ACCESS TO SCIENTIFIC DATA IN THE 21ST CENTURY: RATIONALE AND ILLUSTRATIVE USAGE RIGHTS REVIEW

James Campbell

Spatial Informatics Program, School of Computing and Information Science, University of Maine, Orono, Maine, U.S.A.
Email: campbell@spatial.maine.edu

ABSTRACT

Making scientific data openly accessible and available for re-use is desirable to encourage validation of research results and/or economic development. Understanding what users may, or may not, do with data in online data repositories is key to maximizing the benefits of scientific data re-use. Many online repositories that allow access to scientific data indicate that data is "open," yet specific usage conditions reviewed on 40 "open" sites suggest that there is no agreed upon understanding of what "open" means with respect to data. This inconsistency can be an impediment to data re-use by researchers and the public.

Keywords: Data access, Data copyright, Database copyright, Open access, Data licenses, Database licenses, Usage rights, Public domain

1 INTRODUCTION

Data has been, and remains, the lifeblood of science. For nearly 400 years, the scientific method has depended on access to data to move knowledge and society forward. That tradition stalled to some degree in the second half of the 20th century and the beginning of the 21st. For a variety of economic, legal, and, to some extent, professional reasons, access to scientific data today is not nearly as open as many wish. That situation has been changing in recent years due to a variety of societal and scientific forces, yet obstacles to open access to scientific data still exist, especially in the area of clearly delineated legal rights and restrictions.

This paper reviews some of the forces pushing toward more open access to scientific data in the 21st century. The focus is primarily, though not exclusively, on publicly funded, geospatially-related data in a U.S. context although in today's connected world, data access often transcends political borders, especially in disciplinary contexts. This examination looks at usage policies of a selection of data repositories that are attempting to make scientific data more accessible to determine whether usage policies are clearly understandable and consistent among repositories.

2 THE ROLE OF DATA IN 21ST CENTURY SCIENCE

More than one scientist has used the metaphor of "drinking from a fire hose" to describe the huge amount of scientific data already being generated by large scale data collectors. That "hose" will only get larger as huge data generators such as the Large Synoptic Survey Telescope and the Large Hadron Collider at CERN collect more and more data. Yet "Small Science" projects are an even more important factor in the exponential growth of scientific data generated today, possibly generating two to three times as much data as "Big Science" (Carlson, 2006).

Wireless sensors, increased computing power, higher bandwidth communication, and other increasingly affordable technologies, to say nothing of the increase in the number of researchers around the world, are giving birth to data streams unthinkable even a decade ago. Data mining and analysis are increasingly important in 21st century scientific discovery, so much so that one pop-science observer penned an article entitled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" (Anderson, 2008).

While this may seem an extreme characterization, data mining, database analysis, and other data manipulation tools and processes are now central to the enterprise of science and to new discoveries. Some researchers are even developing algorithmic processes for machine identification of natural laws from data sets without any attempt to

"teach" the machines before the analysis process begins (Anthes, 2009). While this may not signal the end of theory as Anderson postulated, it certainly adds a new method to scientific discovery.

In looking at a specific subset of scientific data, geospatial data, Lance McKee of the Open Geospatial Consortium has listed "Seventeen reasons why geospatial research data should be published online using OGC standard interfaces and ISO standard metadata." Among those reasons were an assertion, based on an analog to network theory first popularly stated in Metcalf's law, that "The value of data increases with the number of potential users" and an observation that "Data are not efficiently discovered through literature searches" (McKee, 2010).

In the U.S., the National Science Foundation has funded the DataNet Federation Consortium, one among an increasing number of efforts to create an infrastructure that will maximize the utility of data to scientists and researchers. Stan Ahalt, one of the team working on the project, in describing that effort asserted that "Data is the currency of the knowledge economy... [By building infrastructure] We'll be more efficient at producing new science, new innovation and new innovation knowledge" (Tuutti 2011).

3 REASONS FOR CALLS FOR OPEN ACCESS TO SCIENTIFIC DATA

Over the past fifteen years, there has been an increasing number of position papers and studies calling for open access to scientific data from governments, professional and academic organizations, citizen groups, and industry. The rationale driving these calls range from adhering to the traditional mores of science to stimulating economic growth to asserting access to scientific data should be considered a basic human right.

Governments and government organizations, e.g., The National Science Foundation, the National Research Council in the U.S., the European Commission and the Royal Society in Europe, have called for better access to scientific data as a means to spur innovation and economic growth because they realize that data generated by governments and made freely available for re-use can have a significant impact on economic activity. In the U.S., for example, at least 500 companies have been identified as building new businesses on freely available data generated by the U.S. Federal Government (GovLab, 2014). One of those companies began in 2004 using openly available NOAA data and sold for a billion dollars a decade later (Kash, 2014).

While the economic benefits of open access are clearly important, in this review we focus on the scientific and, to a lesser extent, social rationales for open access to scientific data.

3.1 Traditional functions: experiment replication and validation

Traditional science is built on replication and validation. Without access to the original data that a scientific conclusion is based on, it is almost impossible to perform that replication. In an age when data are increasingly the starting point for discovery, access to data becomes even more essential for carrying out the traditional process of science.

To enable access, storage and retrieval are essential: so is knowing what can be done with the data once they are discovered. Confusion over intellectual property rights, or outright refusal to provide access to data, is more common in science than many imagine. In a 2006 AAAS survey of academic and industry bioscience researchers, 35% of academic and 76% of industrial researchers said that their research had been adversely affected by intellectual property restrictions of one type or another. The same survey indicated that even obtaining publicly funded data often presented difficulties. Twenty-four percent of respondents who indicated they had tried to obtain data from publicly funded sources reported difficulty in obtaining such data, and this was especially true in the fields of engineering, math, and computer sciences. Seventy percent of those who had difficulty obtaining data reported it had "some negative effects" on their research, and 10% experienced "serious negative effect." Perhaps even more distressing, 16% of those denied access to data from publicly funded sources were denied access to data for which results had already been published, and 44% received no reason for the denial of access (Agres, 2006).

Reports such as this one have been one impetus for the introduction of legislation in the U.S. that would make published articles in peer reviewed journals based on research funded in whole or in part by the federal government

freely available after an embargo period. The Federal Research Public Access Act of 2010 was one early example. The Fair Access to Science and Technology Research Act of 2013 introduced in the 113th Congress (2013-2014) is the most recent example. This bill would make journal articles freely available six months after publication.

The Frontiers in Innovation, Research, Science, and Technology Act of 2014 (FIRST Act) would extend that hold period to 24 months with a possible additional 12 month embargo, a bill more to the liking of publishers of scientific journals. Interestingly, the FIRST bill provides that, unlike the published article itself which may be embargoed for 24 months, "in the case of data used to support the findings and conclusions of such article, not later than 60 days after the article is published in a peer-reviewed publication." Journal publishers widely supported the Research Works Act (HR 3699 in 112th Congress), which would have prohibited open access mandates altogether.

None of these bills have passed in the Congress. However, a provision in the Consolidated Appropriations Act of 2014 requires federal agencies in Labor, Health and Human Services, and Education with research budgets of over $100 million to provide public access within 12 months of publication in a peer-reviewed journal to research resulting from projects they fund. While these requirements do not specifically refer to data *per se*, an increasing number of publishers are endeavoring to include data as part of the publication process.

For example, publishers such as The International Association of Scientific, Technical and Medical Publishers; The Association of Learned and Professional Society Publishers; the Public Library of Science as well as individual journals, e.g., *Nature*, *The American Naturalist*, *Evolution*, the *Journal of Evolutionary Biology*, *Molecular Ecology*, *Heredity*, have all established policies requiring that data that are the basis of articles must be made publicly accessible as part of the publication process.

