

8-2003

Public Commons for Geospatial Data: A Conceptual Model

Chakravarthy Namindi Sharad

Follow this and additional works at: <http://digitalcommons.library.umaine.edu/etd>



Part of the [Databases and Information Systems Commons](#), and the [Geographic Information Sciences Commons](#)

Recommended Citation

Sharad, Chakravarthy Namindi, "Public Commons for Geospatial Data: A Conceptual Model" (2003). *Electronic Theses and Dissertations*. 577.

<http://digitalcommons.library.umaine.edu/etd/577>

This Open-Access Thesis is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DigitalCommons@UMaine.

PUBLIC COMMONS FOR GEOSPATIAL DATA

A CONCEPTUAL MODEL

By

Chakravarthy Narnindi Sharad

B.Tech. Jawaharlal Nehru Technological University (JNTU), Hyderabad - India, 2000

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

August, 2003

Advisory Committee:

Harlan J. Onsrud, Professor of Spatial Information Science and Engineering, Advisor

Kate Beard-Tisdale, Professor of Spatial Information Science and Engineering

Anthony Stefanidis, Assistant Professor of Spatial Information Science and Engineering

PUBLIC COMMONS FOR GEOSPATIAL DATA

A CONCEPTUAL MODEL

By Chakravarthy Narnindi Sharad

Thesis Advisor: Dr. Harlan J. Onsrud

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Spatial Information Science and Engineering)
August, 2003

A wide variety of spatial data collection efforts are ongoing throughout local, state and federal agencies, private firms and non-profit organizations. Each effort is established for a different purpose but organizations and individuals often collect and maintain the same or similar information. The United States federal government has undertaken many initiatives such as the National Spatial Data Infrastructure, the National Map and Geospatial One-Stop to reduce duplicative spatial data collection and promote the coordinated use, sharing, and dissemination of spatial data nationwide. A key premise in most of these initiatives is that no national government will be able to gather and maintain more than a small percentage of the geographic data that users want and desire. Thus, national initiatives depend typically on the cooperation of those already gathering spatial data and those using GIS to meet specific needs to help construct and maintain these spatial data infrastructures and geo-libraries for their nations (Onsrud 2001).

Some of the impediments to widespread spatial data sharing are well known from directly asking GIS data producers why they are not currently involved in creating datasets that are of common or compatible formats, documenting their datasets in a standardized metadata format or making their datasets more readily available to others through Data Clearinghouses or geo-libraries. The research described in this thesis addresses the impediments to wide-scale spatial data sharing faced by GIS data producers and explores a new conceptual data-sharing approach, the *Public Commons for Geospatial Data*, that supports user-friendly metadata creation, open access licenses, archival services and documentation of parent lineage of the contributors and value-adders of digital spatial data sets.

ACKNOWLEDGMENTS

This thesis would have been impossible without the support and guidance of many people. I take this opportunity to extend my sincere gratitude to all.

First, I gratefully acknowledge my advisor, Dr. Harlan J. Onsrud, whose enthusiasm and guidance was crucial for the progress of my thesis. His generous cooperation and encouragement helped me accomplish the successful completion of this thesis.

I would also like to thank my other committee members, Dr. Kate Beard and Dr. Anthony Stefanidis, for their support and guidance. Many thanks also to the staff in the SIE department in their tireless efforts.

Thanks to all my colleagues in the SIE department, especially Anuket, Hari and Chitra for their many stimulating discussions. Thanks to my all friends at the University who made my study here a memorable one especially Anuket, Bitla, Srini, Vibhav, Suresh, Siva, Chitra, Hari, Vijay, Farhan, Amala, Deepu, Ravi, Giri, Chiguru and Vijju.

Funding for this thesis from NURI / FGDC (NURI Grant # NMA202-99-BAA-02) is specially acknowledged.

Finally, I thank all my family members, my parents, and sister for their love and support through all the years and encouraging me in all my endeavors.

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	ii
LIST OF TABLES.....	vii
LIST OF FIGURES.....	viii
Chapter	
1. INTRODUCTION.....	1
1.1 Goal of the Research.....	1
1.2 Background and Objectives.....	2
1.3 The Concept.....	8
1.4 Outline of the Thesis.....	11
2. DIGITAL LIBRARY CONCEPTS.....	12
2.1 Introduction.....	12
2.2 Project Gutenberg.....	13
2.3 Illustrative Digital Open Access Initiatives.....	15
2.4 Creative Commons.....	17
2.5 Geo-spatial Context	21
2.5.1 Alexandria Digital Library Project	22
2.5.2 FGDC Data Clearinghouse.....	24
2.5.3 Geospatial One-Stop.....	25
2.6 Summary.....	27

3. PUBLIC COMMONS FOR GEOSPATIAL DATA.....	28
3.1 Introduction	28
3.2 What is a Public Commons?.....	28
3.3 Legal Framework for Sharing Digital Spatial Datasets.....	31
3.4 Other Highlights of the Use of Public Commons License.....	36
3.5 Advantages of Using Open Access Licensing	36
4. METADATA MODEL FOR PUBLIC COMMONS	38
4.1 Introduction	38
4.2 Designing Web Interfaces for Public Commons.....	39
4.3 Existing Metadata Models and Search Mechanisms.....	43
4.3.1 The Alexandria Digital Library Approach.....	44
4.3.2 The FGDC Data Clearinghouse Approach.....	45
4.3.3 Comparison of Alexandria Digital Library and FGDC Clearinghouse Approaches.....	47
4.4 Proposed Hierarchical Metadata Model for Public Commons.....	49
4.5 Benefits of PC Model over Digital Library and Clearinghouse Approaches.....	53
5. TECHNICAL APPROACH FOR PUBLIC COMMONS.....	55
5.1 Introduction	55
5.2 The Need for Technical Approach for Datasets with Public Commons.....	55
5.3 Suggested Technical Approach.....	57

5.3.1 Steganography for Identifying Contributor in	
Raster Spatial Datasets.....	59
5.3.2 Attaching an Invisible Number to Standard GIS Files.....	61
5.3.2.1 Why in the Header?.....	61
5.3.2.2 Illustrating with an Example.....	62
5.4 Other Potential Technical Approaches.....	65
5.5 Conclusions.....	69
6. OPERATIONAL ASPECTS OF PUBLIC COMMONS MODEL.....	71
6.1 Introduction.....	71
6.2 Functionality of the Public Commons Model.....	71
6.2.1. Architecture.....	71
6.2.2 Submitting a Dataset to the Facility.....	75
6.2.2.1 User Friendly Metadata transcripts.....	76
6.2.2.2 Using Open Access Licensing.....	79
6.2.2.3 Operational Characteristics of SFIPCA.....	80
6.2.3 Downloading a Dataset from the Facility.....	83
6.3 Public Commons Identification Software.....	84
6.4 Conclusions.....	84
7. CONCLUSIONS AND FUTURE WORK.....	86
7.1 Summary.....	86
7.2 Conclusions.....	86

7.3 Future Work.....	91
REFERENCES.....	93
APPENDIX A Common File Formats for Spatial Data.....	100
APPENDIX B Comparison of Different Organizational Metadata Transcripts.....	102
BIOGRAPHY OF THE AUTHOR.....	112

LIST OF TABLES

Table 5.1: Technical description of the ESRI Shape File Format.....	63
Table A.1: Common File Formats for Spatial Data.....	101
Table B.1: Comparison of Metadata Templates of Different Organizations.....	111

LIST OF FIGURES

Figure 1.1: Conceptual approach of Public Commons Data Sharing Model.....	9
Figure 2.1: Metadata record for a copyrighted work with Creative Commons.....	20
Figure 2.2: Alexandria Digital Library with a Centralized Metadata Catalog.....	23
Figure 3.1: A Sample Public Commons Copyright Notice.....	35
Figure 4.1: Centralized metadata-database structure in Digital Library approach.....	45
Figure 4.2: NSDI's Geospatial Data Clearinghouse Search Form.....	46
Figure 4.3: The mechanism of Querying multiple FGDC's clearinghouse nodes.....	48
Figure 4.4: Displaying Search Results from FGDC's Metadata.....	47
Figure 4.5: Public Commons Metadata Access and Search Mechanism.....	50
Figure 4.6: 3° X 3° minimum grid laid on U.S. and bounding range for Penobscot query.....	52
Figure 5.1: Steganography applied to a Vector ESRI Shape dataset to embed information in the header.....	64
Figure 5.2: The pictorial representation of linking metadata and licenses in a database perspective handled by SFIPCA.....	65
Figure 5.3: Handwriting technique.....	66
Figure 5.4: Demonstrating Embedding Technique in a vector polyline spatial Dataset.....	68
Figure 6.1: Flow diagram of the Operational aspects of Public Commons Model.....	72
Figure 6.2: Operational aspects of Spatial File Identification Automated System.....	73

Figure 6.3: Example of a Value added Copyright license.....	82
Figure 7.1: A Scenario of evolving Spatial Information resources.....	88
Figure B.1: Screen Shot-1 of Metadata elements in Public Commons Minimized Metadata Transcript.....	109
Figure B.2: Screen Shot-2 of Metadata elements in Public Commons Minimized Metadata Transcript.....	110
Figure B.3: Screen Shot-3 of Metadata elements in Public Commons Minimized Metadata Transcript.....	110

Chapter 1

INTRODUCTION

1.1 Goal of the Research

Digital spatial archives are being implemented at the federal and local level through efforts such as the National Spatial Data infrastructure (NSDI). A primary objective of the NSDI is to facilitate sharing and provide access to digital spatial data for all levels of government, the commercial sector, the non-profit sector, academia and citizens in general. Efforts such as the Content Standard for digital Geospatial Metadata and Dublin core are attempts to standardize dataset labeling i.e. *metadata* to improve access to available spatial data and promote its reuse (FGDC April 1997).

In spite of these efforts, spatial data producers still are unable to effectively *distribute* or *share* geospatial datasets in digital formats due to many reasons that include inefficient search and access mechanisms, inefficient means for documenting data, insufficient technical protections, lack of appropriate legal frameworks, and fears of unauthorized copying and illegal ownership claims (Duker and Vrana 1994). An intuitive hypothesis suggests that some form of legal and technical protection for spatial datasets can be evolved so that data producers can openly contribute their works to *open access centralized archives* without undue fear of losing credit for their contributions or gaining increased liability exposure. The goal of this research is to explore an explicit *conceptual open access spatial data sharing* model, a *Public Commons for Geospatial*

data, with substantial potential for providing incentives to data producers for their contributions to a public commons and overcoming impediments in wide-scale spatial data sharing. A sub hypothesis is that one or more known steganographic methods in combination with cryptographic methods can be applied successfully to most standard GIS file formats in support of the envisioned open access spatial data sharing model. A further goal is to execute proof of concept tests for various operational aspects of the model.

1.2 Background and Objectives

About 80 percent of all government information has a geospatial data component, such as an address or other reference to a physical location (OMB Jan 2003). In many cases, agencies independently collect and maintain data that, while not identical, is similar and potentially duplicative in many respects. For example, both U.S departments of Housing and Urban Development and Census Bureau maintain separate GIS systems for storing and analyzing essentially the same geospatial data regarding congressional districts, city boundaries, rail roads, interstate highways and state highways. There is a huge cost involved in collecting this duplicative spatial data (Dept. of Environment 1986). According to a recent study by OMB, up to 80 percent of GIS costs are related to the collection and management of spatial data (Center of Technology in Government 2001; D.Koontz June 2003).

Many national governments throughout the world are involved in developing Spatial Data Infrastructures (NSDI's) and Digital Geo-Spatial libraries that will better

facilitate the availability and access to spatial data for all levels of government, the commercial sector, the non-profit sector, academia and citizens in general. A key premise in most of these initiatives is that *national governments will be unable to gather and maintain more than a small percentage of the geographic data that users in their nations want and desire*. Thus, the national initiatives are depending typically on the cooperation of those already gathering spatial data and those using GIS to meet specific needs to help construct and maintain these spatial data infrastructures and geo-libraries for their nations (Onsrud 2001). However, there have been many difficulties in achieving cooperation of local governments, the private, commercial sector, and individual GIS users in regard to their willingness and ability to contribute geographic data. Involved are such issues as compatibility, interoperability, legal, economic and organization culture issues.

Some of the impediments to widespread spatial data sharing are well known from directly asking GIS users why they are not currently involved in creating datasets that are of common or compatible formats, documenting in the standardized metadata format or making their datasets more readily available to others through NSDI's or geo-libraries. Most of these impediments are unrelated to a need for increased funds. For many organizations, even if their budgets were doubled they still would not use the increased funds to make their geographic datasets more accessible to their own communities or the rest of the world. They are inhibited by further impediments that money alone is unable to address. Common wisdom suggests that intellectual property laws and the markets they protect create the only practical environment for producing and sharing useful

information. That is, profit motivations drive all major resource development. Yet the history of the web shows us otherwise. We now have numerous examples of massive voluntary resource production and sharing. In some instances, tens of thousands of individuals have worked collaboratively or as independent contributors in creating new knowledge resources or producing new software e.g. Linux (Onsrud 2001).

A general incentive premise of our emerging spatial data sharing model is that, as individuals, *most of our conduct in daily life is not driven by profit motives*. From past surveys conducted by academics and private sector companies (Pluijmers 1998, Mason 1998), many creators have indicated they would be more than willing to share their spatial data sets with such infrastructures or geo-libraries if, among other reasons, *it was much easier to do, creators could reliably retain credit and recognition for their contributions to the public commons, creators could obtain other non-monetary benefits and creators could acquire substantially increased liability protection* from use of the data they make available to the public (Onsrud 2001, Johnson et al 1995) . We will discuss these reasons and objectives of such a commons infrastructure.

1) it was much easier to do

Many GIS data users do not have the needed relevant information about existing datasets possessed by others that could be appropriately used for their applications; or how to get access to these datasets elsewhere or on the Internet. Therefore, they may unnecessarily create new datasets for their applications incurring heavy investments. Such problems arise due to poor documentation (i.e. insufficient metadata to determine

the fitness and purpose of the dataset), inefficient search and data access mechanisms, and the distributed nature of information sources. Duplication of effort could be minimized if a *centralized information system* for metadata with easy *metadata creation and data-upload mechanisms* could be built. This information system should be able to provide user-friendly metadata transcripts to the data producers for documenting their datasets and an upload tool to conclusively tag datasets and potentially store them at a centralized database location on a remote server. Alternatively, the tagged spatial datasets might be stored at distributed locations. Online centralized metadata catalogs for digital datasets could readily make the existence of information about spatial datasets available to anyone in the world at one single location instead of the user searching at multiple distributed locations. For those willing to make their spatial datasets available through public commons license or by dedicating their datasets to the public domain, the system could also provide direct access to the datasets through links to the datasets at distributed sites and to the same datasets in a centralized archive.

2) creators could reliably retain credit and recognition for their contributions to the Public commons infrastructures

There is a major misconception that profit motivations drive all major resource development. However, this is not true; we now have numerous examples of massive voluntary resource production and sharing. In some instances, tens of thousands of individuals have worked collaboratively or as independent contributors in creating new knowledge resources or producing new open source software (e.g. Linux, ArcGIS Viewer etc) and making them available for free or sharing them through information networks

such as the Internet. Similarly, some data providers are willing to share their spatial data without charge with no limitations imposed on the further use of the data. Others are willing to freely share but only if certain limitations are met. In most cases, dataset producers would like to retain credit and recognition for their contributions. Thus many would like to obtain *visible* credit (i.e. the world should know the origin of the dataset) for not only their direct contributions but also be acknowledged in the derivative works that originated from their contributions. In the case of multiple contributions contributing to a derivative work, the list of all contributors should be acknowledged. Similarly a string of derivative works should acknowledge the string of subsequent contributors at each level. We envision an easy legal mechanism by which any individual may affirmatively and permanently mark their 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. Therefore, an automated identification mechanism that can embed and retrieve information in a spatial dataset such as the name of the author, date of creation, purpose, place, legal status, acquisition, instruments used, and accuracy or alternatively an identification number in the dataset that could provide a link to such information would be an ideal solution for intellectual property protection and providing credit and recognition for the contributors to these commons infrastructures.

3) creators could acquire substantially increased liability protection from use of the data they make available to the public, and

Spatial datasets are often relied upon to make decisions. Users may incur damages by improper use of data relying on data that contains blunders or is incomplete, or by using

data for a purpose for which it is not fit. The potential for damages raises the possibility of liability for those who created the data. Whereas liability for the producers of data files containing literature, music, and art is inconsequential, liability for making spatial data files available is more likely (Onsrud 1999). As a result liability exposure is a real concern for spatial data producers and should be dealt with by any system promoting and enabling the widespread use of spatial datasets.

4) creators could obtain other non-monetary benefits.

Incentives to contribute to creation of broadly available information resource are not necessarily monetary nor are monetary incentives necessarily the most effective incentive under many circumstances. Any system for enhanced sharing of spatial data should incorporate non-monetary benefits that many data producers might value. By example such a system might offer ease of creation of meta-data, a permanent archival service, a tagging and identification service for GIS data files, increased search and retrieval capabilities, and increased visibility for the contributions that producers contribute to a commons infrastructure.

Therefore, the main objective is to develop a conceptual framework for such a spatial data sharing infrastructure to address the primary impediments to data sharing and leverage individual spatial data efforts so that data can be exchanged by government agencies, commercial organizations, and individuals choosing to openly share their data with others. The system should aim at providing basic geographic data in common encoding and make them discoverable through a *catalogue* through which GIS users with

varying knowledge levels can participate. By finding and downloading spatial datasets and using advanced GIS technologies, users should be able to perform, develop or create new value-added data and applications instead of duplicating data production efforts. The system should support also cross jurisdictional and cross organizational analyses and operations. Given a carefully assembled framework on which to base their work, GIS users will be better able to create new data sets that can fit together and be used in conjunction with other data sets.

1.3 The Concept

The goal of the conceptual model is to enable and entice the non-expert GIS user to preserve his created spatial dataset in a *public geolibrary-like system* and make it accessible to the rest of the world. Under the model, the non-expert GIS user can access a website that allows the creation of a metadata record in response to a web interview transcript (i.e. series of questions with limited choice responses). As part of the series of responses, the contributor agrees to apply one of a limited selection of “open access” licenses or dedicate the file to the public domain. The transcript responses (i.e. metadata file) and the actual data file (i.e. spatial data file) are submitted to an automated processing system, the *Spatial File Identifier and Public Commons Archive* (SFIPCA). SFIPCA affixes an *identification number* permanently to the data file such that a click on the dataset (dataset icon on the desktop) would display information identifying the originator and the licensing status of the spatial data file either locally or through a web link. This number is *invisible* to the typical user but can be retrieved on request to provide evidence of ownership. Depending on the implementation approach either a single

identifying number or more extensive metadata may be embedded in encrypted code and hidden in the data file using *Steganography* techniques (Schneier 2000, Katzenbeisser and Peticolas January 2000). In the event that someone takes the file, adds value to it, and resubmits the updated/improved file with new metadata to SFIPCA, clicking on the dataset would identify any value-adder who has added to or altered the file. Thus, the originator and the string of value-adders (up to a practical limit) would always be maintained with a file processed in this manner. The contributor's lineage, intellectual property status, applicable licensing provisions and other metadata could be exposed on request for any files that had been previously processed through SFIPCA. Figure 1.1 shows a pictorial representation of the concept adopted for our Public Commons Sharing model.

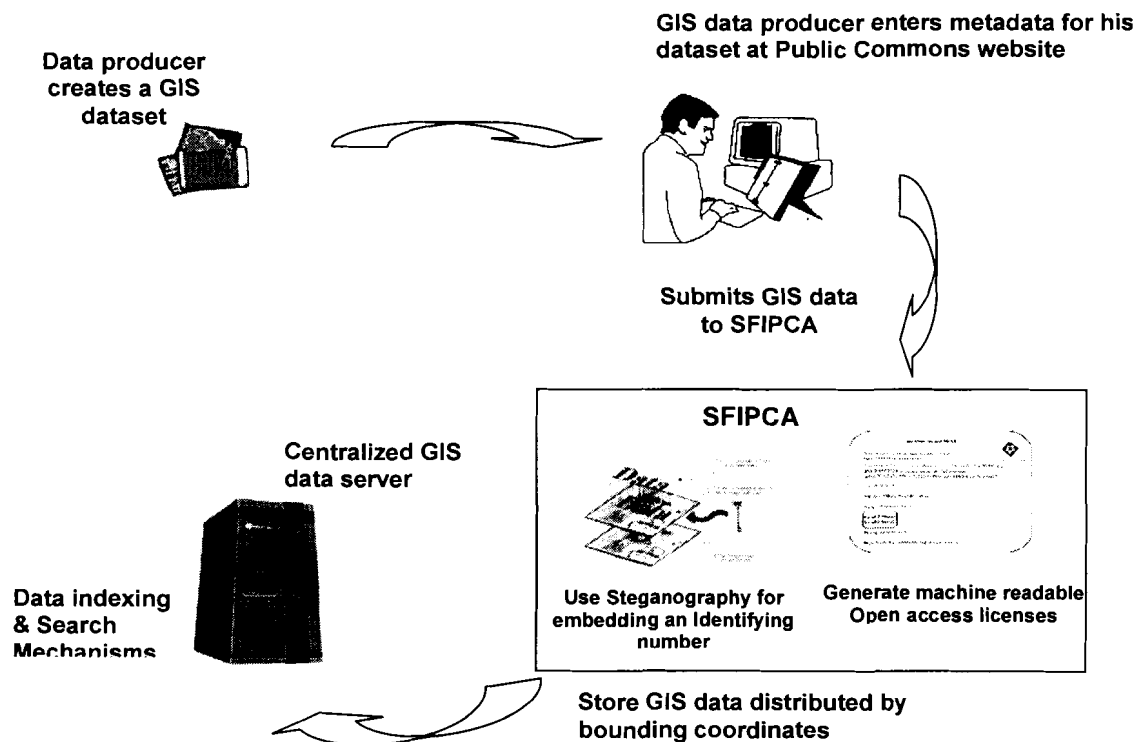


Figure 1.1: Conceptual approach of Public Commons Data Sharing Model

With the appearance of the unique public commons dataset icon and the invisible encrypted identification number, there would be strong legal evidence that a user is allowed to use freely and liberally the data in accordance with the open access license provisions with little danger of impinging on the intellectual property rights of others. For files marked in this manner, there would be little reason to remove the invisible identification since the file would already be available for free use. The primary result of stripping away identification information would be to establish grounds for a lawsuit against the infringer, who could never be certain that there might not be additional hidden identifiers in the image or vector data files (Onsrud 2001).

