III International Conference on Geographical Information Science (GIScience 2004). Washington, 2004.
Lecture Notes in Computer Science, vol. 3234, 2004.

Public Commons of Geographic Data: Research and Development Challenges

Harlan Onsrud[1], Gilberto Camara[2], James Campbell[1], and Narnindi Sharad Chakravarthy [3]

[1] Department of Spatial Information Science and Engineering
5711 Boardman Hall, University of Maine
Orono, ME 04469-5711
onsrud@spatial.maine.edu, campbell@spatial.maine.edu
http://www.spatial.maine.edu/geodatacommons
[2] Image Processing Division, National Institute for Space Research
Av. Dos Astronautas, 1758 –12227-001
São José dos Campos , SP, Brazil
gilberto@dpi.inpe.br
[3] GlaxoSmithCline
North Carolina, USA
narnindisharad@hotmail.com

Abstract. Across the globe individuals and organizations are creating geographic data work products with little ability to efficiently or effectively make known and share those digital products with others. This article outlines a conceptual model and the accompanying research challenges for providing easy legal and technological mechanisms by which any creator might affirmatively and permanently mark and make accessible a geographic dataset such that the world knows where the dataset came from and that the data is available for use without the law assuming that the user must first acquire permission.

1 Introduction

Geospatial data analysts require as much data as possible about geographic features to make informed judgments about their "meaning" in a particular frame of reference. While automated systems may queue satellite images or sensor data and identify potentially interesting selections for analysts to focus on, the analyst must take those queued images and put attributes to them in order to make sense of them and place their meaning in a larger context.

No matter how elegant an aerial or satellite image might be, it can only show, for example, the physical presence of power lines, not what the attributes of those lines are in terms of age, carrying capacity, interconnection links, where they run underground, or other non-visual data. An aerial photo may show a house but won't show its assessed value, the age of the roof shingles or the number of inhabitants. In short, geographic imagery requires geographic attributes to become fully useful.

How does the analyst quickly find the “on the ground” geographic attribute data corresponding to an area depicted in an aerial or satellite image that enables the analyst to complete an assessment? Obtaining access to data appropriate to the question at hand is often difficult, at best. Yet, across the globe, local governments, small companies, non-profit organizations and individuals often generate detailed local geographic information. These parties, however, seldom expend the significant effort required to make that information available to others. It often sits on a local server, unknown outside of the organization and effectively hidden from anyone else.

The goal of the Public Commons of Geographic Data, using open-source and open-access technology, is to remove technical and legal barriers facing the tens of thousands of GIS users (e.g. researchers, local government agencies, nonprofit organizations, field scientists, and individual citizens) that wish to contribute, access, and use locally generated geographic information. This approach has the potential to help free up currently unavailable information generated by non-federal and non-professional sources, and make it available to the widest possible range of potential users. Although not all local governments, private companies, non-profits or individuals will want to provide access to any or all of their geographic data files in a “commons licensing” environment, more people will participate once a user-friendly capability is available. The historical development of the web itself demonstrates that fact.

The “public commons” incorporates both *public domain* and *open access* works. The body of scientific and technical data currently within the *public domain* is significant and the ability of researchers and others to freely use this material has contributed to the economic, social, cultural, and intellectual vibrancy of the entire world [1], [2], [3], [4], [5]. Geographic resources in the public domain are comprised of geographic data and information provided by U.S. federal government agencies which cannot, by law, hold copyrights; information which may have once been copyrighted but on which the copyright has expired; information which is not subject to copyright, e.g., facts; and material affirmatively placed in the public domain by its creators which would otherwise have been subject to copyright. Works within the public domain are completely free of any intellectual property restrictions.

Open access works, while still copyrighted, also allow use without obtaining prior permission since a general license is granted ahead of any specific use, provided any attached conditions of use are met. Open-access works typically invoke copyright law and licensing restrictions to help ensure that they remain freely available. Thus, software, data files, and journal articles, for example, distributed under open-source or open-access licenses contribute to the "public commons" but are not by typical legal definition within the "public domain." Examples of such licenses include the General Public License (GPL) (http://www.gnu.org/copyleft/gpl.html), Creative Commons licenses (http://www.creativecommons.org), and the Public Library of Science Open-access License (http://www.publiclibraryofscience.org).

A primary goal of the Public Commons of Geographic Data is to create a broad and continually growing set of freely usable geographic data and information products (i.e. no monetary charges for data use) similar in effect to the public domain data sets and works created by federal agencies. The overarching objective is to provide easy legal and technological mechanisms by which any creator may affirmatively and permanently mark and make accessible a geographic dataset such that the world knows

where the dataset came from, and that the data is available for use without the law assuming that the user must first acquire permission.

National governments throughout the world are involved in developing spatial data infrastructures that will better facilitate the availability and access to spatial data for all citizens. A key premise in most of the initiatives is that national governments, including the government of the United States, will be unable to gather and maintain more than a small percentage of the geographic data that users want and need in the digital age. Thus, it will become increasingly important to overcome obstacles and construct ways for non-federal geographic data providers who wish to do so to make their data available to the public.

