BIG OPPORTUNITIES IN ACCESS TO "SMALL SCIENCE" DATA

Harlan Onsrud[1*] and James Campbell[2]

[1] Department of Spatial Information Science and Engineering, University of Maine, Orono, ME 04469*
Email: onsrud@spatial.maine.edu
[2] Department of Spatial Information Science and Engineering, University of Maine, Orono, ME 04469
Email: campbell@spatial.maine.edu

ABSTRACT

A distributed infrastructure that would enable those who wish to do so to contribute their scientific or technical data to a universal digital commons could allow such data to be more readily preserved and accessible among disciplinary domains. Five critical issues that must be addressed in developing an efficient and effective data commons infrastructure are described. We conclude that creation of a distributed infrastructure meeting the critical criteria and deployable throughout the networked university library community is practically achievable.

Keywords: Small science, Commons, Licenses, Metadata, Provenance, Archiving, Peer review

1 INTRODUCTION

Data is the lifeblood of science. Researchers, government institutions, not-for-profit organizations, schools, commercial organizations, and individual citizens all need the widest possible access to data from all sources to explore, experiment, test, create new knowledge and new products, and, ultimately, to increase understanding of our world. Access to data, put simply, is crucial to the future development of science and society.

There are a number of initiatives underway to make scientific data accessible to other scientists and, in some cases, to the public, without the user needing to obtain prior permission from the data owner/provider. To date, these initiatives generally have been discipline specific and have in many cases developed their own tools for storing, searching, and accessing data relevant to the discipline.

They have also generally been focused on what some have referred to as "big science" rather than "small science" research endeavors, performed by a single investigator or small group. To scientific data generated by these "small science" researchers, we would add data generated for purposes "besides science" that would be of value to scientists if available to them. "Besides science" data refers to data collected by those who are not usually seen as professional academic or institutional scientific researchers but who generate technical or observational data of scientific value. Examples of such data gathering might include a high school class project that maps the location of all alcohol related car accidents involving teenagers in their county; a homeowners' association that monitors the water quality and plant growth in the lake on which their property is located; or a wheelchair-bound man who has mapped every wheelchair accessible street crossing in his city - information the city itself did not have - so that he could plan his routes to shop, go to the library, and carry out his personal business.

2 INVESTIGATOR-FOCUSED SMALL SCIENCE

Small group and single investigator science is still at the core of knowledge advancement in most scientific domains. Science often advances through the corpus of thousands of individual efforts that depend upon the past efforts of peers. Further, some of the greatest advances in science have occurred when the knowledge from one scientific domain becomes useful and accessible to researchers in another. In today's scientific environment, it is important to provide efficient capabilities within and across disciplinary domains to readily allow scholars to archive the data used in producing their intellectual output and to allow and enable others to efficiently access such data for verification, critique, or uses that the original investigator did not foresee.

Research efforts that build on the past intellectual works of others ideally require the efficient ability to check on the

logic, procedures, processes, and data used in producing that intellectual work. Each individual investigator may draw from many databases, non-digital sources, and the investigator's own observations, and it is important to capture and document the details of such efforts so that they may be revisited, confirmed, or challenged by other scientists.

For example, where does the academic geologist archive the twelve critical data files upon which the geologist's latest published investigation was based in a form that will be findable, accessible, and usable by others twenty years in the future? Where may students completing master's and doctoral theses archive the most important digital data sets created through their research in a manner that will allow the data to be revisited by other students and scholars in future years? A system allowing efficient access to the data sets used by individual and small groups of scientists in accomplishing and reporting their research work would be an important and valuable step forward.

3 LOCALLY GENERATED "INVISIBLE" DATA

Most large-scale scientific data efforts focus on collection and maintenance of data by scientists for scientists. Yet across the globe, individuals and organizations are gathering detailed data at a level that could be of immense value to scientists but which is, for all practical purposes, hidden from their view.

Such locally-collected, detailed data is accumulated through on-the-ground direct observations or interpretation. For example, individuals, small companies, not-for-profit organizations, civic groups, small local government agencies, or others may have had reason to discover and record the diameter and use of a buried pipe, the types of plant life in a meadow, the use of a building, the types and amounts of pollutants in the air, the number of people suffering from a particular illness in an area, or any number of other types of data. College or high school students in science classes may gather important data in assorted data files about a biological or social community. That data could be of significant use to others in the future, but typically is never reused and is often deleted from local servers at the end of the academic year.

