Contributing to NIH-wide Initiatives of Big Data and Data Sharing: Building a Web Portal of NIH Data Sharing Repositories and Data Sharing Policies

Project Leader

Kevin Read, NLM Associate Fellow

Project Sponsor

Jerry Sheehan, Assistant Director for Policy Development, Office of the Director

December 2011 – January 2012

Table of Contents

Abstract 3

Introduction 4

Procedures 5

- Defining Selection Criteria 5
- Selecting NIH Data Sharing Policies 6
- Selecting NIH Data Sharing Repositories 7
- Presentation to BMIC Committee and Request for Submissions 8
- Creating Website Wireframes for the Data Sharing Policy and Data Sharing Repository Inventories 9

Outcomes 9

Discussion 10

Recommendations 12

- Recommendation 1: Maintain the sustainability of the inventories 12
- Recommendation 2: Create categories for the data sharing repositories inventory to improve retrieval 12
  - Image 1. NIH Data Sharing Repositories Data Access Policy 13
  - Image 2. – NIH Data Sharing Repositories Metadata Identifiers 14
- Recommendation 3: Develop promotional strategies to increase access to the inventories 14
- Recommendation 4: Seek out data sharing communities or repositories that are not funded by the NIH 15

References 17

Acknowledgements 19

Appendices 20

- Appendix A. User Interface Mockup for BMIC Web Portal 20
  - Image 3. BMIC Home Page Mockup 20
  - Image 4. NIH Data Sharing Policies Inventory Web Portal 21
  - Image 5. NIH Data Sharing Repositories Inventory Web Portal 22
- Appendix B. NIH Data Sharing Policies Inventory 22
- Appendix C. NIH Data Sharing Repositories Inventory 22

Abstract

OBJECTIVE: To address the current NIH-wide discussions of Big Data and the NIH Data Sharing Policy, a project was undertaken to develop two inventories of data sharing information: the first inventory includes all NIH data sharing repositories that are open to receiving data from any researcher – funded by the NIH or not. The second inventory is comprised of all the data sharing policies that exist within the NIH that assist researchers in developing a plan to share their research data. The goal is of this effort is to host both inventories on the Biomedical Informatics Coordinating (BMIC) Committee website where any researcher can visit the site and access this information.

METHODS: To develop inventories of NIH data sharing policies and repositories, the following tasks were performed: searching every NIH Institute and Center to find data sharing policies and data sharing repositories that met the selection criteria; searching systematically through the BMIC Catalog of NIH Databases, Disease Registries, and Biomedical Information Resources to find more results; exploring other data sharing repository registries to increase results; adding revisions and suggestions made by BMIC members after initial drafts were presented at their December meeting; and crafting website wireframes for the NLM Web Information Management (WIM) unit so that the inventories could be transferred onto the BMIC website.

RESULTS: Comprehensive inventories of both NIH data sharing policies and data sharing repositories were developed. Both inventories along with their respective website wireframe designs were sent to the WIM unit where they will be translated into web portals that will be hosted on the BMIC website.

CONCLUSION: This project contributed to the NIH-wide data sharing initiative by compiling all of the NIH data sharing policies and repositories in one place for the first time. Positioning these inventories on the BMIC webpage will encourage researchers to share their data by providing them with access to NIH repositories that support data submissions and NIH data sharing policies that guide and support the data submission process.

Introduction

The Trans-NIH Biomedical Informatics Coordinating (BMIC) Committee is a group devoted to improving the communication and coordination of issues related to clinical- and bio-informatics at the NIH. The Committee also provides a forum for sharing information about NIH informatics programs, projects, and plans. At present, BMIC has been involved in discussions related to the Big Data phenomena, as well as improving data sharing within the NIH. To address the current NIH-wide discussions of Big Data and the NIH Data Sharing Policy, an effort was made to develop inventories of NIH funded data repositories and data sharing policies.

In collaboration with members of BMIC, a comprehensive inventory of NIH-funded data repositories from each Institute and Center (IC) was compiled to include repositories that are open to receiving data from NIH-funded extramural researchers as well as researchers who have *not* received funding from the NIH. An example of the type of data sharing repository selected for inclusion is the National Database for Autism Research (NDAR). This repository is open to receiving data submissions from NIH-funded researchers, but it also welcomes data from researchers who are not funded by the NIH. The goal of this repository is to increase collaboration and share data across the entire Autism Spectrum Disorder (ASD) field. The NDAR repository is a good example of the type of data sharing repositories that were selected for inclusion.

Similarly, an inventory of NIH led data sharing policies was also assembled. In addition to the general NIH Data Sharing Policy – which applies to NIH funded projects that receive more than $500,000 in any given year – the NIH and its various ICs have established a number of other data sharing policies that apply to specific types of funded research, such as genome-wide association studies, autism research, and traumatic brain injury research. These policies include considerations, directions, expectations or obligations from the NIH or an IC that a researcher will deposit their data into a specific repository once they have finished their research. Policies of this kind are created to encourage researchers to share their findings in a location that will be accessible where other researchers can view their data. For example, the National Institute of Neurological Disorders and Stroke (NINDS) have created a data sharing policy specifically to address traumatic brain injury research; their policy expects researchers to a) provide descriptive information about their studies, and b) submit coded genotypic and phenotypic data to the Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System. This specific policy places an emphasis on data sharing, and reflects the types of policies from each IC that were collected for inclusion to the inventory.