For researchers, nonprofit organizations, citizen groups, local government bodies, and others who collect and use geographic information, the implementation of a Public Commons of Geographic Data could remove many obstacles they currently face in sharing the geographic data and information they have produced, and in gaining access to the information others have produced.

2 Implementation Objectives

Many who generate digital geographic information would be more than willing to make their spatial data sets and information freely available, if (1) creating metadata was much easier to do, (2) creators could reliably retain credit and recognition for their contributions to the public commons, (3) creators could acquire substantially increased liability protection from uses by others, (4) creators could reap benefits such as having their data evaluated by peers and made "visible" and widely available to potential users, and (5) creators could have their data stored in a long term archive they would not have to maintain. We propose a conceptual model for a public commons that addresses all of these impediments.

The envisioned system, implemented through open access content support and open-source software, should:

- enable simple straightforward construction of contributor-defined open-access licenses using a check-box system suitable for use even by non-professionals,
- enable simple, straightforward construction of machine readable, standards compliant metadata using a menu driven system suitable for use even by non-professionals,
- allow non-removable identity information to be embedded in contributed files,
- track data lineage and improve the ability to find data meeting specific criteria, and
- provide access to a powerful peer-based evaluation system that is simple to use.

From the perspective of a searcher for data, the commons database and search software should be designed to allow a user to (a) locate data for a spatial region and content of interest (b) avoid or solve data formatting and semantic translation prob-

lems, and (c) obtain detailed explanatory information about found data. Ultimately the system should also support users in extraction of subsets of information from files contained in the Commons.

To illustrate how the Public Commons of Geographic Data could interact with a person desiring to contribute data to or search the Commons, a mock-up of an interaction session with the Commons is available at http://www.spatial.maine.edu/geodatacommons (See Figure 1). Steps that users would go through in contributing data to the Commons are also summarized in Table 1.

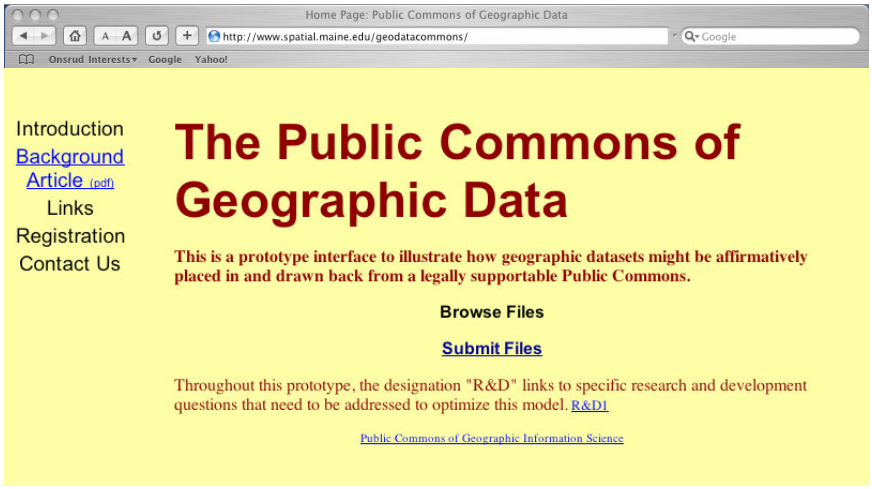

Fig. 1. Public Commons of Geographic Data Web Mock-Up

In the conceptual model we assume that the typical data contributor, although perhaps an expert in another domain such as epidemiology or ecology, is unlikely to ever become an expert in geographic information technologies or geographic metadata creation. Further, we assume that the typical contributor has gathered digital data from several existing databases or other digital sources, collected some of her own field data, completed an analysis, and now desires to make the resulting digital geographic work product available to others.

Table 1. Conceptual Model for a Public Commons of Geographic Data: Operational Characteristics

1. A non-expert user creates a GIS data set or a dataset locatable in space that he or she wants preserved and accessible to the rest of the world.

2. The user accesses a web site that automatically generates an open access license and facilitates the creation of a metadata record in response to a web interview transcript.

(a) *Open Access License Creation* - In responding to the transcript, the contributor agrees to (1) dedicate the file to the public domain or (2) choose among a limited selection of "open access" license provisions to apply to the data set. The basic concept of an "open access" license is that any subsequent user may freely use the data file without asking for permission yet the license also can ensure that (1) the originator and all value-adders have a legally enforceable right to credit for their work, (2) liability exposure for the data contributor may be substantially reduced, and (3) the efforts of the originator and value-adders may be protected from capture as the intellectual property of others.

