5-2001

Access to Geographic Scientific and Technical Data in an Academic Setting

Bastiaan Van Loenen

ACCESS TO GEOGRAPHIC SCIENTIFIC AND TECHNICAL DATA IN AN ACADEMIC SETTING

By

Bastiaan van Loenen

M.Sc. Delft University of Technology, 1998

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

May, 2001

Advisory Committee:

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

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

Max J. Egenhofer, Professor of Spatial Information Science and Engineering

ACCESS TO GEOGRAPHIC SCIENTIFIC AND TECHNICAL DATA IN AN ACADEMIC SETTING

By Bastiaan van Loenen

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)
May, 2001

Data availability is a key issue affecting society's social well being. Information technology has increased the availability of and improved access to data. The academic community that uses spatial data is one of the groups that has taken advantage of fast and inexpensive opportunities to share data and knowledge in a relatively unfettered fashion across digital networks.

However, pressure by the private sector to increase protection for databases through database legislation, self-help measures (contracts, licensing and technological methods for limiting access) and movement by some local governments towards revenue generation from sales of data are decreasing or threatening to decrease access to information for academics. This research explores current and potential access to information principles having substantial potential for promoting sharing and openness for scientific exploration. Current laws and policies on intellectual property and access to information are explored in the context of such principles. A literature study and a

questionnaire are used to investigate the access to data environment of academia using geographic data in accomplishing academic research. Current problems are assessed, and legal constraints are analyzed. Whether or not adhered to, an assessment is made in each project of the productivity of scientists compared to the actual principles followed and the extent and nature of problems encountered. Productivity is assessed on a dataset level. It is measured in terms of satisfaction by scientists with the principles imposed upon scientists for accessing that dataset, the extent of problems encountered by scientists when confronted with the specific access principles, and the accomplishment of research goals under the constraints imposed. This research has resulted in new knowledge that should help inform policy makers and scientists themselves of the means by which a satisfactory environment for accessing data might be maintained or accomplished. Ultimately the results are used to supply evidence of academic community practices that would be supported or not supported by a range of legal options for protecting databases, some of which are currently before Congress.

Acknowledgments

I would like to thank my advisor, Harlan Onsrud for giving liberally of his time and his enthusiasm for my research. I also would like to thank Max Egenhofer and Kate Beard for their suggestions to improve my thesis.

Further, I thank the 300 respondents for giving their time to the questionnaire: without them I would not be able to perform this research. I also would like to thank the National Center for Geographic Information and Analysis and the VSB foundation. This adventure would have been impossible without their financial support.

Thanks to fellow ((inter)national) students, who brightened my graduate school experience with their (Greek) sport qualities, humor and friendship.

Special thanks to Gill, Barbara, the fat guy from Germany, Eva, Dan and Maria, Christian, Victor, and my BIG brother. I enjoyed every moment of being together with you, hiking, talking, discussing, playing whatever, living together, cooking, partying, having pot luck parties but utmost allowing me to be myself and to enjoy Maine.

I am also grateful to all the people on the other side of the ocean helping me out when needed and providing updates on local issues. Special thanks go to my family.

But most thank yous go to Saskia for her encouragement, ideas and her supporting and continuing love at any time and any place.

Maine thank you: I hope you will be there when I return…

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Data availability is a key issue affecting society's social well being. With widespread availability of information on the Internet and other media, abundant opportunities have come to search for scientific and technical gold in the ore of factual elements. The possibilities for discovery of new insights about the natural world with both commercial and public interest value are extraordinary (NRC 1999B, 21-22). Information constitutes the building blocks of knowledge (Reichman and Franklin 1999, 886) and unfettered access to scientific and technical data has allowed knowledge to advance (Reichman and Samuelson 1997, 64-65). The academic community has taken advantage of the fast and inexpensive opportunities to share data and knowledge across digital networks. The segment of the academic community using geographic data also benefits from the opportunities of the new medium.

Geographic data may be described as all data related to (the surface of) the earth. Geographic data have the characteristics of a public good; that is, geographic data are non-rival and are typically non-excludable in consumption. A good is non-rival when its use by one person does not interfere with its use by others. Non-excludable refers to the availability of the good to all, including those who do not help produce it, once the good is provided (extracted from Schmidtz 1991, 55, Cornes and Sandler 1986, 6, and Onsrud 1999).

Some of the most prevalently used tools for processing geographic data, are geographic information systems (GISs). A geographic information system (GIS) is often described as a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e. data identified according to their locations. The capabilities of a GIS depend on its database. "Bits of Power" (NRC 1997, 198) describes a database as a collection of interrelated data, often with controlled redundancy, organized according to a schema to serve one or more applications. The data often are stored so that they may be used by different programs with little or no restructuring or reorganization of the data. A systematic protocol is used to add new data or modify and retrieve existing data.