The concern with these data sharing repositories and data sharing policies was that they were spread across a range of websites hosted by the aforementioned NIH ICs. The goal of this project was to compile all of the data sharing repositories and data sharing policies in one place to provide the scientific community with access to this information through a web portal to facilitate a system for researchers to track what repositories they will be able to use to deposit their data, and what policies they can follow to share their data. This effort is intended to promote data sharing at the NIH as well as to a larger audience, and encourage extramural researchers to share their data by submitting datasets to one of the NIH data sharing repositories. It is important to note that at these data sharing policies and repositories do not currently include intramural researchers. However, intramural researchers at NIH could still submit data on the grounds that these resources were compiled with the intention that they are open to receiving data from anyone. This project report outlines the procedures taken to create both inventories and describes the benefits and outcomes of this endeavor. Finally, recommendations are made to promote the inventories and ensure the sustainability of the web portal over time as new data sharing policies and repositories are created.

Procedures

Defining Selection Criteria

The first step in developing the inventories of data sharing repositories and NIH data sharing policies was to clearly define the selection criteria for inclusion. Working with the project sponsor on the interests and priorities of BMIC and the NIH, clear definitions were created to guide the selection process.

The selection criteria for data sharing repositories was limited to those that are supported by the NIH; are open to submissions of data from NIH-funded investigators as well as potential researchers who are not funded by the NIH; and expect to be supported for an extended/indeterminate period of time (e.g., beyond completion of a single project). Furthermore the decision was made to include some resources that are aggregators of resources rather than data repositories themselves (e.g., Neuroscience Information Framework [NIF] and the Neuroimaging Informatics Tools and Resources Clearinghouse [NITRC]) because they also promote data sharing through the distribution of data tools that assist researchers in manipulating their datasets. Registries of data were excluded for this project because they do not allow researchers to submit datasets and therefore failed to meet the selection requirements.

The process of defining our selection criteria for the NIH data sharing policies was a greater challenge. Because there are data sharing policies that exist within Request for Application (RFA) grants and program announcements (PA), a decision had to be made on how to select policies that would truly reflect the openness of sharing and submitting data. As a result, the selection criteria was dedicated to collecting data sharing policies that are promulgated at the NIH, Institute, Center, division, or program level and apply to broader sets of investigators and data than those involved in or resulting from a particular RFA or narrowly defined project (e.g., a particular clinical trial). IC policies that were restatements of the general NIH Data Sharing Policy were also excluded to avoid redundancy as the examples found did not add supplementary information to the existing policy.

Selecting NIH Data Sharing Policies

Searching for data sharing policies was a two part process. The first part involved searching through every NIH IC website to look for potential data sharing policies. The first iteration of this process involved examining the 'Research' and 'Funding' section of each webpage; this was a logical approach because each IC website included a section devoted to this topic. If nothing was found in either section, the 'About' section of each site was analyzed to determine if a link to a data sharing policy was included. Once these exploratory options were exhausted, a search strategy was used to locate the policies within each IC website:

*((data shar*OR "data sharing" OR data) AND polic*)*

If a data sharing policy was found that met the selection criteria, it was entered into the inventory. The pertinent information for the data sharing policies included the following attributes: a link directly to the policy; a detailed description of the data sharing policy; which repositories the data sharing policy suggests that data be submitted to; and links directly to the aforementioned repositories.

The next phase of the selection process was to review the BMIC Catalog of NIH Databases, Disease Registries and Biomedical Informatics Resources [1]. This catalog contains a list of approximately three hundred NIH resources with links to their web content. Once registries were eliminated from the catalog based on our selection criteria discussed earlier, the next step was to systematically work through each resource in the catalog and search through their website for a data sharing policy. Similar to the approach taken when searching through each IC website, the first step for looking through each resource involved finding a 'Data' section to see if a data sharing

policy was available. If the resource did not include a 'Data' section, the next step was to find a 'Research' or 'Funding' section. If a data sharing policy could not be found in these three sections, a search within the website was completed using the same strategy outlined above. If a data sharing policy was found, it was recorded into the inventory.

A final sweep of data sharing policies was completed using a Google search to ensure that no policy was overlooked. Using the same strategy for the IC websites and the BMIC catalog, a search was completed using the *site:nih.gov* feature in Google. Based on the first 100 results, no additional data sharing policies were found. This finding concluded the search for data sharing policies.

Selecting NIH Data Sharing Repositories

Following a similar strategy used for searching for data sharing policies, the BMIC Catalog of NIH Databases, Disease Registries and Biomedical Informatics Resources [1] was used to search for data sharing repositories. Again working systematically through each resource in the catalog, each resource was searched for a data submission policy. Finding a data submission policy was essential because it indicated whether or not the repository supported data submissions from NIH-funded investigators or investigators from outside the NIH; and if it was expected to support the data for an extended or indeterminate period of time as defined in the selection criteria section of this report. Finding the 'Data' section of the website was the most crucial component when searching through each repository website. This section would indicate whether the repository had a data submission policy at all, and outline who was permitted to submit data, and the procedures necessary to carry out the process.

After finalizing the analysis of the BMIC catalog the next step was to extend the search to current data repositories that may have been missed since the catalog was created in 2009. It was agreed upon with the Project Sponsor that Databib [2] would be an excellent resource for finding additional NIH data sharing repositories. Databib is a tool that is designed to help identify and locate online repositories of research data. Users and bibliographers from the site create and curate records that describe data repositories so that they are easily searchable. The benefit of Databib is that it is geared towards data users, data producers, publishers and professional societies, librarians and research funding agencies. The advisory and editorial boards consist of senior staff members from academic institutions, as well as international and government agencies – Databib is

hosted by Purdue University Libraries. When looking for NIH data sharing repositories within this site, the most efficient way to search was to limit the repositories by subject, and systematically work through both the 'Biological Sciences' and 'Health and Medical Sciences' subject headings to uncover which repositories were NIH funded. After finding suitable NIH funded data sharing repositories under each subject heading, they too were reviewed for data sharing or submission policies and then added to the inventory.