(b) *Metadata Creation* - The metadata record is created in semi-automated fashion. The user is walked through a series of questions with limited choice responses. Portions of the transcript form are automatically filled in based on previous responses provided through the user registration and license creation processes. Other portions of the transcript automatically change depending on responses to initial questions. That is, the system guides the user by asking the data contributor to select among a fixed set of definitions for some of the terms the user selects. Those definitions, along with ontologies appropriate to the primary theme designated for the data file by the contributor, are then used to predict and simplify subsequent metadata selection choices.

This is a very different approach from current metadata approaches that have been designed for flexible use by specialists. Non-expert users probably will never take a course in how to create metadata for geographic data files, nor are they likely to have familiarity with many technical geographic terms. Therefore, open-ended questions with free-form responses need to be minimized. Yet the system also needs to be responsive to a variety of disciplines using the language and classification schemes of those disciplines.

3. The transcript responses and the actual data file to which the responses apply are submitted to an automated processing facility. An encrypted identifier is automatically embedded in the geographic data file. The identifier does not interfere with the file nor is it stripped from the file through standard GIS operations. Through the availability of freely downloadable client software, any user may readily determine the status of legal rights and metadata for any standard format geographic data file that the user has in her possession. This approach varies from the current commercial approaches in which metadata is maintained separate from the data file. If properly designed, the originator and the string of value-adders to a data set would always be known when a file is processed in this manner. Appearance of the identifier information would provide legal evidence that a user is allowed to use the file in accordance with the license provisions without impinging on intellectual property rights.

4. The system would return a copy of the "marked" geographic data file back to the originator incorporating the embedded metadata information. In an optimal system, all files so processed also would be permanently and publicly archived. Whether maintained on the open web or maintained in a long-term commons electronic archive, anyone would be able to search for, access, and legally download and use any such data sets.

3 Research and Development Challenges

Several research and development challenges must be addressed in order to move from concept to effective implementation.

3.1 Intellectual Property

Open-access seeks to clarify the legal status of digitally available works by enabling creators who choose to do so to make an affirmative statement that they are allowing access to and use of their work under only some (or no) conditions of current copyright law without requiring further permission for use on the part of the user.

Today, there are many gray areas in what a user may or may not do with copyrighted works in the digital domain. Current U.S. Copyright law does not make it possible to copyright facts or even obvious arrangements of facts such as an alphabetical listing in a telephone directory (Feist Publications, Inc. v. Rural Tel. Service Co., 499 U.S. 340 (1991)). However, the threshold for “originality” that would make arrangements of facts (such as raw geographic data) copyrightable is very low. Thus the typical user must assume that an interest of another may exist in the vast majority of geographic data compilations openly found on the web or elsewhere, and the law thus presumes that permission must be acquired even if the compilations are mostly factual. This presumption will become even stronger if, as many are predicting, the U.S. moves closer to the database protection regimes enacted in the European Union and advocated by the World Intellectual Property Organization [6]. Providing a way for local geographic information originators to affirmatively state that their work is open-access will eliminate any present or future doubts as to its status in the eyes of potential users.

The geographic data commons concept extends from the open-source licensing model. Currently, the law assumes that geographic data creators have all proprietary rights (e.g. copyright) in the data sets they produce. A common alternative to this approach is to place the data in the public domain with no rights reserved. Emerging open access licensing approaches, derived from the open-source licensing model, provide a middle ground that allows access and use of data for wide-ranging productive purposes but with “some rights reserved.”

The most prevalent current open access license approach, which is the one we advocate for use with the geographic data commons, is that developed by the Creative Commons project (http://www.creativecommons.org). Through this middle ground approach, data producers are able to specify whether future users must provide attribution, are allowed to modify the data, are allowed to modify the data as long as the users apply the same license to any derivative works, or may not use the data for commercial purposes. With the exception of these possible constraints, users are granted affirmative permission to use data drawn from the commons.

An interesting and perhaps critical aspect of the open access license model is that data and product producers have the option, if they so choose, of charging for the service of transferring their work to others and charging for support services. That is, many parties are generating substantial revenues by making their works available through open access and open source licensing approaches. This licensing approach is viewed as supporting a relatively new mode of economic production where individual contributors are organized neither in response to price signals nor by explicit firm managers [7]. However, Adam Smith's more traditional notion of "enlightened self-interest" aptly describes the motivations of many businesses and individuals contributing to open source and open access efforts. Thus such licensing models may be viewed as supporting basic free-market principles whereby claims of property right are used to distribute work efforts in furtherance of competition, creativity and enterprise. For certain products and parties this new form of production works well and even better than price signal or hierarchical management arrangements. In other instances, traditional means of marketplace production are likely to remain more efficient.

One core area for research investigation is whether the current semi-automated licensing options used by the Creative Commons project might be improved to be more responsive to the needs of the scientific and technical community. Further, what are the conditions and limits under which open access economic models for supplying geographic data succeed or fail relative to competing models? Answering this second research question will require first the development of an operational Public Commons of Geographic Data.