Connecting underlying data sets to articles in which they appear is not a trivial undertaking. Organizations such as NISO/NFAIS (2013) in the U.S. and the Digital Curation Centre in the UK (Ball & Duke, 2011) have issued standards for citing and connecting data sets to the articles in which they appeared so that the data is findable and permanently linked to published journal articles.

The National Science Foundation has made inclusion of a Data Management Plan (DMP) that indicates where data is located and how it can be shared a required part of research grants it funds (National Science Foundation, 2011). The University of California has created a web site with "easy-to-use" tools to develop those required DMP plans (University of California 2014).

In short, the traditional functioning and, in fact, the traditional *mores* of science since the Enlightenment require the ability to find data, to access them, and to be able to use them to both verify scientific claims and to extend discovery. Funding agencies and publishers alike are beginning to take steps to ensure that data discovery and access are possible.

3.2 Avoidance of duplication

In an era of tight research funding and limited resources, an important reason to make scientific data available for widespread use is the wasted cost of duplication of effort, particularly when it occurs simply because researchers do not know what other work has been undertaken if data are not openly accessible. Mounting expensive expeditions to places such as Antarctica to gather what turns out to be essentially duplicative data are obvious examples of expensive and avoidable duplications of effort.

In short, reproducibility of experiments for the purpose of validation is essential to the practice of science. The practice of duplicating efforts, however, is wasteful science, and timely access to data can help to reduce such wasteful activity in an era of limited resources.

3.3 Access to data as a human right

In the 21st century, science and technology will continue to have an enormous impact on standards of living around the world as well as on freedom and governance. This is one reason why there is an increasing interest in the claim of access to information, including scientific data, as a human right.

For some, e.g., Shaver (2009), that claim finds its source in Article 27 of the Universal Declaration of Human Rights: "Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits" (United Nations, 1948).

Others, e.g., the New York Law School/Healthcare Information for All 2015 Human Rights and Healthcare Information Project (2009), focus on a particular "right," in this case a right to health, and this claim finds its basis in another section of the Universal Declaration of Human Rights: "Everyone has the right to a standard of living adequate for the health and well-being of himself and his family, including food, clothing, housing and medical care."

In preparation for the UN's 2016-2030 development agenda to succeed the UN's Millennium Development Goals, the International Federation of Library Associations (IFLA) submitted the "Lyon Declaration on Access to Information and Development" to the UN. The Declaration includes the statement:

> We, the undersigned, therefore call on Member States of the United Nations to acknowledge that access to information, and the skills to use it effectively, are required for sustainable development, and ensure that this is recognised in the post-2015 development agenda by:
>
> a) Acknowledging the public's right to access information and data, while respecting the right to individual privacy…." (International Federation of Library Associations, 2014)

As of December, 2014, that language is included in the UN Secretary General's Draft of Sustainable Development Goals for the next decade and a half.

Evaluating the validity of these claims that access to information is a human right is not within the scope of this review. The import here is that the assertion that access to information and data is a human right has reinforced calls for open access to scientific data from still another perspective. Some recent initiatives, while not specifically speaking to the rights claim, have seemed to support it by providing immediate open access to both reviewed papers and raw data when an emergency threatened.

A good example is one of the efforts to provide real time open access to research into the science and spread of H1N1 flu in 2009-2010 via the *PLoS Currents–Influenza* web site (Olson et all, 2011). In this case, there was an immediate emergency which this initiative responded to, and the open sharing of data became almost an imperative. Similar efforts by the general public as well as professional researchers using Google Maps or other online technology have taken place in several cases to follow the spread of a contagious disease.

None of these efforts would be possible without open access to data. Open data advocates point to examples such as these in arguing for increased access to data in the service of the health and well being, both physical and economic, of all people, often pointing to international agreements such as the UN Declaration on Human Rights as a justification.

3.4 Data preservation and archiving

Today, a tremendous amount of scientific data is "born digital," and that fact is a source of much unease in the scientific and public policy communities. A huge amount of digital data is essentially "endangered data" and, in many cases, once it is gone, it can never be replaced (Murillo, 2014).

Examples of new discoveries being made based on existing data that the original authors had no idea about are common in scientific history. "Many classic results in science have come from the analysis of existing knowledge already available in the open literature" (Murray-Rust, 2007). With the "data deluge" today, that is likely to be even more true as machine algorithms mine ever expanding data sets in ways and at speeds that no human can match. As

one researcher responding to a European Union survey on data preservation put it: "The most important reasons for preservation are the ones we do not see now" (van der Hoeven et al, 2010).

Agreement on the need for the preservation of digital data is widespread. In the U.S., the Committee on Science, Engineering and Public Policy of the National Academies of Science put the rationale for preservation very simply: "Research data should be retained to serve future uses. Data that may have long-term value should be documented, referenced, and indexed so that others can find and use them accurately and appropriately… In some research areas, accessible databases have become essential parts of the research infrastructure, comparable to laboratories, research facilities, and computing devices and networks" (2009). This type of thinking is mirrored in reports or position papers or grant funding requirements by the National Science Foundation (2006, 2010), by the European Commission (2013), and NSF/Jisc (Arms & Larson, 2007).

While the motivation and justification for effective archiving of scientific data are widely acknowledged to be valid, what is actually happening on the ground, especially in "Small Science," often fails to capture data for archiving and re-use. In some disciplines, the estimate is that as much as 80% of data developed by individual researchers or small teams is not captured in a public way and is often simply lost over time (Murray-Rust, 2007). The National Science Board (2005) has noted that at the level of what it refers to as *Research Collections* "Authors are individual investigators and investigator teams. *Research collections* are usually maintained to serve immediate group participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards."

Kansa and Bissell (2010) have proposed a web syndication approach for sharing primary data in "Small Science." This approach, if implemented by researchers, would make distribution of data sets more widespread. Yet this approach does not specifically address preservation.

In an effort to capture data from "Small Science" *Research Collections*, as well as from larger research endeavors (both *Resource Collections* and *Reference Collections*, in the National Science Board's terminology), universities are establishing institutional repositories that can handle data as well as publications; disciplinary repositories are being established; and some publishers are setting up data repositories to house data related to articles published in their journals.

With this flurry of activity over the past decade, the questions naturally arises: What characteristics should scientific data repositories have in order to be effective in ensuring that data will be "readily available, accessible, and usable" (Arms & Larson, 2007) and can be "easily consulted and analyzed by specialists and non-specialists alike"? (National Science Foundation, 2006)

4 DESIRABLE CHARACTERISTICS OF DATA COLLECTION AND STORAGE SYSTEMS

Although goals and aspirations can be expressed in general terms, operational characteristics of an effective repository environment need to be more specific. A number of workshops and reports over the past decade have endeavored to outline functions that are desirable in a data storage and access system. In the U.S., examples include Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content (Institute of Museum and Library Services, 2003), Licensing Geographic Data and Services (National Research Council, 2004), and To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering (Association of Research Libraries, 2006).

While those sets of recommendations differ in some ways, reports share common characteristics that the authors see as important for the preservation of scientific data for use by both current and future generations of users. Characteristics include: access; clear use conditions; findability; interoperability; evaluation capability; and the technical issues of ensuring data integrity, scalability, and life cycle management for preservation through time.

While some of these reports are focused on very large data sets, they are readily applicable to data repositories for data of any size. We briefly describe these desirable characteristics in turn, clustering related characteristics together where appropriate.

4.1 Access

The first step in being able to benefit from scientific data is being able to get access to it in the first place. Data that are not online, or are hidden behind paywalls or other restrictive barriers online are not readily accessible to researchers or to the public. Part of The National Science Foundation's *Cyberinfrastructure Vision for 21st Century Discovery*, for example, describes an environment in which data "are openly accessible while suitably protected" so that they may be "regularly and easily consulted and analyzed by specialists and non-specialists alike" (2006).

