United States - Open science country note

Open science and the national context

In February 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum directing federal science agencies that spend more than USD 100 million per year in research and development to develop plans for increasing public access to the results of federally funded research, in particular peer-reviewed scientific publications and digital scientific data. The policy recognises that data provided by the federal government catalyse innovative breakthroughs that drive the economy and provide the basis for progress in areas such as health, energy, the environment, agriculture, and national security. Agency policies are intended to accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation. They are also intended to increase the impact and accountability of federal government research investments, and to allow companies to focus resources and efforts on understanding and exploiting discoveries.

See www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [1].

Open science research and innovation actors

As mentioned above, the OSTP Memorandum applies to federal agencies with more than USD 100 million in annual spending on research and development. This includes agencies within the Department of Agriculture; Department of Commerce; Department of Defense; Department of Energy; Department of Health and Human Services; Environmental Protection Agency; National Aeronautics and Space Administration; and National Science Foundation, but other agencies may voluntarily comply with the directive. The requirements of the memo apply to research conducted by these agencies, or research funded by these agencies and conducted by researchers in universities and other private research organisations. In developing their policies, agencies are instructed to consult with relevant stakeholders, including federally funded researchers, universities, libraries, publishers, users of federally funded research results, and civil society groups. Funding for implementation and operation of policies will come from existing agency budgets. Some agencies, including the National Institutes of Health (NIH), already have policies for increasing public access to publications and digital data that will form the basis of their plans; others will create new policies. Inter-agency working groups aim to improve the consistency of agency policies to the extent practicable.

Open science and business sector actors

Businesses and non-profit organisations play an active role in open science. For data, a growing number of established and startup businesses are developing products, tools and platforms to facilitate data storage, sharing, discovery and analysis. For publications, increased public access builds on the long-standing relationship among federal science agencies, researchers, and publishers in supporting the creation and dissemination of peer-reviewed journal articles. A number of scientific publishers are collaborating with federal science agencies to enhance access to peer-reviewed publications through existing public repositories, such as PubMed Central; others are collaborating to develop new private sector platforms to provide public access to journal articles. Many private and public universities and research libraries have established institutional repositories for publications and data resulting from their research, and are exploring ways to collaborate in support of public access beyond their existing communities. Several private foundations that fund research, including the Howard Hughes Medical Institute, Autism Speaks, and the Health Research Alliance have also

established public access policies that apply to their funded research and require the resulting publications to be made publicly accessible. The OSTP memo explicitly encourages collaboration with the private sector, including through public-private partnerships, to increase public access to federally funded scientific research results.

Policy design - Open data

The OSTP Memo specifies that agency plans should maximise public access without charge to digital scientific data created with federal funds, while protecting confidentiality and personal privacy, recognising proprietary interests and intellectual property rights, and preserving the balance between the relative costs and benefits of long-term preservation and access. The OSTP memo specifies additional objectives for agency plans:

· Ensure that all federally funded researchers (intramural and extramural) develop data management plans describing how they will provide for the long-term preservation of, and access to, digital scientific data resulting from federally funded research, or explaining why long-term preservation and access cannot be justified.

· Ensure appropriate evaluation of the merits of the data management plans submitted.

· Include mechanisms to ensure compliance with data management plans and policies.

· Allow the inclusion of appropriate costs for data management and access in proposals for federal funding for scientific research.

· Promote the deposit of data in publicly accessible databases, where appropriate and available.

· Encourage co-operation with the private sector to improve data access and compatibility, including through public-private partnerships.

· Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan.

· Co-ordinate with other agencies and the private sector to support training, education, and workforce development related to scientific data management.

· Provide for the assessment of long-term needs for the preservation of scientific data, and outline options for developing and sustaining repositories for digital scientific data, taking into account the efforts of public and private sector entities.

Two specific examples are provided of data-sharing policies of the National Institutes of Health. A more complete listing of NIH data-sharing policies (for different types of data) is available at (www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html [2]).