3.2 Metadata Generation

One of the major barriers for non-specialists who wish to offer their data to a larger audience is the generation of metadata. There are currently many competing metadata standards in use [8] and even professionals have difficulty staying current with them. In addition, using any current metadata system, for example the ISO 19115 Metadata Standard (ultimately ISO 19139 in XML), requires systematic study and practice. Non-professionals in geographic information, no matter how competent in their own areas of expertise, hesitate to wrestle with any of the current geographic metadata systems, and even many GISci professionals find metadata generation burdensome and do as little as possible. Currently, in fact, metadata fields typically are minimally populated, and there is a lack of depth in the meanings of the data submitted.

The ISO 19115 standard will soon replace the Federal Geographic Data Committee Metadata Content Standards [9], one of the most widely used metadata systems. Both the FGDC and ISO standards are geared toward professionals. Historically the FGDC system has been not fully utilized by local professionals and has almost never been used by non-professionals.

The Public Commons of Geographic Data speaks directly to this problem by creating a minimal metadata set and options for extending it to the full ISO 19115 standard. The Dublin Core specifies 15 elements that must be included to conform to its standard [10]. Most of these also map to a subset of the current FGDC standard and the new ISO 19115 standard (see http://www.spatial.maine.edu/geodatacommons/metadatadublin.html). Using the Dublin Core elements as a minimal set for metadata generation, it would be possible to provide sufficient metadata for geographic information to make it accessible to today's search engines as well as to the semantic web search tools of the future which will be based on XML, and which will recognize and parse Dublin Core elements.

With this in mind, the Public Commons of Geographic Data should incorporate ways to generate ISO compliant metadata that meets Dublin Core standards. This should be accomplished using pull down menus and other user-interactive "choose-one" techniques that will make it reasonably simple for non-professionals in geographic information to generate usable metadata without taking a course in how to do so. Professionals may also choose to utilize the proposed system in generating metadata for ease of use in populating all of the ISO 19115 fields.

As users of the system select terms and affiliated definitions appropriate to their data from pull-down menus, the Public Commons of Geographic Data system will need to be able to provide subsequent branching menus based on the "meaning" of the previous user input. We hypothesize that selected responses in a pull-down menu by a respondent may be used along with formal specifications for potential domains of interest (i.e. ontologies) to predict and simplify metadata choices. That is, existing ontologies may be affiliated with each ISO 19115 data topic category (e.g. MD_DataIdentification.topicCategory). Provision of menu choices that change based on earlier choices should speed up metadata creation for infrequent contributors and make the typical completion of metadata much more comprehensive. If thousands of users make pull-down menu choices according to an initial ontology, e.g., for "transportation," we hypothesize that those responses may be used to automatically develop an improved ontology that reflects the primary understanding and usage of the community, as opposed to reflecting the logic of classification specialists [11]. This adaptive ontology then may be used to continuously optimize the system for future metadata submissions by the community.

To make the commons of greatest use across a variety of domains, occasional users should be encouraged, but not required, to complete more comprehensive metadata fields corresponding to the data topic categories they have selected. For instance, if contributors selected "biota" under the ISO19115 data topic category, they might be led to complete the remaining ISO19115 elements using a broad vegetation metadata profile (e.g. FGDC-STD-005). However, if they further selected under "biota" a subcategory such as "wetlands" they might be led to a different selection of pull-down menu options based on the terms and classifications used in developing metadata, for

example, by the U.S. National Wetlands Inventory. Similarly, if they had chosen under "biota" a further subcategory of "flora" they might be led to complete the remaining ISO19115 elements using the classifications established by the "Darwin Core" element set. The goal is to federate the system across disciplines, and thus be responsive to the widest range of potential contributors of geographic data files. The additional depth of documentation and meaning included in the metadata should then contain sufficient text to allow inferences to be made in future semantic web environments. To be effective, the system must be designed in such a way that each additional request for information should extend for no more than a page, and take only a few minutes for the typical user to complete.

In order for a pull-down menu system to work, there will need to be a powerful dictionary underpinning it. At present, there is no standard dictionary for geographic information suitable for this use, although there are efforts underway that may be adaptable. In Scotland, the Association for Geographic Information [12] in collaboration with the University of Edinburgh maintains an online GIS Dictionary. Several commercial providers, such as ESRI, also offer dictionaries. The Alexandria Digital Libraries [13] project has developed a Thesaurus with 210 preferred terms and 946 non-preferred terms (non-preferred terms refer the user back to the preferred term e.g., "ditch" refers the user to "canal"). Use of controlled vocabularies such as WordNet (http://www.cogsci.princeton.edu/~wn/) may have application when terms are specified by contributors that are not contained in the standard geographic dictionary. All of these initiatives offer elements that may be able to be adapted for a commons environment, and the limited scope of the ADL Thesaurus indicates that the scope of this task is manageable. Today's methods for finding and using information on the web are often insufficient. Yet, if semantic web methods are to be able to draw inferences from text, such as the text in metadata, that metadata must exist in the first place, and be at a level of detail far greater than is currently being provided. Further, for the "Spatial Semantic Web" to reach its full potential, automated searches must be able to reach and explore actual geographic data sets [14], [15], [16] as well as their metadata. Without full access to the data set, data semantics cannot be used to find and assess the suitability of a geographic data file for an explicit need. Additionally, searches that rely on data similarity assessments require access to the data rather than just metadata.

