Thu 1/12/2012 9:08 PM
RFI response

Public interests in data from federally funded research

A response to the Request for Information on Public Access to Digital Data Resulting from Federally Funded Research.

John Hawks, Ph.D.
Department of Anthropology
University of Wisconsin-Madison

Introduction

The United States provides grant funding to scientists through many federal programs. This funding advances work of public interest that might not happen without federal assistance.

The creation of scientific knowledge may serve the public interest directly by enabling useful inventions or supplying actionable information on issues of public importance. A funded project may also serve the public interest indirectly, by (1) finding negative results that prevent wasted effort or public harm; (2) building the scientific infrastructure that enables future discoveries and advances; (3) training new and established scientists in effective research techniques; (4) enhancing international cooperation and public/private partnerships.

Congress and the Executive Branch have recognized that access to the published results of scientific research is not sufficient to advance the direct and indirect public interests served by federally funded projects. Facilitating the indirect benefits of research is a major aim of federal agencies' "Broader Impacts" and data access rules. These policies have been a qualified success since their implementation, limited mainly by the exceptions carved out by programs and agencies to avoid requiring certain kinds of data to be reported along with research reports.

I argue that open public access to digital data should be a requirement for all federally funded scientific research. Digital data can be maintained by federal agencies as a part of the reporting requirement of federal grant funding. Doing so will advance the interest of the public and ensure that today's science generates a continuing heritage of research excellence.

Data access and transparency

Transparency is essential to public trust. Scientific conclusions are formed by observation and replication, and for this process to be transparent, all data must be available for independent inspection. The possibility of such inspection should not be limited to qualified researchers, because the very existence of special access requirements blocks transparency of the scientific process.

Changing technology has shifted the public's expectations about transparency. Digital technology enables most research data to be shared rapidly and at low cost. If data are produced in digital form, and digital data can be shared at low cost, researchers and agencies cannot credibly claim that the difficulty of reproducing and disseminating data is a sufficient reason to restrict access. Where no competing interest argues for restricted access (such as human subjects protections), a lack of access to digital data itself can now be a compelling reason for public distrust.

Therefore, federally funded researchers should release digital data to the public by default. Federal agencies should facilitate this public reporting by requiring digital data to be supplied as part of final project reporting.

Data access has a well-established record of success

The recent history of human genetics demonstrates that open access to data has unforeseen benefits that can spawn innovation, support more effective education, and catalyze new discovery. In genetics, both federal and journal policies require release of data; raw data from federally funded projects are often available as they are generated, long before publication.

My own laboratory has no federal research funding to date, but is actively engaged in research using data from federally funded projects. Today my laboratory trains undergraduate students in genetics with new data from ongoing federally funded genetic projects such as the 1000 Genomes Project. We use open access data from archaic human genomes to investigate the variation of ancient people and their relationships to living humans. This kind of work would be impractical without clearly established open data access policy.

The open access to data from the Human Genome Project facilitated the rapid development of microarrays that are now used on a broad scale in human genetics to investigate the genetic correlates of human health and disease. Access to data from these studies has enabled other scientists to independently replicate many genetic associations. More important, meta-analysis of such data has shown that many associations cannot be replicated, while also showing some cases in which nonsignificant results across different samples give rise to a significant finding when pooling those samples. Access to negative results and raw data is necessary, in other words, to establish the facts in subsequent research. This goes beyond access to published research results and requires open access to unpublished digital data.

Intellectual property protections and data access

Research data are somewhat distinct from the intellectual property issues relating to research publications. Some kinds of data do not meet the standard of originality necessary for copyright protection, such as sequence data, CT or MRI data, or data from measurement instruments. For raw data from instruments, there is no intellectual property reason why federal agency should not maintain an open archive for the public.

Much research data is unquestionably subject to copyright protection, such as lab notebooks, written descriptions, photographs, and original reconstructions. Yet there is still a substantial

public and scientific interest in inspecting such data. For example, photographic documentation of archaeological sites and specimens are of particular scientific value and are today routinely produced by digital technologies and stored in digital form. Some primary digital records are unique products that cannot be recreated at another time and place: for example, in situ photographs of specimens, photographs and records of sites before excavation, and digital reconstructions. The scientific record would be incomplete without such contributions, and maintaining an archive of such data over the long term is a difficult task for a single investigator, beyond the scope of a grant term.

In cases where it is impracticable to obtain Creative Commons or other open licenses to such content, a funding agency should at a minimum require that a copy of all such archival information be deposited along with the final project report and a limited-use non-commercial license permitting electronic dissemination of these materials to the public as part of the report.

Metadata and data access

Many have noted that raw data may be useless in the absence of additional information about how the data were obtained. Such information is known as "metadata". Researchers generate instrumental data using particular instrument settings and recording standards. They gather observational data under particular research protocols. These standards are may change quickly as instrumentation, technology, and scientific results themselves demand new practices.