In short, there is a wealth of "small science" and "besides science" data that might be valuable to other researchers or to the general public, but is not made available to them.

4 TOWARD A COMMONS OF SCIENTIFIC AND TECHNICAL DATA (CSTD)

Is there a way to efficiently "reveal" this data – always at the option of the data producers – and make it publicly available? Put another way: What kind of infrastructure would be sufficient to enable those with "small" and "besides" science data who wished to do so to contribute it to a digital commons environment so that the data could be preserved and made available to others? (A digital commons environment, in the simplest terms, is one in which users of materials "located" in the commons do not have to seek prior permission for use of the materials, either because they are part of the public domain, or because owners of materials contributed to the commons which qualify for intellectual property protection have granted permission for their use, as long as users respect whatever conditions the owner has placed on use of the materials.)

Although the implementations of infrastructure designs may vary, we believe that any successful commons infrastructure for contributed "small" or "besides" science data will have to deal with at least five critical issues:

- Making explicit the intellectual property licensing conditions that the contributor has established for use of the contributed data and binding that licensing information to the data;
- Generating metadata semi-automatically: providing a mechanism for creating standards-based metadata without contributors having to become metadata specialists;
- Tracking file provenance: ensuring that data provenance is always traceable through generations of digital copying and re-use, both to ensure credit for contributors who wish to have it and to be able to trace the source of any data used in subsequent combined datasets;
- Archiving: ensuring data integrity over time, because one incentive for data creators to contribute to a commons is long-term storage of, and accessibility to, their datasets on reliable media other than their own, either as a primary means of storage (e.g., for students finishing their studies) or as a reliable secondary safety net;

- Peer-review evaluation systems: enabling users to evaluate data for its suitability to meet users' needs, which also provides feedback to the contributors and to future users.

In the remainder of this article, we briefly outline the challenges involved in dealing successfully with each of these issues in the context of "small" and "besides" science and suggest some characteristics of infrastructure systems that can meet those challenges.

5 INTELLECTUAL PROPERTY MANAGEMENT IN A CSTD

The first question that arises in discussing intellectual property management in a commons of scientific and technical data is whether the data residing in the CSTD qualifies for copyright or other types of legal protection. Unfortunately, the answer is not as clear as we might like it to be. Here, we use the situation in the U.S. to illustrate our points. Intellectual property regulations in other jurisdictions may be somewhat different and the implementation of the conceptual framework we suggest would need to conform to local laws.

U.S. Copyright law does not extend copyright protection to facts per se nor, in the words of the copyright statute, to "any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work."[1] Justice Sandra Day O'Connor went to great lengths in writing for the 9-0 Supreme Court majority in the Feist case[2] to demonstrate that while "sweat of the brow" in compiling a collection of facts might legitimately be called an investment, that investment alone does not meet the criteria for copyright protection in the U.S. In her words:

> This case concerns the interaction of two well-established propositions. The first is that facts are not copyrightable; the other, that compilations of facts generally are… The key to resolving the tension lies in understanding why facts are not copyrightable. The sine qua non of copyright is originality. To qualify for copyright protection, a work must be original to the author… To be sure, the requisite level of creativity is extremely low; even a slight amount will suffice...[3]

A simple compilation of facts alone, absent originality in selection or arrangement, does not merit copyright protection in the U.S. (or most other countries). There is, Justice O'Connor pointed out, a clear difference between "discovery" and "authorship." A census taker, for example, who records how many people live in a building is simply discovering and recording a fact – there is no originality hence no protectable authorship involved. The only source of originality, from a copyright perspective, lies in the way that those facts are creatively selected or arranged.

But where is the line that separates the absence of originality from the required level of even a "slight amount" of creativity that will trigger copyright protection? A CSTD infrastructure cannot know. It must leave the choice of whether to assert copyright to the contributor and should offer the contributor a limited set of options for making his or her work available if asserting copyright. Further, if reasonable people might disagree as to whether a slight amount of creativity might exist in the selection, coordination or arrangement of data in a dataset, the need to resolve the question is eliminated by creators conveying to the world any rights they may have in the dataset through a public domain or open access license (e.g., creative commons licenses).