A further concern is that metadata entry must be extremely efficient for the occasional contributor of data sets. For example, the typical user of local level geographic data does not know the bounding latitude and longitude coordinates of the data set they are using. Many local geographic data sets throughout the world are not tied to universal coordinate systems. Therefore an efficient tool should be supplied to provide the approximate bounding coordinates for the data file. For metadata and search purposes, the bounding coordinates do not need to be precise. The user should be able to type in the name of the location of concern, be presented with an image centered on the coordinates of the place, and then zoom in or out on a high resolution image to allow a box to encompass the area of concern as precisely as possible. The bounding coordinates of the drawn box should then be used to automatically populate the required metadata fields for coordinates. This online process must take less than a minute to avoid frustrating contributors. The National Map Viewer (http://nmviewogc.cr.usgs.gov/viewer.htm) and the Alexandria Digital Library Gazet-

teer (http://fat-albert.Alexandria.ucsb.edu:8827/gazetteer) partially illustrate this capability.

Research questions include: (1) Will responding to the Dublin core set of elements as the minimum set take too much time to elicit widespread responses from the broad user community? Would this element set or fewer elements provide insufficient information for effective future searches? (2) We hypothesize that provision of menu choices that change based on earlier choices should speed up metadata creation for infrequent contributors and make the typical metadata entry far more comprehensive. Will this prove true in practice? (3) We hypothesize that initial ontologies for specific domains or data themes may be automatically revised through thousands of submissions to reflect the primary understanding and usage of the community, and that this adaptive ontology might then be used to continuously optimize the system for future metadata submissions by the community. What specific approaches might be used to promote enhanced efficiency for individual users, users reporting metadata within a specific domain, and for all users on average?

In summary, once users are familiar with the commons interface, it should take them only a few minutes to create a license, complete an accurate and sufficient metadata script, and submit their geographic data file. An initial user interface mock-up is available at http://www.spatial.maine.edu/geodatacommons.

3.3 Tracking Data File Lineage.

In a commons environment, tracking of license provisions makes more sense than controlling access by the methods of more traditional Intellectual Property Rights Management Systems [17], [18], [19]. A unique encrypted identifier would be embedded in each submitted file but should not interfere with subsequent use of the file nor should the identifier be stripped from the file through standard GIS operations. Current commercial GIS software systems do not provide this capability. The goal is to discourage license breakers, but not ban them. Getting credit "most of the time" is probably sufficient for most contributors to the commons. There is little incentive for those downloading to strip unobtrusive IDs, even if software becomes available to do so, since users may use the files for free anyway and license infringers may still be identified if contributors additionally use the more traditional methods of embedding false objects or watermarks in their files.

A range of methods have already been developed for embedding encrypted IDs in the most commonly used file formats, including raster files [20]. At least one vector steganography approach shows great promise as well [21]. To make the tracking system operational, open-source software should be developed to embed identifiers in all of the primary formats of files likely to be delivered to the commons (see http://www.spatial.maine.edu/geodatacommons/toload.html for a sample listing). Using known techniques, it is typically possible to embed numerous copies of an identifier throughout a single geographic data file so that even if only a small portion of a large file is extracted for use, that small part will continue to carry the ID in most instances. Thus, such methods would be used to automatically generate ID's, encrypt them, and embed them in any file delivered to the commons.

Software would also need to be developed for identifying data files that have been processed through the Public Commons of Geographic Data. If a hidden commons identifier is detected in a file on a person's desktop through use of the free software, the core license provisions are exposed and a link is provided to the complete metadata file and license in the archive.

Similarly, when a file is uploaded by a contributor to the central server, the system checks to see if there is one or more hidden identifiers in the submitted file. If found, this means the submitted file is a derivative of other files previously processed by the system. Metadata fields would be populated automatically for the new file showing that it is derived from those other files and direct links will be provided to the parent files. In this manner, any file may 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 certain license provisions, such as the "share alike" or "copyleft" provision, through successive generations of use.

A frequently suggested alternative to the identifier tracking system just described is the technique of "hashing." Hashing transforms a string of characters into a shorter fixed length value or key that represents the original string. While most frequently used as a technique for increasing the efficiency of recalling information from a database or in implementing encryption processes, hashing may also be used as a check on similarities among files. Hashing is insufficient for the system envisioned since, for the most part, all hashing can indicate is that a file is likely to have a derivative or ancestry relationship to another file but cannot adequately resolve which of the files came first in time. Hashing might be used however to make the file lineage system more robust. In light of the large number of standard file formats for geographic data and the range of useful steganographic approaches available, the most optimal technical means for file lineage tracking is still an open question.