1) NIH Data Sharing Policy (http://grants.nih.gov/grants/policy/data_sharing/ [3])

The 2003 NIH Data Sharing Policy expects NIH-funded researchers requesting funding of USD 500 000 or more in direct costs in any year to include in their application a data-sharing plan that describes how they will share their final research data for use by other researchers in a timely way (i.e. no later than acceptance of the main findings from the final data set for publication). NIH plans to revise this policy to bring it into line with the objectives outlined in the OSTP memo.

2) NIH Genomic Data Sharing Policy

(http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html [4])

The National Institutes of Health (NIH) Genomic Data Sharing (GDS) Policy, issued 27 August 2014, is an expansion of the 2007 NIH policy for genome-wide association studies (GWAS), and sets forth expectations to ensure the broad and responsible sharing of genomic research data. The GDS Policy applies to all NIH-funded research (grants, contracts and intramural research) that generates large-scale human or non-human genomic data, as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies, single nucleotide polymorphism (SNP) arrays, and genome sequence, transcriptomic, metagenomic, epigenomic, and gene expression data. Under the GDS Policy, investigators should register all studies with human genomic data in the database of Genotypes and Phenotypes (dbGaP), maintained and operated by the National Library of Medicine, by the time data cleaning and quality control measures begin. Data are to be submitted to the appropriate NIH-designated data repository, e.g. dbGaP, Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), and Cancer Genomics Hub. For non-human genomic data, current databases remain the standard mechanism for sharing (e.g. GEO, SRA, WormBase, GenBank). The effective date of the GDS Policy is 25 January 2015.

· *Carrots* – NIH provides funding for the projects to which the data-sharing requirement applies. In addition, the databases in which applicable studies are registered and to which resulting data are submitted assign unique study accession numbers, so that they may be properly cited. Investigators who use controlled-access human data for secondary research are expected to acknowledge in all oral or written presentations, disclosures or publications the specific data set(s) or applicable accession number(s) and the NIH-designated data repositories through which the investigator accessed any data.

· *Sticks* – Compliance with the GDS Policy will become a special term and condition in the Notice of Award or the Contract Award. Failure to comply could lead to enforcement actions, including the withholding of funding. Additionally, NIH will take appropriate action if there are violations regarding the use of controlled-access data.

· *Enablers* – NIH has established and continues to support several databases for archiving genomic data, and has established a governance structure for the oversight of the GDS Policy. Data access committees review requests for secondary use of controlled-access human data for conformity with data use limitations based on the informed consent of study participants. NIH has also invested in the development of standards and consensus measures for phenotype and exposure data that are collected in genome-wide association studies, e.g. the PhenX measures (www.phenx.org [5]).

The policy was developed and implemented by the National Institutes of Health, with considerable input from the relevant stakeholder communities via a public comment process, webinars, and other mechanisms.

The policy applies to all NIH-funded investigators. These include NIH intramural scientists, as well as the larger number of extramural scientists who receive NIH funding for research conducted in universities, academic medical centres, and other research organisations in the United States and abroad.

Sharing research data supports the NIH mission, and is essential to facilitate the translation of research results into knowledge, products and procedures that improve human health. This policy aims to make genomic data sets accessible for secondary research purposes, to accelerate scientific discovery, and to improve the return on the NIH investment in science.

Respect for and protection of the interests of research participants are fundamental to NIH's stewardship of human genomic data. The informed consent under which the data or samples were collected is the basis for the submitting institution and Institutional Review Board (IRB), to assure the appropriateness of data submission to NIH-designated data repositories and to determine whether the data should be available through unrestricted or controlled access. Requests for controlled-access data are reviewed by NIH Data Access Committees (DACs), whose decisions are based primarily on conformity of the proposed research with the data use limitations established by the submitting institution through the Institutional Certification. Investigators who receive approval to

download controlled-access data from NIH-designated data repositories and their institutions are expected to abide by the NIH Genomic Data User Code of Conduct (http://gds.nih.gov/pdf/Genomic_Data_User_Code_of_Conduct.pdf [6]). The Code includes provisions to protect against attempts to identify individual human research participants from whom the data were obtained; and to promote acknowledgement in all oral or written presentations, disclosures, or publications the specific data set(s) or applicable accession number(s) and the NIH-designated data repositories through which the investigator accessed any data.