The content within the NIH data sharing repositories inventory was organized into the following attributes: the NIH Institute or Center that funds the repository; the name of the repository; a detailed description of the data repository; a link to the repository website; and a link to the data submission policy. The goal in creating these attributes was to make all the pertinent elements of the data repository easily accessible and readily transferable to a webpage.

Presentation to BMIC Committee and Request for Submissions

Once the material for both inventories was fully assembled, it was reviewed in order to present the findings to the BMIC Committee meeting on December 5, 2012. During the presentation a request was made for each IC member from BMIC to investigate their current funding projects to find other data sharing repositories and data sharing policies that may have been missed during the initial analysis. Because the IC members are best acquainted with the repositories that they fund and policies they create, it was decided that their own investigation of what is available should be the next logical step. The criteria for selection was outlined in detail during the meeting and then sent to each BMIC member via email as a reminder. Members were given the rest of the month (three weeks) of December to accrue suitable policies and repositories and submit them for inclusion. As policies and repositories were submitted to the project leader, they were added to the inventory if they met the selection criteria. If the submissions were not appropriate, there was a discussion with the respective BMIC member to inform them why their suggestions did not fit the scope of this project.

It was also suggested at the BMIC meeting that another attribute be added to the data sharing repository; for the sake of data access, members felt that it would be useful to include an attribute that provided a link to a repository's data access policy to accompany the data submission policy attribute. In order to gather this information each repository had to be searched again for their data access policy. This information was located in the same section where the data

submission policies were found under the 'Data' section of each repository – as a result it was not difficult to add this attribute to the inventory.

Creating Website Wireframes for the Data Sharing Policy and Data Sharing Repository Inventories

After every BMIC member had an opportunity to add their submissions and make corrections to both inventories, each document was finalized. The final set of repositories and data sharing policies were accrued, and then prepared for the web portal on the BMIC website.

Quality assurance was completed on all of the entries to ensure that the links were functional and that the repository and data sharing policy descriptions were free of errors. Each description was also reviewed to ensure that the function of each policy and repository was appropriately defined and provided enough detail. After these changes were approved by the project sponsor, the inventories were declared ready for the BMIC website.

Before the inventories could be submitted to the Web Information Management (WIM) unit for web development, a website wireframe was created to provide the best options for the design of the content. Using PowerPoint, several iterations of the inventories were built and presented to the project sponsor. These mockups were evaluated based on their location within the BMIC web site; the readability of the tables; the descriptions of the inventories; and whether they met 508 compliance regulations. The wireframes were developed according to the existing BMIC website schema. After several iterations the final draft was confirmed [Appendix A], and the information was sent to the WIM unit to be created.

Outcomes

The task of creating inventories of NIH data sharing policies and repositories has produced two main outcomes: the first outcome represents the fact that this is the first time the NIH has had all of its data sharing policies and repositories organized in one place. Because these policies and repositories were scattered across several locations within a large number of NIH IC websites, it was difficult to ascertain how many of these policies and repositories exist. Now that this information has been collected in one place it is much easier for NIH staff to locate these policies and repositories and point researchers to this information. Furthermore these inventories can now

be monitored and reviewed so that additional material can be added as they are created over time. This guarantees the sustainability of the inventories and solidifies the idea that this data sharing information has a set location where it can continue to grow.

As a second outcome, this project has provided the groundwork for the creation of a web portal. Because both inventories provide valuable information to researchers on how and where they can share their data, the next logical step is to post this information on the web. Since the work was completed as part of BMIC, the web portal will be hosted on their webpage under the heading *BMIC Resources*. Both the data sharing policies and data sharing repositories inventory will have their own separate page within the site. As a result, the BMIC web page will serve as a single point of contact for data sharing information. This information will benefit the research community at the NIH and beyond as it serves to promote data sharing and provide avenues for researchers from a range of disciplines to share their datasets.

Lastly, this project has advanced the NIH-wide discussion to promote data sharing and continue the effort to attract researchers to the benefits of data sharing. The information gathered for this project will serve as the foundation for another project that will analyze the common metadata elements of each data sharing repository listed in the inventory. The purpose will be to see if common themes emerge from the metadata – this effort will consequently serve as the first step towards building an NLM data catalog.

Discussion

This project is an example of the efforts undertaken by the BMIC Committee to address the current discussions and growing interest in data sharing at the NIH. Over the past several years, sharing research data has become a hotbed of discussion within the scientific research community. Emerging from this discussion are advocates who emphasize the need for researchers to openly share their data. Van den Eynden et al. [3] from the UK Data Archive suggests that sharing research data encourages scientific enquiry and debate; promotes innovation and potential new data uses; leads to new collaborations between data users and data creators; and maximizes transparency and accountability among several others. Borgman [4] echoes the idea that data sharing increases the potential to collaborate with other researchers. She states that sharing data should incentivize researchers to have their data open to peer review, and receive feedback that would ultimately improve their research and foster a cohesive relationship with other scientists. An article by Piwowar et al. [5] claims that sharing research data is associated with an increased citation rate. In

a study of 85 cancer microarray clinical trials published from 1999 and 2003, the authors found that 48 percent of trials that shared their data received 85% of the total citations (5334 of 6239). Although the sample size in this case is small and from a very specific type of study, the result is an indication that sharing research data can have positive results. The benefits of data sharing are being promoted internationally, and the BMIC data sharing web portal is an opportunity for the NIH to promote data sharing.