The concept here is to create a broad and continually growing set of freely usable and accessible data at local level scales similar in effect to the public domain data sets created by federal agencies. By investing the extra time, effort and expense of creating metadata, creators obtain something in return. Through this approach they obtain visible credit when their data set is used in the products or services of others. Those sharing under this approach obtain a level of liability protection never acquired when data is simply released and they obtain a potential archiving service. These benefits are in addition to the benefits gained from current Internet sharing environments, such as through the FGDC clearinghouse node network and the Geography Network. The model addresses some of the most frequent reasons given by scientists and local governments for not making their data sets available to others.

1.4 Outline of the Thesis

Chapter 2 is dedicated to a literature review of different digital library initiatives both in the case of conventional *digital* media and *geo-spatial* datasets. An understanding of the pros and cons of these initiatives is crucial for supporting our argument of developing an alternative geolibrary-like data sharing model in the successive chapters.

Chapter 3 focuses on the need for an effective legal frame work that can allow contributors to place their dataset affirmatively in the public domain or into an Open Access Licensing status in a legally supportive manner.

Chapter 4 describes a new hierarchical metadata model that has improved metadata access and search mechanisms with advanced features such as prioritized query results, and a centralized metadata database.

Chapter 5 focuses on the need for a strong technical approach to protect digital datasets and provide visible credit for the contributions made by the data producers. It explains the use of Steganographic techniques to hide metadata information in the dataset for copyright protection and data authentication.

Chapter 6 discusses various features and operational aspects of the public commons data sharing model.

Chapter 7 finally presents the conclusions and future work of this research work.

Chapter 2

DIGITAL LIBRARY CONCEPTS

2.1 Introduction

Libraries in the new millennium have evolved into technologically driven repositories of digitized information in contrast to the usual traditional library with collections of books, articles and microfilms. Digital libraries play a key role in providing a knowledge network that equips individuals with the necessary resources to tap into the information gateway and transform raw information into items of commercial value (NREC 1999). The role of digital libraries has been extended beyond not just bringing information close to the general public but also building the resources, maintaining interoperability standards, protecting intellectual property with copyrights, preserving the resources and providing effective access to any number of people through out the world (Nimmer and Patricia 1992). Examples of digital libraries include collections of published research articles, journals, books, electronic medical records, multimedia files (audio and video), government documents, spatial datasets and scientific knowledge bases.

As repositories of information and ideas, digital libraries enable the sharing of knowledge and facilitate lifelong learning - a vital component for success in this ever-changing world (Litman 2001; Bauer and Joroff 1969). Digital libraries not only make data available for use, but also provide technical and organizational means of capturing new data for the research community, thereby, provide a live, growing, and evolving

resource. It is also very important to properly protect as intellectual property, the results of this creative activity and to construct a social and economic system for their effective utilization as the future resource of value-added products and services. Construction of a system balancing protection and exploitation is therefore indispensable for the establishment of an intellectual creation cycle.

A thorough discussion of various Digital library initiatives both in the case of conventional *digital* media (text, audio and video etc) and *spatial* datasets that are being distributed on the Internet is presented here as a part of the literature review. An understanding of the concepts employed to *enable easy creation of metadata, declaration of use rights or establishment of licensing provisions, support of catalogues and search mechanisms, and support of download and sharing mechanisms in the library examples* explained here is crucial for supporting our argument of developing an alternative geolibrary-like data sharing model for spatial datasets in the successive chapters.

2.2. Project Gutenberg

The greatest value created by computers would not be computing, but would be the storage, retrieval, and searching of what was stored in our libraries - These are the words of Michael S. Hart, founder of the Project Gutenberg. Project Gutenberg is one of the early on-line efforts to convert massive amounts of public domain printed material into digital text. He believed that it would be a really good idea if famous and important texts were freely available to everyone in the world. Since then, he has been joined by hundreds of volunteers who share his vision. They started converting famous copyright expired books, articles and treatises into digital text with the intent that they would be

available in a long term archive freely and readily accessible to anyone. The Project Gutenberg Philosophy is to make information, books and other materials available to the general public in forms, that a vast majority of computers, programs and people can easily read and search (Hart 1997). The e-texts should cost so little that no one will really care how much they cost and should be easily usable that no one should ever have to care about how to use, read, quote with any fear of ever infringing copyright.

Although highly beneficial, the project mostly targets public domain literature and books or copyrighted material with authorized permissions for reproduction, distribution and transmission. The text in the hosted books were either painstakingly typed or laboriously scanned digitally and then stored in plain Vanilla ASCII text files (such that the format can be accessed with any computer of any advanced operating system at anytime in the future). These e-text files are placed in the public domain and are readily available for download at <http://gutenberg.net/> (a web-based online archival service) for free or for minimal fees towards dissemination costs. It maintains metadata or detailed information such as name of the author, title, genre, date of publication, date of release into public domain, language and availability. The project website maintains an indexing service that can retrieve the text (or zipped) files based on text-based search fields such as author, title or genre from a database containing thousands of similar records. The project has also encouraged many volunteers to find public domain books, convert them into digital files and upload them to their centralized archive before the books become extinct.

Project Gutenberg has inspired academia, research communities and the corporate sector to develop something similar for their own internal reference purposes. They have developed and maintained these archives for storing and sharing their new findings

within their communities, and are providing access to this wealth of information to their people within the context of security and trading restrictions, and some organizations are even making such works open to the public.

2.3 Illustrative Digital Open Access Initiatives

Many public digital libraries, mostly disciplinary specific, have adopted Project Gutenberg's philosophy to store, share and provide access to their information. A few examples of these libraries include: Public Library of Science (Publiclibraryofscience.org) - a public resource for scientific and medical literature, NEC CiteSeer (citeseer.nec.com/cs) - a digital library for scientific literature predominantly for research papers in Computer science and engineering, Perseus (perseus.tufts.edu) - resources for the study of the humanities, and the National Geospatial Data Clearing House (fgdc.gov) - sharing of geo-spatial datasets. Individual digital libraries of research materials such as at MIT (Bass, Stuve et al 2000), UC Berkeley and Tufts share scientific research papers. More specialized digital libraries make available imagery, folk literature, computational tools for digital morphology, and so on.

Recognizing their immense potential for the establishment of intellectual creation and benefits to the nation's intellectual property interests, the Library of Congress had initiated many digital library programs to preserve digital information and the policies that govern them. The initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks all in user-friendly ways.

Many national research organizations and private sector companies have invested much money and time to develop information archives and provide access to the general

public on a large scale. However, one major difficulty continuously confronted is who is to retain the copyrights or the intellectual property rights of the material that are posted in these libraries. In the case of Project Gutenberg, most of the material posted had already entered the public domain. Therefore no one can claim copyright on those materials and people can do what ever they wish to do with them. What about the research papers which are copyrighted to the author by default in most jurisdictions the minute they are completed? What should be the extent of use restrictions on these research findings? Should use be restricted to only a few people or should access be more universal?

The open access initiatives brought out by Open Society Institute (OSI) at Budapest 2001 explored the concept of *Open Access*, and recommends that the information available through or affiliated with their archive is freely available on the public internet, and is permitted to be downloaded, copied, distributed, printed, searched, or linked to the full texts, crawled for indexing, passed as data to software, or be used for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself (Budapest Open Access Initiative 2001). *The only restriction to this type of access to the information is that the author of the work has whole control over the integrity of his work and the right to be properly acknowledged and cited.* The goal is that removing access barriers to the literature with this kind of licensing approach, will accelerate research, enrich education, share learning among different sects of the world, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge.

There are two complementary approaches currently being implemented on a widespread basis to successfully achieve open access to the scholarly material. The first

is to encourage *Self-Archiving* by providing the scholars with the tools and assistance to deposit their documents in standard open electronic archives; thereby retaining control over their work and the right to be properly cited. Secondly, authors can *donate their copyrights to library-like institutions committed to open access*, which have their own copyright terms (i.e. they use copyright and other tools to ensure permanent open access to all the articles they publish) and turn to other alternatives to cover their expenses. Either way the document is stored in a public archive with open access to the world. The first approach is exemplified by Cite Seer's Research Index and the Creative Commons Project who have enabled self-archiving with an efficient mechanism for supplying tools to upload, process, serve, and search. The second approach is exemplified by publications such as URISA Journal, PubMed Central, Public Library of Science, SPARC and the individual digital libraries of research collections found at MIT, UC Berkeley and Tufts.

Open access to vast amount of literature online gives immense opportunity to scholars for accessing many research articles and findings all over the world which were inaccessible and expensive before. The material that should be freely accessible through these facilities should not only include peer-reviewed journal articles, and the substantiated datasets which the research communities contribute to the world, but also any unreviewed preprints and intermediate or unfinished findings that they might wish to put online for comment (Mary Feb 2002). This open access approach gives readers extraordinary power to find and make use of relevant literature, and it gives authors and their works vast and measurable new readership, review, visibility, and impact.

2.4 Creative Commons

Creative Commons, founded in 2001 and led by cyber law and intellectual property experts, is one of the organizations dedicated to raise the momentum towards realizing the objectives of open access. It encourages self-archiving with sufficient assistance to scholars to make things much easier for licensing, search, labeling and retrieval of required documents.

Creative Commons in Dec'02, released a set of copyright licenses, inspired by the Free Software Foundation's GNU General Public License (GNU-GPL), free for public use (GNU-GPL June 1991). Unlike the GNU-GPL, Creative Commons licenses are not specifically designed for software, but rather are designed for other kinds of creative content such as websites, music, video films, photography, books, articles, literature, and courseware. There are a total of eleven Creative Commons licenses possible that might be compiled from the conditions elaborated below *individually or combined* in addition to a public domain license (Creative Commons May 2001).

Others can copy, distribute, display, and perform a copyrighted work and generate derivative works based upon it –



but *only if they give you credit.*



but *for noncommercial purposes only.*



but *should not generate any derivative works based upon it.*



Others can distribute derivative works only *under a license identical to the license that governs the work (often known as Share-Alike or Copyleft)*. The Share Alike requirement applies only to derivative works.

Every Creative Commons license carries with it a full set of other rights in addition to the general allowances specifically made above. Every license will help the contributors to retain their copyright and declares that other people's fair use, first sale, and free expression rights are not affected by the license. Every license applies worldwide; lasts for the duration of the work's copyright; and is not revocable. Every license requires licensees to get the contributors permission for any actions that are restricted. For example, the author might restrict the ability to make a commercial use or create a derivative work; to keep the copyright notice intact on all copies; to link to contributors license from copies of the work; not to alter the terms of the license; and not to use technology to restrict other licensees' lawful uses of the work.

Creative Commons developed a Web application that helps people dedicate their creative works to the public domain or retain their copyright while licensing them as free for certain uses, on certain conditions. It collects information such as the type of license the author wishes to impose on the use, contact information and the details of the work that can be uploaded as metadata for the work to the online catalog. Depending on the author's choice, a digital commons license is generated which states all the restrictions imposed on the use of the work. A hyperlink to the registered copy of the digital machine-readable license and a small snippet of HTML code that can be included on the contributor's webpage is sent to the author by e-mail. Therefore, a link to the license would apparently give the licensing information to the user. Unlike Digital Rights Management (DRM) technology (Iannella June 2001), which tries to restrict use of digital works, Creative Commons is providing ways to encourage permitted sharing and reuse of works.

Creative Commons has also developed metadata catalogs (available at creativecommons.org/works/) that can be used to associate creative works with their public domain or license status in a machine-readable way. This will enable people to use search engines to find particular data (for example, music that are free to use; provided that the original composer and musicians are credited). An example of a record is shown in Fig 2.1, which shows the title, description of the format, author, the date created and the licensing status of the work. By the license agreement one can copy, distribute, display, and perform the copyrighted work provided that the author is acknowledged and the derivative works should be placed under the same licensing agreement.


TITLE:	<u>Stifled Love</u>	LICENSING 
DESCRIPTION:	Video	
CREATOR:	People Like Us	
DATE CREATED:	January 3, 2003	

Figure 2.1: Metadata record for a copyrighted work with Creative Commons

One important thing that should be mentioned here is that Creative Commons maintains the web-based catalog of different media works only in records, such as the one shown above in Fig 2.1 and provides an efficient technological mechanism to better find ones work online. It neither deals with the intellectual property rights of the original work nor does it directly host the original work. This means that the organization does not hold any direct responsibility or liability for the work that was mentioned by the person who has filled the metadata. It only provides metadata and licensing information

at one location for different contributor works residing at different locations. Therefore, a person who wanted some photos for his website could visit creative commons web-catalog and place a query for photos. The results of the query would show different works by many people. Then depending upon the description of the work and the licensing restrictions one could decide and contact the person or their website for the work.

The Creative Commons Project is working to build an *intellectual works conservancy* that protects works of special public value from exclusionary private ownership and from obsolescence due to neglect or technological change. They continue to encourage people to donate their copyrights to be held in public trust.

CreativeCommons's ultimate goal is to develop a rich repository of high-quality works in a variety of media, and to promote an ethos of sharing, public education, and creative interactivity by providing contributors sufficient technological means to better find their works online with ease and evidentiary licensing methods.

2.5 Geo-Spatial Context

Similar efforts in the context of disseminating geospatial data has led to the establishment of National Spatial Data Clearinghouse nodes in the US and many other nations throughout the world. There are a few organizational efforts with goals similar to Creative Commons in the geo-spatial context encouraging data producers to share spatial datasets at large. These examples include FGDC Clearing Houses, Geography Network, Alexandria digital library project, MIT & Harvard Geospatial Library and Free GIS depot.

Data clearing houses and digital spatial archives are being developed and maintained in the U.S. to find, share and exploit information across jurisdictions to reduce duplication, improve technical support to users, and better coordinate the activities of different agencies engaged in surveying, mapping and related GIS functions (Office of the President Oct 1990). They provide a basis for spatial data discovery, evaluation, and application for users and providers within all levels of government, the commercial sector, the non-profit sector, academia and by citizens in general. SDI's and geo-libraries combined host geographic data and attributes, and sufficient documentation in the form of metadata: a means to discover, visualize, and evaluate the data, and some method to provide access to the geographic data. Additional services and software applications are also provided to support the many applications of these data.

The next section discusses three of such initiatives based on different metadata search models (explained in detail in Chapter 4) in the field of GIS.

2.5.1 Alexandria Digital Library Project

The University of California, Santa Barbara based Alexandria Digital Library (ADL) project began in 1995 with the development of a working digital library with *centralized* collections of geographically referenced materials and services for accessing those collections. It was developed by a consortium of researchers, developers, and educators, spanning the academic, public, and private sectors, exploring a variety of problems related to a distributed digital library for geographically-referenced information. ADL is in the process of loading significant collections of geospatially-referenced information and its metadata for datasets such as Digital Elevation Models (DEMs), Digital Raster Graphics (DRGs), Scanned Aerial Photographs, Landsat TM,

Seismic datasets and technical reports. They have been collecting varied datasets all over the nation for research purposes as well as for an operational digital library, which makes their collections relatively large with diverse extents and density of coverage.

Fig 2.2 shows a web-interface developed by the project which takes minimal information such as geographical location information and data formats from the user and queries a database that includes extensive metadata of original spatial data that might be sought.

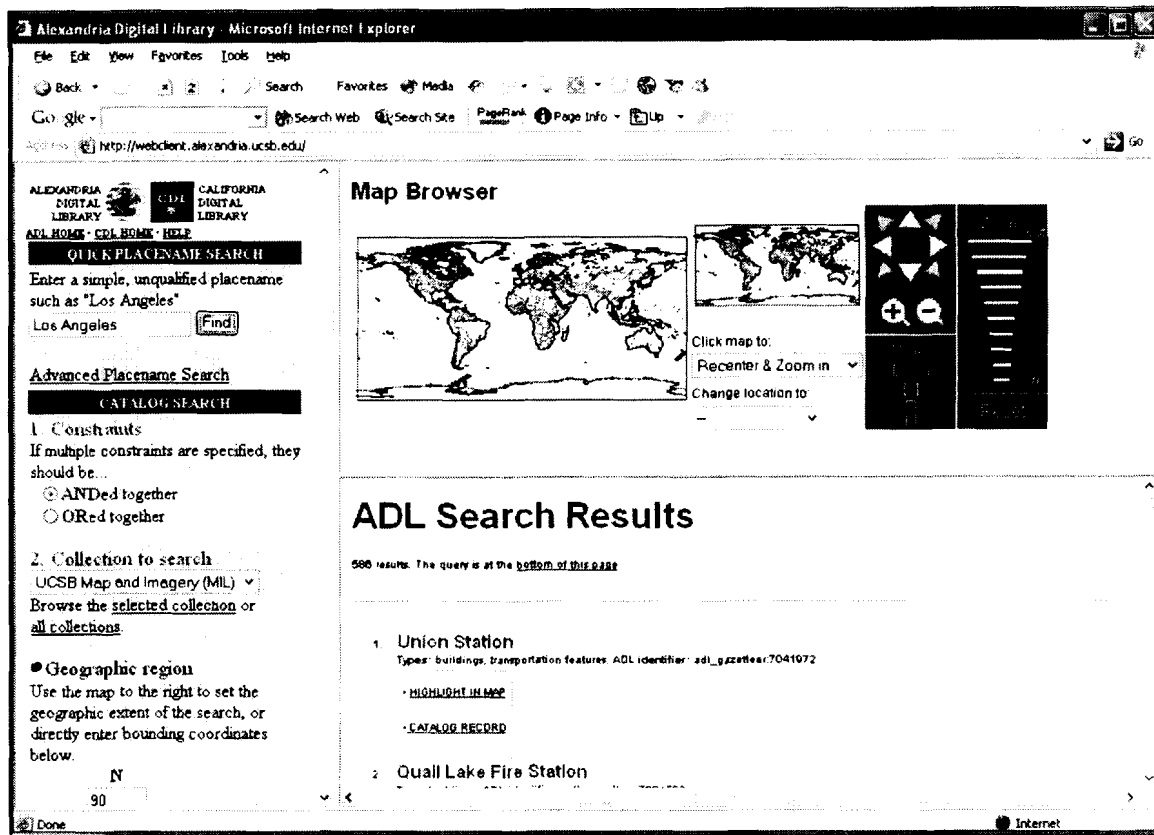


Figure 2.2: Alexandria Digital Library with a Centralized Metadata Catalog

The matching results are shown on the screen with links to the data required. This method simplifies the task of searching for data and is quite similar to the web application developed by the Creative Commons. One more important feature of the interface is that

the geographical location can be chosen by spotting the location in a Java enabled applet depicting the world map (<http://webclient.alexandria.ucsb.edu/>).

Unlike Creative Commons, Alexandria digital library project offers library services only and does not provide any data or metadata upload mechanism. Most of the data that is available through their library are either data from the federal agencies (which are in the public domain) or data from third party agencies for which they have a mutual contract in order to deal with the legal ownership issues. ADL often provides metadata information and the contact links for some datasets not in their possession.

2.5.2 FGDC Data Clearinghouse

Clearinghouses provide a primary data dissemination mechanism to traditional and non-traditional spatial data users. By promoting the accessibility, quality, and requirements for digital data through a searchable online system a Clearinghouse facility can minimize duplication of effort in the collection of expensive digital spatial data and improve cooperative digital data collection activities.

The FGDC Clearinghouse is a decentralized system of servers located on the Internet, which contain field-level descriptions of available digital spatial data. This descriptive information i.e. metadata, is collected in a standard CSDGM (Content Standard for Digital Geospatial Metadata) compliant format to facilitate query and consistent presentation across multiple participating sites. Clearinghouse sites provide hypertext linkages within their metadata entries that enable users to directly download the digital data set in one or more formats. Clearinghouse node uses readily available Web technology for the client side and ANSI standard Z39.50 server technology for the query

search, and presentation of search results to the Web client. Through this model, Clearinghouse metadata provide low-cost advertising for providers of spatial data, both non-commercial and commercial, to potential customers via the Internet.

The Data Clearinghouses follow the same intellectual property method as the Alexandria Digital Library. Most of the data available are from federal agencies i.e. data are already in the public domain or the system provides direct access to only metadata. The commercial datasets are protected by copyrights imposed by their companies. Unlike ADL, the clearinghouse system provides a metadata upload mechanism known as “Metadata Lite” (<http://130.11.52.178/metaform.html>), where non-federal data supplying agencies can fill in *metadata for their datasets* and upload it to the FGDC site. Not only does this advertise the existence of the datasets but also provides direct access to the datasets.

2.5.3 Geospatial One-Stop

More recently, the E-Government Act of 2002 initiated *Geospatial One-Stop* (GOS) - an initiative to promote coordinated geospatial data collection and maintenance across all levels of government. GOS plans to build on and accelerate federal spatial data collaboration initiatives by (1) developing a portal for seamless access to geospatial information, (2) providing standards and models for seven geospatial data themes, (3) creating an interactive index to geospatial data holdings at federal and nonfederal levels, and (4) encouraging greater coordination among federal, state, and local agencies about existing and planned geospatial data collections.

The Geospatial One-Stop Project proposes to support “one-stop” access i.e. citizens and government will only have to go to one location for Federal government and other spatial data assets. GOS proposes to develop a Portal (i.e. website scheduled to be complete by the end of September 2003) as a virtual repository for spatial data and web services to support local, state and federal programs, and decision making. The vision of the Geospatial One-Stop Portal is to allow users to view and obtain desired data for a particular part of the country, without needing to know the details of how the data are stored and maintained by independent organizations. Seven geospatial data themes, commonly known as *Framework Data*; Elevation, Ortho-imagery, Hydrography, Transportation (including Road, Air, Transit and Rail sub-themes), Government Units (administrative boundaries), Cadastral (property boundaries), and Geodetic Control are considered to be of fundamental importance to many applications. Framework data content standards are now under development by the Geospatial One-Stop initiative (OGC Dec 2002).