The characteristics of digital data(sets) and collections of data (databases) that make them easy to share help to advance science but also may provide disincentives for collecting data; "If [information] can be infinitely reproduced and instantaneously distributed all over the planet without cost, without our knowledge, without even its leaving our possession, how can we protect it?" (Barlow 1994, 85). The reverse question is raised by people on the other side of the access to data issue: If access to data is overly constrained through legal or technological methods, how can we realistically use the data in advancing the well-being of society?

Some foresee that current relatively open access to data for academia will continue to exist and expand because "information wants to be free" (Stewart Brand’s slogan cited in: Barlow 1994, 89 and Boyle 1997). Others contend that the real future of the information age lies "in metering every drop of knowledge and charging for every sip" (Okerson

1996, 80). Most suggest models that balance between the two extremes (see e.g. Varian 1995, 201, Reichman and Samuelson 1997, Pluijmers 1998B).

Many scholars and organizations suggest that database producers need a new form of legal protection (Reichman and Samuelson 1997, 55, Perritt 1999A, 460, Goldstein 1994, 211, Library of Congress 1997, ix, x, D'Andrea 1997, 1, Reichman and Uhlir 1999, 837, Pelman 1998). Thus, there is an indication that the rights of the owner of a database and the rights reserved for the public are unbalanced.

However, pressure by the private sector to shift the legal balance by increasing the protection for databases through legislation (HR 3531, HR 2281, HR2652, S 2291, H.R. 354) and self-help measures (contracts, licensing and technological methods for limiting access) is threatening the ability of the scientific community to access data.

Pressure by some local governments towards revenue generation from sales of data (Onsrud 1998, D'Andrea 1997, 18 (section 5), NRC 1997, 6, Reichman and Samuelson 1997, 68), private funding of academic research (Nelkin 1984, 97, NRC 1997, 111, 132) and pressure by university administrators to generate royalties from the products of faculty (Reichman and Samuelson 1997, 68) are other developments decreasing or threatening to decrease access to data for academics using geographic scientific and technical data.

However, empirical data about academic access to the scientific and technical data environment is scant. We have little empirical evidence validating the extent to which various access policy environments do or do not contribute to the satisfaction of academic researchers or to the accomplishment of their project goals. Economic and legal scholars have argued that the current broad access to data environment is beneficial to

advancing knowledge and the economy. This work attempts to evidence support or lack of support of these broad conventions in the context of access to and use of geographic data for knowledge advancement purposes within the university research environment.

1.1 The Thesis

This research has five objectives: (1) to gather information on the policies and administrative processes confronted by university researchers using geographic data in gaining access to data for their research, (2) to develop a set of recommended principles for accessing geographic scientific and technical data, drawn primarily from the literature, (3) to determine in an objective manner which principles have been adhered to in gaining access to geographic information for specific research projects, (4) to determine the degree of satisfaction with the access policies imposed on the researcher, and (5) to test whether hypothesized recommended principles result in greater degree of satisfaction and productivity on the part of researchers than adherence to competing access principles.

Although addressed only in part and for a small subset of scientists, the central question guiding this research has been as follows:

Based on theory and evidenced through empirical testing, which specific access principles appear to best enable scientists that use geographic data to achieve success in advancing knowledge and in meeting their research objectives?

A set of recommended access to data principles has been synthesized from recommendations set forth in various study reports issued by the National Research Council or recommended in the academic literature that relate to policies for providing access to scientific and technical data. They are presented and discussed in chapter 2. Whether or not these specific principles are adhered to, an assessment is made in each project of satisfaction by scientists with the principles actually followed in gaining access to specific datasets and whether goals were achieved. We hypothesize that geographic data sharing relationships are more productive for science if the recommended principles are followed. Productivity is measured in terms of (1) satisfaction by scientists with the actual principles followed, (2) the extent of problems encountered by scientists with the actual principles followed, and (3) the accomplishment of research goals under the constraints imposed by the various policies.

The results may be used to supply evidence of the likely ramifications on research if various legal options for protecting databases are actually implemented or passed into law.

1.1.1 Scope of the Thesis

The project addresses all data acquired or accessed for use in GIS projects in the academic community and not just geographic data. The research is not directed at a special academic discipline. Legal issues, policy issues and technical issues affecting access to scientific and technical data were all addressed to some degree by this research. For instance, in regard to technical issues, questions were asked about compatibility of the software, quality of the records about the data, and reliability of the data.

The sample we strove for was members of the academic community who are employed by a university, either public or private, and who are conducting academic research using digital geographic data or a GIS in their work.

1.1.2 Sampling Group

The research explores current access policies imposed on researchers in U.S. universities that affect geographic scientific and technical data. Because a broad spectrum of disciplines use geographic data in scientific research, one would suspect that the data provided by our sample may be indicative of the responses across many research domains due to the cross disciplinary nature of our sample.