NIH encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs without impeding research. However, naturally occurring DNA sequences are not patentable in the United States. Therefore, basic sequence data and certain related information (e.g. genotypes, haplotypes, p-values, allele frequencies) are pre-competitive. Such data made available through NIH-designated repositories, and all conclusions derived directly from them, should remain freely available, without any licensing requirements. NIH encourages broad use of NIH-funded genomic data that is consistent with a responsible approach to management of intellectual property derived from downstream discoveries, as outlined in the NIH Best Practices for the Licensing of Genomic Inventions, and Section 8.2.3, Sharing Research Resources, of the NIH Grants Policy Statement. NIH discourages the use of patents to prevent the use of or to block access to genomic or genotype-phenotype data developed with NIH support.

NIH experience with genome-wide association data from dbGaP under the NIH GWAS Data Sharing Policy is illustrative. As of 1 December 2013, NIH had received almost 17 750 requests from more than 2 000 investigators (and almost 7 000 collaborators) for access to dbGaP datasets. Sixty-nine per cent of these requests have been approved. Secondary use of these data resulted in over 900 publications, many in top-tier journals, and resulted in significant discoveries in a wide range of fields. See National Institutes of Health Genomic Data Sharing Governance Committees (2014), “Data use under the NIH GWAS Data Sharing Policy and future directions”, *Nature Genetics*, Vol. 46, pp. 934-938, 27 Aug. 2014. www.nature.com/ng/journal/v46/n9/full/ng.3062.html [7]>.

Policy design - Open/increasing access to scientific publications

The OSTP Memo establishes the following principles for agency policies to increase access to scientific publications resulting from scientific research funded by the federal government.

To the extent feasible – and consistent with law, agency mission, resource constraints, US national, homeland, and economic security, and the objectives listed below – the results of unclassified research that are published in peer-reviewed publications directly arising from federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyse in ways that maximise the impact and accountability of the federal research investment.

In developing their public access plans, agencies shall seek to put in place policies that enhance innovation and competitiveness by maximising the potential to create new business opportunities and that are otherwise consistent with the principles articulated in Section 1 above.

Agency plans must also describe, to the extent feasible, procedures the agency will take to help prevent the unauthorised mass redistribution of scholarly publications.

Further, each agency plan shall:

· Ensure that the public can read, download, and analyse in digital form final peer-reviewed manuscripts or final published documents within a time frame that is appropriate for each type of research conducted or sponsored by the agency. Specifically, each agency: i) shall use a twelve-month post-publication embargo period as a guideline for making research papers publicly available – however, an agency may tailor its plan as necessary to address the objectives articulated in this memorandum, as well as the challenges and public interests that are unique to each field and

mission combination; and ii) shall also provide a mechanism for stakeholders to petition for changing the embargo period for a specific field by presenting evidence demonstrating that the plan would be inconsistent with the objectives articulated in this memorandum.

· Facilitate easy public search for, analysis of, and access to peer-reviewed scholarly publications directly arising from research funded by the federal government.

· Ensure full public access to publications' metadata without charge upon first publication in a data format that ensures interoperability with current and future search technology. Where possible, the metadata should provide a link to the location where the full text and associated supplemental materials will be made available after the embargo period.

· Encourage public-private collaboration to: i) maximise the potential for interoperability between public and private platforms and for creative reuse to enhance value to all stakeholders; ii) avoid unnecessary duplication of existing mechanisms; iii) maximise the impact of the federal research investment; and iv) otherwise assist with implementation of the agency plan.

· Ensure that attribution to authors, journals, and original publishers is maintained.

· Ensure that publications and metadata are stored in an archival solution that: i) provides for long-term preservation and access to the content without charge; ii) uses standard, widely available and (to the extent possible) non-proprietary archival formats for text and associated content (e.g. images, video, supporting data); iii) provides access for persons with disabilities consistent with Section 508 of the Rehabilitation Act of 1973,1**;** iv) enables integration and interoperability with other federal public access archival solutions and other appropriate archives.

Repositories may be maintained by the federal agency funding the research; through an arrangement with other federal agencies; or through other parties working in partnership with the agency, including but not limited to scholarly and professional associations, publishers and libraries.