The OpenGIS Consortium (OGC) has been contracted to conduct a test bed portal project with the major focus on defining reference architecture and interoperability specifications and standards. OGC will work with selected member organizations to build a portal; integrating commercially available components and reusable software components within an Application Integration framework that ties them into a coordinated portal. When the development is complete, OGC shall demonstrate the working portal. ESRI has been tasked with quickly building the first version of an operational portal based on standards based COTS technology with features including interoperability, allowing choice of databases, hardware, GIS software, networks and web

browsers. The portal would be able to incorporate the current and new standards and functions resulting from the OGC test bed.

Geospatial One-Stop builds on federal efforts to develop a National Spatial Data Infrastructure (NSDI) through the FGDC. Therefore, they are quite similar with the FGDC data clearinghouses in terms of metadata search and access mechanisms as well as the use and licensing conditions of the spatial datasets available with them. However, in the near future, the efforts of Geospatial One-Stop may facilitate improved geospatial data access and collaboration; develop interoperable web GIS interfaces and services within the set geospatial standards for GIS data sharing communities.

2.6 Summary

In this chapter we have discussed different initiatives that have highlighted the importance of data sharing and the benefits they bring to society. The initiatives cited are preserving digital information, protecting certain author rights through the use of copyright, and providing broad access in the case of conventional *digital* media (text, multimedia) and *geo-spatial* datasets. We have seen some examples of digital libraries which provide metadata upload mechanisms which make the finding and searching for digital data easier for anyone.

Chapter 3

PUBLIC COMMONS FOR GEOSPATIAL DATA

3.1 Introduction

This chapter explains a new digital library concept in the field of geographic information that addresses the need for an effective legal framework that can allow contributors to place their work within an open access digital spatial archive while retaining some authorship control over the works and minimizing liability exposure. The chapter defines open access licensing and highlights the advantages of using it for digital spatial datasets in contrast to placing datasets in the public domain. A point to remember through out this thesis is that we are concerned about *the reasons why data producers are reluctant to share their datasets* and **not** *the need for stringent copyright protection of digital datasets*.

3.2. What is a Public Commons?

The Introduction chapter provided a brief conceptual overview of a data sharing model that can be undertaken to overcome impediments to wide-scale data sharing faced by federal agencies, the private sector and individual GIS users. Before exploring implementation approaches for the model we discuss the Public Commons and its features.

We define the *Public Commons for Geospatial Data* to be “a data-sharing facility that automatically supports user friendly metadata creation, open access licenses, archival services and the documentation of parent lineage of the contributors and value-adders of newly submitted digital spatial data sets” (Narnindi and Onsrud Sept 2002). The Public Commons concept may be viewed as an extension of the current search capabilities of the linked FGDC clearinghouse nodes, where anyone would be able to search for, access, and legally download and use any data sets found with the capability. It’s approach is similar to data sharing environments like Napster, Limewire, Kazaa and many other P2P (peer-to-peer) sharing efforts in the computer science community. Registered users can have access to audio, video and other files shared by different users in the network. The major difference being that in our approach people would instantly know the licensing restrictions on the files that are being shared.

Public Commons might consist of a centralized online data sharing facility where any expert or non-expert GIS user can create standard metadata for their datasets and share their spatial data with potential users. Users can freely copy or download GIS datasets instantly with complete metadata records and with any restrictions on free use readily known. Additionally, services such as identification software, the tracking of contributors’ lineage and permanent archiving of datasets are proposed as extra benefits to all those data producers who sign up and register their datasets under the open access license or public domain arrangements. The title “Public Commons for Geospatial data” derives from the fact that the spatial datasets from this data sharing facility are available to all the public (i.e. for the common good) and are legally free to use (abiding by any

open access licensing restrictions) just like public domain data for music, videos and literature. The important features that are supported under this data sharing model include

-

- *Open Access licensing Approach* - enables contributor and value-adder credit recognition, free distribution of digital spatial datasets, and potential minimization of liability exposure,
- *Advanced User-Friendly Web-Interfaces* for web transcripts to document metadata for datasets and data upload mechanisms,
- *Enhanced Metadata Model* which allows indexing, rapid access and search of spatial datasets,
- *Embedding Copyright Information* into the data using Steganographic techniques enabling identification and documentation of contributor lineage, and
- *Potential long-term archiving* of spatial datasets.

Thus, the model addresses the primary impediments to sharing spatial datasets presented in section in 1.2 and provides additional benefits for those choosing to participate. We will discuss each of these implementation concepts in successive chapters and the full operational aspects of the Public Commons model in Chapter-6. The next section will discuss the legal framework that is adopted in our Public Commons model that minimizes liability exposure, offers automatic credit recognition, and ensures free distribution of digital spatial datasets to all its users.

3.3 Legal Framework for Sharing Digital Spatial Datasets

Public agencies and GIS user communities are witnessing increasing demand for geospatial data and decreasing budgets, stimulating the desire to share geospatial data in order to reduce the costs of data capture, update and management. Moreover, national initiatives such as NSDI and Geospatial One-Stop are trying to make it easier to access and share existing geospatial information across the nation in order to help leverage investments and reduce duplication of data (Office of the President October 1990). Many GIS databases have been developed over the years among government agencies, private companies, academic institutions and other GIS users. Many of those data can be re-used by other users for different applications selectively or as a whole. Some of these data providers are willing to share their data with *centralized digital libraries* or *sharing environments* in anticipation that others also do the same for mutual benefits (Onsrud and Rushton 1995). Some users are interested in reusing spatial data for value-adding activities, if the data are shareable and accessible. Value added utilization of spatial information in this manner by profit and non-profit organizations will stimulate the growth of number of the available datasets in such centralized libraries. But there are still tens of thousands of spatial data producers who are skeptical about open data sharing as they fear liability issues, loss of attribution, unfair competition, and illegal ownership claims.

In the Introduction chapter we discussed that many data producers have indicated that they would be willing to share their spatial datasets with others if they could acquire substantially increased liability protection, and reliably retain credit and recognition for

their contributions to the public commons. Here we outline the expectations of data producers and then discuss the legal concepts that will likely aid in addressing them.

Assume that a user has downloaded a freely available dataset. The user has no contractual relationship with the original contributor. Let us also assume the user uses the dataset for a purpose for which it was not intended. Economic and physical injury damages result. The contributor under some circumstances may be liable for damages even though the data were freely available (Onsrud 1999). Therefore, producers want to guard against liability exposure under such circumstances. This can be accomplished through use of disclaimer language made evident whenever someone downloads a desired file. One option is to have licensing language that disclaims any express or implied warranties.

Liability for breach of copyright is also a significant concern for users of freely available spatial datasets. Creators are assumed by the law to have copyright in their works. While facts are not copyrightable, the selection, coordination and arrangement of facts such as in a dataset often are protected (Berne Convention 1967). Thus one must typically assume under the law that someone has a copyright interest in datasets even if those datasets are freely available on the web. The model being proposed addresses this issue since the producer of a dataset in the envisioned system is known and that person has affirmatively granted the right to others to use their dataset without further permission.

Liability concerns are very real for the producer and users of spatial datasets found openly available on the web. These concerns may be addressed through a simple process of licensing data to the general public as part of the metadata creation process. With these concerns addressed producers are more likely to come forward and share their datasets with other people in the world.

We discuss the *design of the technical method* of linking the original dataset with the licensing restrictions and waiver of warranties in Chapter 6: Operational Aspects of Public Commons at pages 79-82. We propose the use of Commons Identifier software that actually links the dataset with the licensing information and metadata information. It should be assumed for now that users can instantly access the licensing information through an Internet link.

A major consideration is that the Public Commons should provide a legal framework that allows all GIS users of varying knowledge levels to freely download and use the files. A seemingly ideal solution to this situation is to convince the contributors to place their datasets under one of a limited number of open access licenses. In the literature review chapter, we discussed the open access initiatives and an example of Creative Commons licensing approach. Specifically highlighted were the different licensing options that would allow contributors to decide on the amount of control that they might give to users.

Our Public Commons data sharing model follows the open access concepts to work towards an equivalent initiative in the field of spatial datasets. The model encourages data producers to place their datasets with an *Open Access* arrangement. By this we mean, that any dataset available within the archive is freely available on the public internet, permitting any users to search, download, copy, distribute, print, or link to the full contents of these datasets, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself (Budapest Open Access Initiative 2001). The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. In addition to this, it applies a license provision, similar to the Creative Commons Share-a-Like license provision (Creative Commons 2001), to require value-adders to maintain copyrights on all copies and derivative works originated from the datasets downloaded from the archive. Thus, the license requires that *all copies and derivatives retain the same permissions and licenses identical to the original work*, generally referred to as “Copyleft licensing” in the literature. People are allowed to download and use these datasets without any further restrictions when they accept the terms and conditions of the Public Commons model. Linux operating system software is one such example of how an open access license has provided the legal framework for maintaining and expanding a project in the public commons over an extended time period.

A sample open access copyright notice that might be used for datasets available with Public Commons arrangement shown in Fig 3.1. The actual license would likely have far greater detail than the notice. The detailed language of the full license and options for alternative open access licenses are not addressed by this thesis.

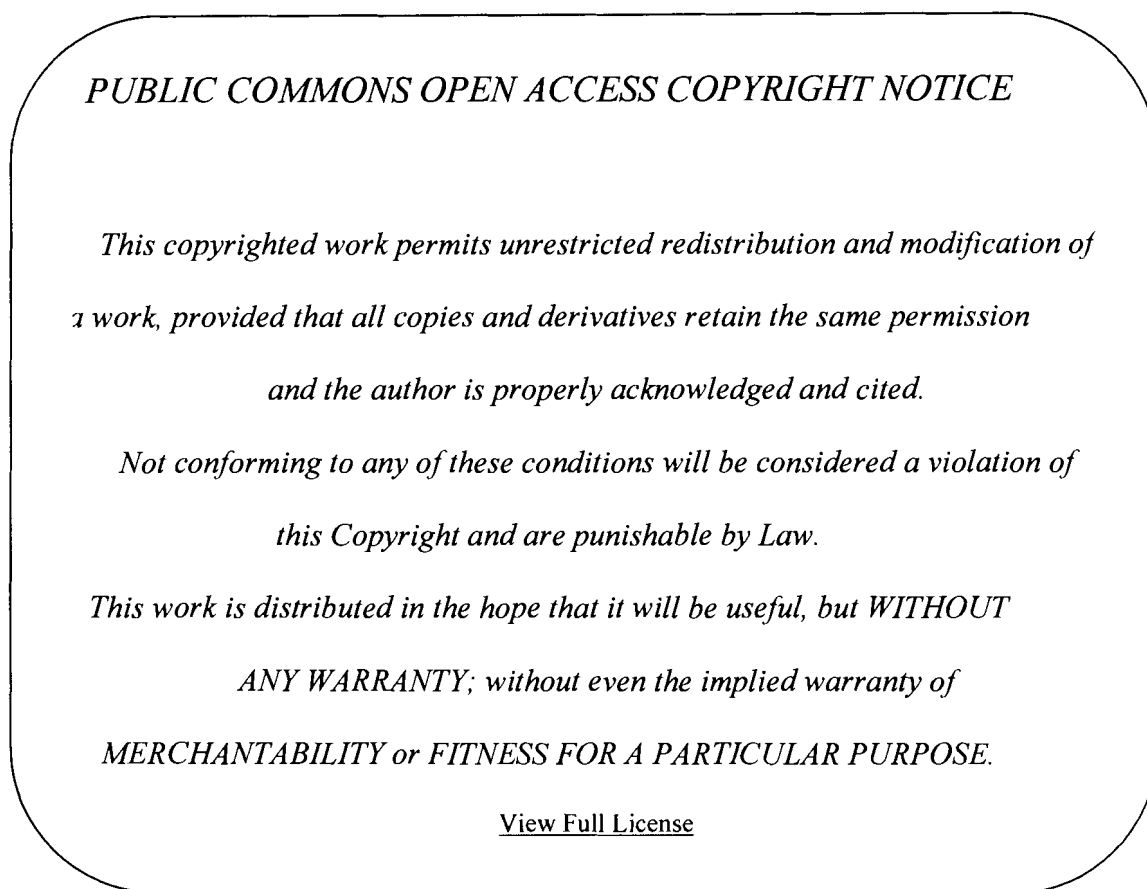


Figure 3.1: A Sample Public Commons Copyright Notice

An alternative to the above license approach is to follow the open access license options being promoted by the Creative Commons (creativecommons.org/works/).

3.4 Other Highlights of the Use of Public Commons License

The Public Commons License is intended to guarantee the user freedom to share and exchange any free dataset available within its archive. Each time a user downloads a dataset he automatically receives a link from the Public Commons displaying the licensing information. The user has the freedom to distribute copies of this dataset (and charge for dissemination services if he wishes); he can change the dataset or use pieces of it in new datasets assuming that the derivative dataset is also made available for free and adheres to the other license provisions.

The user may copy and distribute verbatim copies of the dataset as he receives it, in any medium, provided that any downstream recipient of the dataset is also put on notice of the license conditions. The user may not impose any further restrictions on the downstream recipients' exercise of the rights granted by the license. The user is not responsible for enforcing compliance by third parties to the license. Rights granted to the public are irrevocable.

3.5 Advantages of Using Public Commons Open Access Licensing

The basic concept of an *open access* license is that any subsequent user may freely use the data file. The advantage of using open access licenses in our model are-

- liability exposure may be substantially reduced through the license provisions,
- the originator and all value-adders have a legally enforceable right to credit for their work,

- the license can prevent the efforts of the originator and value-adders from being captured by a company with a large market share or otherwise being removed from an open sharing arrangement, and
- Commons Identification software (discussed at page 83 in Chapter 6) can provide instant access to the detailed licensing language through an Internet link.

Chapter 4

METADATA MODEL FOR PUBLIC COMMONS

4.1 Introduction

Metadata, usually defined as *data about data* (D.Nebert 2001), describe the content, quality, condition and other characteristics of the data. Metadata is the information which can facilitate users or computer systems to access and archive centralized or distributed information services such as datasets, software components and web services (Tsou September 2002). With increasing amounts of geospatial data being created and stored (but often unorganized) there is a real need to document the data for future use. Complete metadata descriptions of the content and the accuracy of a geospatial dataset will encourage appropriate use of the data and avoid duplication of data collection efforts. Such descriptions also may provide some protection for the data supplier if conflicts arise over the misuse of the data.

By making metadata available through digital libraries, data catalogs and clearinghouses, organizations can find data to use, partners to share data collection and maintenance efforts, and customers for their data. Many studies have established that although the value of spatial data is recognized by both government and society, the effective use of spatial data is inhibited by poor knowledge of the existence, origin of data, insufficient documentation on the datasets, and inefficient access and search mechanisms (Zaslavsky 2000; Beard 1996; Timpf, Raubal et al. 1996). One of the possible solutions to overcome all these problems is to develop a *Hierarchical Metadata*

repository model that has web based user interfaces to *collect metadata* (manually or automatically) and improved metadata access and search mechanisms with advanced features such as prioritized query results, *keeping track of subsequent contributors* in instances of value-addition etc. This chapter focusses on the details of the web interface including the metadata elements presented for expert and non-expert GIS users and an improved metadata search mechanism used in the proposed Public Commons data sharing library-like facility.

4.2. Designing Web Interfaces for Public Commons

Metadata available for any spatial data generally means the *What, Who, Where, Why, When and How* of the data (FGDC April 1997). The major difference that therefore exists from many other non-spatial metadata sets being collected for libraries, professions, research and elsewhere is the emphasis on the spatial component - or the *where* element. The descriptions given below are the minimum amount of information that needs to be provided to convey to the inquirer the nature and content of the dataset -

<i>What</i>	gives the title, description, legal status, and administrative information for the dataset
<i>Why</i>	an abstract detailing reasons for the data collection and its uses.
<i>When</i>	time period in which the data set was created and the update cycles if any.
<i>Who</i>	details of the originator, subsequent value-adders, archive, and possibly intended audience.
<i>Where</i>	the geographical extent and location based on latitude / longitude, bounding coordinates, geographical names or administrative areas.

How tools and software etc used for preparation and how to access the data.

With sufficient metadata, users can become familiar with the dataset and be able to make good judgments about its proper use and whether it is appropriate for their applications. The metadata is either gathered during the data collection process itself or some time later. Often data producers do not gather this detailed information during data collection. Documenting metadata later requires considerable effort on the part of the data collector and all information might not be available at that time. Therefore, the accuracy of these details often depends on the metadata editor's skill in documenting these datasets. Consistency in metadata content and style is recommended to ensure that comparisons can be made quickly by data users as to the suitability of the data from different sources. Thus, creating detailed metadata for some spatial datasets while providing only brief descriptions for other datasets does not support consistent and comprehensive cataloguing. Standards for documenting metadata are required to resolve this situation both nationally and globally.

Detailed metadata standards that provide for an exhaustive definition of all aspects of various types of geospatial data suitable for domestic as well as international use are being developed by a number of organizations such as FGDC, ISO (ISO TC 211 Metadata Standard) and the OpenGIS Consortium etc. FGDC's Content Standard for Digital Geospatial Metadata (CSDGM 1998) represents one effort in the United States focused specifically on spatial data with the objective to provide a common set of terminology and definitions for documenting digital geospatial data (FGDC 1997; Office of the President 1994). The CSDGM identifies seven major metadata components -

identification, data quality, spatial data organization, spatial reference, entity and attribute, distribution information and metadata reference information. The standard establishes the names of data elements and compound elements (groups of data elements) to be used for these purposes, the definitions of these compound elements and data elements, and information about the values that are to be provided for the data elements.

Specialized web user interfaces adopted by Data clearinghouses and geo-libraries can be of significant help to users who are already familiar with the subject area and know specifically what they are looking for. But, metadata records based on standards such as FGDC's CSDGM and others tend to be extremely complex, and difficult to read and understand for all but the creator. At the same time, they can be very confusing for users who are unfamiliar with metadata and the FGDC standards. Experts in other domains may be using GIS in their work and creating valuable files. These non-experts in GIS will never take a meta-data course nor will they ever have familiarity with many technical terms. This raises the following question:- *Alternatively, What is the minimum set of metadata information required for documenting a dataset would allow many people of varying knowledge to find spatial datasets meeting their needs? What are the metadata elements that could best describe the dataset and the operational environment that could aid any individual GIS data producer, including novices, to easily fill a metadata form as opposed to a FGDC standard form?*

Historically, the top priority for designing NSDI interfaces has been GIS specialists and government agencies followed by scientific researchers, educators and students at the second level and non-specialist businesses, the general public, commercial

and non-commercial organizations at the third. However, the design of our Public commons conceptual model assumes some members of the general public (*non-expert GIS users*) will also be able to make contributions of valuable spatial datasets. Therefore, the web-interfaces developed under this model should be designed to accommodate the knowledge levels of all groups and encourage people to document their datasets to the maximum level of detail reasonably attainable.

Therefore a Public Commons capability should have user-friendly web interfaces for documentation and both upload and download of datasets. Design considerations for the web-interface are dependent on two different perspectives-

Contributor Perspective: If someone is contributing spatial data to the geo-library, the novice contributor would like to provide only the minimum metadata information required to make their spatial data generally useful to others. If metadata creation is too arduous they are unlikely to complete the web form and therefore their data won't become available. Therefore, the web-interfaces should provide separate transcripts for both non-experts and experts in the field. Non-expert GIS users should be presented with a minimal version of the metadata transcript (discussed in detail in the Chapter 6 on page 75) that *automatically changes subsequent questions intelligently depending on responses to previous questions*. Experts should be presented with the standard CSDGM metadata transcript (sample of elements in Appendix-A)

A comparison of the metadata elements pursued by FGDC, NOAA, Geography network is placed in Appendix-B for further review. Other options for the metadata interface that should be provided are – (1) easy access to previously saved profile

information and some other repetitive technical information would save time instead of re-typing and (2) software that can automatically generate metadata from a dataset. For example Arc Catalog, Metadata collector v2.0, and Meta Lite from USGS collect metadata information from the datasets automatically.

User Perspective: If users wanted to download a dataset from the geo-library, they should have instant access to mainly two things (1) metadata for assessing the fitness of the dataset for their application, and (2) the license and use restrictions on the dataset. Developing a Public Commons Identification software application (discussed in detail in the Chapter 6 on page 83) that can directly retrieve metadata about the purpose of the dataset and licensing restrictions automatically on a click from a remote site on the Internet would be an appropriate solution to meet this need.

From both contributor and user perspective, interfaces should be more interactive, user-friendly and intelligent in order to accommodate different needs and knowledge levels of spatial data producers and users under one common model.

4.3 Existing Metadata Models and Search Mechanisms

One of the important performance conditions in the design of online digital libraries is the pace at which one can access, search and find the required data among the database collections. It mostly depends on the *search engine* implementation that has been adopted for retrieving metadata records and the *way the search results are presented* to the end-use (Walsh and Pancake 2002). Under a traditional meta-database framework, there are two types of metadata *database* implementation approaches. One approach is to

create a *centralized* metadata database and the other is to establish *distributed* metadata repositories that can be accessed through an information gateway server. The search mechanism in either case can be both text-based and interactive map based (i.e. clicking on a location on an interactive map) or a combination of both. The main objectives of both the approaches is to help patrons to index, archive and search distributed geographic information and services.

In this section we compare two different implementation approaches adopted by Alexandria Digital Library and the FGDC data clearinghouses and then put forward our Hierarchical Metadata repository approach - an *enhanced* version over the two.