Data sharing has also been largely promoted by government agencies. In Europe, the Royal Society compiled a report on data sharing called *Science as an open enterprise* [6]. This report stresses the importance of including published conclusions and associated data together to increase comprehension. It stresses that publishers are also required to play a role, where they should require datasets associated with published papers to be accessible in an electronic format. Finally, the report suggests that datasets should be deposited in a subject-appropriate repository that can be linked via a DOI or accession number via the published article [6]. This last point closely aligns with BMIC's decision to collect data sharing policies so that researchers are aware of which repositories they are encouraged to deposit their data into and what steps they need to take to do so. By making these policies more readily available, BMIC can promote data sharing to extramural researchers beyond the NIH and intramural researchers from various ICs. The web portal also contributes to the Royal Society's set of data stewardship principles with respect to providing clear routes for researchers to access data to facilitate openness among the scientific community [6]. Including the inventory of data sharing repositories provides an opportunity for researchers to easily locate open data and potentially submit their own.

Similar to the report from the Royal Society, the U.S. National Science Board at the National Science Foundation created a report in 2011 that addressed the key challenges government agencies face with research data sharing [7]. These challenges address the importance of broad stakeholder involvement of the scientific community when it comes to sharing data. Active researchers in a variety of disciplines, universities, research libraries, publishing companies, scholarly societies and funding agencies can play critical a role in sharing and managing data. The challenge is to get these stakeholders to communicate with each other and be open to sharing data. Data sharing needs to be fully accepted as a common, beneficial practice by all science communities [7]. Creating policies that acknowledge and provide disciplinary nuances is another challenge that needs to be addressed in order to foster a culture of interdisciplinary sharing across scientific research communities. This project attempts to meet these challenges as the web portal is meant to

promote data sharing by providing access to policies and repositories that allow researchers to contribute and access data that is relevant to them. Recommendations from both the Royal Society and the National Science Foundation indicate that it is important for Federal agencies to provide leadership in the area of data sharing. In collaboration with BMIC, the web portal will promote the Institutes and Centers at the NIH who are currently involved in data sharing while serving as an example of leadership for the scientific community.

Recommendations

Recommendation 1: Maintain the sustainability of the inventories

The first recommendation to the BMIC committee is to maintain the sustainability of the data sharing inventories. In order to avoid the issue of having these data sharing policies and repositories spread throughout the NIH ICs it is crucial that these inventories are reviewed regularly and new material is added accordingly. At present, the BMIC committee holds quarterly meetings; it is recommended that revising the NIH data sharing policies and repositories inventories is brought up at every second meeting of the year. Reinvestigating the data sharing initiatives of each IC every six months will ensure that all new data sharing policies and repositories will be accounted for. At the aforementioned meeting BMIC members should be encouraged to seek out new policies and data sharing repositories that may have been created within their IC. A time frame that is appropriate will be given to members for providing their submissions. For example this phase of the project gave BMIC members three weeks to find their material – this could be a suitable length of time to allot. Once this information is collected it should then be immediately added to the existing inventories with all the attributes described in this report.

Recommendation 2: Create categories for the data sharing repositories inventory to improve retrieval

The first iteration of the data sharing repository web portal is designed to simply list all the data sharing repositories and have make them sortable by NIH IC and by the name of the repository (Image 1).

Image 1. NIH Data Sharing Repositories Data Access Policy

U.S. National Library of Medicine
National Institutes of Health

Contact NLM
Search
The World's Largest Medical Library

Databases | Find, Read, Learn | Explore NLM | Research at NLM | NLM for You

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

NIH Data Sharing Repositories

This table contains NIH-supported data repositories that are open to submissions of appropriate data from NIH-funded investigators (and possibly other investigators) and are expected to be supported for an extended/indeterminate period of time (e.g., beyond completion of a single project). Also included are resources that are aggregators of resources rather than data repositories themselves. This table can be sorted according to the NIH Institute or Center so that you can find repositories more relevant to you.

To view more information about data sharing at NIH and data sharing policies, please see the BMIC page data sharing tools.

NIH Institute/Center	Repository Name	Repository Description	Data Submission Policy	Data Access Policy
NCI	The Cancer Imaging Archive (TCIA)	The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as "Collections", typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.	How to Submit Data to TCIA	How to Access TCIA Data
NCI (NHGRI, NIGMS)	PeptideAtlas	PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.	How to Submit Data to PeptideAtlas	How to Access PeptideAtlas Data
NHGRI	FlyBase: A Drosophila Genomic and Genetic Database	Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BAC's.	How to Submit Data to FlyBase	How to Access FlyBase Data

Copyright, Privacy, Accessibility, Site Map, Viewers and Players
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Health & Human Services
Freedom of Information Act, Contact Us

USA.gov

Last reviewed: 04 January 2013
Last updated: 04 January 2013
First published: 26 June 2012
Metadata | Permanence level: Permanence Not Guaranteed

These sorting functions are useful because the list contains forty five different repositories, but in the future it would be useful to create categories for the data access policies included in the list. For example, under data access policies (highlighted in yellow) it would be useful to categorize each policy according to who has access to the data from that particular repository. Because some repositories are open to the general public, while others only provide access to researchers and are subject to approval from the repository Principal Investigator (PI), it is recommended that this attribute is categorized according to what type of access each repository allows. This would afford each user of the portal the opportunity to search based on the type of access, and as a result would improve their retrieval.