4.2 Clear use conditions

Accessibility by itself, as the NSF's words above suggest, does not guarantee the ability to re-use data. To be maximally useful to others, data sets must carry with them information about how they may be used, e.g., through clear licenses. While facts *per se* are not copyrightable in many political jurisdictions around the world, including the United States, it is often difficult to tell whether an arrangement of facts is original enough to afford copyright protection and could restrict or limit entirely the uses to which data can be put. In a world of increasingly internationalized repositories, data originating outside of the U.S. may have other legal or restrictions on use, e.g., *sui generis* provisions in the EU. Absent a clear indication by those who produce data sets indicating to what uses the data sets may be put and conditions on their use, if any, the data are essentially useless to others. The uncertainty about possible consequences of misuse will deter most present and future potential users from employing the data for new purposes.

4.3 Findability

In a time of ever-increasing growth of scientific data, being able to find what a user is looking for in a sea of data becomes critically important. Findability depends upon being able to search for data in a consistent manner and in having that data be identified in a consistent manner over time so that they are always discoverable. Both finding particular data across time and space and then being able to access that data depend heavily on standards-based metadata. Data must also have a consistent and permanent identity and location identifier over time. Put simply, if a user cannot find data s/he is seeking, they will never get used.

4.4 Evaluation capability

Science, for at least the past 400 years, has been based on peer review. Repositories or other data collection structures help to make data more valuable when those interested in the data can, if not formally review them, at least comment upon them and discuss their usefulness for particular purposes. In the case of irreproducible data (e.g., time series data, data gathered on expeditions that are not likely to be repeated, etc.), discussions about the data themselves, methods of collection, and so on are critically important. Using the data for other purposes, applying new tools in the future with which to analyze data or even re-analyzing samples collected and stored that are made visible through the metadata in repositories all benefit from having access to the comments of prior users.

4.5 Technical characteristics

For data sets to be useful for future research purposes, users must have confidence in the integrity of the data set. Life cycle management will require data being transferred from one storage medium to another on a routine basis over time to ensure accessible preservation. If any corruption of that data takes place in the process, or for any other reason, the data become suspect, at best. Repository sponsors are naturally concerned with ensuring data integrity, and research is under way to develop standards and best practices for data handling and preservation. From developing unique hash based identities to keeping redundant copies, efforts are underway to ensure users that the data they access are an exact copy of the data that were contributed to the repository or collection. "Lots of Copies Keeps Stuff Safe" (LOCKSS), for example, is software that not only allows institutions to keep redundant copies of information but also regularly audits files at the byte and bit level and repairs them on an ongoing basis (LOCKSS, 2014).

In addition to managing data integrity over time, effective preservation of scientific data in today's world requires scalability, the ability to grow storage and access capabilities and still operate reliably and efficiently. Computer scientists and database designers are constantly working to reduce uncertainty in system performance while dealing with exponential growth of the data to be preserved. At present, a new focus is developing on decentralized and virtual storage and access facilities, often run by large commercial organizations such as Google and Amazon. "Cloud-based" storage offers institutions, especially smaller ones, the opportunity to have both scalable repositories and redundancy without building physical infrastructure themselves.

And wherever data reside, interoperability is a key challenge. Data file structures and layout often differ from one data set or data base platform to another. Metadata are often inconsistent when they exist at all. Searching disparately formatted data sets is a huge challenge. Designing ways to enable a user to search across file structures and types of scientific data and come up with comprehensive and accurate results is the subject of ongoing research. While existing data sets may never be fully interoperable, efforts such as DataNet in the U.S. are working to build structures that may help future data interoperate more effectively.

5 A BRIEF OVERVIEW OF RECENT INITIATIVES TO PROVIDE OPEN ACCESS TO SCIENTIFIC DATA

5.1 Open access data repository growth

The calls for access to scientific information are being heard and acted upon in many quarters today. There are now hundreds of data repositories available online. Some were established and operated for a while but no longer seem to be maintained although they are still accessible, e.g., GlycomeDB or antbase. Some have merged with others in the same domain to provide more efficient operation, e.g., ORegAnno. Many others are still current and vibrant.

In this ever changing environment, finding online data repositories is becoming increasingly difficult unless the URL is already known. Not surprisingly, this challenge has given rise to the creation of a number of data repository cataloguing and search sites. These sites provide lists of repositories and offer various ways to search for particular types of data.

The Open Access Directory (http://oad.simmons.edu/oadwiki/Data_repositories), for example, lists over a hundred directories or repositories in over a dozen different disciplines in which there is at least some open access to data. DataBib (http://databib.org) lists almost a thousand research data sites as of this writing, as does re3data (www.r3data.org). In an effort to provide a more centralized access point and more complete search service for data repositories throughout the world, DataBib and re3data have agreed to merge their catalogs by the end of 2015.

While these catalog sites are operated by organizations, some sites that offer data search and access capabilities are maintained by individuals. One very useful such directory of geographic data sets, Freegisdata (http://freegisdata.rtwilson.com), includes a list of over 300 sources of "free as in free beer" geographic data sets sorted by the type of data they contain although information varies as to whether particular repositories are also "free as in free speech," i.e., what usage rights are. Governments, too, are endeavoring to provide access points to data repositories they provide. Some U.S. examples are discussed in the following section.

Even a cursory look at repository sites confirms that science data repositories include a wide range of capabilities and coverage, ranging from small prototypes to sites containing access to great stores of data from, for example, space probes (e.g.,http://nssdc.gsfc.nasa.gov/), automated astronomical telescopes (e.g., http://tdc-www.harvard.edu/), or the Large Hadron Collider (http://opendata.cern.ch/).

Few of these sites are interoperable in terms of shared metadata schema or data formatting; few have anything resembling a life cycle management plan; few have a commenting or evaluation capability. Still, their existence demonstrates that there is a widening realization that providing access to, and preservation of, scientific data is a valuable and worthy endeavor. The challenge is to make generated data more widely available. Such a goal brings with it many challenges, especially with Big Data, and organizations are currently trying to clearly identify the spectrum of challenges involved and ways to deal with them (e.g., CODATA/ICSU, 2014).

It is not surprising that data that require a huge financial investment to generate, such as astronomical data from the Hubble Space Telescope, are often funded by government bodies. In the U.S. and in many other countries such data are made freely available for anyone's use although that is not the case in every jurisdiction worldwide. In large multinational efforts such as the Global Earth Observation System of Systems (GEOSS), for example, which includes 84 countries and 54 additional Participating Organizations, settling on common usage licenses for data made available through www.geoportal.org by many different countries and agencies remains a significant challenge (Onsrud et al, 2010).

5.2 Access to U.S. government generated data

The U.S. federal government collects and generates enormous amounts of publicly funded data useful to science as well as to industry and the general public. In recent years, the federal government has been attempting to make the data it collects available for research and for simple daily use by anyone. The same is true to different degrees for governments in other parts of the world.

In the U.S., the recently launched Data.gov web site is one example. It provides access to data collected by 18 federal agencies, currently containing well over 100,000 data sets. Because the U.S. government cannot hold copyright on materials it generates (U.S. Code, Title 17, S.105), there is no claim of copyright on any of the data sets, even if they might qualify for copyright protection if generated by non-federal sources.

The U.S. federal government makes both data and tools available for use by anyone who wishes to access them. Sites such as The National Map (http://nationalmap.gov/) provide a starting point for geographic information. The U.S. also makes life science data of various kinds available through the National Institutes of Health for both professional researchers (e.g., PubChem: http://pubchem.ncbi.nlm.nih.gov/) and for lay users (e.g., MedLine Plus: http://www.nlm.nih.gov/medlineplus/); geologic data through the U.S.G.S: http://www.usgs.gov/); and so on.

While the importance of access to data is mirrored at the state and local level in the U.S., access to that data and re-use conditions are much more mixed than on the federal level.