4.3.1 The Alexandria Digital Library Approach

The digital library approach creates a centralized metadata catalog or database containing standardized metadata records for millions of original spatial datasets. These records are generally stored in MS Access or Oracle databases which can be retrieved with simple SQL statements. Advanced search features with keywords such as names and location or short descriptions of the spatial data may serve the purpose of retrieval. Web based user interfaces designed for on-line digital libraries facilitate the remote access of centralized metadata records. Fig 2.2 (ADL metadata search figure in Chapter 2) illustrates an interface example from Alexandria Digital library (ADL). Figure - 4.1 shows the technological model showing the client / server sides and the database for retrieving metadata information. For a query with spatial location “Los Angeles” and keyword “water” the system would simply query the centralized metadata-database and retrieve many results. In order to narrow down the search response the user should refine

his query with appropriate keywords or zoom to highest resolution of the bounding area at a particular location.

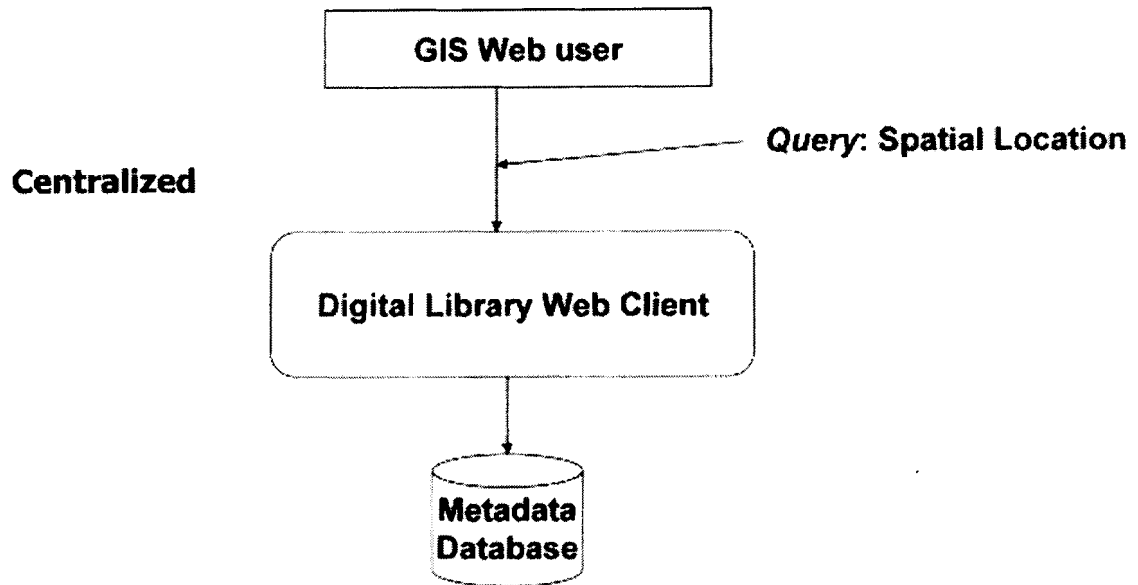


Figure 4.1: Centralized metadata-database structure in Digital Library approach

4.3.2 The FGDC Data Clearinghouse Approach

The FGDC Clearinghouse uses readily available Java Servlet Web technology for the client side and ANSI standard Z39.50 protocol on the server side to index and access multiple metadata repositories placed remotely. The client sends a request to the database server to identify records that meet specified criteria, and later retrieves those records. Figure 4.2 illustrates the web-based interface of NSDI's clearinghouse search form. Clearinghouses use ISITE software, developed by the Center for Networked Information Discovery and Retrieval (CNIDR, www.cnidr.org), to enable querying multiple metadata repositories simultaneously via Z39.50 protocol. ISITE has a built-in search engine

(Isearch) for indexing metadata files. Figure 4.3 shows the mechanism of querying multiple FGDC's clearing house nodes. The metadata database in this case is decentralized and distributed at different physical locations.

ISITE software in each local clearinghouse node (level 2 in Fig 4.3) indexes the metadata records on a regular basis. When each clearinghouse node receives the request from the gateway, their local ISITE Isearch program is initialized to search their metadata index records and then the combined results of all of the houses are sent back to the browser. Users have an option to select different clearing house nodes registered with the FGDC entry gateway.

The screenshot shows a web browser window titled "Clearinghouse Gateway Search Page - Microsoft Internet Explorer". The address bar displays "http://130.11.52.154/remote/FGDC/remote/". The main content area is titled "National Spatial Data Infrastructure Clearinghouse Search Form". It contains three main sections: "Define the Geographic Area of Coverage" with radio buttons for "United States" (selected) and "International", and a list of US states (Alabama, Alaska, Arizona, Arkansas) with a "Reset to Globe" button; "Select Data Servers to Search" with a list of servers including "Africa Data Dissemination Service", "Alaska Area Council of Government", "Alaska Geospatial Data Clearinghouse", "Alaska State Geospatial Data Clearinghouse (ASGDC)", "Anchorage Alaska Geospatial Data Clearinghouse Node", "Argentina - IGM - Instituto Geografico Militar", "Arizona Clearinghouse Node for Spatial Data", "Arkansas Geolibrary", "Australia - WAUS Interrogator - Environmental Impact Statements", "Australia - WAUS Interrogator - Spatial Data", "Australia - ACT Spatial Data Directory", "Australia - AUSLIS Data Directory", "Australia - Australian Hydrographic Service - Product Metadata Directory", and "Australia - BPS - Incorporating Other Commonwealth Data"; and a "Maximum number of records to show on each results page" set to "35 Records". At the bottom are "Search Now" and "Reset this form" buttons.

Figure 4.2: NSDI's Geospatial Data Clearinghouse Search Form

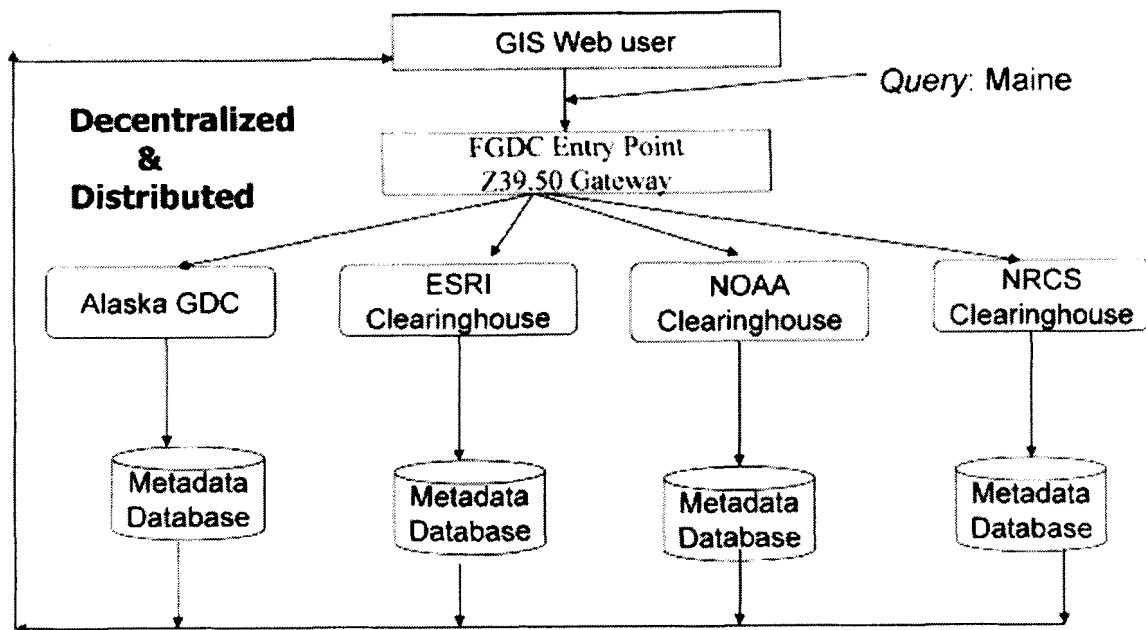


Figure 4.3: The mechanism of Querying multiple FGDC's clearinghouse nodes

For example, the same query of "Los Angeles" (described before in the Alexandria Digital Library approach) is sent to all the selected clearinghouse nodes simultaneously. Each metadata-database is searched for a match simultaneously and finally the combined search results are sent back to the browser.

4.3.3 Comparison of Alexandria Digital Library and FGDC Clearinghouse

Approaches

The Alexandria digital library approach is a straight forward approach that can execute a query quickly when compared to the clearinghouse approach. The clearinghouse approach looks complex with querying of multiple metadata repositories. There are many problems which are associated with the clearinghouse approach. First, the *clearinghouse approach places all distributed clearing nodes on the same level*

without any classification (i.e. not based on the type of the data served). GIS users have difficulty in deciding which clearinghouse nodes may contain metadata they seek. To simply query all nodes requires a lengthy response time.

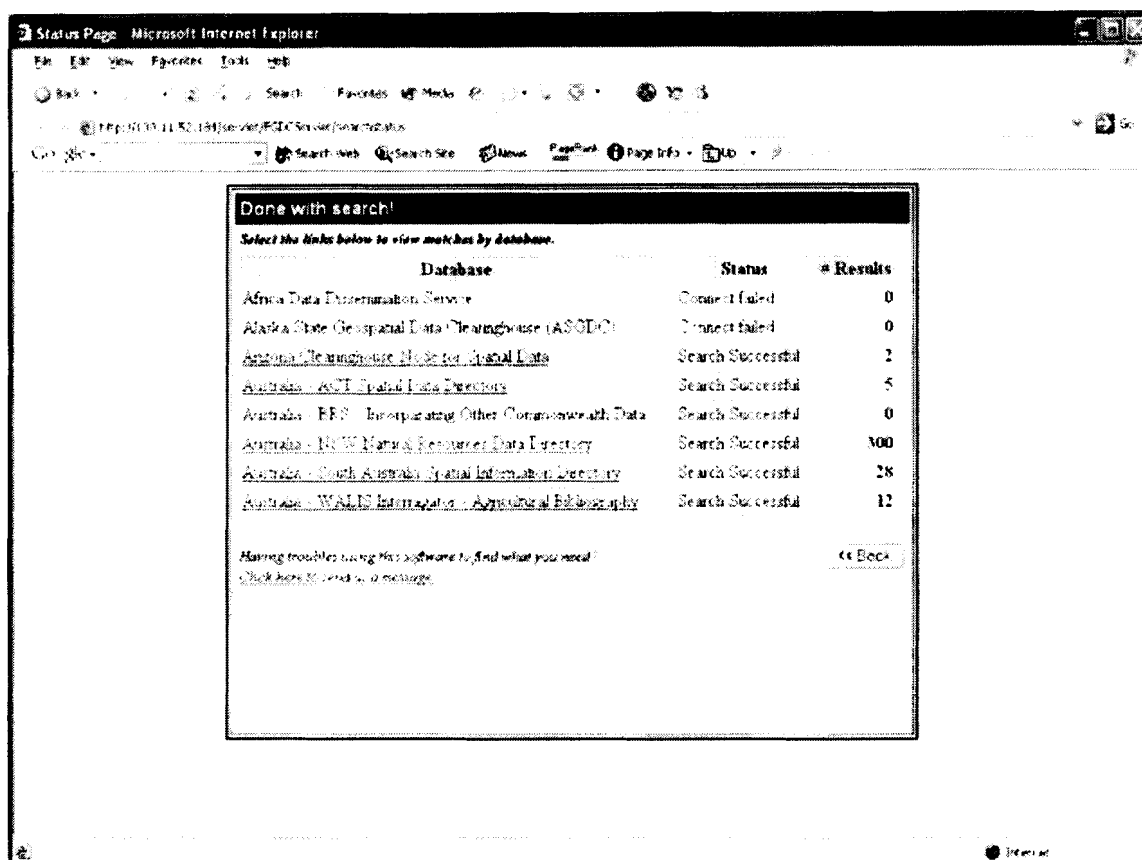


Figure 4.4: Displaying Search Results from FGDC's Metadata

Second, the results of a metadata query are retrieved by individual clearing house servers and not an integrated list ranked by their suitability of content. Therefore the user needs to explore all the individual servers for relevant data which is a painstaking process. If the numbers of clearinghouses increase in the future, the query results could contain hundreds of metadata records (including duplicates) which a user may not be able to process or evaluate. For example, Fig 4.4 shows the results of a simple query that has

resulted in 300 records in one of the clearinghouse nodes and 28 in another node. The data seeker has no option other than refining his query or go through each hit one at a time. A third problem is dealing with duplicate *metadata registrations across clearinghouses*. Many GIS data producers register their works with multiple clearinghouses in order to gain publicity for their datasets over a wide region. These duplicate metadata registrations create inefficiencies for both the contributor and the data seeker.

Similarly, the digital library approach becomes *complex when the numbers of metadata records reach a point where the protocol cannot handle many clientele queries simultaneously*. These problems of metadata implementation frameworks require reconsideration of fundamental metadata model design and index service architectures. The next section introduces a proposed hierarchical metadata repository architecture which promises a more efficient solution for indexing spatial metadata.

4.4 Proposed Hierarchical Metadata Model for Public Commons

The registration framework of FGDC's current data clearinghouse is horizontal and inefficient. As hundreds of clearinghouse nodes are registered at the same level (level 2 from the bottom in fig 3.3), GIS users have difficulty when specifying required nodes from the hundreds of possible selections without prior knowledge. One possible solution is to develop a hierarchical metadata repository as shown in Fig 4.5. Metadata of geographical datasets can be grouped or organized *initially by their spatial locations (North, East, West and South bounding coordinates)* and further *optionally by their theme or data type* under this framework. These meta-databases can be placed at different

physical locations (i.e. decentralized) but still can be accessed from a central information gateway server location as in the case of the FGDC Clearinghouse approach. The major differences being that – (1) for a query for a particular location, *only meta-databases closest to the requested geographical location are accessed and searched* (Geoffrey 1989 and 1999), (2) the results returned would be displayed *by their ranks determined by*

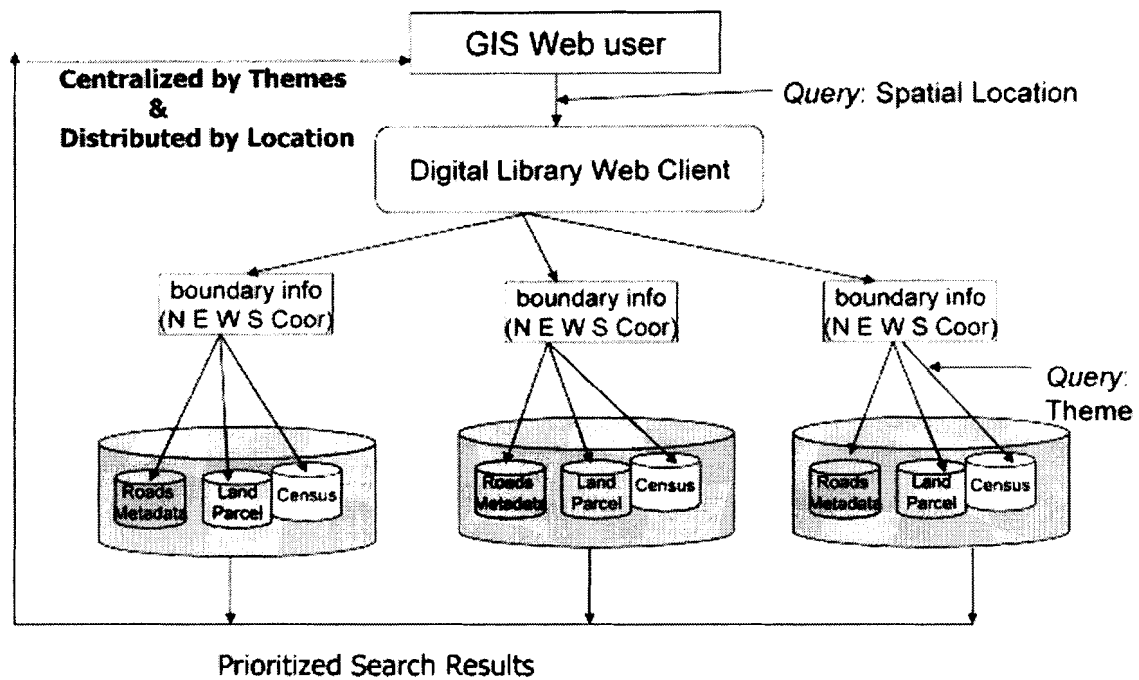


Figure 4.5: Public Commons Metadata Access and Search Mechanism

considerations such as best matches to search criteria, matches in extent of coverage, and the number of times it was browsed or downloaded previously. By adopting this metadata repository model, GIS users can more efficiently access, search, index and distribute datasets and services on the Internet by geographic location.

By example, assume that a grid covers the United States (extent 30-60° N, 60-128° W) at every 3° of the latitude and 3° of longitude and this is the finest level at which we decide to store metadata records. The geographic extents of some contributed spatial datasets might fall within the bounds of the 3° cells. The extents of other contributed spatial datasets might fall within regular nominal grid cells of 5°, 10°, 30°, 90° or 180° (e.g. global datasets). Metadata for any spatial dataset would be stored in affiliation with the smallest cell within which the spatial data is completely bounded.

Now assume that a student has created a vector road dataset for Penobscot County (e.g. bounding box for the project in Penobscot county is (55° 40' N, 55° 45' N, 65° 40', W 65° 46' W) during a school project and submitted metadata information at the Commons website. This metadata record, according to our model, would be stored automatically at the 3° grid cell level in affiliation with the grid boundary by 54°-57°N and 63° W-66° W (Figure 4.6). Any dataset falling within this same extent of 3° cell in Maine would have its metadata stored in affiliation with the same cell. Therefore, a spatial search query for a dataset in California would not retrieve any dataset of Maine area. However, a raw search (i.e. with no other keywords) for spatial datasets of whole Maine might retrieve all datasets falling within all the 3° cells extending over Maine. Thus, a query for any dataset within a bounding box would reach that single metadata cell level most appropriate for the extents and location being sought. The metadata of spatial datasets within cells subsumed within the larger grid cell would also be returned but further down on the lists of hits. Further, datasets can be grouped according to a theme such as roads, river, water, and census. This arrangement eliminates the need to query

multiple clearinghouses. Thus, the entire metadata database of a state or globe can be broken down *recursively* into many bounding regions in a *hierarchical fashion*, be it state-wide or county wide or even the tiniest break down depending upon the amount or demand of datasets at particular geographic levels. Thus the more specific the query in terms of spatial location, the more refined is the query and the closer the user gets to his required dataset.

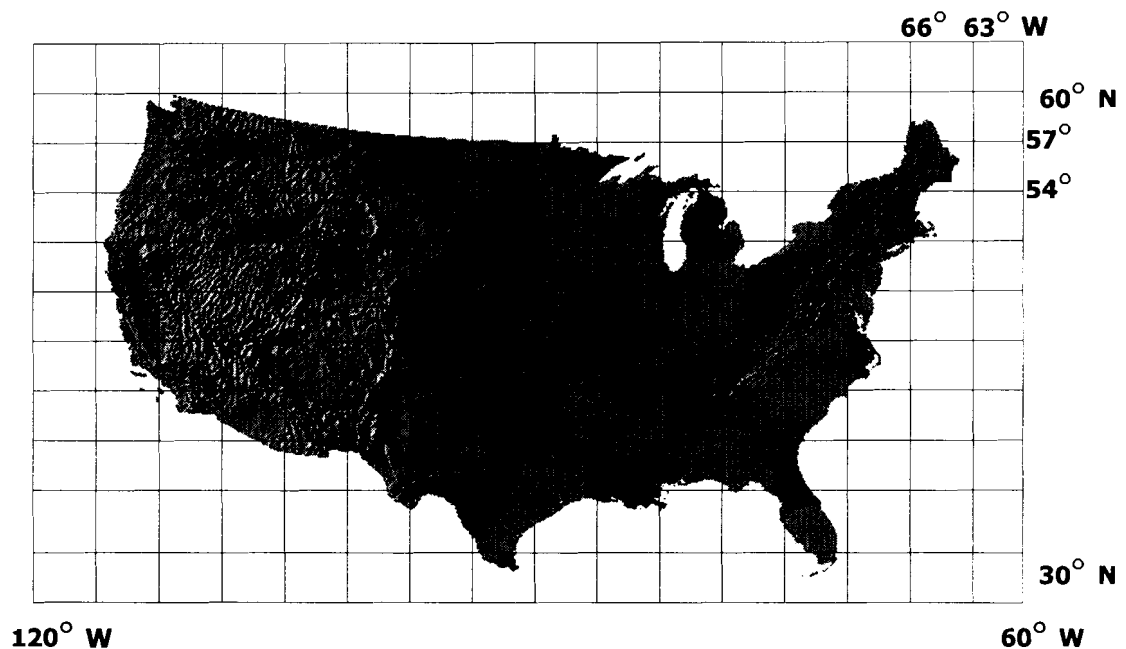


Figure 4.6: 3° X 3° minimum grid laid on U.S. and bounding range for Penobscot query

Note that metadata could be delivered to and accessed from a centralized server (similar to the Geospatial One-Stop concept) or could be implemented across distributed nodes (similar to the Federal Geographic Data Committee clearinghouse concepts). The spatial datasets themselves would likely be retained on the site where they were created but might also be cached or archived on the central server (e.g. similar to Citeseer).

4.5 Benefits of PC Model over Digital Library and Clearinghouse Approaches

The proposed metadata delivery, search and retrieval model seems to be a more *meaningful metadata archive structure* than the current FGDC's approach. The current approach makes it difficult for the normal GIS user to search for datasets with multiple clearinghouses having metadata records of multiple regions. If implemented in a distributed fashion, each parent metadata repository can relay a user's request to its child node (in fig 4.5) and send results back to users. Another advantage is that *duplicate metadata registrations can be eliminated on multiple server locations* i.e. each spatial dataset registered at one clearinghouse need not be registered at any other place on the distributed servers as any query for it would be directed to that particular clearinghouse only. Therefore, new datasets, metadata and services that are added later are only stored at one particular location on these distributed servers. This approach ensures that only records of the requested location are retrieved. This is a significant improvement over the FGDC's metadata model because it does not require querying multiple metadata repositories at different locations simultaneously, thus saving a huge amount of computer processing time, web space, bandwidth and memory storage. Another important advantage with this model is that *each lower level metadata repository can function independently on its own while sharing the same database with the upper level* (in fig-4.5). For example, a person who is searching online for a dataset in Maine at publiccommonsUS.org might be directed to publiccommonsME.org and so on. The most important advantage is that the *results sent back from a query are listed by their ranks* based on matches in geographic extent of coverage, theme of data and other metadata matches, and number of times a dataset was selected previously. This is similar to the

algorithmic approach used by the Google search engine, where web pages with higher numbers of hits are listed which has the most number of hits is listed higher in the query results.