Our sample of researchers using geographic information was developed and drawn from three primary sources. The first group consists of 619 academics listed as having interests in GIS on the web site of the University Consortium for Geographic Information Science (UCGIS). UCGIS is a non-profit organization of universities and other research institutions dedicated to advancing understanding of geographic processes and spatial relationships through improved theory, methods, technology, and data (website 1). A list of member universities of which its employees were asked to participate in the survey is provided in Appendix M.

The second group consists of 33 additional academics drawn from a URISA list of individuals with interests in geographic information science. URISA is a non-profit international association of information professionals with specific emphasis on applications in state and local government (website 2).

The third group consists of 53 academic researchers with National Science Foundation (NSF) support that indicated an intent to use a GIS in their research work. These individuals were identified through a key word searches of the NSF website (website 3). Only those researchers were selected whose research proposal was accepted in 1994 or more recently.

The total sampling group includes 705 academia using geographic data in their work.

1.1.3 Outline of the Thesis

A literature study was used to explore existing and promising models dealing with access to data issues and, from the models, principles of successful access to geographic data were extracted. The principles are described and discussed in chapter two. A questionnaire was developed that allowed us to gain sufficient information to determine whether recommended principles were adhered to in the acquisition of each specific dataset and whether scientists were successful in their use of each dataset. In chapter three the questionnaire is presented and discussed. Chapter four provides an evaluation of the questionnaire. Chapter five evaluates the hypotheses. The evaluation sets forth indications of satisfaction and accomplishment of goals for when a recommended access principle was or was not followed and also discusses the extent that current GIS use environments in the U.S. adhere to the recommended principles as set forth in chapter two. Chapter five also presents guidelines to improve access environments that are not sufficiently meeting the satisfaction and goals of scientists. Finally, the conclusions and recommendations are presented in chapter six.

1.2 Data and Information

There does not exist one uniform interpretation of the definitions of data and information. We provide two different interpretations found in the literature.

The International Standard Organization defines data as: "A representation of facts, concepts or instructions in a formalized manner suitable for communication, interpretation or processing by human beings or by automatic means" (ISO 2382/1 01.01.01). Information is defined as: "the meaning that a human being assigns to data by means of the convention applied to that data" (ISO 2382/1 01.01.02). Information arises through someone recognizing it as such (Couclelis 1998, 211). The location of a river (data) might mean to a tourist a place to swim (information 1) or it might mean a source of hydro energy (information 2) to an energy company. A useful operational distinction between data and information is that data can be automatically manipulated and processed by a machine, whereas information presupposes the involvement of a cognitive agent (Couclelis 1998, 211).

Alternatively, others (Crawford and Gorman 1995, 5, NRC 1999A, 42) define data as "facts and other raw material that may be processed into useful information", and information as "data processed and rendered useful". Including the human mind in the categorization leads to the introduction of knowledge. Knowledge can be defined as information transformed into meaning. It can be recorded and transmitted but the computer is by no means the ideal medium for such transmission (Crawford and Gorman 1995, 5).

"Data" in the first interpretation finds its equivalent in "data and information" in the latter and "information" in the first definition is interchangeable with "knowledge" in the other. In this work data and information are interpreted according to the first (ISO) definitions.

The value of geographic data comes from its use. Sharing of geographic data is important because the more it is shared, the more it is used, and the greater becomes society's ability to evaluate and address the wide range of pressing problems to which such information may be applied (Onsrud and Rushton 1995, xiv). Maximizing uses of data in society is of course inconsistent with the frequent goal of individuals or corporations to maximize profits. The laws of society should seek a balance between the interests of the public and private entities.

Chapter 2 Controlling Access to Data

2.1 Introduction

This chapter discusses legal and technical methods for controlling data, introduces a data flow opportunities model for the academic research environment, provides an overview of the role of data producers in the information economy, and proposes principles in regard to access to scientific and technical data that may be advantageous to academic researchers conducting research with geographic data. It is the derived principles, drawn primarily from the literature, which are used in chapter 3 to construct a questionnaire to test whether adhered to principles make a difference for researchers using digital geographic data.

2.2 Current Legal Controls over Data(sets)

The means used to protect a dataset or provide access to it depends on the owner of the dataset. Ownership of data implies having rights to control the data. It implies a complex set of rights: rights to use, sell, rent, give away, abandon, consume, or even destroy (Boonin 1987, 253). In broad terms these rights may be categorized as: “rights of access and beneficial use” and “rights to exclude others from its use without permission”.

Ways to protect or provide access to data from legal and technical perspectives are well documented in the literature (For instance see NRC 2000 and NRC 1999A). For similar discussions in a geographic data context, see Lopez (1996) and less comprehensive Pluijmers (1998B). Legal protection can be found in intellectual property rights (e.g. copyright) and in self help means like contracts or licensing approaches. Other self-help measures may be technical in nature like technical means to control access and versioning of the data.