Some scientists note the problem of incompatible metadata, using it as an argument against to delay the establishment of open public access to data. In their view, the public are likely to misunderstand or misuse scientific data where metadata are not clearly indicated. Meta-analyses combining data from multiple research projects are an important secondary use of digital data, and such meta-analyses are impossible when data cannot be reconciled into common observational or instrumental frameworks. Performing original work with data collected in heterogeneous contexts is a research speciality of its own, and is itself sometimes targeted by federal grants.

However, meta-analysis is only one purpose of data access. Transparency, replicability, and education are central public interests that do not require the reconciliation of data collection methods from multiple studies. They require only clear description of the methods under which data were obtained. At a minimum, final research reports on federally funded projects must describe the standards of data collection with sufficient detail to allow independent replication, including all unpublished results and data.

Data access in paleoanthropology

I am an anthropologist, and am most familiar with the scientific data relating to human evolution. These data include genetic observations on living and skeletal samples of humans. They also include fossil and archaeological evidence such as photographs, CT scans, isotopic records, anatomical measurements and descriptions.

Successes of data access in paleoanthropology

For many years, nearly all genetic data resulting from federally funded research have been made available for public download. Much genetic data generated by non-federally funded research programs, including foreign and domestic institutes, has also been free for public download. These data have resulted in a massive acceleration of research on recent human evolution and human origins. They have also led to unexpected discoveries and a burgeoning contribution of other disciplines to understanding our evolution.

Data from radiocarbon dating and other isotopic sampling has also been made available to the public. Human occupation sites are among the best sources of evidence about past climates. The investment of federal resources in human evolution research has generated a temporal record that is now essential to studying changes in the faunal and plant compositions of past environments. Free access to records has enabled stronger calibration of radiocarbon dates, the development of a more secure chronology, and a more highly replicable scientific record correlating different regions of the world. Our understanding of such events changes is vastly stronger when data are made public.

Institutions and data access in paleoanthropology

By contrast, CT scans and photographs pertaining to human origins are typically not made accessible by the public. The United States funding agencies are not the only parties with an interest in such data. In particular, museums and institutes that curate specimens often permit data collection under agreements that restrict the dissemination of the resulting data. Such agreements may be equated to "non-disclosure agreements" with respect to scientific data.

An institution has a legitimate interest in controlling the public use of images and access to curated materials. Nevertheless, the lack of access to digital data results in reduplication of effort, overapplication of destructive sampling and measurement techniques, and unnecessary handling of precious and fragile specimens. Where it is practical, the United States should facilitate agreements with institutions that allow the release of digital data produced by public funding. Where release is not possible, funding should be granted only for those activities that will result in the release of data under a limited-use non-commercial license. Non-disclosure of data from instruments such as CT scanners, electron microscopes, mass spectrometers is incompatible with scientific replication.

Scientific careers and data access in paleoanthropology

The economy of federal funding for scientific production sometimes leads to perverse incentives for high-ranking researchers that prevent public access to research data. Some scientists believe that their own future research will require exclusive access to data. Others want to impede research achievements by their academic rivals, or to maintain prestige and future funding opportunities.

Scientific data in some areas may constitute "trade secrets" until they are protected by patents. Even in noncommercial research, federally funded scientists sometimes claim exclusive

ownership over data that they plan to use in future research. In my own field of paleoanthropology, data secrecy supports a clandestine "quid pro quo" economy among researchers, in which established researchers and institutions allow furtive looks at unpublished data, to support and consolidate their power and influence.

This is a game that the United States should simply decline to play. When federal research supports scientific results that are not subject to independent replication, it betrays the public interest in science.

Established collaborations and centers of scientific research will always exert a strong influence the future of science irrespective of federal data access policies. But established players should not use federal funding to construct barriers to open inquiry.

Conclusion

Open public access to data is one indication that a research project is following scientific principles. Making digital data available to the public would be good practice for any researcher, irrespective of funding source. Data access mitigates the risk that negative data will be unreported. Data access facilitates broader stewardship of research projects, in particular where collaborations create data that are distributed across many institutions. Data access and reporting standards enable other researchers to fill in for those who cannot complete scientific project due to health or other personal reasons.

Federal grant agencies already have successful repositories for many kinds of digital data. Such data are shared with the public at minimal cost relative to the overall budget for federal research grants. Supporting digital data repositories has itself been an important granting aim for several federal agencies and continues to be an active part of scientific infrastructure. Limiting such repositories for the exclusive use of a small cadre of researchers is enormously wasteful of resources, when they can be opened to an interested public for a small incremental cost.

The public has repeatedly invented surprising uses for digital data that can complement or enhance the scientific record. But much more important, open access to digital data serves the scientific values of transparency and independent replication, essential to maintaining public trust and investment in the research enterprise.

John Hawks
Department of Anthropology
University of Wisconsin-Madison
http://johnhawks.net/weblog