Through this approach GIS users and applications can utilize hierarchical metadata repositories to search for the datasets they need under specific categories rather than search through thousands of items from unorganized data clearinghouses.

Chapter 5

TECHNICAL APPROACH FOR PUBLIC COMMONS

5.1 Introduction

This chapter focuses on the need for a strong technical approach to maintain identification of the contributors of spatial datasets and value-adders, and provide visible credit for their contributions to a spatial data “discovery-access-sharing” system. It will discuss the potential use of Steganography in raster GIS datasets, and a new data hiding technique in the case of vector GIS datasets as a potential technical approach for providing visible credit to the contributors registered with the system.

5.2 The Need for Technical Approach for Datasets with Public Commons

Intellectual Property rights management poses one of the greatest challenges for digital communities in this digital age (Litman 2001). Traditional rights management of physical materials is benefited from the materials’ physicality itself, as this provides some barrier to unauthorized exploitation of content. But this situation is very different for digital datasets. Local governments, research communities, organizations and individuals involved in creation and innovation of different spatial data products fear to bring them open to share with other communities for many reasons such as loss of attribution, liability, plagiarism, and inappropriate ownership claims etc. Once a copy of a dataset is out, it opens doors to a number of unauthorized or inappropriate copying. Many would be more willing to share their spatial data if *some type of technical method was available that would permanently mark their name or other information on the dataset.*

Chapters 1 and 3 have highlighted that at least some data producers have indicated that they would be willing to freely share their spatial datasets with others if they could acquire substantially increased liability protection, and reliably retain visible credit and recognition for their contributions to the public commons. Our recommendation is to provide a *technical option that links the original dataset with a Public Commons Copyright notice and license such that the users of contributed datasets have instant access to the contributor's license terms and embed a hidden identifier in the spatial data file.*

Similarly, data producers and value-adders can be accommodated by providing them with visible credit and recognition for their contributions in all derivative works. The solution is to *somehow design a technical method that can automatically update the list of contributors in the metadata.*

In Chapter 1, we discussed that metadata for a typical spatial dataset is provided in most commercial systems through a separate text or html format files and this metadata is not linked directly with the main spatial data file. Thus, the metadata may be lost when placed separately. Chapter 4 indicated that complete metadata descriptions of the content and accuracy of a geospatial dataset is necessary to determine reasonably the appropriateness of use of the data and to avoid duplication of data collection efforts. Therefore, the recommendation was to *link the separate metadata file to the main dataset*

such that the metadata of the dataset may be retrieved for review instantly from either a local disk or through a link on a remote server maintaining the metadata archive.

The resulting research challenge was to develop a conceptual model and proof of concept whereby one could *link the licensing and metadata information to the spatial dataset, permanently mark identification information directly into the dataset, and automatically update the list of contributors in the metadata.* The next section discusses the different options available for achieving each of these data sharing problems mentioned.

5.3 Suggested Technical Approach

For identifying the originator and assessing ownership, the practice previously was to place visible proprietary logos, copyright notices and some type of false identifiers or information at seemingly unidentifiable locations in the original work (Singh Sept 2000). These methods are since then being employed for copyright protection and data authentication in a wide range of digital media and documents (Craver, Memon et al. May 1998). In a spatial data context, visible proprietary logos and copyright notices work out well on printed maps but they are still vulnerable to “scan - edit - print” attacks i.e. scanning the map, deleting the identifications and printing them again. Moreover, inclusion of such logos and copyright notices in digital spatial datasets would not allow the processing of datasets in many GIS processing systems. This is one of the major reasons why at one time placing tracer data had become popular among cartographic and GIS communities (Lopez 2002). The author could place fictitious objects or false

identifiers such as a road or a street in a map that does not exist in reality or misspell names in a database as a means of identification and proof of their ownership. But these methods destroy the integrity and veracity of the dataset at the expense of the users trust. For example, in some serious instances, a commercial location based company (LBS) using such datasets might lead a LBS user astray. Therefore, such identification methods are not ideal technical approaches for data authentication but may still be used to complement more rigorous and less intrusive approaches.

Basically, in order to identify digital datasets, two types of technical conditions must be accommodated. First, the dataset must be assigned a *unique label or identifier*, which identifies it uniquely as property of the contributor. Second, the dataset should be *permanently marked* in a manner that allows its distribution to be tracked as well as link to the source information at any time. This does not limit the number of copies allowed, but provides a mean to track the data set back in time. In order to catch violations of the licensing provisions, the label must be irremovable and unalterable, and furthermore survive GIS processing operations such as re-projection, and re-sampling. This requires that first the label must be secretly stored (hidden) in the dataset. Thus, the location for embedding the label should be kept as a secret (i.e. invisible) or made inaccessible to the user. Second the label must be robust even if the dataset has been processed incidentally or intentionally. That is the label will remain even after extensive processing of the dataset. These methods can be further categorized into vendor-dependent methods (i.e. each GIS data vendor has their own strategic method of embedding hidden information) and vendor-independent methods (i.e. universal method for all types of GIS datasets).

Here, we will focus on vendor-independent methods rather than on vendor-dependant methods as these methods bring all types of GIS datasets under one uniform identification method and enhance accessibility and automation.

5.3.1 Steganography for Identifying Contributor in Raster Spatial Datasets

Digital watermarks have been proposed recently as a means for copyright protection of multimedia data and seem to be a promising technical approach for our model for identifying contributors of raster datasets. Steganography or Watermarking is the art of hiding extra information in multimedia data in ways that prevent the detection of hidden messages (Zhao and Koch 1995). The extra information (or the watermark) might be an small image or textual matter that can be included in a file and embedded into a carrier file without being noticed. A watermarked image is expected to be indistinguishable from the unwatermarked; original one. Generally, extra information is encoded into the least significant bit of every byte in an image using the most popular Least Significant Bit (LSB) encoding method (Cox and Linnartz 1998). By doing so, the value of each pixel is changed slightly, but not enough to make significant changes to the image except for a small increase in file size and decrease in quality of the data. In contrast to cryptography, steganography does not immediately arouse suspicion of something being present that is secret or valuable (A.P.Petitcolas, J.Anderson et al. April 1998). Further, if the extra information is encrypted then it would be highly impossible for even a seasoned hacker to see what information might be placed there. However, the watermarked image might be susceptible to heavy compression techniques, geometric transformations, format transformations (e.g. shape to DXF and back to shape file).

Through the watermark, extra information such as a small image or text can be embedded in GIS datasets that can identify the originator and metadata information, making it possible to trace the dataset back to its source without destroying its usefulness for the intended application (Craver, Memon et al. May 1998). The data producer can recover the embedded information on request in order to produce evidence of ownership.

LSB Watermarking is readily demonstrable for digital raster imagery involving DRG's (distributed in TIFF format), JPEG's, GIFF's, and IMG's. Multiple software vendors offer watermarking solutions for digital imagery, formatted text, and 3D meshes. Popular software companies such as Steganos, Invisible Secrets by Neobytes Solutions, Datamark Technologies and similar companies are using these methods and provide wide support for digital images in JPEG, PNG, BMP, GIF, PDF, TIFF, and TGA formatted files. Datamark Technologies, Singapore, uses both spread spectrum coding and frequency hopping methods (DigiMarc Tech July 2002) to scatter the watermark over pixels through out the image. They claim that their watermarking methods can also survive "print + scan" attacks.

Steganography is a very complex subject and is an ongoing research focused predominantly in the multimedia arena. Limited applications are yet available in the GIS area. While steganographic methods for image data has limitations, those limitations do not appear to be substantial in the context of placing raster files in a public commons for geographic data where some free riding is tolerable.

5.3.2 Attaching an Invisible Number to Standard GIS Files

Despite the large costs associated with the collection and preparation of spatial datasets, the ‘copy protection means’ has not been to date of particular interest to the GIS research community. Least Significant Bit Steganographic (LSB) methods cannot be universally employed for vector datasets and many raster GIS datasets. Thus, there is no universal procedure existing to date that has been developed that can actually aid in identification across all GIS data formats. In this research work we attempt to develop some methods which apply the core concept of Steganography (i.e. embedding extra information into the dataset) to achieve our objective. A major challenge was to determine *where can the extra information be embedded in digital dataset such that it does not interfere with the processing applications of the file while allowing distribution of the dataset to be tracked?* The first potential solution was to explore placing this information in the *header space* of the digital file format.

5.3.2.1 Why in the Header?

Generally, any file under any operating system has associated with it a header space (or equivalent bytes at some location in the file) where the files attribute information such as name of the file, size or length of the file (in bytes or KB), the file code, version number and other information may be stored. The operating system reads this information every time it needs to access the file and displays the information when requested. Users generally are not able to change some of this information as these are internal to the programming of the operating system. For example, a word format file

(say) “thesis.doc” authored on a computer whose operating system (OS) is registered to (say) X would still display the author as X even when transferred to or modified or copied on any computer whose OS is registered to (say) Y, unless Y copies the contents and saves it as a new file. That means that the other computer’s OS (Y) does not have permissions or access to change the author’s attribute information. This is possible because there are pertinent software programs associated with the computer’s OS (Y) that actually disables the OS’s permission to access that author’s field for a file authored by X. Thus any operations performed on the file would affect the contents of the file but not this particular location in the header. One more example is the PDF file. One cannot print or copy content when protected by a master password.

Using the header space it is possible to embed encrypted messages in the file; if we can programmatically shield the OS’s access to this attribute information in the header (i.e. encrypted ID similar to the author name as discussed above) and make modifications in the header of the dataset such that the OS inserts this ID into each and every copy the user makes. This is a computer science problem and programs can be developed to achieve this. We in our data sharing model attempt to formulate a universal GIS vendor-independent method based on this concept for attaching an invisible number to standard GIS files.

5.3.2.2 Illustrating with an Example

One of the popular GIS vector data formats is ESRI’s Arc Shape file (ESRI July 1998) and is used here for illustration. Table 1 shows the header information of the main

Table 1
Description of the Main File Header

Position	Field	Value	Type	Byte Order
Byte 0	File Code	9994	Integer	Big
Byte 4	Unused	0	Integer	Big
Byte 8	Unused	0	Integer	Big
Byte 12	Unused	0	Integer	Big
Byte 16	Unused	0	Integer	Big
Byte 20	Unused	0	Integer	Big
Byte 24	File Length	File Length	Integer	Big
Byte 28	Version	1000	Integer	Little
Byte 32	Shape Type	Shape Type	Integer	Little
Byte 36	Bounding Box	Xmin	Double	Little
Byte 44	Bounding Box	Ymin	Double	Little
Byte 52	Bounding Box	Xmax	Double	Little
Byte 60	Bounding Box	Ymax	Double	Little
Byte 68*	Bounding Box	Zmin	Double	Little
Byte 76*	Bounding Box	Zmax	Double	Little
Byte 84*	Bounding Box	Mmin	Double	Little
Byte 92*	Bounding Box	Mmax	Double	Little

* Unused, with value 0.0, if not Measured or Z type

Byte 32 signifies the shape of the elements in the file.

Currently, shapefiles are restricted to contain the same type of shape as specified in the table.

In the future, shapefiles may be allowed to contain more than one shape type. If mixed shape types are implemented, the shape type field in the header will flag the file as such.

Value	Shape Type
0	Null Shape
1	Point
3	PolyLine
5	Polygon
8	MultiPoint
11	PointZ
13	PolyLineZ
15	PolygonZ
18	MultiPointZ
21	PointM
23	PolyLineM
25	PolygonM
28	MultiPointM
31	MultiPatch

Table 5.1: Technical description of the ESRI Shape File Format

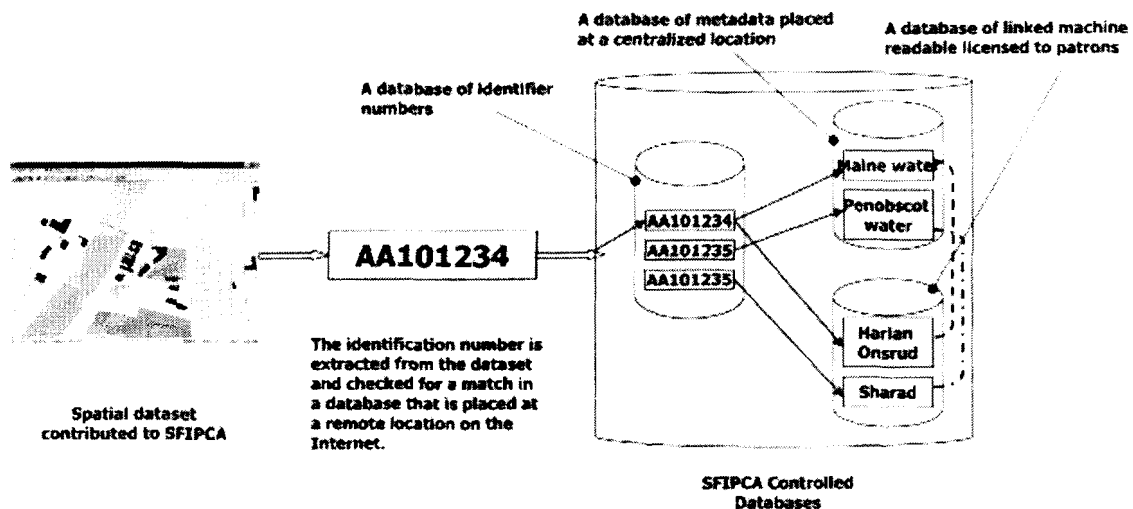


Figure 5.2: The pictorial representation of linking metadata and licenses in a database perspective handled by SFIPCA

fills the metadata web transcript when submitting a dataset to SFIPCA (discussed in detail in chapter 6 on pages 79-82).

5.4 Other Potential Technical Approaches

There are some other technical approaches that are discussed in the literature for copyright protection and data authentication in digital media many of which are computationally intensive or vulnerable to simple attacks (Thoen April 2002). Very few people are working in the GIS arena. William A. Huber of Quantitative Decisions, PA discusses three interesting approaches for the challenging problem of reliably hiding copyright messages or signatures within vector datasets in his article “Vector Steganography” (A.Huber April 2002).

First, *Jittering* (Thoen April 2002) consists of making tiny changes in the vector coordinates. Extra digits of false information for copyright protection can be added to one

of the coordinates, for e.g. a sequence of coordinate numbers like 3.142, 2.783, -1.000, and then 5.9265358979324. The information is contained in the extra digits. Because those digits have low numerical significance--in the example they would not change any single value by more than 0.001--their introduction does not alter the accuracy of datasets considerably. The limitation, however, is that geo-referencing operations often move figures around, rotating them, changing their scale, projecting them (from the earth's surface to a flat map), and un-projecting back again. These processes usually introduce changes in coordinates, thereby destroying any information contained in their least significant digits and making the identification technique unreliable. The method has validity as a backup for files that do not undergo change.

Second, the *Hand writing technique* (Thoen April 2002) is accomplished by adding extra points to the description of a vector figure. Since the points lie on the figure itself, they do not change how it looks; they only change its internal representation.

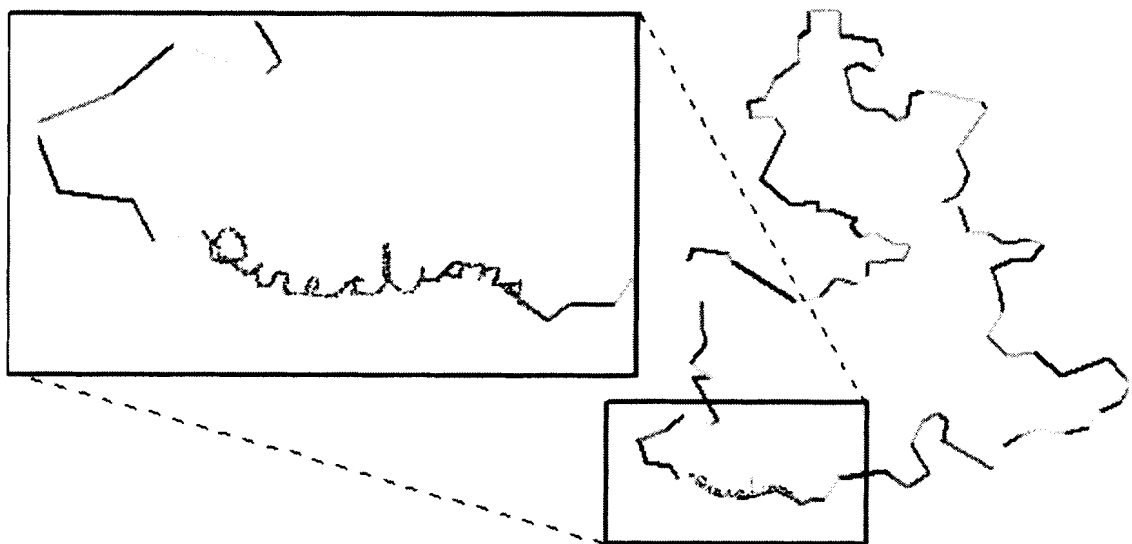


Figure 5.3: Handwriting technique

Fig 5.3 gives an example how it is done. Messages can be hidden by making them extremely small compared to the rest of the figure. The benefit of this method is that the handwriting message will typically survive multiple transformations and processing of the dataset.

There are many limitations to this technique of which the major ones are as follows:- (1) It is inefficient. A large number of new vertices must be introduced to transmit each character, (2) It ruins the shape by introducing self-intersections; this can be a problem for subsequent geographic analysis carried out in software, and (3) the identifier is easily detectable and therefore readily removable.

Third, *Embedding* (Thoen April 2002) consists of inserting a sequence of points along one or more line segments that form a vector figure. The first point establishes a reference length called the *strength of the embedding*. The distances to subsequent points will either equal or be less than the reference length, or exceed its length by some factor. These lengths can be represented as bits in a signal: a long length for a 1, a short length for a 0. The first length is interpreted as the starting 0. Subsequently, any large increase in the next segment length encountered is interpreted as a 1 and any large decrease in length is interpreted as a 0. By focusing on increases and decreases, the decoder does not depend on the exact preservation of relative lengths. In order to send a message consisting of an ASCII "A" (binary value 0100 0001) we must encode the sequence 00100 0001 into one of the lines in the dataset. The longest non-intersecting line segment in the dataset is an ideal startup point. Therefore we can divide that segment into lengths in such a manner that it interprets our binary A. Figure 5.4 demonstrates the way it is

done. If this figure gets distorted in any of the geo-referencing operations, the message can still be read provided the relative lengths along the message do not change by much. The higher the strength of the embedding, the more resistant the message becomes to such distortions. There are technical problems with the simple method just shown, but most of them can be overcome with programming techniques. The major problems are (1) it limits the size of the message to be encoded (2) finding the beginning of the encoded message bits may be complicated to program (3) the method may reduce accuracy and cartographic quality (4) addition of a new line segment to the coded line could destroy the number , and (5) the method is suitable for polyline datasets only (some vector datasets have only point or polygon features).

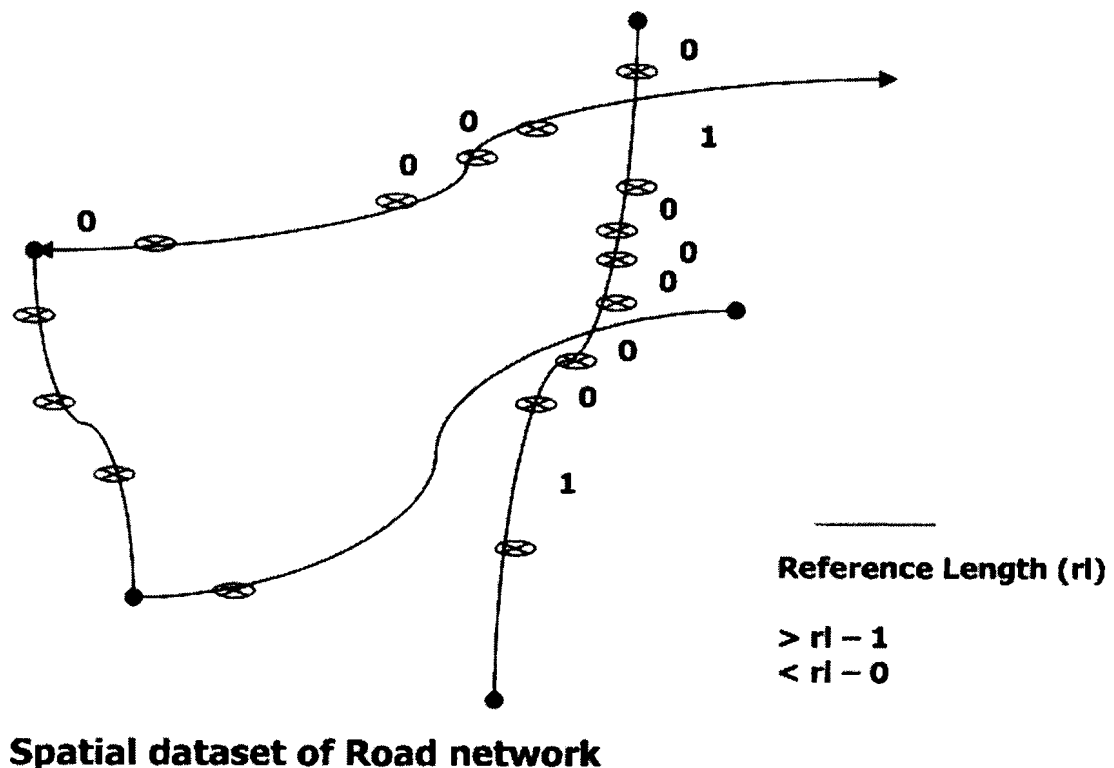


Figure 5.4: Demonstrating Embedding Technique in a vector polyline spatial dataset

For hiding a single identifier number in a typical vector dataset, the embedding method is one of the most promising techniques and trial software is already available. The same identifier number can be inserted along several straight lines or splines to provide redundancy in case one of the encoded numbers is lost. The Embedding method seems to be an efficient technical method for identifying vector files since the method typically survives processing and is largely undetectable.