3.4 Archiving

Archiving ensures a back-up for commons licensed data files and is a major benefit for contributors since contributors will always be able to find and copy their previously submitted files from the long term archive. Data files for the commons would otherwise be distributed among thousands of machines that inevitably are subject to broken links and lost data over time. Similar to CiteSeer (www.citeseer.com), we envision that the 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 data sets over time as the popularity of some data formats wane.

The Creative Commons web site already has a reference repository in place in which creators may list their work that is available under a Creative Commons license. Sites such as Geospatial One-Stop (http://www.geo-one-stop.gov) and GIgateway (http://www.askgiraffe.org.uk/metadata) provide repositories and search capabilities for metadata for geographic data files. The Public Commons of Geographic Data system should automatically submit references of all files contributed to it to the Crea-

tive Commons repository and provide a link back to the Public Commons of Geographic Data metadata archive. Similarly the commons should automatically forward its metadata files, if allowed, to other major geographic metadata file depositories. Thus, in addition to accessing directly the Public Commons of Geographic Data metadata and geographic data repositories, potential users of open-access geographic information could have several other entry points for finding datasets available in the commons.

3.5 Data Storage and Search Optimization

Primary contributors to the commons are envisioned as individuals and organizations that produce geographic data sets only occasionally or for limited areas. As such, storage capacity limitations and file transfer times across networks should be less problematic than faced by many current data centers handling very large files. It is assumed that organizations managing large databases would continue to manage and archive them in-house. Further, while organizations like NASA have a strong need for capabilities that would allow extraction and transfer of portions of large files, such a need would be less critical for a commons environment. If such capabilities eventually emerge, they should be incorporated in the commons system environment but are unlikely to be critical to its success.

Assuming that an efficient location specification tool can be supported for contributors in creating their metadata, this opens the possibility of organizing and storing both geographic metadata and geographic data sets by their spatial extents on one or numerous servers. For example, Figure 2 illustrates the current decentralized and distributed FGDC Clearinghouse Metadata Model Approach. Under this configuration metadata is not stored by location and therefore a typical query must search all nodes on the network, which may number in the thousands, in order to be comprehensive. That is, a server in Sao Paulo Brazil might contain metadata and associated geographic data files pertaining to a location in Maine and vice versa. Figure 3 suggests an alternative arrangement for storing metadata gathered through the Public Commons of Geographic Data system. Depending on the bounding coordinates provided in the metadata and the size of geographic area encompassed by those coordinates, metadata could be automatically categorized and stored by geographic location (e.g. assume organization through equal size grid cells covering the globe), extent of coverage (i.e. this would result in approximate groupings of files containing images or maps of similar scales), and by primary topic (i.e. theme) of the file. Searches germane to any specific bounded region might be made much more efficient through this arrangement. A further alternative might be to maintain distributed servers similar to the current