5.3 Access to data in the U.S. generated by non-federal government bodies

State and local governments in the U.S. may hold copyright to datasets that they generate that qualify for copyright protection. Some states and some local governmental bodies are making conscious efforts to make their spatially referenced data available with no or minimal conditions on its use. Maine and Montana are good examples on the state level. Both provide significant collections of spatial data available to users, in Maine through the Maine Office of GIS (http://www.maine.gov/megis/catalog/) and in Montana through the Montana Geographic Information Clearinghouse (http://geoinfo.msl.mt.gov/). MetroGIS (http://metrogis.org/) in the Minneapolis/St. Paul area is a good example on a local/regional level.

Some states, and particularly local government bodies, view their data as a source of income and resist efforts to make it accessible at no cost and under minimal reuse restrictions. This is particularly true for spatially-referenced deed, tax, and other information associated with real estate and real property. Even in states with strong Freedom of Information laws, some municipal and county governments seek to hold onto control over access to data, especially when it is in electronic form, out of concern that the income potential for the government body will be reduced if other entities get access and then make the information available at low or no cost (e.g., the case of Brick Township, NJ: http://www.rcfp.org/news/2005/0712-foi-utilit.html**).**

There are also other motivations for limiting access to information collected by state or local government bodies. Locations of endangered species, for example, are often not made public or exact information about locations of certain types of conservation easements granted to towns out of respect for the privacy of the donors.

Whatever the justification, access to locally generated data at the non-federal level in the U.S. is much more varied than access to data generated by the federal government.

5.4 Private and corporate initiatives

While the focus of this review is primarily on publicly funded data, it is important to note that although private companies usually view their data as proprietary, there are cases in which they make that data available for use at no charge even though they retain ownership.

In the area of spatially-referenced data, Google Earth, Google Maps, and related services by providers including Rand McNally, Mapquest, and others offer access to various types of spatially-referenced information through both computer and mobile devices that are now a part of everyday life for many people. While widely used, including in academic and government contexts, these services lack important features that dependable open access and archival services should include.

First and most simply, these services are proprietary, and even if a company's public goal is "Don't be evil" (as Google's is), there is not and cannot be any guarantee that policies in private companies, especially publicly traded stock companies, will not change when shareholder value demands it. Company policies and practices can change abruptly, as any of Facebook's billion users or the millions of users of Google's Gmail service or even Google Maps well know. Building access to scientific information on proprietary foundations is risky as far as guaranteeing access to, and preservation of, data into the future is concerned.

In addition, even though services such as Google Earth allow contributions of spatially-referenced information from users, questions about usage rights and provenance of posted information abound, and there are no metadata standards in use for contributed information. While keyword search mechanisms have considerable power, they are simply inadequate for scientific search and retrieval purposes, and this is particularly true in the case of spatially-referenced data. In addition to these considerations, the question of the quality of Volunteered Geographic Information (VGI) is also an unsettled one (Flanagin & Metzger 2008).

Private companies may also offer access to a subset of their tools and data for a combination of public service and quasi-promotional purposes. Often these are educational endeavors such as ESRI's ConnectEd Initiative (http://connected.esri.com/) which, while providing students and teachers with classroom tools, also introduce students to the company's products.

In dealing with medical data, as another example, private companies sometimes find it in their interest to make some of their data publicly accessible. When private companies do so, they often, as in the case of clinical trial data made available by some pharmaceutical companies to the Yale Open Data Access Project (http://yoda.yale.edu), retain proprietary ownership of their data and are free to remove them from public sight at any time.

In short, private and corporate initiatives can be welcome supplements to, but at present are unlikely to be major contributors of, openly available scientific data.

5.5 Non-U.S. access efforts

While the primary focus of this review is on U.S. policies and access efforts, in today's international environment, it is impossible to ignore the access to primarily publicly funded scientific data in other countries. Many large repositories, especially disciplinary repositories, include data originating from different countries. In some cases, those repositories have a single policy regarding access and re-use, but in many other cases, access and re-use policies are tied to the laws in the countries from which the data originates. Countries around the world, most of which are able to hold copyright on data, have varying policies on access and re-use. An overall review of those policies is not appropriate here, but it is worth noting that many countries are making efforts to make government generated data, especially geodata, more widely open and available. Examples include UK Location (http://location.defra.gov.uk) in the United Kingdom, the Atlas of Canada (http://atlas.gc.ca/site/english/index.html), and Geoscience Australia (www.ga.gov.au), all of which provide open access to some government-generated spatially-referenced data. In the brief review below, we include some sites that include non-U.S. data and/or are non-U.S. based for illustrative purposes.

5.6 Usage rights and data repositories: a brief review

As the discussion so far suggests and as the examples in the next section illustrate, there already exist numerous disciplinary and government run repositories, particularly those designed to provide access to collections of large scale data. In "Small Science," the picture is much less encouraging, whether those small science data gathering efforts are university or institution based or are the results of sporadic efforts to enable individuals or small local groups with locally generated data of their own to expose them and make them available for others to use.

One absolutely critical component to the reuse of data in repositories of any scale is a clear description of usage rights and conditions for data access and re-use. In some cases, repository sites simply do not even post license information or usage conditions. In others, terms like "free" and "open" are used with a variety of meanings that are sometimes only discernible by drilling deep into the site or in some cases are not specified at all.

Data repositories are usually made up of data that, even if collected on one site, originate from many different sources and often different countries. Some repositories are "federated" in that they provide links to sites where data sets actually reside but do not collect or store data themselves. In either case, data sets may have a variety of usage rights and/or conditions attached to them, and sorting those rights and conditions can be a difficult task.

Absent a definition of terms, repository search engines or catalogs may provide information on usage rights in similar language, but whose usage rights may be very different from other sites using similar language. While there is, as yet, no universally accepted definition of "open" in the context of scientific data, there are efforts underway to create a definition that can be used generally. The Open Definition, offered under the auspices of the Open Knowledge Foundation, asserts that *"A piece of data or content is open if* ***anyone*** *is* ***free to use, reuse, and redistribute*** *it — subject only, at most, to the requirement to attribute and/or share-alike."* (Open Definition, 2014)

Very few repositories specifically reference this Open Definition. One that does is Open Street Map (http://www.openstreetmap.org) which licenses its data under the Open Data Commons Database License (http://opendatacommons.org/licenses/odbl), which in turn depends upon the Open Definition.

In reviewing the status of usage rights and conditions in the context of scientific data repositories, 40 repository sites were examined. This list includes many U.S. based sites, but because of the international nature of data today, especially data located in disciplinary repositories, some reviewed sites are based outside of the U.S. Some, such as re3data.org, are operated as collaborations of organizations located in the U.S. and in Europe. Whether accessible through U.S. government, disciplinary, or even privately operated sites in the U.S. or beyond, the great majority of open data listed below are the result of publicly funded research.

In these 40 sites, 13 different sets of usage terms and conditions for reuse of the data were identified. Summary descriptions of usage rights and conditions are listed below, followed by Table 1 identifying which usage information applied to the 40 sites. A fuller description of the sites and the conditions of use and re-use are attached in Appendix A.

The list below contains simple language descriptions of usage information based on conditions available on the listed repository sites as of December 15, 2014. The numbers are referred to in the "Usage Rights" column of Table 1 below.

1. All U.S. government sites use a similar usage message: data produced by U.S. government workers are Public Domain. However, sites may contain data, datasets, or databases provided by others that may be subject to copyright use restrictions. Such material will be labeled.
2. Data, where copyright restrictions are applicable, are available under a Creative Commons license.
3. Access to the data is available to the public at no charge. The author was not able to find any information about use restrictions.
4. Site asserts copyright in all copyrightable materials including the database itself but makes data free to use for personal, scholarly, or private research purposes. Source attribution requested or required.
5. Data are free of charge, but some data sets may have Conditions of Use.