5.5 Conclusions

These technical methods described above can be used by SFIPCA to embed an identification number into the dataset once or multiple times. Software programs can be developed to read the encrypted identification numbers, decrypt them and then link to the licensing or metadata information from a database stored remotely on a server via the Internet. Whenever a dataset is submitted to SFIPCA with appropriate metadata, SFIPCA executes the programs and automatically embeds a unique identification number into the dataset. In the event that someone takes the file, adds value to it, and resubmits the updated spatial dataset with new metadata to SFIPCA, retrieving the ID would identify the parent file or files. SFIPCA adds a new ID for the updated dataset, and updates the contributor's list to include both the value-adder and originator. This new ID is then linked to the new metadata (a link to the old metadata file is also placed under it as a backup in providing value-adder lineage back to the originator file) and the machine-readable license. The new license is enforced by both the value adder and the original contributor with licensing conditions in force as specified by the originator (the value adder is not given any choice in licensing terms as he has to accept to contributor's open

access share-a-like license - discussed in Chapter 3). Thus, the identity of the originator and the string of value-adders (up to a practical limit) would always be maintained with the succession of files processed in this manner. By using this technical approach we are able to *link the licensing and metadata information, permanently mark identification information directly into the dataset, and automatically updating the list of contributors in the metadata and license.*

By adopting this technical approach for identification, we are able to protect the property interests of the contributors whose goals are to keep the spatial data available in the public commons. While the steganographic methods are not inviolable, they are sufficient for public commons protection since some free riding is acceptable and license breakers who can use the dataset for free anyway, have little incentive to strip the identifiers. These methods need not be fool proof, because it would be easier to follow the license than breach it for the typical user. Using this also method would expose and embarrass license breakers through hidden identifiers and existence of previous similar files by an earlier submitter.

For files marked in this manner, there would be little reason to remove the invisible identification since the file would already be available for free use. The primary thing that stripping away ID information would accomplish would be to establish grounds for a lawsuit against the infringer, who could never be certain that there might not also be hidden identifiers in the datasets.

Chapter 6

OPERATIONAL ASPECTS OF PUBLIC COMMONS MODEL

6.1 Introduction

In the previous chapters we discussed the conceptual approaches used in the model including open access licensing, improved spatial data search and access mechanisms, and data embedding techniques that are required for developing an efficient data sharing facility that can support easy sharing of spatial datasets in a legally supportive manner. This chapter will focus on the internal implementation and operational aspects of the Public Commons data-sharing model that supports user-friendly metadata creation, open access licenses, and documentation of contributor's lineage of spatial datasets. By implementing key elements of the operational system, evidence of proof of concept for the model in entirety is provided.

6.2 Functionality of the Public Commons Model

This section discusses the functionality of the Public Commons Model. A visual representation of the functionality in the form of a flow diagram is shown in the figures 6.1 and 6.2.

6.2.1 Architecture

The Internet is the gateway for information or data sharing in recent times and is becoming increasingly popular in all parts of the world. The Internet can provide the

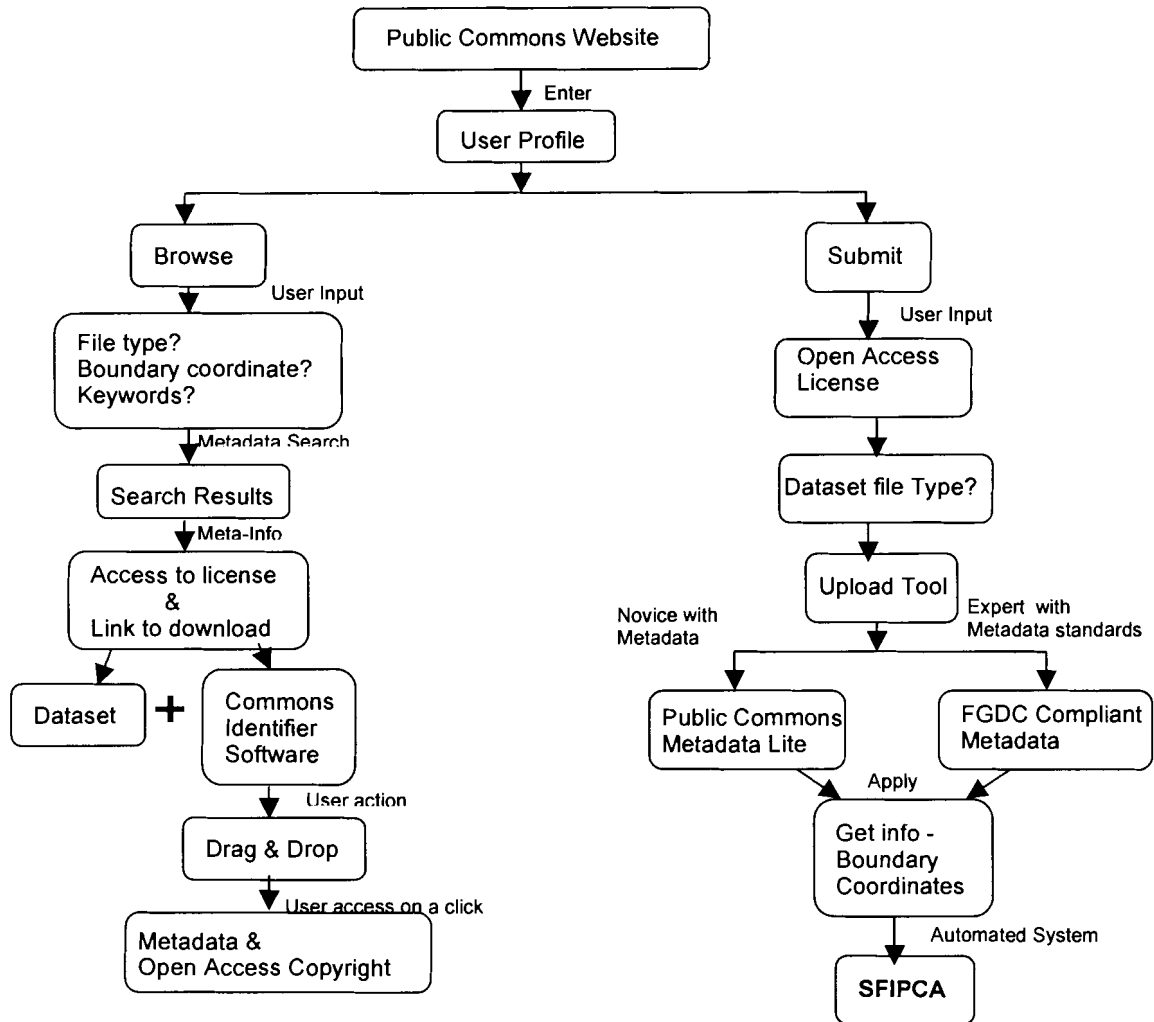


Figure 6.1: Flow diagram of the Operational aspects of Public Commons Model

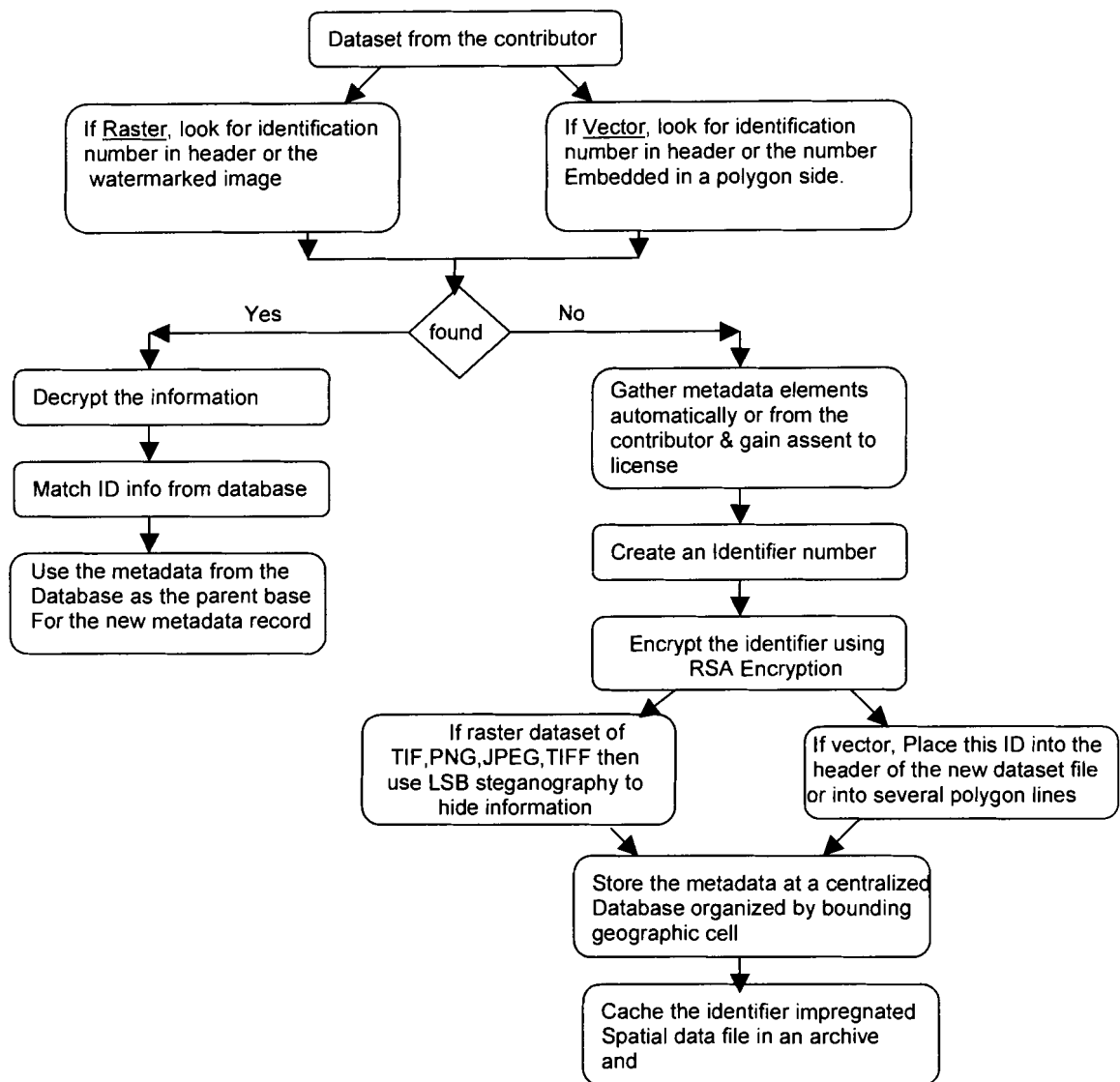


Figure 6.2: Operational aspects of Spatial File Identification Automated System

effective medium for people to share their GIS datasets in digital form with anyone in the world with minimum hassle and expertise.

The Public Commons for Spatial Data recognizes the immense potential of the Internet. The design incorporates an Internet *portal* with centralized metadata database and mechanisms for people to easily upload and download GIS datasets. The Public Commons for data sharing architecture primarily constitutes client-server architecture similar to many other file transfer or data upload mechanisms found on the Internet. In this architecture, the client side consists of a Website using dynamic HTML pages to gather user information and metadata information that are used later for spatial data indexing and searching datasets. A contributor logged on the web site can upload his dataset file to the remote Public Commons server location on the Internet. The server side consists of an automated system (SFIPCA) that automatically determines the storage location and other functionalities such as identification and verification of newly submitted datasets. The server side functionality is quite complex given the amount of pre-processing to be done (i.e. embedding identifiers, updating contributor information etc) before it stores the metadata and the ID embedded spatial dataset at a centralized data base to make it readily available for download. Maintaining such a web site would centralize all metadata and potentially all spatial datasets at one location, which makes it easier for people to search at one location instead of searching at multiple locations (as discussed in Chapter 4).

The website is designed with highly informative, interactive and dynamic web pages that can act intelligently based on user responses to previous questions, to gather

user profile and metadata information for the datasets. For example, a person who is not sure about the bounding coordinates of his dataset is provided with an interactive map of the world where he can identify the extent of his coverage by drawing a rectangle around the place that is transformed later into coordinates by the system.

For such a website involved in data sharing, there would be two types of users: a data contributor and a data user. A data contributor is one who is willing to share his dataset and would like to utilize the services of the system to make their contributions known and accessible to others. A data user is one who wants to download datasets for his application purposes. The public commons website offers two different tracks of services for these users and contributors which can be seen in Fig 6.1 as two different rectangles for Browse and Submit. The later subsections of the functionalities of the model are divided into separate discussions that will concentrate on each of these tracks.

6.2.2 Submitting a Dataset to the Facility

Under this model, any user who creates a GIS dataset can preserve and make their work accessible to the rest of the world by just uploading the dataset to the proposed web portal with appropriate metadata. Before uploading, our prototype web interface requires information about the contributor; a HTML form is provided that includes fields such as name of the organization, name of the contact person, postal address, e-mail and URL for contact. This information can be used for identification and for auto fill in some fields of the metadata transcript. Once the contributor registers as a member to the system, he is

asked if he is going to contribute a dataset to the public commons that day, if yes, he is provided with further instructions and the license agreement for review. The contributor is asked to choose from the pull down menu, the format of the dataset (for example, shape, DXF, DLG, TIFF or other) and is then provided with an upload tool to upload the respective files to a location on the server.

6.2.2.1 User Friendly Metadata transcripts

Under this model, we assume that prototypical contributors might be university researchers and students from a range of disciplinary areas. Examples might be a geology professor who has created numerous GIS datasets related to a research project, or perhaps a junior high school class student that has mapped all the tree species in their community, etc. Non-expert users will never take a Metadata course nor will they ever have familiarity with many technical terms. Therefore open ended questions with free form responses need to be minimized. Thus the model should be able to accommodate the different expertise levels of both GIS novices and experts by providing different metadata transcripts. The non-expert GIS users should be presented with a minimal version of the Metadata transcript with many *user friendly pull down menus, extensive information on mouse-over and auto-fills*. Experts in metadata documentation should be provided with the option of completing the entire standard CSDGM Metadata transcripts.

The minimal version of metadata information as required by the initial suggested public commons metadata transcript is –

- a) File reference ID (default added by the SFIPCA system)
- b) Details of the originator (the system auto-fills the information from the information provided on login)
- c) Title of the content
- d) Presentation form (ex: map, aerial map, base data, shape files)
- e) Abstract or Extensive information for the files above with details such as the details of the data used, what platform is used, what he has worked on, what purposes can it be used for etc.
- f) Time period of the content i.e. the data used was of which year?
- g) Status of the work? (i.e. completed, ongoing, left incomplete)
- h) Information about maintenance work.
- i) Spatial Extent Info (i.e. North, East, West, South bounding Coordinates) with options:
 - i. Do you know the latitude / longitude of North, East, West, South limits? YES NO
 - ii. Do you know limits of the maps or database in any other coordinate system? YES NO
 - iii. Zoom in and draw a box around the approximate extent of your map or database.

Note: This information might be generated automatically or through a bounding rectangle on a map interface.
- j) Data Theme Info

- k) Keywords for the content as well as the place of work, so a search engine can easily identify it.
- l) Spatial Data Info:(1) Data type: Raster / Vector (2) Data format .
- m) Access Constraints: Open Access Licensing protection / Limited rights / None (can be viewed by clicking the link on License agreement)
- n) Use Constraints: Free / Permission required / can or cannot be used for commercial (can be viewed by clicking the link on License agreement).
- o) What type of licensing contributor would insist on? Full description of the licenses and copyright information is explained with strong recommendation for Public Domain.
- p) He is provided with an option of additional distribution of the files from his server apart from hosting from this archive.
- q) If he wants to additionally serve the dataset from his server, then the form asks the contributor to enter a valid URL.
- r) Liability Information

These fields of the public commons version are fundamentally a subset of the FGDC's CSDGM but are standardized in agreeable fashion such that all-important information is included and are easily comprehensible for experts and non-experts. Moreover, some of these Metadata elements are automatically filled by SFIPCA (with specialized software) using processed information directly obtained from the contributor's dataset.

The elements as selected are only illustrative and a first good pass at the minimal set of information required. Experience might show that requiring all these elements causes contributors to not contribute their data. If so, much smaller set of the most critical metadata elements should be required. Some national organizations (NOAA, Geography network, FGDC) and software companies (ESRI, USGC) in these fields are pursuing similar interview approaches and automated population of some metadata fields internal to their software. However, an open access non-proprietary capability (such as SFIPCA) able to process any proprietary data format (E.g. ESRI's shape files, USGC's DLG, AutoCAD's DXF etc.) might allow greater creation of metadata, uniformity and accessibility.

A comparison of characteristics and the number of metadata elements strictly required by the FGDC standards and other organizations with the Public Commons metadata transcripts (See Appendix B) reveals the amount of efficiency, productivity, flexibility achieved with the minimal recommended in this thesis.

6.2.2.2 Using Open Access Licensing

As a part of the series of responses of the Public Commons Metadata transcript, the contributor agrees to (1) apply one of a limited selection of open access licenses to the dataset or (2) dedicate the file to the public domain. Since the public commons model requires the use of an open access license or dedication to the public domain on all works placed under it, any user has unrestricted rights to copy, reproduce, distribute and modify the work, provided the contributor is properly acknowledged and that all copies and

derivatives retain the same license that governs the original work. The advantage of placing a dataset under an open access license over dedication to the public domain has been explained in Chapter 3 (Page-36). SFIPCA automatically includes the previously collected contributor's information into the license agreement and metadata information directly thus providing visible credit for the succession of all contributors (explained in detail in next section).

6.2.2.3 Operational Characteristics of SFIPCA

The dataset uploaded to a web location on the Internet is then processed by the automated system i.e. SFIPCA. Fig 6.2 shows the flow diagram of the operational characteristics of SFIPCA. The system checks if the dataset uploaded to its location is of raster or vector format. For raster format datasets, the system first attempts to find the possibility of an embedded identification number in the header or a watermarked image so as to check if it was a previously contributed dataset. Similarly for vector format datasets, the system attempts to find an embedded ID number or any number embedded in a polygon side. Finding an ID in either case would establish that the dataset was a previously contributed dataset to the system and further updates are done to include original contributor's information. The operations performed in these cases are explained here -

Case I:

Dataset uploaded for the first time: In this case the system has not found any identification information in the dataset. The system gathers the bounding coordinate's

information, and specific metadata and other identification information from the dataset and metadata transcript respectively. Based on the format of the dataset and technological approach discussed in Chapter 5 (pages 58-68), the automated system creates an identification number for the dataset that can serve as a pointer to a metadata record in the database (see figure 5.2 in Chapter 5 on page 64). SFIPCA then encrypts it using RSA Public key Encryption algorithm, and embeds this encrypted identifier number into the header of the dataset file or into several polylines and as watermarked text or image in the case of vector and raster spatial datasets respectively. SFIPCA creates a metadata record and a machine-readable license agreement (which has the contributors name and a brief descriptions of its use, see figure 3.1 in chapter 3 on Page 35) for the dataset and then stores it at a centralized metadata database depending on the bounding coordinates (as accordingly discussed in the proposed hierarchical metadata model in Chapter 4).

Case-II:

Value-added dataset re-submitted to SFIPCA: In the event of value-addition (i.e. someone downloads the file, adds value to it, and resubmits the updated/improved file with new metadata) on a dataset that was previously submitted to the public commons, the system extracts the encrypted information from the header of the dataset file or from those ID embedded polylines or watermarked text or image in the case of vector and raster spatial datasets. This information is then decrypted and checked for a match in the database of identifiers (see figure 5.2 in Chapter 5 on page 64). On a match, the metadata pointed by the identifier is pulled out for previous contributor's information. This original contributor information is appended or hyper-linked to the end of the contributor's list of

the new file in the metadata record as well as well as in the license agreement. Once this process is completed, SFIPCA completes the rest of the process of embedding identifier and metadata record generation for the new data file as discussed in the previous subsection. If required, a mouse-click on the names of the previous contributor's would retrieve the metadata records of their respective original works. For example, Fig 6.3 shows an example where the name "Narnindi Sharad" is added to the contributor's list of the dataset created by "Harlan Onsrud" placed under open access within the Public Commons. Thus, the originator and the string of value-adders (up to a practical limit) would always be maintained with a file processed in this manner.

LICENSE AGREEMENT

This work is protected by the Open Access License as defined by the Public Commons for Geospatial Data

This work is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY, without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

Title of the Work


Analysis of Maine Water Resources

Names of the Contributors

Harlan J. Onsrud
Narnindi Sharad

Description of the work

Digitization of 1:24000 scale map of State of Maine



Public Commons for Geospatial Data

Figure 6.3: Example of a Value added Copyright license

This automated mechanism is developed as a part of the different functionalities of the SFIPCA and provides a solution to the second impediment to data sharing discussed in Chapter 3, in which creators would like to retain credit and recognition for their work by permanently marking their datasets with metadata information.

6.2.3 Downloading a Dataset from the Facility

Under this model, any user who has access to the Internet can easily download GIS datasets archived at the Public Commons geo-spatial data repository. Fig 6.2 Browse track in the flow diagram shows the features available for a data user. The user is asked to choose from a pull down menu the format of the dataset (for example shape, DXF, DLG, TIFF or other), enter a few keywords of his choice and the spatial location that he is interested by selecting a bounding coordinate rectangle on an interactive map. Many examples and combinations of frequently used keywords are provided in the smart menus that can automatically change depending on user responses.