There is another important consideration in deciding how to offer license options in the CSTD in addition to the question of whether a particular dataset does, or does not, qualify for copyright protection. A license is a contract. Even though a dataset may not qualify for copyright protection, if a user accepts the license that the contributor offers, there is quite likely a contractual obligation created, even if copyright does not apply. In that case, according to Feist, a contract might bind the first user, but once that user made the data publicly available, subsequent third-party users could simply use the data as they pleased with no need to enter into a contract with the original owner. In the Feist case, once the names and addresses in the phone book were published, because they consisted of an obvious – and thus non-copyrightable – collection and arrangement of facts (i.e., names and addresses in alphabetical order), anyone was free to copy and use elements of or the dataset in entirety. In the U.S., therefore, licenses for non-copyrightable datasets would be essentially useless. A license might legitimately bind the first user if that user accepted the license, but any subsequent user would be free to use the dataset without entering into a contract with the original disseminator.

Given this framework, what types of licensing infrastructure make the most sense for data contributed to a CSTD? Because a CSTD is a commons environment, even if contributors assert copyright protection for their datasets, they will nonetheless make their data available under "some rights reserved" style licenses that inform users of any stipulated conditions for use. As long as those conditions are respected, they may use the datasets without obtaining prior permission.

Our research suggests that the management system that will best achieve the overall goals of a CSTD – making data available for the widest possible use, while also encouraging contributors to place their work in the CSTD environment – will have the following characteristics:

- A limited number of clear licensing choices so that a contributor can choose and attach a license quickly and easily to the data file.
- License choices that have the least impact on future re-use, while preserving a contributor's essential interests. To accommodate this characteristic, we suggest offering contributors only three license choices:
    1) dedicating the datasets to the public domain, which removes any potential copyright or other restrictions on subsequent use;
    2) requiring attribution if the dataset is subsequently used by others; or
    3) requiring that creators of any derivative works that use data from a contributor's dataset release their work under the same license that the contributor has chosen.[4]

The restrictions imposed by the second and third license choices may of course be overcome by obtaining permission from the licensor to allow use of the dataset without those restrictions. By imposing the second or third restriction above, or both, the permissions burden currently imposed on scientists by publishers is reversed.

For example, scientists currently are usually required by their publishers to seek the permission of other publishers and/or scientists if they use datasets from previously published works in a subsequent publication. Many scholarly publishers take a conservative position and assume that some creativity in the selection and coordination of the dataset in the previous work may have been copied into the new work and therefore demand of scientists publishing a new work that they obtain permission for every such dataset that they may include in their new papers. It has been the general practice of the corporate counsels of publishers to apply a one-size-fits-all rule and to "require permissions for all copied material" rather than to assess each situation on a case-by-case basis. This appears to be the status quo across much of scholarly publishing.5

However, if scientists impose either the second or third licensing conditions above on their data sets, the situation is reversed. The subsequent scientist – and the scientist's publisher – would not need to obtain permission since it has already been granted, provided they comply with the attached conditions.

Of course, a subsequent scientist or publisher might not want to abide by those conditions. In that case, the subsequent scientist would be placed in the awkward position of asking permission of the previous scientists who chose license option 2 above to allow him to use their work without citing their contributions. If the previous scientist had chosen option 3 above, and the subsequent scientist did not want to abide by that condition, the subsequent scientist would, in effect, have to seek permission to make the previous scientist's work less accessible in the subsequent new work by not releasing the new work under the same "some rights reserved" license chosen by the previous scientist.

Our suggested approach provides flexibility. A subsequent user, whether a commercial concern or another scientist may pursue alternative licensing approaches through negotiating individual permissions, and the previous scientists may choose to grant such permissions, perhaps for a price. For most subsequent "small science" scholars and publishers, however, it would be far easier and less expensive to simply adhere to the conditions proposed here. Under the licensing choices above, the shift in legal burden promotes a general environment of openness to the underlying data used by scientists in pursuing and documenting their research.

While there are arguments for offering licenses with the additional restriction of non-commercial use only, we believe there is a serious drawback to this approach when applied to scientific datasets and that this requirement actually obstructs rather than forwards the goal of making currently invisible scientific data more widely available. The problem arises primarily in the ability to define what is or is not a commercial use. This restriction imposes a

need to evaluate every circumstance to determine whether a use might arguably be for a commercial purpose, and thus this requirement would strongly impede the general open use of the datasets.[6]