6. Data are free of charge but some data sets may have Conditions of Use, and those may require user registration.
7. Data and other material remain property of original contributing organization and should be available at no cost.
8. License for use granted under Open Canada License - attribution required.
9. Data available for public use with attribution.
10. Database available under Open Database License. Any protectable content is licensed under an Open Contents License.
11. Majority of material is Public Domain. Some data provided by others may be subject to copyright use restrictions. Such material is labeled.
12. Data have been placed in the Public Domain by contributors.
13. Data available under the Open Data Commons Open Database License. Other material available under a Creative Commons license.

Table 1. Data repository sites with usage rights referenced to the list above.

Site Name	URL	Usage Rights
Scientific Earth Drilling Information Service - SEDIS	http://sedis.iodp.org/front_content.php	3
Data.gov	http://www.data.gov	1
PubChem	http://pubchem.ncbi.nlm.nih.gov/	1
Online Mendelian Inheritance in Man (OMIM®)	https://www.ncbi.nlm.nih.gov/omim	4
Montana Geographic Information Clearinghouse	http://geoinfo.msl.mt.gov/	3
MetroGis	http://metrogis.org/	3
BOLD	http://www.barcodinglife.org/views/login.php	3
ChemSpider	http://www.chemspider.com/	4
Freebase	http://www.freebase.com/	2
Sage Bionetworks	http://sagebase.org/	3
uBio	http://www.ubio.org/	3
ICDNS	http://www.icdns.org/	3
ZooBank	http://www.zoobank.org/	3
OneGeology	http://www.onegeology.org/home.html	7
Gateway to Scientific Data	http://cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/gateway-scientific-data.html	8
Europeana	www.europeana.eu	2
DOE Data Explorer	http://www.osti.gov/dataexplorer/	1
NEXTBIO	http://www.nextbio.com/	10

ChemBank	http://chembank.broadinstitute.org/	3
EnvBase	http://envgen.nox.ac.uk/cgi-bin/envbase.cgi	3
LTSRF	http://ghrsst.nodc.noaa.gov/	1
ZINC	http://zinc.docking.org/index.shtml	4
ADS	http://ads.ahds.ac.uk/	4
GeoGratis	http://www.geogratis.cgdi.gc.ca/	11
NARCIS	http://www.narcis.info/index	3
GBIF	http://www.gbif.org/	12
LinkedGeoData	http://linkedgeodata.org/About	13
dbGaP	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html	6
Open Context	http://opencontext.org/	2
RRUFF	http://rruff.info/	3
PCL Map Collection	http://www.lib.utexas.edu/maps/	14
COD	http://www.crystallography.net/	16
PCOD	http://www.crystallography.net/pcod/index.html	16
Biozon	http://www.biozon.org/	3
ORegAnno [latest entry 2008]	http://www.oreganno.org/oregano/Index.jsp	3
antbase [latest entry 2009]	www.antbase.org	2
AntWeb	www.antweb.org	2
OpenStreetMap	http://www.openstreetmap.org/	17
TOXNET	http://toxnet.nlm.nih.gov/	1
GlycomeDB [latest entry seems to be 2102 - copyright notice is 2007]	http://www.glycome-db.org/	1

OBIS	http://www.iobis.org/home	8
ChEMBL	https://www.ebi.ac.uk/chembl/	5
GeoNames	www.geonames.org	2
Dryad	datadryad.org	16
WorldWideScience.org: The Global Science Gateway	worldwidescience.org	14
National Historical Geographic Information System	www.nhgis.org	12
GlobalSoilMap.net	www.globalsoilmap.net	3
FreeGISData	http://freegisdata.rtwilson.com	6
The National Map	http://nationalmap.gov	1

As the information in this table indicates, there are a wide variety of meanings attached to the term "Open" in terms of use and re-use of scientific data. In a few cases, "Open" conforms to the Open Definition mentioned above. But in far more cases, there are actually conditions on re-use which, if not discovered and adhered to by subsequent users, could cause significant reputational, and/or legal or financial, risks. These "non-obvious" conditions placed on the use of data that are labeled "Open" could create impediments to wider use of such data in science research.

6 CONCLUSION

There is strong, though not universal, support for open access to publicly funded scientific data among governments, the research community, business and industry, and private users. While there are many challenges to overcome to make scientific data findable, technically accessible, and to preserve them effectively through time, even if these challenges are met, there is still a very significant question of whether and under what conditions users may re-use data in online repositories. At present, usage conditions vary widely, and a user's ability to even find what usage conditions are in effect also varies widely, even in the somewhat focused domain of spatially-related data. Absent use of specific, accepted licenses, terms like "Open" can give rise to different interpretations.

As a first step toward making scientific data really open, repositories could select from one of the currently available widely recognized and standardized data licenses that promote open access and use, such as Creative Commons licenses or Open Database Licenses. Repositories could, as some do now, make accepting the conditions of the repository's chosen license a requirement for contributing data to the repository. Users would then clearly know what they could and could not do with data found in the repository.

Having to deal with a variety of such standardized licenses, even if that variety is limited, is not ideal from a user perspective, but it is far better than having dozens of variations on usage and imprecise use of terms like "open" or "free." Ultimately, the ideal would be to have a common set of usage licenses for all repositories of scientific data to help realize the significant benefits to science and society of truly open access to, and use of, scientific data.

7 REFERENCES

Agres, T. (2006) *The Scientist 20*(1), p 77. Retrieved January 7, 2015 from the World Wide Web: http://www.the-scientist.com/article/display/18850

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired 16*(7). Retrieved January 7, 2015 from the World Wide Web: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Anthes, G. (2009) Deep Data Dives Discover Natural Laws. *Communications of the ACM*, *52*(11), pp 13-14. Retrieved January 7, 2015 from the World Wide Web: http://cacm.acm.org/magazines/2009/11/48443-deep-data-dives-discover-natural-laws/fulltext

ARL Workshop on New Collaborative Relationships: the Role of Academic Libraries in the Digital Data Universe, Friedlander, A., Adler, P., & National Science Foundation (U.S.) (2006) *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*. Washington: Association of Research Libraries.

Arms, W. Y., & Larson, R. L. (2007) *The Future of Scholarly Communication: Building the Infrastructure of Cyberscholarship*. Washington: National Science Foundation.

Association of Learned and Professional Society Publishers, International Association of Scientific, Technical and Medical Publishers (2006) Databases, Data Sets, and Data Accessibility–Views and Practices of Scholarly Publishers. Retrieved July 15, 2011 from http://www.stm-assoc.org/2006_06_01_STM_ALPSP_Data_Statement.pdf

Ball, A. & Duke, M. (2012) How to Cite Datasets and Link to Publications: a Report of the Digital Curation Centre. Retrieved December 12, 2014 from the World Wide Web: http://codata2012.tw/sites/default/files/text/slide/codata2012-Duke%20and%20Ball-How%20to%20Cite%20Datasets%20and%20Link%20to%20Publications.pdf

Boseley, S. (2009) Drug Giant GlaxoSmithKline Pledges Cheap Medicine for World's Poor. Retrieved July 15, 2014 from the World Wide Web: http://www.theguardian.com/business/2009/feb/13/glaxo-smith-kline-cheap-medicine

Carlson, S. (2006) Lost in a Sea of Science Data. *Chronicle of Higher Education 52*(42). Retrieved January 7, 2015 from the World Wide Web: http://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

CODATA/ICSU (2014) Big Data for International Scientific Programmes: Challenges and Opportunities. Proceedings from *Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities*, Beijing.

Committee on Science, Engineering, and Public Policy (U.S.) (2009) *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington: National Academies Press.