The query on submission returns results ranked based on criteria as discussed in the previous chapter. The user can decide on the use of the datasets by checking the metadata and licensing information provided. Once decided, he can download the dataset from the Public Commons site or from the URL provided there. The user is provided with Commons Identification software (discussed in the next section) that can be installed on any machine. With this software, the user can instantly access metadata and licensing information by just dragging the spatial dataset onto the software icon and letting the

software download this information to the local disk or temporary cache. Thus the user can always have access to this information by using this software either locally or by connecting to the Internet.

6.3 Public Commons Identification Software

The Public Commons model provides freely downloadable software programmed to retrieve the encrypted identification information. In the prototype the assumption is that the ID is drawn from the header of the digital GIS dataset. The software then decrypts this information and transmits this information to the remote online server as a query. Upon request the server sends the metadata information as well as the license language. This Commons identification software acts as client software which will be provided free to anyone who wishes to review the metadata information of datasets downloaded from the public commons digital library at his convenience, provided he is connected to the Internet. Dragging the dataset file on to the software should open a new browser window presenting metadata and the liability information on the use of the dataset. This software solution is developed as a part of the different functionalities of the SFIPCA and seems to be a viable solution to the third problem for data sharing discussed in Chapter 3 where creators would like to minimize their liability through the licensing agreement.

6.4 Conclusions

Through this approach the GIS dataset contributors obtain visible credit when their dataset is used in the products or services of others. By going to the extra time, effort and expense of creating metadata, creators get something in return. Those sharing

through this system obtain a level of liability protection never acquired when data is simply released. Further, they obtain a potential archiving service. The system would allow one's work product to be archived for longer than if one simply left it, for instance, on one's computer or on a web server at a university. Anyone would be able to search for, access, and legally download and use GIS data sets with this system. Thus, the concept has substantial benefits over the metadata and sharing systems currently in operation on the web.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Summary

In this thesis we addressed the problems of wide-scale spatial data sharing faced by the GIS data producers and the need for a supplemental Internet based spatial *data discovery-access system* on a national basis, to better facilitate the availability and access to spatial data to all levels of government, commercial sector and general public. We discussed that information infrastructure building programs such as NSDI, and the Geography Network heavily depend on active participation and contributions from government agencies, the academic community, the private sector, and the non-profit sector in developing shared spatial data resources. Further, some members of these communities have indicated that they would be more willing to share spatial data sets with national infrastructures, if they were provided with user-friendly metadata creation interfaces, improved search and access mechanisms, and techniques that can protect their authorship and retain visible credit and recognition for their contributions. We have discussed a conceptual framework, the *Public Commons for Geospatial Data*, for sharing and discovering GIS data and services on the Internet. It basically provides mechanisms for GIS users to easily publish and access GIS data and services worldwide.

7.2 Conclusions

The conceptual framework of the Public Commons for Geospatial Data is based on the framework of a geospatial data clearinghouse developed by government organizations around the world aimed at facilitating the access, re-use and utilization of

geographic information. The main objective of the Public Commons approach is to provide a variety of non-monetary incentives to people who want to share spatial datasets. The minimized metadata transcripts, identifier embedding, author identification methods and improved search and access mechanisms addressed in the conceptual model are vital components in providing a solution for those tens of thousands of individuals, who are creating GIS datasets with few incentives and little ability to effectively share with the world.

The Public Commons for Geospatial Data conceptual model as outlined is one of several possible approaches in meeting the needs for sharing within and among governments, non-profit and science sectors throughout the globe. This approach cultivates a positive interaction by encouraging individuals, local and federal agencies, private, commercial and non-profit sectors to utilize these raw data resources to add value and create better spatial products for improved decision making and growth of the GIS industry. Continuous value-added contributions of spatial data by these communities to public information infrastructures such as NSDI, the National Map, the Geography Network, Geospatial One-Stop and Public Commons will stimulate the growth and availability of raw data sources from which all sectors of the nation may draw. For example, Fig 7.1 depicts a scenario where value-additions to public domain and public commons GIS datasets could continuously grow and provide an expanding source of freely accessible raw GIS data (i.e. growing shaded area in the figure).

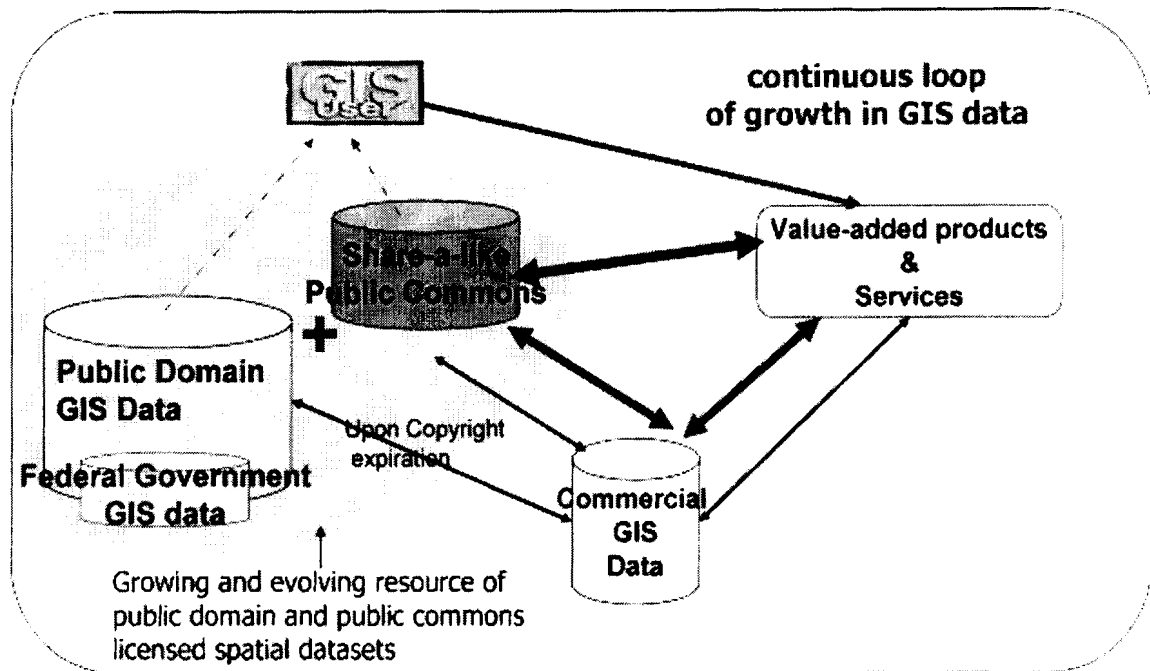


Figure 7.1: A Scenario of evolving Spatial Information resources

By developing such a Public Commons model that allows the effective sharing of spatial data files among expert and non-expert GIS users, the large data collections of private companies and government agencies would become all the more accessible and valuable to society. Moreover, the growing availability of government and private information resources on the Internet, including software and datasets, along with the continued development of data exchange standards and metadata standards all promise to boost the use of geographic information systems.

Here our approach assumes that every spatial data collection effort is valuable and can be used as a starting step for some other project in the future. However, an important issue that has to be mentioned here is that people could be contributing low quality, incomplete or non-reliable datasets to the Public Commons, which would diminish the

appeal of the Commons approach to the user community. This situation would largely depend on the choice and decisiveness of the people who upload and download datasets from such a data sharing facility. We assume here that people who download datasets would at some point have an ability to rate the datasets they download, such as through the methods used by Slashdot.com or e-Bay.com. We have previously discussed that the results returned by the search mechanism (as discussed in Chapter 4) would be ranked by their suitability of the content. Poorly rated datasets would be further down on the retrieved list thereby lessening the likelihood of further downloads.

The model proposes a centralized data sharing facility. That is, all the datasets are stored at one central location. The storage, query and retrieval of the voluminous datasets at a central location might look like an overwhelming task. But considering the current computing and storage technology adopted in large data-warehouses employed by digital libraries, military, credit card companies and national space research agencies, we can imagine that storage would not be a significant problem. The objective here is to develop a permanent archiving facility so that people do not need to worry about losing their datasets to a computer crash and the system always retains a copy for reference in the future. By maintaining such repositories we can secure the datasets at one place that will provide a vast resource of GIS data to different organizations and the general public.

The model would not be very attractive to commercial companies attempting to generate profits from the sale of data. However, commercial companies are moving rapidly towards online intellectual property management systems for their digital files

and are readily able to take care of their own needs. Further, the public commons approach envisioned would provide raw material from which the commercial sector likely would extend, particularly those private companies who view their future as being in the delivery of services and solutions as opposed to the delivery of raw data (Onsrud 2001)

Because the commercial sector would not have a substantial economic interest in the initial development of the outlined conceptual model, the tools and experimentation needed to implement this or similar public data commons models will not arise through marketplace dynamics. Just as the local library did not arise from nor does the marketplace maintain it, an online library for public geographic data will arise only through support by government and taxpayers of the needed research, development, and institutional frameworks to make it happen. Ultimately, such a public commons data sharing model might be best hosted by a federal government agency. We believe that the various capabilities, facilities, and incentives suggested by our conceptual model should be assessed and pursued as one of the possible approaches to promote the coordinated use, sharing, and dissemination of spatial data nationwide.

Finally, we conclude that *given easy ability to create metadata, declare use rights, support upload and sharing mechanisms, and provide visible credit for contributions and access spatial data by way of the Public Commons approach, a significant number of individuals in local to federal government agencies, private*

companies and non-profit organizations would share their spatial datasets through such a system.

7.3 Future work

This research work focused on providing solutions to the impediments of data sharing problems faced by individuals and organizations involved in GIS data creation. Suggestions for future research include-

- In our model, we discussed that the search mechanism is either a text search depending primarily on keywords or a spatial search with selection of location on an interactive map. Further research might develop ontological dictionaries and associate them with the metadata such that the creation of metadata and searching for data becomes more logical and therefore easier for the user.
- Recently, the E-Government Act of 2002 initiated Geo-spatial One-Stop (GOS) to promote coordinated geospatial data collection and maintenance across all levels of government. Being developed is an Internet portal for one-stop access to geospatial data for all levels of government in the U.S. A comparative analysis between the GOS Internet portal architecture and features with those of our Public Commons model might be enlightening.
- The next research questions that may arise here are – How can we try and accommodate people who would like to share spatial databases in our model? What are the other extra components that we may need to consider and develop in that case?

- In our Public Commons model, we discussed the use of steganographic techniques to embed an identifier that can identify the originator and metadata information. Alternative techniques to achieve this task should be explored in greater depth.
- While we have suggested one general conceptual model, further research might investigate alternative conceptual and technical approaches to creating efficient web interfaces, alternative open access licensing approaches, other archival options and additional documenting parent lineage of the contributors and value-adders of newly submitted digital spatial data sets to such a system.

REFERENCES

Bass, J.M, D. Stuve, and R. Tansley, (2000). *DSpace - Technology and Architecture*, MIT Laboratories.

Bauer, K. G. and M. L. Joroff (1969). *Toward a More Comprehensive Community Health Information System*. Proceedings of the URISA'69 Conference, Washington D.C.

Beard, Kate. M. (1996). *Structure for Organizing Metadata Collection*. Proceedings of the Third International Conference on Integrating GIS and Environmental Modeling, Santa Fe, New Mexico.

Berne Convention. (1967). *Berne Convention for the Protection of Literary and Artistic Works*, Legal Information Institute.

The Budapest Open Access Initiative, (2001). *Call for Open Access Literature*, <http://www.soros.org/openaccess/read.shtml>.

Center for Technology in Government. (2001). *Sharing the Costs, Sharing the Benefits: The New York State GIS Cooperative Project*.

Chronicle of Higher Education. (2003). *New web will Enable Scientists to share Data across disciplines*. The Chronicle Magazine.
<http://www.criminology.fsu.edu/book/Cybercriminology/New%20Web%20Will%20Enable%20Scientists%20to%20Share%20Data%20Across%20Disciplines.htm>

Craver, S., N. Memon, B.L. Yeo, and M. M. Yeung. (May 1998). *Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications*. IEEE Journal on Selected Areas in Communications 16(4): 573-586.

Creative Commons. (May 2001). *Creative Commons Metadata Model and Copyright Licenses*, <http://creativecommons.org/learn/technology/metadata/>

I.J, Cox, Linnartz. J.P. (1998). *Some General Methods for Tampering with Watermarking*. IEEE Journal on Selected Areas in Communications 16(4): 587-593.

Department of Environment (1986). *Handling Geographic Information*. London: HMSO.

Duker, K.J., and Vrana R. (1994). *Data sharing in the GIS Community*, International Journal of Geographical Information Science vol-2: 63-78.

Egenhofer, M.J. (1988). *Designing a User Interface for a Spatial Information System*. Proceedings of GIS/LIS '88.

ESRI (July 1998). A White paper: *ESRI Shapefile Technical Description*. Redlands, CA, ESRI: 40.

FGDC (Federal Geographic Data Committee), (April 1997). A White paper: *A Strategy for the NSDI*.

Flewelling, D. and M. J.Egenhofer (1999). *Using Digital Spatial Archives Effectively*., International Journal of Geographical Information Science 13: 1-8.

Geo InSight International, I. (2000). *Management Issues in GIS Development*. Geo World Magazine June 2002.

Geoffrey, H.D. (1999). *A Hierarchical Coordinate System for Geoprocessing and Cartography*. Springer Publications, NewYork, ISBN 3-540-64980-8.

Geoffrey, H.D. (1989). *Modeling Locational Uncertainty via Hierarchical Tessellation*. The Accuracy of Spatial Databases - Goodchild M.F. and Sucharita G. (Ed's) London: Taylor & Francis. Vol.pp.125-140.

GNU-GPL (1991). *General Public License for Software*, Free Software Foundation, Inc. 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

Hart, M.S. (1997). *Project Gutenberg: Fine Literature Digitally Re-published*, Promo.net. <http://gutenberg.net/>

Huber, A.W. (April 2002). *GIS & Steganography-Part 3: Vector Steganography*. Directions Magazine. http://www.directionsmag.com/article.php?article_id=195

Iannella, R. (June 2001). *Digital Rights Management Architectures*. D-Lib Magazine. vol 7. <http://www.dlib.org/dlib/june01/iannella/06iannella.html>

Koontz, D.L. (June 2003). *Geographic Information Systems: Challenges to Effective Data Sharing*, United States General Accounting Office.

Karjala, D. S. (1995). *Copyright in Electronic Maps*, Jurimetrics Journal 35: 395-415.

Katzenbeisser, S. and F. A. P. Petitcolas (January 2000). *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House Books.

Kitch, E. (1980). *The Law and Economics of Rights in Valuable Information*. Journal of Legal Studies 9: 683.

Litman, J. (2001). *Digital Copyright in the New Millenium*, Prometheus Books.

Longley, A.P., M. F. Goodchild, et al., Eds. (2002). *Geographical Information Systems. Principles and Technical Issues*. John-Wiley Publications.

Lopez, Carlos. (2002). *Watermarking of digital geo-spatial datasets: a review of technical, legal and copyright issues*. International Journal of Geographical Information Science. 16: 589-607.

Lopez, X. R. (August 1996). *The Impact of Government Information Policy on the dissemination of Spatial Data: A North American-European Comparative Study*. M.S. Thesis, University of Maine.

Luce, R. (2001). *Evolution and scientific literature: towards a decentralized adaptive web*. Nature Magazine. <http://www.nature.com/nature/debates/e-access/Articles/luce.html>

Mary M. Case, (Feb-Mar 2002). *Promoting Open Access: Developing New Strategies for Managing Copyright and Intellectual Property*, ARL Bimonthly Report 2002.

Mason, R. (1998). *Spatial Data Provider and Data User Survey*. University of New South Wales. <http://www.gmat.unsw.edu.au/survey/questionnaire/>
<http://www.gmat.unsw.edu.au/survey/report/>

Metadata Ad Hoc Working Group. (1998). *Content Standard for Digital Geospatial Metadata (CSDGM)*. Reston, VA, National Spatial Data Infrastructure: 90.

Narnindi, S. and H. J. Onsrud (September 2002). *Public Commons for Geographic Data: A Conceptual Model*. Second International Conference, GIScience '2002, Boulder, CO.

National Academic Press. (1995). *A Data foundation for the National Spatial Data Infrastructure*. Commission on Geoscience, Environment and Resources (CGER).

National Academy of Public Administration (1998). *Geographic Information for the 21st Century: Building a Strategy for the Nation*. Washington D.C.: 171.

Nebert D. Ed. (2001). *Developing Spatial Data Infrastructures: The SDI Cook Book*. Global Spatial Data Infrastructure.

NRC, Computer Science and Telecommunications Board, (2000). *The Digital Dilemma: Intellectual Property in the Information Age*, National Academy Press.

Nimmer, R. and A. K. Patricia (1992). *Information as a Commodity: New Imperatives of Commercial Law*. Law and Contemporary Problems 55 (Summer): 103-115.

NRC, Summary of National Research Council Workshop, Ed. (1999). *Distributed Geolibraries: Spatial Information Resources*. Washington, DC, National Academy Press.

Office of the President, (April 1994). *Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure*. Edition of the Federal Register 59(71): 17671-17674.

Office of the President, (October 1990). Circular No. A-16, *Coordination of Surveying, Mapping, and Related Spatial Data Activities*.

OpenGIS Consortium Inc. (July 1997). *Data Sharing*. Geo Info Systems Magazine: 32.

OpenGIS Consortium Inc. (May 2002). *Piecing It Together: Interoperability Enhances Planning and Response*. Geo World Magazine.

OpenGIS Consortium Inc. (December 2002). *Request for Quotation in the OGC Geospatial One-Stop Portal Initiative*. Bluomington, IN.

OMB (Office of Management and Budget), Ed. (January 2003). *Geospatial One-Stop, Office of Management and Budget Capital Asset Plan and Business Case* (Exhibit 300).

Onsrud, H. J., Ed. (1999). *Liability in the Use of Geographic Information Systems and Geographic Data Sets*. Geographic Information Systems: Principles, Techniques, Management, and Applications. New York: Wiley.

Onsrud, H. J. (2001). *The Public Commons of GIScience: A Funding Proposal*, University of Maine: 12.

Onsrud, H. J. (March 2000). *Global Survey and Analysis of National Spatial Data Infrastructure Activities*. 4th Global Spatial Data Infrastructure Conference, Cape Town, South Africa.

Onsrud, H.J. and X. Lopez (1998), *Intellectual Property Rights in Disseminating Digital Geographic Data, Products, and Services: Conflicts and Commonalities among European Union and United States Approaches*. In Masser, Ian and Francois Salge, eds., *European Geographic Information Infrastructures: Opportunities and Pitfalls* (London: Taylor and Francis), pp. 153-167.

Onsrud, H. J. and G. Rushton (1995). *Sharing Geographic Information*. Center for Urban Policy and Research, Rutgers University.

Petitcolas, F.A.P., R. J. Anderson, et al. (April 1998). *Attacks on Copyright Marking Systems*. Lecture Notes in Computer Science, Portland, Oregon, USA. 1525: 218-238.

Plewe, B. and S. Johnson (1999). *Automated Metadata Interpretation to Assist in the Use of Unfamiliar GIS Data Sources*. Interoperating Geographic Information Systems. <http://www.ncgia.ucsb.edu/conf/interop97/program/papers/plewe.html>

Pluijmers, Y. (1998). *Protecting Intellectual Property in Private Sector Spatial Datasets*. M.S. Thesis, University of Maine, Orono: 104.

Rogers, E. (1995). *Diffusions of Innovations*. New York: Free Press.

Samuels, E. (1993). *The Public Domain in Copyright Law*. Journal of the Copyright Society of the USA 41(Winter): 137-182.

Schneier, B. (2000). *Secrets and Lies: Digital Security in a Networked World*, Wiley Computer Publishing.

Singh, S. (September 2000). *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, First Anchor.

DigiMarc Technologies, (July 2002). *Protecting Images and Video in the Digital Age*. Tualatin Oregon, USA, Digimarc: 32.

Thoen, B. (April 2002). *GIS & Steganography - Part 1: Hidden Secrets in the Digital Ether*. Directions Magazine,
http://www.directionsmag.com/article.php?article_id=189.

Timpf, S., M. Raubal, and W. Kuhn (August 1996). *Experiences with Metadata* 7th International Symposium on Spatial Data Handling, SDH'96, Delft, Netherlands: pp 12B.31 - 12B.43.

Tsou, Ming-Hsiang (September 2002). *An Operational Metadata Framework for Searching, Indexing, and Retrieving Distributed Geographic Information Services on the Internet*. Second International Conference, GIScience 2002, Boulder, CO.

U.S. Congress, (1986). *Intellectual Property Rights in an Age of Electronics and Information*, OTA-CIT-302. Washington D.C: U.S. Government Printing Office, Office of Technology Assessment.

A.Walsh, K., C. M.Pancake, et al. (September 2002). *"Humane Interfaces to Improve the usability of Data Clearinghouses."* Geographic Information Science, Second International Conference, GIScience2002 LNCS 2478: 333-345.

Zaslavsky, I. *Archiving Spatial Data: Research Issues*, San Diego Super Computer Center.

Zhao, J. and E. Koch (1995). *Embedding Robust Labels into Images for Copyright Protection*. Proceedings of the Intl. Congress on Intellectual Property Rights for Specialized Information, Knowledges and New Technologies, Vienna.

Zhong-Ren P, Nebert D. (1997). *An Internet-Based GIS Data Access System*, URISA Journal, Vol 9 Spring'97.