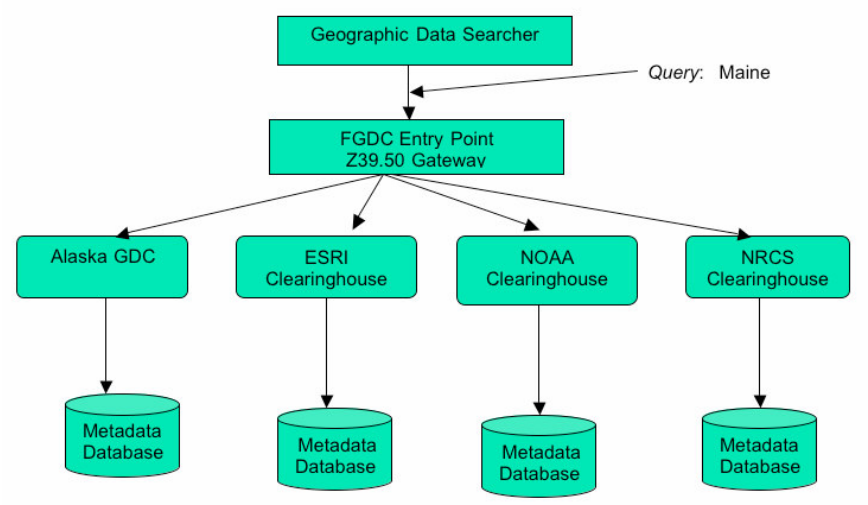

Fig. 2. FGDC Clearinghouse Metadata Model Approach

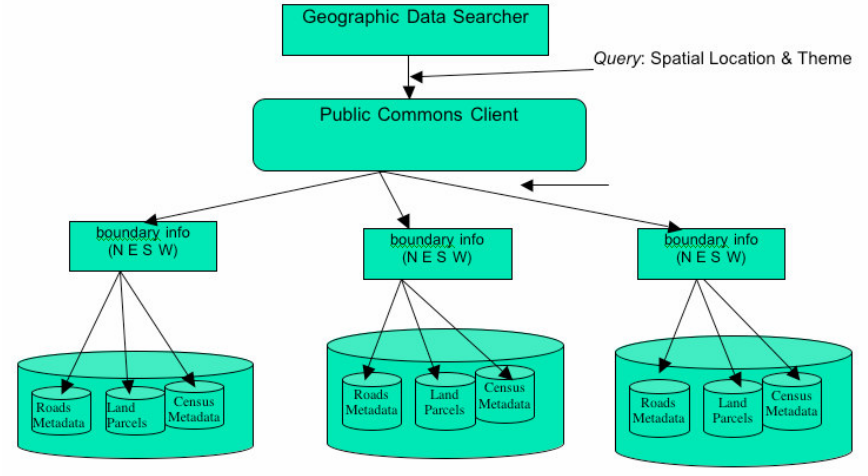

Fig. 3. Potential Public Commons Metadata Model

FGDC Clearinghouse node arrangement but provide a comprehensive centralized metadata server capability that mines metadata regularly and efficiently from all other metadata nodes and mirrors back the comprehensive collection to selected distributed full metadata sites around the globe. Determining which distributed architecture would be most efficient for serving metadata as well as the actual geographic data files with embedded IDs is a significant research question.

3.6 Peer Review and Evaluation

Metadata reported by local geographic data originators must, of course, be "truthful." Otherwise, the concept of gaining access to their data through metadata becomes dysfunctional or useless. The data, too, must be suitable for a user's purposes.

The same problem faces a number of web services today that aggregate original information from many sources, and a variety of procedures for evaluating submitted information have been developed. These range from statistical methods [22], to pledges of "neutrality" in contributing information [23], to review by founders [24], to post-publication peer review systems [7]. One promising peer-review method for assessing the reliability of reported data and for ferreting out inaccurate metadata reporting, whether purposeful or otherwise, is to use peer review methods similar to those developed by the Open-source Development Network which operates the web site www.slashdot.org. In this model, rather than hand pick or financially support editors or other "quality control evaluators," everyone in the entire community of data users becomes a potential evaluator in the quality control effort. This general methodology for quality control has worked well in online endeavors with users who are literate in the subject matter. The approach has promise as a good starting point for development of a Public Commons of Geographic Data system peer evaluation mechanism.

A further quality control issue relates to responsibility. Registration should be required for contributors in order to identify those purporting to have ownership or authority to place a specific geographic data file into the commons licensing environment. In the event of a conflict over rights in a specific file, the dispute would be primarily among claimants but administrators of the system would need to be able to be responsive to requests to remove a file until the dispute is resolved. The semi-automated license creation process includes a liability disclaimer clause and should sufficiently accommodate the liability exposure concerns of most data contributors. Thus, the primary remaining research challenge is to determine which of several alternative methods for assessing quality of data in on-line environments would work best for ensuring quality control of submissions contributed to the Public Commons of Geographic Data.

3.7 Governance

Several possibilities exist concerning governance and hosting of the proposed capability. The governance structure is likely to be closely linked to the technical design. With a design focused on centralized processing and storage facilities, the primary operations might be funded and administered by a single government agency or a non-profit organization set up for the purpose. More decentralized designs might rely more heavily on the server, network, and storage facilities already being provided and supported by public libraries, data archives, and government agency GIS operations spread across the globe. There is also the possibility that a parallel global marketplace in geographic data and services offering similar licensing, metadata, and tagging capabilities could be developed and tapped to support the ongoing operations and con-

tent expansion of the commons. The means for governing, supporting and expanding the public commons in geographic data is itself a source of numerous research questions.

4 Summary

One key fact emerges throughout all of the open-source and open-access work going on at present: in the early stages, projects have to be initiated, nurtured, and managed by a central team until a project is ready for release to the open-source/open-access community where it can effectively take on a life of its own [25], [26]. Software application models abound, including highly complex discipline specific endeavors that parallel the level of functionality of the proposed Public Commons of Geographic Data system. One such example is the Koha library automation system [27], and a number of others exist. The key to their success and, we believe, the key to the proposed Public Commons of Geographic Data system's success, is building the system to the point at which others can effectively expand upon a working model.

The intent of the Public Commons of Geographic Data is to be responsive to the tens of thousands of individuals currently creating geographic data on their desktops but who have little incentive or ability to effectively share their data sets. These individuals may want to create metadata and contribute files to a common pool only occasionally. As such, once they are familiar with the interface it should take contributors only a few minutes to create a license, complete an accurate and sufficient metadata script, and submit their geographic data file.

The significance of the conceptual model and its practical implementation is that ultimately it will provide a means to make visible a substantial body of geographic information that now exists but is effectively hidden from the view of geographic scientists and researchers, researchers in a wide range of other fields, agency analysts, nonprofit organizations, and private citizens. Looking forward, the Public Commons of Geographic Data will provide a vehicle for those who generate detailed local-level information in the future to provide access to the information they generate in a simple, "business as usual" way. The goal is to enable the sharing of locally generated geographic information to become the norm, rather than the exception that it is today, and to expand the amount and quality of locally generated data available for analysis and public use.