European Commission (2013) Guidelines on Data Management in Horizon 2020. Retrieved July 15, 2014 from the World Wide Web: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission (2013) Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Retrieved July 15, 2014 from the World Wide Web: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Flanagin, A. J. & Metzger, M. J. (2008) The Credibility of Volunteered Geographic Information. *GeoJournal 72*, 137-148. Retrieved December 12, 2014 from the World Wide Web: DOI 10.1007/s10708-008-9188-y

GovLab (2014) Open Data 500. Retrieved July 15, 2014 from the World Wide Web: http://www.opendata500.com

Herper, M. & Langreth, R. (2007) Biology Goes Open Source. Retrieved July 15, 2014 from the World Wide Web: http://www.forbes.com/2007/02/12/novartis-genes-diabetes-research-biz-cz_mh_0212novartis.html?partner=rss

Institute of Museum and Library Services (U.S.) (2003) *Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content*. Washington: Institute of Museum and Library Services.

Interagency Working Group on Digital Data (2009) *Harnessing the Power of Digital Data for Science and Society*. Washington: National Science and Technology Council, Executive Office of the President.

International Federation of Library Associations (2014) Lyon Declaration on Access to Information and Development. Retrieved July 15, 2014 from the World Wide Web: http://www.ifla.org/node/8559

Kansa, E. C. & Bissell, A. N. (2010) Web Syndication Approaches for Sharing Primary Data in "Small Science" Domains. *Data Science Journal 9*. Retrieved January 7, 2015 from the World Wide Web: https://www.jstage.jst.go.jp/article/dsj/9/0/9_009-012/_article

Kash, W. (2014) Why Free Government Data Remains a Tough Sell. *Information Week*. Retrieved January 8, 2015 from the World Wide Web: http://www.informationweek.com/government/open-government/why-free-government-data-remains-a-tough-sell/d/d-id/1113604

LOCKSS (2014) What Is the LOCKSS Program? Retrieved July 15, 2014 from the World Wide Web: http://www.lockss.org/about/what-is-lockss

McKee, L. (2010) Seventeen reasons why geospatial research data should be published online using OGC standard interfaces and ISO Standard Metadata. Open Knowledge Foundation. Retrieved December 2, 2014 from the World Wide Web: http://blog.okfn.org/2010/06/21/open-geoprocessing-standards-and-open-geospatial-data/

Murillo, A. P. (2013) Data at Risk Initiative: Examining and Facilitating the Scientific Process in Relation to Endangered Data. *Data Science Journal 12*, pp 207-219. Retrieved January 7, 2015 from the World Wide Web: http://dx.doi.org/10.2481/dsj.12-048

Murray-Rust, P. (2007) Data-driven Science–A Scientist's View. *Proceedings from NSF/JISC Repositories Workshop*, Phoenix.

National Information Standards Organization, & National Federation of Advanced Information Services (2013) Recommended Practices for Online Supplemental Journal Article Materials. Retrieved November 6, 2014 from the World Wide Web: http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf

National Research Council (U.S.) Committee on Licensing Geographic Data and Services (2004) *Licensing Geographic Data and Services*. Washington: National Academies Press.

National Science Board (2005) *Long-Lived Digital Data Collections: Research and Education in the 21st Century*. Washington: National Science Foundation.

National Science Foundation (U.S.) (2011) Grant Proposal Guide (Chapter II). Retrieved Dec 12, 2014 from the World Wide Web: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

National Science Foundation Cyberinfrastructure Council (2006) *NSF's Cyberinfrastructure Vision for the 21st Century (v5.0)*. Washington: National Science Foundation.

New York Law School Institute for Information Law and Policy (2009) *Access to Health Information Under International Human Rights Law (Draft)*. New York: New York Law School. Retrieved January 7, 2015 from the World Wide Web: http://www.hifa2015.org/hifa2015-and-human-rights

Olson, D.R., Paladini, D. R., Lober, W.B., & Buckeridge, D.L. (2011) Applying a New Model for Sharing Population Health Data to National Syndromic Influenza Surveillance: DiSTRIBuTE Project Proof of Concept, 2006 to 2009. *PLOS Currents*, September 11, 2011. Retrieved January 7, 2015 from the World Wide Web: http://currents.plos.org/influenza/article/applying-a-new-model-for-sharing-261w1jjdm6zrb-5/

Onsrud, H., Campbell, J., & van Loenen, B. (2010) Towards Voluntary Interoperable Open Access Licenses for the Global Earth Observation System of Systems (GEOSS). *International Journal of Spatial Data Infrastructures Research* 5, pp 194-215. Retrieved January 7, 2015 from the World Wide Web: http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/168

Royal Society of Chemistry (2014) ChemSpider Terms and Conditions. Retrieved July 15, 2014 from the World Wide Web: http://www.rsc.org/help/termsconditions.asp

Shaver, L. (2010) The Right to Science and Culture. *Wisconsin Law Review 2010*(1), pp 121-184.

Szalay, A. & Gray, J. (2006) 2020 Computing: Science in an Exponential World. *Nature 440*, pp 413-414.

Tuutti, C. (2011) NSF Seeks Cyber Infrastructure to Make Sense of Scientific Data. Retrieved January 7, 2015 from the World Wide Web: http://fcw.com/Articles/2011/10/04/NSF-taps-UNC-researchers.aspx?Page=1

United Nations (1948) Universal Declaration of Human Rights. Retrieved Jul 15, 2014 from the World Wide Web: http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=eng

University of California (2014) DMPTool. Retrieved July 15, 2004 from the World Wide Web: https://dmp.cdlib.org

van der Hoeven, J., Mele, S., Guidetti, V., & Schrimpf, S. (2010) What We Learned from PARSE.Insight. *Proceedings from PARSE.Insight Symposium*, Paris, France.

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., & Moore, A. J. (2010) Data Archiving. *The American Naturalist 175*(2), pp 145-146.

Yale University (2014) Yale University Open Data Access (YODA) Project. Retrieved July 15, 2014 from the World Wide Web: http://medicine.yale.edu/core/projects/yodap/index.aspx

8 APPENDIX A

Site Name	Description	URL	Usage Rights
Scientific Earth Drilling Information Service - SEDIS	The Integrated Ocean Drilling Program (IODP) is developing a web based information service SEDIS - to facilitate access to all data and information related to scientific ocean drilling, regardless of origin or location of data. SEDIS will be designed to integrate distributed scientific drilling data via metadata.	http://sedis.iodp.org/front_content.php	No mention of usage rights. Data sets can be downloaded right from site
Data.gov	Data.gov is the official portal for open data from the U.S. government. It is a public domain website	http://www.data.gov	U.S. Federal data available through Data.gov is offered free and without restriction. Data and content created by government employees within the scope of their employment are not subject to domestic copyright protection under 17 U.S.C. § 105. Non-federal data available through Data.gov may have a different licensing method as noted under "Show more" at the bottom of the dataset page. Non-federal data can be identified by name of the publisher and the diagonal banner that shows up on the search results and data set pages. Federal data will have a banner noting "Federal" and non-federal banners will note "University", "Multiple Sources", "State", etc."
PubChem	PubChem, released in 2004, provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative.	http://pubchem.ncbi.nlm.nih.gov/	"Information that is created by or for the US government on this site is within the public domain… This site contains resources such as, but not limited to, PubMed Central (see PMC Copyright Notice), Bookshelf (see Bookshelf Copyright Notice), OMIM, and PubChem which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws."