Finally, we take note that our public and university libraries are filled with innumerable "orphaned" works. While a researcher might find an interesting selection and arrangement of empirical observations in a table in a book from 1948 in a library, current copyright law across the globe states that as a general proposition you may not reproduce that table in your current publication without permission of the copyright owner. However, finding the valid rights holders in the work may be very difficult and expensive. The book may be out of print and no longer generating any publication income, yet current copyright law holds that you may still not extract creative content from it without permission. Making a reasonable effort to find the rights holders is no defense, and dependence on "fair use" ("fair dealing" in some other legal systems) is limited and unreliable in digital publication environments. If a scientist reproduces and publishes such a table without permission, the scientist could face very substantial statutory damages if a valid claimant later surfaces. This is a potential problem with the print content libraries. The licensing approach proposed above avoids perpetuating the orphaned works problem into our digital future.[7]

6 METADATA GENERATION

Generating even partial sets of metadata that conform to the standards in any scientific discipline usually requires substantial amounts of time and expertise. Most potential contributors to a CSTD will have neither. Yet the presence of standards-based metadata is essential for potential users of the data files to be able to find the data in the first place and then to make a judgment about whether the data may be suitable for the user's purposes. In addition, metadata makes it possible for data stored in one location to become visible to researchers around the world today through, for example, Open Archives Initiative (OAI) harvesting tools, and will be essential for the Semantic Web of tomorrow.[8]

Attaching standards-based metadata to data files is therefore crucial in a CSTD environment. The challenge is to design a way to enable contributors to attach metadata to their contributed files without forcing the contributors become experts in the metadata systems of the relevant discipline (e.g., ISO 19115 for geographic data). Many potential contributors would find such a barrier burdensome, and their potential contributions to the CSTD would therefore be lost.

A first step to generating standards-based metadata is to define a limited core metadata set that any file should have attached to it. The typical "small science" data contributor should be able to complete all required metadata fields for a data contribution in about ten minutes. The Dublin Core serves this purpose well.[9] While the 15 required fields in the Dublin Core standard will not be of great help for users seeking information about data in the context of a specific scientific discipline, the Dublin Core fields do capture basic information such as the creator of the data, relevant dates, a brief narrative description, etc. OAI harvesters are set up to capture data from these fields, and many discipline-specific metadata systems have created crosswalks between the Dublin Core and their own fields.

Dublin Core elements would be common to all datasets in a CSTD and would be automated, as much as possible. For example, the system would automatically generate the date the file was contributed as well as filling in other fields from information known to the system, such as previous contribution or registration information for the contributor or information that could be extracted from the data file itself, e.g., file type. Any automation would lighten the burden on contributors and reduce input errors and thus facilitate achieving the goal of making it possible for a contributor to complete all required metadata fields quickly (e.g., in ten minutes or less).

Another important step is to create a semi-automated infrastructure in which contributors would make initial selections of the type of data their files contained, and the system would then narrow the subsequent choices presented to the user. To use the example above of the ISO 19115 international standard for geographic metadata, a user would choose one of the 19 top level-themes (e.g., "inland waters," "transportation"), and the system would then present the user with subsequent choices relevant to that theme.

A third step is to make it possible for contributors to describe their data in their own words and then have the system do the crosswalk to controlled vocabulary terminology. After the contributor makes an initial selection of the type of data in the dataset and, if appropriate, a high-level theme relevant to the specific discipline, the contributor's choices

would trigger a "behind the scenes" selection of a relevant thesaurus by the system. The contributor would then be asked to submit common language terms that the contributor thinks best describe the data in the file. These terms could be any that the contributor chooses. If the contributor's submitted term matches a controlled vocabulary term, the contributor would be prompted with a definition to confirm that the mapping is accurate, i.e. matches the contributor's meaning. If no such controlled term exists in the relevant thesaurus, the user would be shown a cluster or "cloud" of possibly relevant controlled terms and definitions much as a user of popular tagging sites such as Flickr or del.icio.us or Connotea might be shown. The contributor would select the closest controlled term that reflects the contributor's meaning.[10]

Over time, the "cloud" of controlled terms shown to the contributor would become more precise based upon the system's experience with former contributors so that the mapping process would grow increasingly able to map contributor terms to controlled vocabulary terms accurately. Naturally, if a contributor were knowledgeable about standards based metadata for his or her discipline, that contributor could choose to bypass this semi-automated metadata mapping system and could simply choose a thesaurus and enter controlled vocabulary terms directly.

