good set of data licenses is the one defined by the Open Data Commons: http://opendatacommons.org/licenses/ . Among this set, the recommended license is the Public Domain Dedication License: http://opendatacommons.org/licenses/pddl/, which simply states that the data is in the public domain. Placing data in the public domain makes sense because scientific data must not contain any glimpse of creative work. Instead, scientific data must be factual, and facts are not copyrightable.

### A.3 Compliance

Once a set of acceptable licenses are defined for data, funding agencies should require that researchers and institutions use such licenses when delivering data for dissemination, or for storage in external repositories. All such licenses must allow for redistribution, reorganization, and repackaging of the data.

It is reasonable to demand attribution of data sources. Attributions will cascade when data has been passed through multiple stages of processing from one institution to another. In order to prevent the attribution process from becoming a heavy burden, federal agencies should adopt the policy that attribution must be done by citing the URI (Uniform Resource Identifier) of the datasets used. This form of attribution has the properties of being machine readable, searchable, indexable, unique, and compact.

**B. Pilot Educational Projects**

Create streams of funding for pilot projects that will demonstrate how to systematically access public data repositories and generate concise representations of the data. The goal of the pilot projects will not be to innovate by themselves, but to educate the larger public on how to harvest data. Empowered with skill, citizens and institutions will have a lower barrier of entry into the practice of taking advantage of public datasets.

In parallel, funding agencies should spur educational programs for researchers to provide training on the management of data and data collections. Libraries, archives and repositories will be the organizations with the proper background to compose such training programs.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

**Response:**

**A. Tagging Data with Attribution MetaData**

A commonly defined set of metadata annotations will facilitate tagging data with identifiers that point to the funding source, researcher name, research lab, institution, and other key attribution information.

When considering articles for publication, publication venues should require researchers to disclose if they used data from third parties, and if so, to provide the proper attribution using the standard annotation identifiers corresponding to that third-party data source.

As with the rest of the scientific publishing practices, this will be a combination of an honor system, with a light-weight verification system for publishers and funding agencies. The whole becomes effective if it is done in an open and transparent manner, given that any other third party, and in particular any other researcher who suspects that the data she/he disseminated has been used without proper attribution, could raise concerns and trigger corrective measures.

**B. Promoting the Creation of Self-Regulating Governance Bodies**

The problem of proper attribution to the providers of FFSR data is equivalent to the socioeconomic problem of governing the use of common pools of resources (CPRs). As described by Elinor Ostrom, 2009 Nobel Laureate in Economics, such governance models are successful when they have the following characteristics, among others:

- Collective-choice arrangements that allow most resource appropriators to participate in the decision-making process. For the purpose of discussing attribution in this RFI, a “resource appropriator” will be any person or institution who takes FFSR produced data and uses it to further their own mission and goals.
- Effective monitoring by individuals who are part of or accountable to the appropriators. In the case of this RFI, both monitors and appropriators are the researchers who produced and used data.

- A scale of **graduated sanctions** for resource appropriators who violate community rules. This system makes possible to actually apply sanctions when needed, given that the first scale of them will mostly be used as a "call to order," so that researchers who inadvertently broke rules have a chance to fix their omissions without dramatic consequences. At the same time, those who dismiss the "calls to order" can be progressively exposed to increasingly serious sanctions.
- Mechanisms of conflict resolution that are cheap and easy to access.
- Self-determination of the community recognized by higher-level authorities.

The Funding agencies should foster, but not control, the creation of researchers' managed **Data Attribution Tribunals**, perhaps with a less dramatic name, in the image of the "Water Tribunals" that have been used for centuries to successfully manage common water resources. This is one of the practical examples of Governance of Common Pools of Resources from which Ostrom deduced the governance principles listed above. Note that these tribunals are not government organizations; on the contrary, they are community groups composed by the same researchers who have a stake in the process of generation, dissemination, and attribution of scientific data. They would operate on the same honor system and volunteer bases that the current peer-review process operates, but whilefully public and transparent.

Given the sensibilities of researchers, a name less dramatic than "Tribunal" will certainly be more conducive to engage them in the process. For example, **"Open Data Attribution Arbitration Group"** could be a better name.