Additionally, US law requires the National Institutes of Health – as well as its parent agency, the Department of Health and Human Services – the Department of Education and the Department of Labor to have public access policies for peer-reviewed scholarly publications resulting from funded research.

· Section 217 of Public Law 111-8 (Omnibus Appropriations Act, 2009) states:

The Director of the National Institutes of Health ("NIH") shall require in the current fiscal year and thereafter that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication: Provided, that the NIH shall implement the public access policy in a manner consistent with copyright law.

· Section 527 of Division H – Public Law 113-76 (Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriations Act of 2014) states:

Each Federal agency, or in the case of an agency with multiple bureaus, (or operating division) funded under this Act that has research and development expenditures in excess of $100,000,000 per year shall develop a Federal research public access policy that provides for: (1) the submission to the agency, agency bureau or designated entity acting on behalf of the agency, a machine-readable version of the author's final peer-reviewed manuscripts that have been accepted for publication in peer-reviewed journals describing research supported, in whole or in part, from funding by the Federal Government; (2) free online public access to such final peer-reviewed manuscripts or published versions not later than 12 months after the official date of publication; and (3) compliance with all relevant copyright laws.

As an example, the NIH Public Access Policy is described below.

Under the NIH Public Access Policy (http://publicaccess.nih.gov/ [8]), the Director of the National Institutes of Health requires that all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine's PubMed Central repository an electronic version of their final, peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication.

· *Carrots* – NIH's PubMed Central repository is crawled regularly by major search engines, making deposited papers more easily discoverable. With more than 3.2 million full-text articles (as of August 2014), it is also a major destination site for users interested in biomedical research. On a typical weekday, more than 1 million different users download more than 1.65 million different articles from PubMed Central. Deposited papers are also indexed in PubMed, one of the most highly used databases for biomedical researchers, clinicians and the public.

· *Sticks* – The requirement to deposit publications into PubMed Central is a term and condition of award for NIH grantees and contractors. In July 2013, after five years of education, outreach and administrative reminders, NIH began to delay processing of annual award renewals until publications arising from those awards are brought into compliance with the public access policy.

· *Enablers* – NIH has developed a repository for publications that are subject to the policy (PubMed Central) and a bibliography management system (My Bibliography) to facilitate tracking and reporting of papers arising from NIH funds, and the development of a personal curriculum vitae (Science Experts Network Curriculum Vitae or ScENcv, http://rbm.nih.gov/profile_project.htm/home.shtml/ [9]). NIH also engages in active outreach and communication with universities and research organisations to make them aware of the policy and to offer information around mechanisms for compliance. NIH has also established agreements with more than 1 600 journals that deposit final published articles on behalf of NIH-funded investigators. Other large commercial publishers also submit manuscripts on behalf of NIH-funded investigators.

The cost of implementing the NIH Public Access Policy ranges between USD 4 million and USD 4.5 million per year, depending on the number of articles submitted to PubMed Central. These costs represent a tiny fraction of the annual budget of NIH of approximately USD 30 billion, and helps expand the reach and potential impact of its funded research results.

For the NIH Public Access Policy, the institution involved is NIH – but many different parts of NIH are actively engaged, including the Office of Extramural Research (which establishes policy for NIH-funded investigators) and the National Library of Medicine (which operates the PubMed Central repository). The policy affects funded investigators in research institutions across the country – and around the world. It was developed with considerable input from the stakeholder communities, including public meetings, requests for comment, and input from advisory boards.

The NIH Public Access Policy applies to NIH-funded investigators. The policy is intended to benefit researchers, care providers, entrepreneurs, and members of the general public who seek to learn and apply NIH research findings.

The policy is intended to archive and provide enhanced access to the results of NIH-funded research, and to therefore accelerate research and its translation into practice.

The NIH Public Access Policy is implemented in a manner consistent with copyright. Funded investigators are instructed to retain the right to deposit their manuscripts in PubMed Central when submitting papers for publication. For papers deposited by publishers, the publisher retains the copyright and papers are available in accordance with those rights (e.g. open access publications are subject to fewer restrictions than those from subscription journals).

Between 2008 and 2013, NIH funding is estimated to have helped generate approximately half a million peer-reviewed articles, more than 84% of which have been deposited in PubMed Central.