Another recommendation for improving the retrieval of the NIH data sharing repositories is to create metadata identifiers based on the description information for each repository. In this case, it is recommended that the subject of the data be identified (eg. multi-organism, frog, drosophila, etc.); the type of disease or condition it addresses (eg. cancer, Parkinson's disease,

autism, etc.); and what type of data it is (eg. genomic, genetic, imaging, clinical, etc.). From the example below (Image 2), the subject of the data is highlighted in red, the type of disease in green, and the type of data in blue to give an indication of how this task can be performed.

Image 2. – NIH Data Sharing Repositories Metadata Identifiers

U.S. National Library of Medicine
National Institutes of Health

Contact NLM
Search
The World's Largest Medical Library

Databases | Find, Read, Learn | Explore NLM | Research at NLM | NLM for You

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

NIH Data Sharing Repositories

This table contains NIH-supported data repositories that are open to submissions of appropriate data from NIH-funded investigators (and possibly other investigators) and are expected to be supported for an extended/indeterminate period of time (e.g., beyond completion of a single project). Also included are resources that are aggregators of resources rather than data repositories themselves. This table can be sorted according to the NIH Institute or Center so that you can find repositories more relevant to you.

To view more information about data sharing at NIH and data sharing policies, please see the BMIC page data sharing tools.

NIH Institute/Center	Repository Name	Repository Description	Data Submission Policy	Data Access Policy
NCI	The Cancer Imaging Archive (TCIA)	The Cancer Imaging Archive (TCIA) is a large archive of **medical images** of **cancer** accessible for public download. All images are stored in DICOM file format. The images are organized as "Collections", typically **patients** related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.	How to Submit Data to TCIA	How to Access TCIA Data
NCI (NHGRI, NIGMS)	PeptideAtlas	PeptideAtlas is a **multi-organism**, publicly accessible compendium of **peptides** identified in a large set of tandem **mass spectrometry proteomics** experiments.	How to Submit Data to PeptideAtlas	How to Access PeptideAtlas Data
NHGRI	FlyBase: A Drosophila Genomic and Genetic Database	**Drosophila** Genomic and Genetic database that includes **proteomics data, microarrays and Tiling BAC's**.	How to Submit Data to FlyBase	How to Access FlyBase Data

Copyright, Privacy, Accessibility, Site Map, Viewers and Players
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Health & Human Services
Freedom of Information Act, Contact Us

USA.gov

Last reviewed: 04 January 2013
Last updated: 04 January 2013
First published: 26 June 2012
Metadata | Permanence level: Permanence Not Guaranteed

It is expected that this inventory will continue to grow over time as new data sharing policies emerge from various NIH ICs. As the list becomes larger, it will become increasingly important to make the content easily retrievable. Creating these metadata identifiers will serve to benefit users should the web portal be transferred to a database. These identifiers will enhance the retrieval process by allowing users to search for the subject, disease or condition and types of data used in these repositories.

Recommendation 3: Develop promotional strategies to increase access to the inventories

It is under strong recommendation that both the NIH data sharing inventories be heavily promoted so that the broader scientific research community is aware of its existence. Promotion

can be achieved through a number of different avenues including social media, outreach to existing aggregators of data related information, and academic libraries that provide data management services.

With respect to social media, it would be helpful for each NIH IC that has a Twitter account to tweet about both the NIH data sharing policies and repositories web portal on a consistent basis. Furthermore if there are other organizations with Twitter accounts that would find this information useful it would be prudent to include their Twitter handle within the tweets. Facebook is another useful resource where the inventories could be promoted on both the *NIH News in Health* and *NIH Research Matters* pages. Each page has several thousand likes and would be an excellent way to increase exposure to this information.

For aggregators of data sites, it would be useful to have these inventories promoted on both the data.gov [8] and healthdata.gov websites [9]. These sites receive a lot of attention and the NIH data sharing inventories contribute to their open data initiative. The inventories should also be included on the Databib site as well. Reaching out to members of the advisory board or editorial board could potentially grant the inventories a place within the site and thus serve as another avenue for promotion. DataCite [10] is another popular website that provides access to data repositories and works in collaboration with DataBib. Their repository page would be another excellent place to link out to the inventories.

Finally, a key opportunity for promotion can be found through promoting both inventories on academic libraries' resource guide pages. The emergence of data and data management has encouraged many libraries to create resource guides on the topic. The Massachusetts Institute of Technology [11], California Digital Library [12], University of Massachusetts [13], Simmons College [14], University of Minnesota [15], University of Washington [16], and Purdue University [17] all host resource guides with access to data sharing and data management resources. The data sharing inventories would be perfectly suited for these pages, and librarians would be more than likely to include them on their site to direct patrons to this useful information.

Recommendation 4: Seek out data sharing communities or repositories that are not funded by the NIH

As a result of the recent discussions of data sharing as an emerging trend, a number of innovative data sharing initiatives have materialized within the scientific community. Projects such as HUBzero [18], Cytobank [19], and WebPAX [20] have approached the subject of data sharing

through online communities that encourage the sharing of research data, foster research collaboration, and promote collective data analysis. These online communities have piqued the interest of several NIH ICs as a platform for sharing data as they could serve to enrich their own research communities' data sharing efforts. As a result, it is recommended that online data sharing communities of this kind be added as a supplement to the existing data sharing inventories. If projects like these receive more attention from the NIH ICs, it will be useful to include them in the web portal to provide another opportunity for researchers interested in sharing their data.