Reference: Elinor Ostrom, "Governing the commons: the evolution of institutions for collective action", Cambridge University Press, 1990

### Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

**Response**:
Digital data standards must be open, patent-free, and must require an open source reference implementation. Research communities will each have different standards to suit particular needs, which is perfectly acceptable as long as they are all open.

In some domains, there are organizations or working groups formed from community members to establish data formats, standard descriptions, or common interfaces with open implementation. For instance, the International Neuroinformatics Coordinating Facility (INCF, www.incf.org) has international working groups on standards for data sharing in neuroimaging and electrophysiology. An efficient use of funds would be to promote the established standards and join existing working groups on metadata standards.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

**Response**:
Effective standard definitions are the result of:

**A.** Involving the users of the standard in the definition process. This may require funding to initiate representative working groups for establishing the standards, or continue the work of existing groups.

**B.** Ensuring the full openness of the standard by requiring patent disclosures and royalty-free patent licensing from any institution participating in the definition of the standard.

**C.** Developing free and open source reference implementations of the standard at the same time that the standard is being defined. This ensures the practical applicability of the standard being defined, and also greatly promotes wide adoption of the standard.

**D.** Promoting an ecosystem in which commercial applications are encouraged to provide implementations of the standard without having incentives to create proprietary variations of it.

(12) How could federal agencies promote effective coordination on digital data standards with other nations and international communities?

**Response**:
**A.** Ensuring that internationalization (of language and “locale”) is made an integral part of the standards.

**B.** Starting with simple standards that can progressively be improved, instead of spending a lot of time on top-down design, committees, and long-term procedural approaches to the definition of the standard. In other words, following the Agile methodologies that have proved to be successful in open source communities.

**C.** Working with existing international organizations that have already defined standards in different disciplines. See for example, INCF.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

**Response**:
Adopt the standard use of

- Unique Resource Identifiers (URI)
- Digital Object Identifiers (DOI)

These two mechanisms have been used for several years to refer to digital resources in the literature.

An URI points directly to the location of the data in an unambiguous way.

The interest for DOIs is that, in many cases, researchers and institutions want their data to be addressed through another level of indirection to enable the moving of data from one hosting service to another. Services such as DOI (http://www.doi.org/) enable that level of indirection in a standard way.

## Signatures

| Name | Title | Institution |
|---|---|---|
| Luis Ibanez | Technical Leader | Kitware Inc. |
| Wesley Turner | Technical Leader | Kitware Inc. |
| Berk Geveci | Technical Leader | Kitware Inc. |
| Amitha Perera | Technical Leader | Kitware Inc. |
| Marcus Hanwell | R&D Engineer | Kitware Inc. |
| Matthew McCormick | R&D Engineer | Kitware Inc. |
| David Stoup | R&D Engineer | Kitware Inc. |
| Andy Bauer | R&D Engineer | Kitware Inc. |
| Paul Tunison | R&D Engineer | Kitware Inc. |
| Zack Galbreath | R&D Engineer | Kitware Inc. |
| Jean-Christophe Fillion-Robin | R&D Engineer | Kitware Inc. |
| Katie Osterdahl | Communications Specialist | Kitware Inc. |
| Katie Sharkey | Communications Specialist | Kitware Inc. |
| Steve Jordan | Graphic Designer | Kitware Inc. |
| Arno Klein | Asst. Prof. Clinical Neurobiology | Columbia University |
| Jean-Baptiste Poline | Researcher | UC Berkeley and CEA France |
| Cameron Smith | Computational Scientist | Rensselaer Polytechnic Institute |
| Ziv Yaniv | Principle Investigator | Children’s National Medical Center |
| Raphael Ritz | Scientific Officer | INCF |

## Appendix A - Intellectual Property in Scientific Data.

The term of "intellectual property" is commonly used as an aggregate of the concepts of

- Copyright
- Patents
- Trademarks
- Trade secrets

In order to understand how these concepts apply to the challenge of maximizing access to the results of scientific research funded by the federal government, it is important to analyze the concepts independently.