The NIH Public Access Policy is monitored and enforced in two primary ways. Investigators are required to indicate in annual progress reports any publications that have resulted from funded research, and they must provide the unique identifier assigned by PubMed Central. In addition, the

National Library of Medicine, through its regular abstracting and indexing activities, can identify publications that acknowledge NIH as the funder of research and compare against papers deposited in PubMed Central.

The NIH Public Access Policy does not differentiate between free and *libre* open access model. It supports different licences and is compatible with all publisher business models. The only requirement is that papers be accessible without charge on PubMed Central no later than 12 months after publication. Funded investigators may publish in the journal of their choice, whether open access (author pays) or subscription based (subscriber pays). NIH has clarified that grant funding may be used to pay authors' processing charges and/or other costs of publishing.

Skills for open science and open data

The OSTP Memo directs federal science agencies to co-ordinate with other agencies and the private sector to support training, education and workforce development related to scientific data management, analysis, storage, preservation and stewardship.

Within the NIH, skills development is one of four major focuses of the Big Data To Knowledge (BD2K) initiative. The goal is to develop a sufficient cadre of researchers skilled in the science of big data, in addition to elevating general competencies in data usage and analysis across the biomedical research workforce. Among other activities to achieve this goal, NIH is accepting applications for funding related to projects that aim at: 1) career development for clinicians and Ph.D.-level researchers; 2) courses for skills development in biomedical big data; and 3) open educational resources for biomedical big data.

See http://bd2k.nih.gov/funding_opportunities.html#sthash.CHNNZhBp.dpbs [10]. A number of NIH institutes and centres already offer funding for training in biomedical informatics (e.g. support to Ph.D.-level training programmes in biomedical informatics researchers and short courses for clinicians/physicians interested in biomedical informatics).

Open science and international co-operation

As relates to the NIH Public Access Policy, NIH co-operates with research funders in several other countries/regions. PubMed Central International (PMCI) is a collaborative effort between NIH and organisations in other countries with public access policies. PMC Europe is a host archive for a score of European funders (for a list, see http://europepmc.org/Funders/ [11]). PMC Canada is the host archive for the Canadian Institutes of Health research (http://pubmedcentralcanada.ca/ [12]). These efforts allow for integration of content and reciprocity in public access policies across funders. The long-term goal of PMCI is to create a network of digital archives that can share all of their respective locally deposited content with others in the network.

Other information

The NIH Big Data To Knowledge (BD2K) initiative was launched to enable biomedical scientists to capitalise more fully on the big data being generated by those research communities. The initiative aims to develop the new approaches, standards, methods, tools, software and competencies that will enhance the use of biomedical big data by supporting research, implementation and training in data science and other relevant fields. This will lead to:

· appropriate access to shareable biomedical data through technologies, approaches and

policies that enable and facilitate widespread data sharing, discoverability, management, curation, and meaningful reuse

· development of and access to appropriate algorithms, methods, software and tools for all aspects of the use of big data, including data processing, storage, analysis, integration and visualisation

· appropriate protections for privacy and intellectual property

· development of a sufficient cadre of researchers skilled in the science of Big Data, in addition to elevating general competencies in data usage and analysis across the behavioural research workforce.

Overall, the focus of the BD2K initiative is the development of innovative and transforming approaches as well as tools for making big data and data science a more prominent component of biomedical research (http://bd2k.nih.gov [13]).

Links
[1] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[2] http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
[3] http://grants.nih.gov/grants/policy/data_sharing/
[4] http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html
[5] http://www.phenx.org/
[6] http://gds.nih.gov/pdf/Genomic_Data_User_Code_of_Conduct.pdf
[7] http://www.nature.com/ng/journal/v46/n9/full/ng.3062.html
[8] http://publicaccess.nih.gov/
[9] http://rbm.nih.gov/profile_project.htm/home.shtml/
[10] http://bd2k.nih.gov/funding_opportunities.html#sthash.CHNNZhBp.dpbs
[11] http://europepmc.org/Funders/
[12] http://pubmedcentralcanada.ca/
[13] http://bd2k.nih.gov/