References

1. Biomedical Informatics Coordinating Committee. Catalog of NIH Databases, Disease Registries and Biomedical Informatics Resources. Bethesda (MD): National Institutes of Health; 2009 [cited 2012 Dec 14].
2. Purdue University Libraries. Databib [Internet]. West Lafayette (IN): Purdue University Libraries; 2012 Dec 31 [cited 2013 Jan 4]. Available from: http://databib.org/index.php
3. Van den Eynden V, Corti L, Woollard M, Bishop L, Horton L. Managing and Sharing Data: Best Practices for Researchers [Internet]. 3rd ed. UK Data Archive; 2011. Available from: http://data-archive.ac.uk/media/2894/managingsharing.pdf
4. Borgman C. Research Data: Who will share what, with whom, when, and why? China-North American Library Conference [Internet]. Beijing; 2010 [cited 2013 Jan 11];(September):21. Available from: http://works.bepress.com/borgman/238/
5. Piwowar H, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. Ioannidis J, editor. PLoS ONE [Internet]. 2007 Mar 21 [cited 2012 Oct 25];2(3):e308. Available from: http://dx.plos.org/10.1371/journal.pone.0000308
6. The Royal Society. Science as an open enterprise [Internet]. London: The Royal Society; 2012. Available from: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf
7. National Science Board. Digital Research Data Sharing and Management [Internet]. Arlington, VA; 2011. Available from: http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf
8. Data.gov. Open Data Sites [Internet]. Washington (DC): U.S. Department of Health & Human Services; 2012 [cited 2013 Jan 5]. Available from: http://www.data.gov/opendatasites
9. Healthdata.gov. Data Catalog [Internet]. Washington (DC): U.S. Department of Health & Human Services; 2012 [cited 2013 Jan 5]. Available from: http://healthdata.gov/dataset/search
10. DataCite. Repositories List [Internet]. London (UK): DataCite; 2013 [cited 2013 Jan 5]. Available from: http://datacite.org/repolist
11. Massachusetts Institute of Technology Libraries. Data Management and Publishing: Manage Your Data Resource Guide [Internet]. Cambridge (MA): Massachusetts Institute of Technology; 2013. Available from: http://libraries.mit.edu/guides/subjects/data-management/
12. California Digital Library. UC3 Manage Your Data Resource Guide [Internet]. University of California; 2011 Jul 8 [cited 2012 Dec 29]. Available from: http://www.cdlib.org/services/uc3/datamanagement/
13. Simmons College. Open Access Directory Wiki: Data repositories [Internet]. Boston (MA): Simmons College; 2012 Dec 17 [cited 2013 Jan 6]. Available from: http://oad.simmons.edu/oadwiki/Data_repositories

14. University of Minnesota Libraries. Managing Your Data Resource Guide [Internet]. Saint Paul (MN): University of Minnesota; 2011. Available from: https://www.lib.umn.edu/datamanagement
15. University of Washington Libraries. Data Management Guide [Internet]. Seattle (WA): University of Washington; 2013 Jan 3 [cited 2013 Jan 6]. Available from: http://guides.lib.washington.edu/dmg
16. Purdue University Libraries. Distributed Data Curation Center [Internet]. West Lafayette (IN): Purdue University; 2012. Available from: http://d2c2.lib.purdue.edu/
17. HUBzero Foundation. HUBzero: Platform for Scientific Collaboration [Internet]. West Lafayette (IN): 2013. Available from: http://hubzero.org/
18. Cytobank, Inc. Cytobank [Internet]. Mountain View (CA): 2013. Available from: http://www.cytobank.org/index.html?slider1=1
19. Heart Imaging Technologies, LLC. WebPAX [Internet]. 2009. Available from: https://www.webpax.com/

Acknowledgements

I would like to thank Jerry Sheehan for all of his help with this project as my sponsor. His guidance and insight were invaluable for helping me understand data sharing policies and come to conclusions on selection criteria for the NIH data sharing repositories. I would also like to thank the BMIC members who took the time out of their busy schedules to provide their additions and suggestions to the inventories. Their input was greatly appreciated. A final thank you to Mike Huerta for providing advice on valuable avenues to pursue regarding the NIH data sharing repositories – his assistance helped to guide this project in the right direction.