**Copyright** is a government-awarded monopoly given to the creators of works of art. This monopoly awards creators the exclusive right to (1) reproduce the work, (2) prepare derivative works of it, (3) distribute copies of it, (4) perform it publicly and (5) display it publicly. The duration of copyright is: (a) the lifetime of the authors plus 70 year, (b) 95 years for works created by a corporation, or (c) 120 years for unpublished works created by a corporation. The goal of copyright is to provide an incentive to the creators of works of art by giving them exclusive rights on the exploitation of the works for a limited time

In the context of dissemination of scientific data, the economic bargain of copyright bears very low or no relevance, given that researchers (those who acquire and process the data) do not get paid when publishing that data. Instead, they get funded proactively for performing the research that leads to gathering information that is later published. Therefore, a very concrete economic incentive has already been provided and delivered to the researcher in the form of funding that American taxpayers have invested in the acquisition of the data.

As opposed to a novelist, whose income if purely based on the sale of copies of her/his book, the salary of a researcher is based on their performing the duties of scientific research. Granted, publishing datasets is part of such duties, but it is not equivalent to the creative activity of writing works of art (such as novels, music, or poems). Given that, in the context of FFSR, researchers are already paid by the public beforehand and so there is no need for the economic incentive of copyrights to address any "market failure" on the production of public goods (in the economic sense of non-rival and non-excludable goods), as is the case for novels, poems, and music. On the contrary, once the FFSR data has been acquired, every day that passes without this data being publicly shared is a day in which economic waste takes place and the economy at large performs less efficiently. It is also a day in which American taxpayers do not get anything back from the funds that they provided to the research enterprise.

Additionally, the nature of scientific research requires that the content of scientific datasets must be measurements of facts and should be devoid of any "creative elaborations". In other words, the more "scientific" a dataset is, the less "creative artistic content" it should have in it; therefore, the less it deserves the protection that copyright is intended to provide to creative works of authorship. The creativity of the researchers lies in the definition of the acquisition protocols, the experimental design, and in the specific apparatus or software used during the data acquisition, which sometimes are made especially for a specific dataset. The dataset itself, on the other hand, shall not include any creative content. A high quality scientific dataset must be a concise collection of facts, measurements, and computations on those measurements. Datasets with high levels of "creative content" are by definition not scientific datasets, and should not be produced as the outcome of federally funded research, or any other process that aspires to be called "scientific".

**Patents** are government-awarded monopolies on the commercial exploitation of an invention. This 20-year long monopoly is awarded to the inventors in exchange for the public disclosure of the invention, and its eventual delivery (at the expiration of the patent term) to the Public Domain. Given that public disclosure is a requirement of the patent economic bargain, for awarded patents there is no concern about including information in articles intended for publication. The full information about the invention should already be publicly available at the U.S. Patent Office at the time that the patent is awarded to the inventors. Data is not "patentable subject matter" given that it is not the result of a creative process and is not useful, non-obvious, or novel. Datasets collected in the course of scientific endeavors are expected to be a collection of factual data, and therefore, they are as far as they can get from the type of "creative" work that patents are intended to protect.

**Trademarks** are symbols, designs, and terms that identify a product, service or company in the public marketplace. They are intended to prevent confusion in the marketplace, to protect the reputation of the producers of goods and providers of services, and to reduce the transaction cost that consumers have to invest in finding good and services that satisfy their needs. In the context of dissemination of scientific data, trademarks play a minimal role given that datasets are not supposed to be mechanisms of marketing goods and services. It is actually contrary to ethical standards in the scientific research field to use dataset publication as a venue for promoting goods and services in the context of commerce.

**Trade Secrets** refer to information that organizations keep confidential. For a piece of information to be considered a trade secret, it must have some value and derive part of its value from the mere fact of being secret. Trade secrets are managed via contracts, typically established between organizations in the form of non-disclosure agreements and between organizations and their employees in the form of confidentiality clauses that are incorporated in employment contracts. It is the responsibility of the institution to take affirmative steps to prevent its confidential information from becoming public.

In the event that a piece of confidential information is leaked publicly, there is no legal protection that can prevent the further dissemination of such information, except from forbidding an intruder to make use of data that was acquired illegally (e.g. by trespassing into private property). Therefore, in the context of dissemination of scientific data, trade secrets are only relevant as a context in which institutions should establish policies and verification mechanisms that prevent confidential information from being included in any dataset that is submitted for public release. It is the responsibility of the institution and its employees to protect such confidential information.  Once data is published, the institution has relinquished its claim for such data to be considered a trade secrets.