There are a number of research and design challenges, as well as practical ones, in implementing this type of semi-automated metadata infrastructure. Designing a system that can match meanings when only single terms are involved is no small task. Obtaining permissions to use a large number of thesauri is a practical challenge, as is providing the storage and processing speed necessary to respond quickly to contributor terminology choices.
The reward for meeting these admittedly difficult challenges, however, will be substantial: a growing distributed pool of "small science" and "besides science" data contributions free of use-inhibiting intellectual property conditions with enhanced findability and usability for users.

7 PROVENANCE TRACKING

In a commons environment, tracking of license provisions and of provenance requires marking of data files in a way that will travel with the file. In a CSTD, a unique encrypted identifier must be embedded in each submitted file to enable provenance tracking and license enforcement.

The identifier should not interfere with subsequent use of the file, nor should the identifier be able to be stripped from the file through standard reuse operations. This is not to say that someone focused on removing an identifier would not be able to do so. It simply recognizes the fact that the most realistic and affordable goal is to discourage license breakers rather than to try to enforce compliance completely. Getting credit "most of the time" is probably sufficient for most contributors to the commons. There is little incentive for those downloading contributed files to strip unobtrusive IDs, even if software becomes available to do so, because users may use the files freely anyway, and license infringers may still be identified.

A range of methods has already been developed for embedding encrypted IDs in the most commonly used file formats. To make the tracking system operational, open-source software would need to be employed that could automatically generate identifiers, encrypt them, and embed them in all of the primary formats of files likely to be contributed to a CSTD. Software would also need to be employed for identifying data files that have been processed through a CSTD. If a hidden commons identifier were detected in a file on a person's desktop through use of the free software, the core license provisions would be exposed and a link provided to the complete metadata file and license in the archive.

Similarly, when a file is uploaded by a contributor to the central server or a distributed network of servers, the system would automatically check to see if there is one or more hidden identifiers in the submitted file. If found, this would mean the submitted file was a derivative of other files previously processed by the system. Metadata fields would be populated automatically for the new file showing that it was derived from those other files, and direct links provided to the parent files. In this manner, any file could be traced back in time through the successive generations of other files that were used to construct it. This capability also should allow the automatic enforcement of license provisions through successive generations of use and would allow any future user to trace the provenance of a set of data back to its original files in order to examine how the data set was originally produced, for what purpose, its state of completeness, the geographic location of observations, and other such information.

8 ARCHIVING

Archiving will ensure a backup for CSTD data files and would constitute a major benefit for contributors because contributors will always be able to find and copy their previously submitted files from the long-term archive. A CSTD system should generate and make accessible several standard and interchange formats of each submitted file, all containing the hidden ID, so that future users will not need to accomplish such conversions. Providing several standard formats for a file also lessens the likelihood of loss or obsolescence of datasets over time as the popularity of some data formats changes or fades.

Metadata for files contributed to a CSTD would be made available for harvesting from the CSTD servers and would thereafter reside on many other servers around the world. Copies of the contributed files themselves should also reside on a number of different servers using a LOCKSS[11] approach so that a problem at one site, even at the primary CSTD site itself, would not endanger the collection of files contributed to the CSTD.

9 PEER REVIEW AND EVALUATION

Metadata reported by "small" or "besides" science data originators must, of course, be "truthful" and suitable for a user's purposes. Otherwise, gaining access to that data becomes useless.

The same problem has, of course, been an issue in the dissemination and use of factual data in other venues. It is already being addressed by a number of web services today that aggregate original information from many sources, and a variety of procedures for evaluating submitted information for both accuracy and usefulness have been developed. These include statistical methods, pledges of "neutrality" in contributing information, review by site founders, and post-publication peer-evaluation systems.

In the CSTD environment, the most promising technique appears to be post-publication peer-evaluation of contributed data. Post-publication evaluation systems in addition to their more established use as recommender systems at commercial websites such as Amazon.com, are beginning to appear for papers, publications, and scholarly web sites. For example, the PubMed Wizard at BioWizard allows everyone in the scientific community to rank and discuss all published literature in an open setting. Connotea (from Nature Publishing Group) enables users to tag bookmarks to papers and sites on the Internet and to share those papers or sites with others.[12] This type of post-publication peer-evaluation format could also be employed in a CSTD infrastructure.

One promising peer-evaluation method for assessing the reliability and suitability of data for a particular purpose (and for ferreting out inaccurate metadata reporting, whether purposeful or otherwise) is to use peer evaluation methods similar to those developed by the Open-source Development Network, which operates the web site www.slashdot.org. In this model, rather than formally select or financially support editors or other "quality control evaluators," everyone in the entire community of data users becomes a potential evaluator. This general methodology for quality evaluation has worked well in online endeavors with users who are literate in the subject matter and is currently being explored by scientific publications such as PLoS One.