APPENDIX - A

Common File Formats for Spatial Data

Source: Natural Resources and Environmental Management
<http://www.edc.uri.edu/training/gis&www/formats.htm>

Name of Format	Sample File Name	Description
Vector (& Sometimes Raster) Data		
Arc/Info Export	Wetlands. E00	This is a proprietary file format used to distribute Arc/Info datasets. Topology and attributes are properly maintained in this format.
Arc Shape	Wetlands. SHP Wetlands. DBF Wetlands. SHX Wetlands. SBX Wetlands. SBN	Shape files are used in ArcView. A shape file actually consists of 3-6 separate files and might be delivered as a single ZIP'ped file. Shape files do not have topology.
ArcView Project	Wetlands. APR	Project files contain the names and locations of all pieces of data associated with an ArcView project. They can be viewed in a text editor.
Zip	Wetlands. ZIP	One or more files compressed and consolidated into a single binary file. Use WinZIP to uncompress and extract the contents of the ZIP file.
Self Extracting Zip	Wetlands. EXE	The same as above, but does not require WINZIP to uncompress and extract the contents. Simply double click on the EXE file and it will unzip itself.
MIF	Wetlands. MIF	Map InFo vector format. MapInfo is a popular GIS software product.
DXF	Wetlands. DXF	These are drawing files from CAD systems.
DWG	Wetlands. DWG	These are drawing files from CAD systems.
DGN	Wetlands. DGN	A file format used by some CAD systems (e.g., Design files from Bentley MicroStation)
TIGER		TIGER files contain the data from our national censuses.
VPF	Wetlands. VPF	Vector Product Format is commonly used in military applications.
DLG	Wetlands. DLG	Digital Line Graph file format is common on some government web systems and is a way to move data from one GIS system to another.

SDTS	Wetlands. SDTS	The Spatial Data Transfer Standard format will hopefully replace most others in the next few years. It is designed to be the single, standard file format for distributing spatial data. The USGS uses SDTS as a common format already.
Image Data		
TIFF	Wetlands. TIF	An image file format. The resolution can be very good and the image can be georeferenced (meaning that you can overlay other GIS data on top). The georeferencing often accompanies the TIF file as a second file called the TIFF world file and carries an extension like .tfw. TIFF files can be very large, there is little compression.
GIFF	Wetlands. GIF	A common format for image data. GIF does not support georegistration. Image resolution can be excellent and file sizes modest.
JPEG	Wetlands. JPG	A common format for image data. JPEG can support georegistration. Image resolution can be excellent and file sizes can be quite small. JPEG compresses images nicely but there can be some loss of resolution.
SID	Wetlands. SID	A very efficient compression format for image data. Many GIS data viewers can directly read images compressed using "Mr. SID" compression tools.
Raster Data		
DEM	Elevation. DEM	Topographic data sometimes come as DEM's -- Digital Elevation Models -- a format used by USGS.
LAN	Wetlands. LAN	A file format used by Erdas image processing software.
IMG	Wetlands. IMG	Erdas Imagine uses this format for satellite and other image data.
BIL	Wetlands. BIL	Band Interleaved format is a common format for distributing satellite image data.
BSQ	Wetlands. BSQ	Band SeQuential format is a common format for distributing satellite image data.
Miscellaneous Formats		
VRML	Paris. VRML	Virtual Reality Markup Language -- a web-based format for viewing 3-D animations. Frequently used for displaying fly-over animations in GIS and manipulating 3-D renderings of spatial data.

Table A.1: Common File Formats for Spatial Data

APPENDIX – B

Comparison of Different Organizational Metadata Transcripts

NOAA and FGDC Metadata Standard

The following is the template mixture of NOAA's optional metadata fields with FGDC's Content Standard for Digital Geospatial Metadata (CSDGM). NOAA indicated optional metadata fields are represented with shaded area.

Source: NOAA Website: Page created by Peter Grimm on May 8, 1997.

The first element (key) of each pair is the FGDC paragraph number (referenced to the "Green Book"), and the second element of each pair (argument) is the FGDC paragraph heading, preceded by a two-character code indicating the 'optionality' and 'repeatability' of the element, and then by one or more spaces indicating the level of indentation of the element. The presence of a colon at the end of the second element indicates that a value should be appended; headings without colons are used only to provide context for the headings below.

```
# For the 'optionality' character:
#   '*' indicates "mandatory for NOAA descriptions,"
#   '@' indicates "mandatory if applicable," and
#   '?' indicates "optional,"
#   'A' (any letter) indicates that at least one of the
#       headings with this optionality letter must be included.
#
# For the 'repeatability' character:
#   '.' indicates "one occurrence only,"
#   '+' indicates "may be repeated indefinitely."
#
# Note that both optionality and repeatability are relative to
# their respective superior headings. Thus, '1.1.8.7.1 Series
# Name: ...' and '1.1.8.7.2 Issue Identification: ...' are
# mandatory only if '1.1.8.7 Series Information' is present.
# '1.1.8.1 Originator: ...' may be repeated under '1.1 Citation'
# even though '1.1 Citation' can appear only once in a description.
# Conversely, '1.6.1.1 Theme Keyword Thesaurus: ...' can appear
# only once for each '1.6.1 Theme' heading, but the entire
# '1.6.1 Theme' section may be repeated (presumably citing a
# different thesaurus for each occurrence) several times.
```

```
*. 1 Identification Information
*. 1.1 Citation
*+ 1.1.8.1 Originator:
*. 1.1.8.2 Publication Date:
*. 1.1.8.4 Title:
@. 1.1.8.5 Edition:
@. 1.1.8.6 Geospatial Data Presentation Form:
@. 1.1.8.7 Series Information
*. 1.1.8.7.1 Series Name:
*. 1.1.8.7.2 Issue Identification:
@. 1.1.8.8 Publication Information
*. 1.1.8.8.1 Publication Place:
*. 1.1.8.8.2 Publisher:
?. 1.1.8.10 Online Linkage:
*. 1.2 Description
*. 1.2.1 Abstract:
*. 1.2.2 Purpose:
?. 1.2.3 Supplemental Information # ":" is optional.
```

?. 1.2.4 NOAA Supplemental Information # New.
 ?. 1.2.4.1 Entry ID: # New.
 ?+ 1.2.4.2 Sensor Name: # New.
 ?+ 1.2.4.3 Source Name: # New.
 ?+ 1.2.4.4 Campaign or Project: # New.
 ?. 1.2.4.5 Originating Center: # New.
 ?+ 1.2.4.6 Storage Medium: # New.
 ?. 1.2.4.7 Reference: # New.
 ?. 1.2.4.8 NEDRES Specific Information # New.
 ?. 1.2.4.8.1 NEDRES..GC-GEOGRAPHIC CODES: # New.
 ?. 1.2.4.8.2 NEDRES..LR-LENGTH OF RECORD: # New.
 ?. 1.2.4.8.3 NEDRES..AN-ACCESSION NUMBER: # New.
 ?. 1.2.4.8.4 NEDRES..CC-CATEGORY CODES: # New.
 ?. 1.2.4.8.5 NEDRES..AV-AVAILABILITY CONDITIONS: # New.
 ?. 1.2.4.8.6 NEDRES..PR-PROGRAM SPONSOR, CONTRACT, PROJECT,
 OR EXPERIMENT NAME: # New.
 ?. 1.2.4.8.7 NEDRES..PU-PUBLICATIONS: # New.
 ?. 1.2.4.8.8 NEDRES..DC-DATA COLLECTION DESCRIPTION: # New.
 ?. 1.2.4.8.9 NEDRES..DD-DATA CENTER PROCESSING DESCRIPTION: # New.
 ?. 1.2.4.8.10 NEDRES..DE-ADDITIONAL DATA DESCRIPTORS: # New.
 ?. 1.2.4.8.11 NEDRES..PO-PROCESSING/COLLECTING ORGANIZATION: # New.
 ?. 1.2.4.8.12 NEDRES..DT-DATE ENTERED/UPDATED: # New.
 ?. 1.2.4.8.13 NEDRES..RR-RELATED RECORDS: # New.
 ?. 1.2.4.8.14 NEDRES..GL-GRID LOCATORS: # New.
 ?. 1.2.4.9 NOAA Server URLs # New.
 ?. 1.2.4.9.1 More Information: # New.
 ?. 1.2.4.9.2 Preview: # New.
 ?. 1.2.4.9.3 Obtain: # New.
 *. 1.3 Time Period of Content
 *. 1.3.1 Currentness Reference:
 *. 1.3.9.3 Range of Dates/Times
 *. 1.3.9.3.1 Beginning Date:
 ?. 1.3.9.3.2 Beginning Time:
 *. 1.3.9.3.3 Ending Date:
 ?. 1.3.9.3.4 Ending Time:
 *. 1.4 Status
 *. 1.4.1 Progress:
 *. 1.4.2 Maintenance and Update Frequency:
 *. 1.5 Spatial Domain
 ?+ 1.5.1 Bounding Coordinates
 *. 1.5.1.1 West Bounding Coordinate:
 *. 1.5.1.2 East Bounding Coordinate:
 *. 1.5.1.3 North Bounding Coordinate:
 *. 1.5.1.4 South Bounding Coordinate:
 *. 1.6 Keywords
 ?+ 1.6.1 Theme:
 *. 1.6.1.1 Theme Keyword Thesaurus:
 ?+ 1.6.1.2 Theme Keyword:
 ?+ 1.6.2 Place
 *. 1.6.2.1 Place Keyword Thesaurus:
 ?+ 1.6.2.2 Place Keyword:
 ?+ 1.6.3 Stratum
 *. 1.6.3.1 Stratum Keyword Thesaurus:
 ?+ 1.6.3.2 Stratum Keyword:
 ?+ 1.6.4 Temporal
 *. 1.6.4.1 Temporal Keyword Thesaurus:
 ?+ 1.6.4.2 Temporal Keyword:
 *. 1.7 Access Constraints:
 *. 1.8 Use Constraints:
 ?. 1.9 Point of Contact: # "Investigator" or "Technical Contact".
 A. 1.9.10.1 Contact Person Primary
 *. 1.9.10.1.1 Contact Person:
 ?. 1.9.10.1.2 Contact Organization:

- A. 1.9.10.2 Contact Organization Primary
 - *. 1.9.10.2.1 Contact Organization:
 - ? 1.9.10.2.2 Contact Person:
 - ? 1.9.10.3 Contact Position:
 - *+ 1.9.10.4 Contact Address
 - *. 1.9.10.4.1 Address Type:
 - @+ 1.9.10.4.2 Address:
 - *. 1.9.10.4.3 City:
 - *. 1.9.10.4.4 State or Province:
 - *. 1.9.10.4.5 Postal Code:
 - ? 1.9.10.4.6 Country:
 - *+ 1.9.10.5 Contact Voice Telephone:
 - ?+ 1.9.10.6 Contact TDD/TTY Telephone:
 - ?+ 1.9.10.7 Contact Facsimile Telephone:
 - ?+ 1.9.10.8 Contact Electronic Mail Address:
- @. 2 Data Quality Information
 - *. 2.2 Logical Consistency Report:
 - *. 2.3 Completeness Report:
 - *. 2.5 Lineage
 - *+ 2.5.2 Process Step
 - *. 2.5.2.1 Process Description:
 - *. 2.5.2.3 Process Date:
- @. 4 Spatial Reference Information
 - @. 4.1 Horizontal Coordinate System Definition
 - *. 4.1.1 Geographic
 - *. 4.1.1.1 Latitude Resolution:
 - *. 4.1.1.2 Longitude Resolution:
 - *. 4.1.1.3 Geographic Coordinate Units:
 - @. 4.2 Vertical Coordinate System Definition
 - @. 4.2.1 Altitude System Definition
 - *. 4.2.1.1 Altitude Datum Name:
 - *+ 4.2.1.2 Altitude Resolution:
 - *. 4.2.1.3 Altitude Distance Units:
 - *. 4.2.1.4 Altitude Encoding Method:
 - @. 4.2.2 Depth System Definition
 - *. 4.2.2.1 Depth Datum Name:
 - *+ 4.2.2.2 Depth Resolution:
 - *. 4.2.2.3 Depth Distance Units:
 - *. 4.2.2.4 Depth Encoding Method:
- @+ 6 Distribution Information
 - *. 6.1 Distributor
 - *. 6.1.10.2 Contact Organization Primary
 - *. 6.1.10.2.1 Contact Organization:
 - ? 6.1.10.2.2 Contact Person:
 - ? 6.1.10.3 Contact Position:
 - *+ 6.1.10.4 Contact Address
 - *. 6.1.10.4.1 Address Type:
 - @+ 6.1.10.4.2 Address:
 - *. 6.1.10.4.3 City:
 - *. 6.1.10.4.4 State or Province:
 - *. 6.1.10.4.5 Postal Code:
 - ? 6.1.10.4.6 Country:
 - *+ 6.1.10.5 Contact Voice Telephone:
 - ?+ 6.1.10.6 Contact TDD/TTY Telephone:
 - ?+ 6.1.10.7 Contact Facsimile Telephone:
 - ?+ 6.1.10.8 Contact Electronic Mail Address:
 - @. 6.2 Resource Description:
 - *. 6.3 Distribution Liability:
 - @+ 6.4 Standard Order Process
 - B. 6.4.1 Non-digital Form:

```

B+ 6.4.2      Digital Form
*. 6.4.2.1    Digital Transfer Information
*. 6.4.2.1.1  Format Name:
?. 6.4.2.1.7  Transfer Size:
*+ 6.4.2.2    Digital Transfer Option
C. 6.4.2.2.1  Online Option
*+ 6.4.2.2.1.1 Computer Contact Information
*. 6.4.2.2.1.1.1 Network Address:
*+ 6.4.2.2.1.1.1.1 Network Resource Name:
C. 6.4.2.2.2  Offline Option
*. 6.4.2.2.2.1 Offline Media:
@. 6.4.2.2.2.2 Recording Capacity
*+ 6.4.2.2.2.2.1 Recording Density:
*. 6.4.2.2.2.2.2 Recording Density Units:
*+ 6.4.2.2.2.3 Recording Format:
@. 6.4.2.2.2.4 Compatibility Information:
*. 6.4.3      Fees:

*. 7          Metadata Reference Information
*. 7.1        Metadata Date:
?. 7.2        Metadata Review Date:
?. 7.3        Metadata Future Review Date:
*. 7.4        Metadata Contact
*. 7.4.10.2   Contact Organization Primary
*. 7.4.10.2.1 Contact Organization:
?. 7.4.10.2.2 Contact Person:
?. 7.4.10.3   Contact Position:
*+ 7.4.10.4   Contact Address
*. 7.4.10.4.1 Address Type:
@+ 7.4.10.4.2 Address:
*. 7.4.10.4.3 City:
*. 7.4.10.4.4 State or Province:
*. 7.4.10.4.5 Postal Code:
?. 7.4.10.4.6 Country:
*+ 7.4.10.5   Contact Voice Telephone:
?+ 7.4.10.6   Contact TDD/TTY Telephone:
?+ 7.4.10.7   Contact Facsimile Telephone:
?+ 7.4.10.8   Contact Electronic Mail Address:
*. 7.5        Metadata Standard Name:
*. 7.6        Metadata Standard Version:

```

FGDC Metadata Lite

The following are the metadata elements of FGDC Metadata Clearinghouse On-

Line "Lite" Entry Form

Source: FGDC's website

at <http://dsdnqvarsa.er.usgs.gov/cgi-bin/getpas s.pl>).

This form produces a set of metadata elements whose output meets the minimum data collection requirements of the Content Standards for digital Geospatial Metadata.

Identity of this entry (for future update):

Originator: Publication date (YYYYMMDD):

Title of data set:

Phone Number *:
 E-Mail *:
 Content Name *:
 Citation: Originator:
 Title *:
 Edition:
 Presentation Form:
 Publisher *:
 Publication Place:
 Publication Date YYYYMMDD:
 Online Linkage (URL):

Description: Abstract *:
 Purpose *:
 Supplemental Information:
 Time Period of Content: Beginning Date: YYYYMMDD:
 Ending Date: YYYYMMDD:
 Currentness Reference:
 Status: Progress:Completed Historical archive Obsolete
 On-going Planned Required Under development
 Maintenance and Update Frequency: Continual Daily Weekly
 Fortnightly Monthly Quarterly Biannually Annually As needed
 Irregular Not Planned Unknown

Spatial Domain: West Bounding Coordinate (DDD.XXX) *:
 East Bounding Coordinate (DDD.XXX) *:
 North Bounding Coordinate (DD.XXX) *:
 South Bounding Coordinate (DD.XXX) *:

Data Theme: Primary Theme *:Select a Primary Theme
 Agriculture and farming Biologic and ecologic Administrative
 and political boundaries Atmosphere, climatology, and meteorology Business and
 economic Elevation and derived products Environment and conservation Geologic
 and geophysical Human health and disease Imagery, base maps, and land cover
 Military and intelligence Inland water resources, Locations and geodetic
 networks Oceans and estuaries Cadastral and land planning Cultural, society, and
 demographic Facilities and structures Transportation networks Utility and
 communication networks

Keywords: Theme Keywords:
 Reference:
 Place Keywords:
 Reference:
 Spatial Data Information: Data Type:Vector Raster Text Table
 TIN Stereo Model Video

Data Format(s):
 Data Projection(s):
 Data Scale Denominator:
 (Required for vector data.) 1:
 Data Resolution:
 (Required for raster data.) X & Y Axis Units
 Constraints: Access: Copyright Patent Patent pending
 Trademark License Intellectual property rights Restricted
 Other restrictions

Use: Copyright Patent Patent pending Trademark License
 Intellectual property rights Restricted Other restrictions

Other:

Browse Graphic: Browse Graphic URL:

Browse Graphic Caption:
Browse Graphic File Type:

Order Information: Content Price:
Map Service Username:
Map Service Password:

Would you be interested in seeing this content published through other mapping sites, such as the National Geographic Map Machine? Yes, but please contact me first.
No, not at this time.

Public Commons Metadata Transcript


The following are the metadata elements of the Public Commons minimized Metadata transcript.

SHORT VERSION OF THE METADATA FORM Microsoft Internet Explorer

Address: <http://www.spsai.maine.edu/~narnindi/resweb/shortver.html>

Go to: Search Web Search Site News PageRank Page Info Up

Enter Metadata for the files to be uploaded to the server


Public Commons for Geospatial Data

File Reference ID	1234567890
Details of the Originator	Chakravarthy Narnindi Sharad Department of Spatial Information Science and Engineering Graduate Research Assistant 5711, Boardman Hall
Title of the content	Maine water resources
Abstract	These files have been prepared using a digitizer, basing on 1:24000 scale images of the state of Maine.
Purpose	These shape files can be used for the analysis of the water resources in Maine.
Supplemental Information	These are shape files and can be altered or modified without the permission of the author. Mention of the author would be appreciated

Contains commands for working with the selected items.

Start

Figure B.1: Screen Shot-1 of Metadata elements in Public Commons Minimized Metadata Transcript

SHORT VERSION OF THE METADATA FORM - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print Page Info Up

Address http://www.spatial.maine.edu/~narrandi/resweb/shortver.html

Go Search Web Search Site News PageRank Page Info Up

These are shape files and can be altered or modified without the permission of the author. Mention of the author would be appreciated

Supplemental Information

Beginning date Jan 01 2002

Ending date Jan 01 2002

Currentness reference publication date

Progress Complete

Intended data set maintenance and update frequency Annually

Spatial Extent Information

Do you know the bounding coordinates? Yes No

Data Theme Information Select a Primary Theme

Figure B.2: Screen Shot-2 of Metadata elements in Public Commons Minimized Metadata Transcript

SHORT VERSION OF THE METADATA FORM - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print Page Info Up

Address http://www.spatial.maine.edu/~narrandi/resweb/shortver.html

Go Search Web Search Site News PageRank Page Info Up

Data Theme Information Select a Primary Theme

Spatial data type Point

Theme keywords Type in as many keywords as you please

Place keywords Type in as many keywords as you please

Access Constraints Copyright Licensing imposed

Use Constraints Extra charges with this source required
Expanded distribution is not allowed
Permission of the author required
None

Additionally, Do you wish to distribute from your server?

If yes enter valid URL

Dataset name as known by Distributor

Liability held by distributor

Post Meta Data Clear form

start Internet

Figure B.3: Screen Shot-3 of Metadata elements in Public Commons Minimized Metadata Transcript

Comparison between the different types of Metadata Standard Templates

Table 3 provides a brief count of the number of metadata elements required by the respective organizations.

Organization Templates	Number of Metadata elements (approx.)
FGDC CSDGM	165
NOAA	86
Metadata Lite	41
Geography Network	35
Public Commons	23

Table B.1: Comparison of Metadata Templates of Different Organizations

The table reveals the amount of uniformity and flexibility that can be provided to the contributor by minimizing the number of metadata elements that they have to fill in a typical metadata form. In addition to these, Public Commons model proposes to use software that can automatically extract projection information, geographic extent, bounding information, and data format to bring down the number of metadata fields manually filled by the contributor. Apart from these other options explained on Page 41 would also contribute to decrease in the number of metadata fields. In count and the metadata fields, the Geography Network metadata form was close and similar to the Public Commons transcript. This is due to the fact that they have also included elements that are very critical to the understanding of the fitness of purpose of the dataset. The elements considered in the Public Commons are only illustrative and a first good pass at the minimum set of information required. These elements are also targeted to accommodate different knowledge levels of spatial data producers and users under one common model. It also provides the long form FGDC's CSDGM form for experts and federal agencies to maximize the amount of description available. For further information, the data users are encouraged to contact the contributor directly.

BIOGRAPHY OF THE AUTHOR

Chakravarthy Narnindi Sharad was born in Visakhapatnam, Andhra Pradesh, India on May 3, 1979. He received his undergraduate degree, Bachelor of Technology in Electrical and Electronics Engineering from Jawaharlal Technological University (JNTU), Hyderabad, India, in 2000. Thereafter, he worked at the Computer Maintenance Corporation Limited (CMC Ltd) for a year. At CMC Ltd, he was designated Systems Maintenance Engineer working on a challenging Project for the Indian Railways known as ARTS (Advanced Railway Ticketing System). He then joined the University of Maine's Spatial Information Science and Engineering program as a Master's candidate in the fall of 2001. Here he worked as a graduate research assistant with the National Center for Geographic Information and Analysis (NCGIA). Sharad is a candidate for the Master of Science degree in Spatial Information Science and Engineering from The University of Maine in August, 2003.