Appendices

Appendix A. User Interface Mockup for BMIC Web Portal

Image 3. BMIC Home Page Mockup

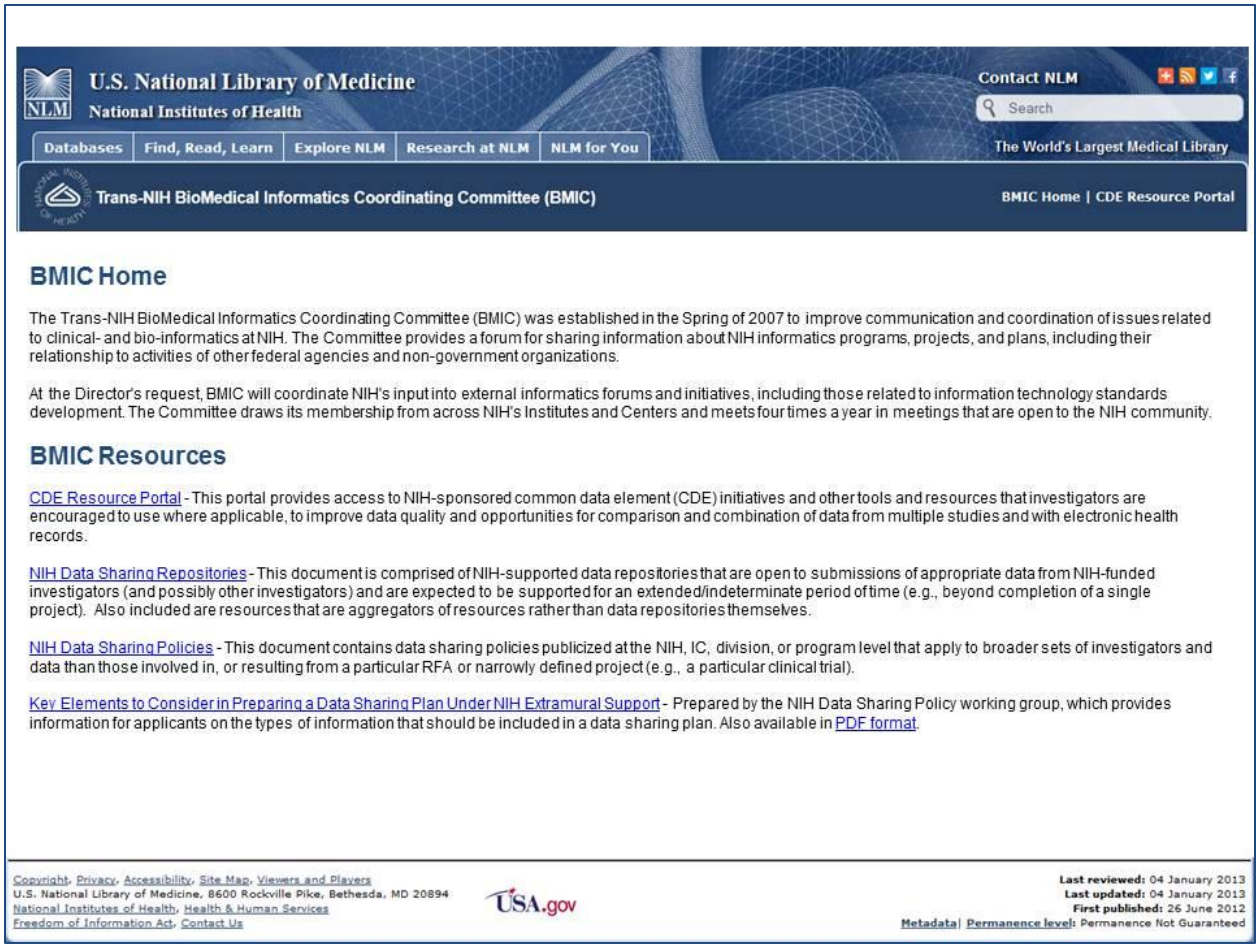


Image 4. NIH Data Sharing Policies Inventory Web Portal

U.S. National Library of Medicine
National Institutes of Health

Contact NLM
Search
The World's Largest Medical Library

Databases | Find, Read, Learn | Explore NLM | Research at NLM | NLM for You

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

NIH Data Sharing Policies

This table contains data sharing policies publicized at the NIH, IC, division, or program level that apply to broader sets of investigators and data. Individual RFAs or PAs may specify other data sharing activities according to their specific project (e.g., a particular clinical trial).