Online Mendelian Inheritance in Man (OMIM®)	OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily.	https://www.ncbi.nlm.nih.gov/omim	"The rights in and to OMIM (excluding information contained therein obtained from third parties) vest in JHU. JHU holds the copyright and trademark to OMIM and OMIM.org, including the collective data therein…Use of OMIM.org is provided free of charge to any individual for personal use, for educational or scholarly use, or for research purposes through the front end of the database."
Montana Geographic Information Clearinghouse	Geographers at the Montana State Library develop, collect, support, deliver, and promote Montana geographic information and data in Montana and beyond.	http://geoinfo.msl.mt.gov/	No usage information stated
MetroGis	The purpose of MetroGIS is to institutionalize the sharing of accurate and reliable geospatial data so user and producer communities can share in the efficiencies of being able to effortlessly obtain the data they need, in the form they need, when they need it.	http://metrogis.org/	"…government data are public and are accessible by the public for both inspection and copying unless there is federal law, a state statute, or a temporary classification of data that provides that certain data are not public."
BOLD	The Barcode of Life Data Systems (BOLD) is an informatics workbench aiding the acquisition, storage, analysis, and publication of DNA barcode records. By assembling molecular, morphological, and distributional data, it bridges a traditional bioinformatics chasm. BOLD is freely available to any researcher with interests in DNA barcoding.	http://www.barcodinglife.org/views/login.php	Incorporates data from GenBank, Canadian Centre, others. Makes what it refers to data as public data available for search or download but does not discuss usage or copyright
ChemSpider [example of a site offering access but not re-use]	ChemSpider is a free chemical structure database providing fast access to over 30 million structures, properties and associated information. By integrating and linking compounds from more than	http://www.chemspider.com/	This web site and any RSC information accessed from this site are protected by copyright. The RSC maintains this site for your information, education, communication, and personal

	400 data sources, ChemSpider enables researchers to discover the most comprehensive view of freely available chemical data from a single online search. It is owned by the Royal Society of Chemistry.		entertainment. You may browse, download or print out one copy of the material displayed on the site for your personal, non-commercial, non-public use, but you must retain all copyright and other proprietary notices contained on the materials. You may not further copy, distribute or otherwise use any of the materials from this site without the advance, written consent of RSC.
Freebase	Initially, Freebase was seeded by pulling in information from a large number of high-quality open data sources, such as Wikipedia, MusicBrainz, and others. The Freebase community along with the internal Freebase team continue to drive the growth of the graph by focusing on bulk, algorithmic data imports, data extraction from free text, ongoing synchronization of data feeds, and rigorous quality management.	http://www.freebase.com/	CC-By license and some under GFDL
Sage Bionetworks	We work to redefine how complex biological data is gathered, shared and used, redefining it through open systems, incentives, and norms.	http://sagebase.org/	Our software is available in Github, and our non-software creative works are licensed under the Creative Commons Attribution 3.0 Unported license except for legacy publications in closed journals. The research projects benefit both the specific collaborators and the larger scientific community because the results will also be accessible in the Sage Bionetworks Commons one year after the conclusion of the research projects.
uBio	Indexing & Organizing 11,106,374 Biological Names. uBio is an initiative within the science library community to join international efforts to create and utilize a comprehensive and collaborative catalog of	http://www.ubio.org/	Many tools and applications. No specific rights info.

	known names of all living (and once-living) organisms.		
ICDNS	To date criteria have been developed by essentially closed groups of interested workers and this may have limited the speed of development and responsiveness of classification schemes. To make such criteria widely accepted many people now believe that there should be an opportunity for any interested worker to participate in their development. Such a democratic forum can now be realised using the internet and the web.	http://www.icdns.org/	"Use, reproduction and intellectual property in the contents of the ICDNS website are assigned according to an 'Open Source' license agreement which is presented in the Discussion Forum. This allows free use of material provided that the user complies with the terms of the license. You must AGREE to the terms of this license before making any use of material on the website." [THIS PAGE NOT AVAILABLE AS OF 7/1/14]
ZooBank	ZooBank provides a means to register new nomenclatural acts, published works, and authors.	http://www.zoobank.org/	Rights usage not specifically noted but see paper on Scientific names of organisms
OneGeology [U.S. not a member but federal and state agencies make data available]	OneGeology's aim is to create dynamic digital geological map data for the world. It is an international initiative of the geological surveys of the world who are working together to achieve this ambitious and exciting venture.	http://www.onegeology.org/home.html	"Map data distributed as part of OneGeology will remain in the ownership of the originating geological survey or organisation, and ideally be available at no cost."
DOE Data Explorer	Use the DOE Data Explorer (DDE) to find scientific research data - such as computer simulations, numeric data files, figures and plots, interactive maps, multimedia, and scientific images - generated in the course of DOE-sponsored research in various science disciplines.	http://www.osti.gov/dataexplorer/	Public domain but... "When using the OSTI website, you may encounter documents, illustrations, photographs, or other information resources contributed or licensed by private individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws. Transmission or reproduction of protected items beyond that allowed by fair use as defined in the copyright laws requires the written permission of the copyright owners."
NEXTBIO	NextBio is the provider of an innovative platform that	http://www.nextbio.com/	"NextBio contains the world's largest repository of curated and

	enables life science researchers to search, discover, and share knowledge locked within public and proprietary data. NextBio's platform seamlessly combines powerful tools with unique correlated content to transform information into knowledge, providing the foundation for new scientific discoveries.		correlated public and private genomic data, including data from multiple public repositories of genomic studies and patient molecular profiles, up-to-date reference genomes, and clinical trial results. Diverse molecular data types from these resources are systematically processed, curated and integrated into our private data center-based platform."
ChemBank	ChemBank is a public, web-based informatics environment created by the Broad Institute's Chemical Biology Program and funded in large part by the National Cancer Institute's Initiative for Chemical Genetics (ICG). This knowledge environment includes freely available data derived from small molecules and small-molecule screens, and resources for studying the data so that biological and medical insights can be gained.	http://chembank.broadinstitute.org/	"The goals of ChemBank are to provide life scientists unfettered access to biomedically relevant data and tools heretofore available almost exclusively in the private sector. We intend for ChemBank to be a planning and discovery tool for chemists, biologists, and drug hunters anywhere, with the only necessities being a computer, access to the Internet, and a desire to extract knowledge from public experiments whose greatest value is likely to reside in their collective sum."
LTSRF	Long Term Stewardship and Reanalysis Facility (LTSRF) for the Group for High Resolution SST (GHRSST), which is routinely delivering individual as well as multi-sensor blended SST products with high accuracy and fine spatial resolution	http://ghrsst.nodc.noaa.gov/	National Oceanic Data Center. "NODC maintains the long term archive and works with the NASA JPL/Caltech Physical Oceanography Distributed Active Archive Center (PO.DAAC) Global Data Assembly Center (GDAC) to provide stewardship of these valuable data sets" [US Gov - public domain]
ZINC	Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF).	http://zinc.docking.org/index.shtml	ZINC is freely available to everyone to use. Significant portions of ZINC may not be re-distributed without express written permission of John Irwin.

GeoGratis	GeoGratis is a portal provided by the Earth Sciences Sector (ESS) of Natural Resources Canada (NRCan) which provides geospatial data at no cost and without restrictions via your Web browser.	http://www.geogratis.cgdi.gc.ca/	"Canada grants to the licensee a non-exclusive, fully paid, royalty-free right and licence to exercise all intellectual property rights in the data. This includes the right to use, incorporate, sublicense (with further right of sublicensing), modify, improve, further develop, and distribute the Data; and to manufacture or distribute derivative products." Attribution is required under Open Government Licence-Canada
GBIF	"The Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments. It allows anyone, anywhere to access data about all types of life on Earth, shared across national boundaries via the Internet.... It provides a single point of access (through this portal and its web services) to more than 400 million records, shared freely by hundreds of institutions worldwide, making it the biggest biodiversity database on the Internet."	http://www.gbif.org/	"The Participants who have signed the MoU have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data. GBIF data sharing should take place within a framework of due attribution."
LinkedGeoData [although not hosted in the U.S., includes VGI data from U.S. contributors]	LinkedGeoData is an effort to add a spatial dimension to the Web of Data / Semantic Web. LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles. It interlinks this data with other knowledge bases in the Linking Open Data initiative.	http://linkedgeodata.org/About	The Linked Geo Data database is made available under the Open Database License. Any rights in individual contents of the database are licensed under the Database Contents License.

dbGaP	The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html	dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.
Open Context	Open Context is a free, open access resource for the electronic publication of primary field research from archaeology and related disciplines. It emerged as a means for scholars and students to easily find and reuse content created by others, which are key to advancing research and education. Open Context's technologies focus on ease of use, open licensing frameworks, informal data integration and, most importantly, data portability	http://opencontext.org/	"Open Context provides a platform for researchers to publish their primary field data and documentation. Because Open Context is a free and open access service, all members of the public are welcome to use and reuse this content." "Open Context licenses all content with Creative Commons, and makes it available in a variety of machine-readible formats."
RRUFF	The RRUFF™ Project is creating a complete set of high quality spectral data from well characterized minerals and is developing the technology to share this information with the world. Our collected data provides a standard for mineralogists, geoscientists, gemologists and the general public for the identification of minerals both on earth and for planetary exploration.	http://rruff.info/	No specific rights info Appears to be OA – funded in part by NSF – also has private contributors.