In sum, no complete analog to the Public Commons of Geographic Data system exists at present. Some proto-elements do exist. The development of the envisioned capability will require combining original research contributions (e.g., needed infrastructure specifics, simple metadata generation and ontologies, identifier embedding, etc.) with adapted processes or initiatives already in place (e.g., Creative Commons) to create seamless access to a multitude of independently developed, heterogeneous geographic data sources open to and usable by any interested citizen.

5 Acknowledgements

Research leading to this publication was supported by NURI grants NMA 202-99-BAA-02 and NMA 201-00-1-2003 with funds supplied by the U.S. Federal Geographic Data Committee.

References

1. National Research Council, Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest 1999, *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Wash, D.C. National Academy Press
2. Commission of the European Communities. 1999, *Public sector information: A key resource for Europe*. (COM(98)585 European Commission, Brussels) <http://europa.eu.int/ISPO/docs/policy/docs/COM(98)585/
3. Pira International Ltd., University of East Anglia and KnowledgeView Ltd. 2000, *Commercial exploitation of Europe's public sector information: Final report for the European Commission Directorate General for the Information Society*. Surbiton, Surrey, UK
4. Weiss, Peter. 2002, *Borders in Cyberspace: Conflicting Public Sector Information Policies and Their Economic Impact*, U.S. National Weather Service <http://www.weather.gov/sp/Borders_report.pdf>
5. Moglen, Eben. 2003, Freeing the Mind: Free Software and the Death of Proprietary Culture. In *Fourth Annual Technology and Law Conference*. Portland, Maine: UMaine Law School
6. Maurer, S., P.B. Hugenholtz, and H. Onsrud. 2001, Europe's Database Experiment. *Science*, 294: 789-790, 26 Oct 2001
7. Benkler, Yochai. 2003, Coase's Penguin, or, Linux and The Nature of the Firm. *Yale Law Journal* 112 (Winter 2002-2003)
8. Hill, Linda, Greg Janee, Ron Doulin, James Frew, Mary Larsgaard. 1999, Collection Metadata Solutions for Digital Library Applications. *Journal of the American Society for Information Science* 50 (13):1169-1181
9. Federal Geographic Data Committee (FGDC), 1999, Metadata Content Standards <http://www.fgdc.gov/metadata/contstan.html>
10. Dublin Core Metadata Initiative (DCMI), 2003 <http://www.dublincore.org/documents/dces>
11. Smith, Barry, and David M. Mark. "Geographical Categories: An Ontological Investigation." *International Journal of Geographical Information Science* 15 (2001): 591–612.
12. Association for Geographic Information (AGI), 2003 <http://www.geo.ed.ac.uk/agidict/welcome.html>
13. Alexandria Digital Library Project (ADL), 2003 <http://www.alexandria.ucsb.edu>
14. Egenhofer, Max. 2002, Toward the Semantic Geospatial Web, *ACM GIS 2002*, November 8-9, 2002, McLean, Virginia
15. Rodriguez, M. Andrea, and Max J. Egenhofer. "Determining Semantic Similarity Among Entity Classes from Different Ontologies." *IEEE TRansactions on Knowledge and Data Engineering* 15 (2003): 442–56.
16. Bateman, John, and Scott Farrar. "Spatial Ontology Baseline [Preliminary]." 18 May 2004. [cited 30 Jun 2004]. Available from <http://www.sfbtr8.uni-bremen.de/project.html?project=I1>.

17. IBM Electronic Media Management System (EMMS) <http:www-306.ibm.com/software/data/emms/features/>
18. Microsoft Windows Rights Management Services (RMS) for Windows Server 2003 <http://www.microsoft.com/windowsserver2003/technologies/rightsmgmt/default.mspx>
19. Freeman, Neil, Tony Boston and Arthur Chapman, 1998, Integrating Internal, Intranet and Internet Access to Spatial Datasets via ERIN's Environmental Data Directory, *The 26th Annual Conference of AURISA*, November 23-27, Perth, Western Australia
20. Sharad, Chakravarthy N. 2003, *Public Commons for Geospatial Data: A Conceptual Model, MS Thesis*, Department of Spatial Information Science and Engineering, University of Maine
21. Huber, W., Vector Steganography: A Practical Introduction, *Directions Magazine* April 18, 2002, <http://www.directionsmag.com/article.php?article_id=195>
22. NASA Clickworkers Project, available at <http://clickworkers.arc.nasa.gov/top>
23. Wikipedia, the Free Encyclopedia, available at <http://www.wikipedia.org>
24. Rivlan, Gary. 2003, Leader of the Free World. *Wired*, November, 2003
25. Raymond, Eric, *Homesteading the Noosphere*. 8/00 [cited 12/26 2003] (http://www.tuxedo.org/~esr>
26. ———. 2000. *The Cathedral and the Bazaar*. 9/00 [cited 12/31, 2003] (http://www.tuxedo.org/~esr>
27. Koha Open Source Library System, available at <http://www.koha.org>