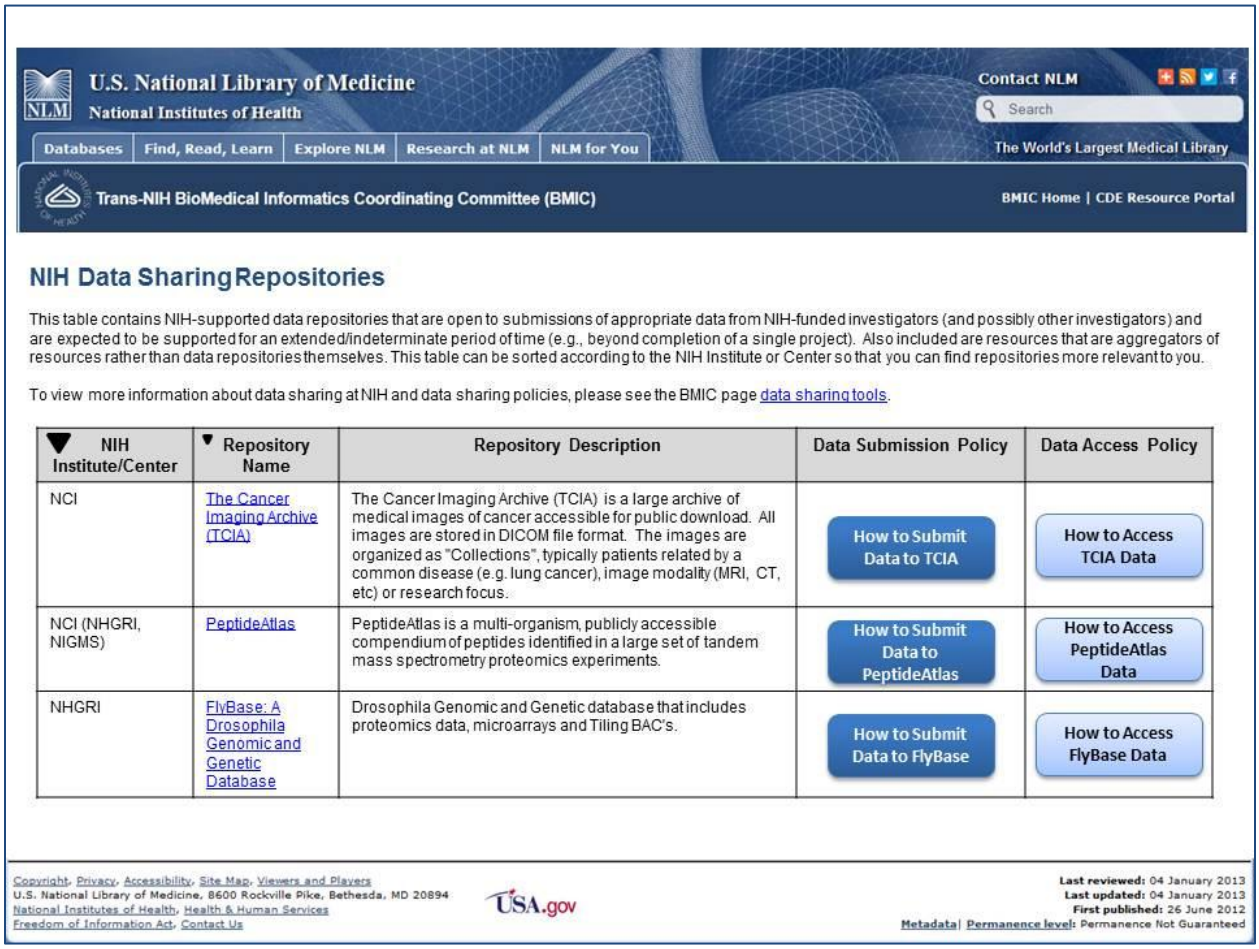
NIH Institute /Center	Data Sharing Policy	Summary of Policy	Associated Repositories
NIH	NIH Data Sharing Policy	Expects investigators seeking more than $500K in direct support in any given year to submit a data sharing plan with their application or to indicate why data sharing is not possible.	None
NIH	NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)	Expects all investigators who receive NIH support to conduct genome-wide analysis of genetic variation in a study population to submit to the NIH GWAS data repository descriptive information about their studies for inclusion in an open access portion of the NIH GWAS data repository. Strongly encourages the submission of curated and coded phenotype, exposure, genotype, and pedigree data, as appropriate, to the NIH GWAS data repository as soon as quality control procedures have been completed at the local institution. These detailed data will be made available through a controlled access process according to the GWAS Data Access procedures.	dbGap
NHGRI	ENCODE Consortia Data Release, Data Use, and Publication Policies	Requires resource producers to release primary data along with an an initial interpretation, in the form of genome features, to the appropriate public databases as soon as the data is verified. Consortia members will also identify validation standards that will be applied in subsequent analyses of the data or with additional experimentation where appropriate.	ENCODE

Copyright, Privacy, Accessibility, Site Map, Viewers and Players
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Health & Human Services
Freedom of Information Act, Contact Us

USA.gov

Last reviewed: 04 January 2013
Last updated: 04 January 2013
First published: 26 June 2012
Metadata | Permanence level: Permanence Not Guaranteed

Image 5. NIH Data Sharing Repositories Inventory Web Portal



U.S. National Library of Medicine
National Institutes of Health

Contact NLM | Search

Databases | Find, Read, Learn | Explore NLM | Research at NLM | NLM for You

The World's Largest Medical Library

Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

NIH Data Sharing Repositories

This table contains NIH-supported data repositories that are open to submissions of appropriate data from NIH-funded investigators (and possibly other investigators) and are expected to be supported for an extended/indeterminate period of time (e.g., beyond completion of a single project). Also included are resources that are aggregators of resources rather than data repositories themselves. This table can be sorted according to the NIH Institute or Center so that you can find repositories more relevant to you.

To view more information about data sharing at NIH and data sharing policies, please see the BMIC page data sharing tools.

NIH Institute/Center	Repository Name	Repository Description	Data Submission Policy	Data Access Policy
NCI	The Cancer Imaging Archive (TCIA)	The Cancer Imaging Archive (TCIA) is a large archive of medical images of cancer accessible for public download. All images are stored in DICOM file format. The images are organized as "Collections", typically patients related by a common disease (e.g. lung cancer), image modality (MRI, CT, etc) or research focus.	How to Submit Data to TCIA	How to Access TCIA Data
NCI (NHGRI, NIGMS)	PeptideAtlas	PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.	How to Submit Data to PeptideAtlas	How to Access PeptideAtlas Data
NHGRI	FlyBase: A Drosophila Genomic and Genetic Database	Drosophila Genomic and Genetic database that includes proteomics data, microarrays and Tiling BAC's.	How to Submit Data to FlyBase	How to Access FlyBase Data

Copyright, Privacy, Accessibility, Site Map, Viewers and Players
U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
National Institutes of Health, Health & Human Services
Freedom of Information Act, Contact Us

USA.gov

Last reviewed: 04 January 2013
Last updated: 04 January 2013
First published: 26 June 2012
Metadata | Permanence level: Permanence Not Guaranteed

Appendix B. NIH Data Sharing Policies Inventory

Please see file supplement: 2012-2013_readkevin_fallproject_datasharinginventories.xlsx

Appendix C. NIH Data Sharing Repositories Inventory

Please see file supplement: 2012-2013_readkevin_fallproject_datasharinginventories.xlsx