A CSTD infrastructure should enable users to download data files, examine them, try them for their specific purposes, and rate and comment on the data in the context of suitability for their purposes, whether for scientific purposes or otherwise. User comments should be linked to the data files, and a potential user should be able to bring up all comments made about a particular data set or file to help him or her decide if the data set would be suitable for the user's purpose. Only registered users who have actually downloaded a file would be able to offer comments to reduce the likelihood of irrelevant comments or spurious evaluations.

As in all aspects of a CSTD infrastructure design, the peer-evaluation system should adhere to the "10 minute rule" – if it takes longer than ten minutes, people will not do it. With that in mind, the peer evaluation system might ask the user/evaluator to provide:

- a numerical rating on a scale of 1 to 5 (or some other commonly used scale);
- the purpose for which the user used the data set, possibly using a drop-down menu; and
- the user's specific positive and negative comments about the data's suitability for the user's chosen application.

This minimal design would make offering evaluations simple enough to be attractive to current users of a particular dataset while still providing future users with enough information for them to decide whether or not to examine the data file for their particular purposes.

10 CONCLUSION

Data generated through "small science" and "besides science" constitute a potentially valuable, currently largely invisible source of future scientific progress. We believe that a commons environment can be designed that will allow scientists and others who wish to do so to make their data available for use by others, without requiring users to seek prior permission for use as long as any conditions placed upon use of the data are adhered to. By designing an infrastructure that: (1) addresses the core concerns of licensing, metadata generation, provenance tracking, archiving, and peer evaluation; and (2) makes it simple for contributors to contribute and users to use data sets, we believe that a treasure trove of currently invisible or partly visible data will become available for others to evaluate and build upon to increase our knowledge and understanding of the world.[13]

11 NOTES

1. USC Title 17, Sec. 102(b).

2. Feist Publications, Inc. v. Rural Tel. Service Co., 499 U.S. 340 (1991).

3. Ibid.

4. Creative Commons has done a great deal of work to design licenses that offer a variety of "some rights reserved" conditions. The license approaches described here are based on that work. See http://creativecommons.org/about/licenses/meet-the-licenses.

5. For an easily digested explanation of why such permissions environments are destructive to the advancement of art and science, see Keith Aoki, James Boyle and Jennifer Jenkins, Bound by Law, 2006 http://www.law.duke.edu/cspd/comics.

6. For a more complete discussion of licensing choices in a commons environment and the drawbacks of restrictions on subsequent use through non-commercial licensing or through restrictions on applying different licenses to subsequent uses of data sets, see the White Paper on Licensing Options at http://geodatacommons.umaine.edu/about.

7. "Orphaned works" have become such a problem that the Registrar of Copyright called for comments on how to address the question and issued a set of recommendations for Congressional action to remedy the problem. See Registrar of Copyrights. 2006. Report on Orphan Works. Washington: United States Copyright Office.

8. For information on the Open Archives Initiative (OAI), see www.openarchives.org. For information about activities to promote standards for the Semantic Web, see http://www.w3.org/2001/sw/.

9. For a list of metadata fields included in the standard, see http://dublincore.org.

10. For an example of how this might work in the case of geographic metadata generation see the White Paper in Metadata Generation at http://geodatacommons.umaine.edu/about.

11. LOCKSS ("Lots of Copies Keeps Stuff Safe") is an approach to preservation of digital data based upon Thomas Jefferson's "multiplication of copies" idea. See http://www.lockss.org.

12. See http://www.biowizard.com and http://www.connotea.org.

13. For an example of how such a commons might be constructed in practice, see http://geodatacommons.umaine.edu, a site at which the initial stages of constructing a Commons of Geographic Data are being pursued with funding from the Institute for Museums and Library Services (IMLS). For a more detailed view of the visions for such a commons, see Onsrud, Harlan, Gilberto Camara, James Campbell, and Narindi Sharad Chakravarthy. Public Commons of Geographic Data: Research and Development Challenges. In *Geographic Information Science*, edited by Max J. Egenhofer, Christian Freska, and Harvey J. Miller, 223–38. Berlin: Springer-Verlag, 2004. Lecture Notes in Computer Science #3234.