PCL Map Collection	Maps digitized by the Univ. of Texas Libraries.	http://www.lib.utexas.edu/maps/	Most of the maps scanned by the University of Texas Libraries and served from this web site are in the public domain. A few maps are copyrighted, and are clearly marked as such.
ORegAnno [latest entry 2008]	AN OPEN ACCESS DATABASE FOR GENE REGULATORY ELEMENT AND POLYMORPHISM ANNOTATION The Open REGulatory ANNOtation database (ORegAnno) is an open database for the curation of known regulatory elements from scientific literature	http://www.oreganno.org/oregano/Index.jsp	This project was funded by Genome Canada, the Michael Smith Foundation for Health Research, the Natural Sciences and Engineering Research Council, and the Canadian Institute for Health Research. It will received ongoing maintenance and support from 2005 through 2007 [now listed in DataBib through Canada's Michael Smith Genome Sciences Centre
antbase [latest entry appears to be 2009]	Antbase now provides for the first time access to all the ant species of the world, one of the ecologically most important groups of animals worldwide.	www.antbase.org	CC – By-NC-SA
AntWeb	AntWeb focuses on specimen level data and images linked to specimens. In addition, contributors can submit natural history information and field images that are linked directly to taxonomic names. Distribution maps and field guides are generated automatically. All data in AntWeb are downloadable by users. AntWeb also provides specimen-level data, images, and natural history content to the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EOL.org), and Wikipedia.	www.antweb.org	AntWeb content is licensed under a Creative Commons Attribution License. We encourage use of AntWeb images. In print, each image must include attribution to its photographer and "from www.AntWeb.org" in the figure caption. For websites, images must be clearly identified as coming from www.AntWeb.org, with a backward link to the respective source page. Photographer and other copyright information is provided on the big image page. Some photos and drawing belong to the indicated persons or organizations and have their own copyright statements. Photos and

			drawings with CCBY, CC-BY-NC or CC-BY-SA can be used without further permission, as long as guidelines above for attribution are followed.
OpenStreetMap	OpenStreetMap is a free editable map of the whole world. It is made by people like you.	http://www.openstreetmap.org/	OpenStreetMap is open data, licensed under the Open Data Commons Open Database License (ODbL). The cartography in our map tiles, and our documentation, are licensed under the Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA).
TOXNET [not listed in DataBIB]	Toxicology Data Network	http://toxnet.nlm.nih.gov/	Government information at NLM Web sites is in the public domain. Public domain information may be freely distributed and copied, but it is requested that in any subsequent use the National Library of Medicine (NLM) be given appropriate acknowledgement. When using NLM Web sites, you may encounter documents, illustrations, photographs, or other information resources contributed or licensed by private individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws. Transmission or reproduction of protected items beyond that allowed by fair use as defined in the copyright laws requires the written permission of the copyright owners. Specific NLM Web sites containing protected information provide additional notification of conditions associated with its use.
GlycomeDB [latest entry seems to be 2102 - copyright notice is 2007]	With this library we have translated the carbohydrate sequences of all freely available databases (CFG , KEGG, GLYCOSCIENCES.de, BCSDB and Carbbank) to GlycoCT, and created a new database (GlycomeDB)	http://www.glycome-db.org/	Database of OA databases so presumably OA although there is a copyright notice on bottom of page

	containing all structures and annotations.		
OBIS	OBIS (Ocean Biogeographic Information System) strives to document the ocean's diversity, distribution and abundance of life. Created by the Census of Marine Life, OBIS is now part of the Intergovernmental Oceanographic Commission of UNESCO, under its International Oceanographic Data and Information Exchange programme	http://www.iobis.org/home	OBIS is committed to keeping its data free and openly accessible for the public. So, if you have sensitive data you probably don't want to publish it through OBIS (or any other publication). OBIS does not claim ownership or rights to the data sets it publishes. All rights remain with the data source, whether distributed directly or mediated, whom may at any time decide to remove their data from OBIS
ChEMBL	The European Bioinformatics Institute is part ofEMBL, Europe's flagship laboratory for the life sciences. EMBL-EBI provides freely available data from life science experiments covering the full spectrum of molecular biology. European Bioinformatics Institute - Funded by the Welcome Trust	https://www.ebi.ac.uk/chembl/	Open - Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends on research consent agreements.
GeoNames	GeoNames contains over 10 million geographical names and consists of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. (more statistics ...). The data is accessible free of charge through a number of webservices and a daily database export.	www.geonames.org	The GeoNames geographical database is available for download free of charge under a creative commons attribution license.
Dryad	The Dryad Digital Repository is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes.	datadryad.org	Repository Users are allowed and encouraged to reuse Content from the Repository in any manner except as described herein under "Prohibited Uses Generally" (Section 8.2) ["unlawful manner"]. To the extent possible under law, Submitters have waived all

			copyright and related or neighboring rights to this data.
WorldWideScience.org: The Global Science Gateway	WorldWideScience.org is a global science gateway comprised of national and international scientific databases and portals.WorldWideScience.org accelerates scientific discovery and progress by providing one-stop searching of databases from around the world (Architecture: What is under the Hood). Multilingual WorldWideScience.org provides real-time searching and translation of globally-dispersed multilingual scientific literature.	worldwidescience.org	“A federation of national science portals where research results are made available by participating nations, providing encompassing coverage of global science and research results across language barriers... Much of the information accessed via this gateway is freely available and open domain.”
National Historical Geographic Information System	The National Historical Geographic Information System (NHGIS) provides, free of charge, aggregate census data and GIS-compatible boundary files for the United States between 1790 and 2012.	www.nhgis.org	Citation and Use of NHGIS Data All persons are granted a limited license to use this documentation and the accompanying data, subject to the following condition: Publications and research reports based on the database must cite it appropriately...In addition, we request that users send us a copy of any publications, research reports, or educational material making use of the data or documentation.
GlobalSoilMap.net	A global consortium has been formed to make a new digital soil map of the world using state-of-the-art and emerging technologies. This new global soil map will predict soil properties at fine spatial resolution (~100 m).	www.globalsoilmap.net	no info on usage found
Free GIS Data	This page contains a categorised list of links to over 300 sites providing freely available geographic datasets - all ready for loading into a Geographic Information System.	http://freegisdata.rtwilson.com/	As you might guess from the title - I only list free GIS datasets here. That word is rather ambiguous (just ask Richard Stallman!), but here I use the meaning of 'free as in beer', and I include sites that provide data free for non-commercial purposes. Funnily enough, given

			the title, I don't link to datasets that cost money!
The National Map	As one of the cornerstones of the U.S. Geological Survey's (USGS) National Geospatial Program, *The National Map* is a collaborative effort among the USGS and other Federal, State, and local partners to improve and deliver topographic information for the Nation.	http://nationalmap.gov	USGS-authored or produced data and information are considered to be in the U.S. public domain. While the content of most USGS Web pages is in the U.S. public domain, not all information, illustrations, or photographs on our site are. Some non USGS photographs, images, and/or graphics that appear on USGS Web sites are used by the USGS with permission from the copyright holder.

(Article history: Received 16 August 2014, Accepted 4 January 2015, Available online 19 January 2015)