This section describes the legal means and self-help methods to protect data and will discuss proposed legislation that may influence access to data environments in the future.

2.2.1 Legal Means to Protect Data

2.2.1.1 Copyright

Federal copyright is the principal form of intellectual property law for protecting "expression". Over the last three centuries it has developed into a constitutionally protected doctrine (9. U.S. Const., Art. I, § 8, cl. 8) "to promote the progress of science and the useful arts" (Goldstein 1994, 19). Copyright extends to "original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced or otherwise communicated, either directly or with the aid of a machine or device" (17 U.S.C 102(a) 1988). *Feist*[1] ruled,

[1] The Supreme Court in *Feist Publishing Inc. v. Rural Telephone Service Co.* (499 U.S. 340 (1991))

consistent with the copyright law, that facts cannot be protected by copyright; only the manner in which the data have been selected and arranged is copyrightable. Facts, data, information, ideas, methods, principles, and systems are directly relegated to the public domain (Reichman and Franklin 1999, 6).

Copyright gives exclusivity to the owner of the work for a limited period of time. Sooner, or later, copyright law directs all protected information goods to the public domain. It is in the U.S. possible to transfer full or partial copyright to someone else (17 USC 201 (d)), unlike some other jurisdictions (e.g. Germany see Hugenholtz 1998, 152). This practice of transferring exclusive rights is well known in the publishing sector (see e.g. Okerson 1996, 80, Guernsey 1998).

Access to information for certain public interest pursuits is guaranteed. Limitations on copyright include fair use, the first sale doctrine, and unfair competition doctrine.

2.2.1.2 Fair Use Provision in Copyright Act

The Copyright Act allows the copying of copyrighted material if it is done for a salutary purpose -news reporting, teaching, criticism are examples- and if other statutory factors weigh in its favor (Goldstein 1994, 20).

The safest course is always to obtain permission from the copyright owner before using copyrighted material. When it is impracticable to obtain permission, use of copyrighted material should be avoided unless the doctrine of "fair use" would clearly apply to the situation or the material otherwise clearly falls outside the ambit of copyright protection. The fair use doctrine is the principal protection of the right of the public, and thus of the scientific community, to have ready, low-cost access to copyrighted material (NRC 1997,

16). Fair use is described in Section 107 of the Copyright Act ~ Limitations on Exclusive Rights. It states literally:

"Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include -

(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes (courts favor non-commercial over commercial use, Goldstein 1994);

(2) the nature of the copyrighted work (scientific works were especially favored in fair use, Goldstein 1994);

(3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole (less is better than more, Goldstein 1994); and

(4) the effect of the use upon the potential market for or value of the copyrighted work (less is better than more, Goldstein 1994).

The fact that a work is unpublished shall not itself bar a finding of fair use if such is made upon consideration of all the above factors" (17 USC 107)

Academic research typically may be characterized as non-commercial, scientific works, with a minimal effect upon the potential market of the owner of the copyrighted work, unless the owner of the copyright is a private publisher. Many uses by academics of

research data reported by other academics will be categorized as fair use. However, the fair use doctrine does not provide an instrument that enforces access to data. It only enforces use of the data once the data is acquired. Private parties, universities, not-for-profit organizations and even state and local government often are free to negotiate conditions of access with potential data-users.

2.2.1.3 First Sale Doctrine

The first sale doctrine states that once a copyright owner sells a copy of his work to another, the copyright owner relinquishes all further rights to sell or otherwise dispose of that copy. The first sale doctrine is found in Section 109 of 17 USC. It provides that the first sale doctrine does not apply to a computer program (including any tape, disk, or other medium embodying such program) for the purposes of direct or indirect commercial advantage (17 USC 109 (b) (1) (A)). However, it continues with "The transfer of possession of a lawfully made copy of a computer program by a nonprofit educational institution to another nonprofit educational institution or to faculty, staff, and students does not constitute rental, lease, or lending for direct or indirect commercial purposes under this subsection"(17 USC 109 (b) (1) (A)). Applying the first sale doctrine to the information age implies that once a document is purchased and downloaded from a website, the copy may be transferred or given to another person as long as only one useable copy remains in existence. Many "clickwrap" and "shrinkwrap" licenses ban the sale or gift of a dataset or software to anyone else. In essence, if the data or software is no longer of use to you as an individual, you are obligated to throw it away rather than give it or sell it to someone else.

The Working Group on Intellectual Property (Green Paper 1994) recommended in 1994 to exclude online distribution of documents from the first sale doctrine. Some scholars took the opposite position (Onsrud 1999, Samuelson 1998). They suggest that we legislate the first sale doctrine or a similar right in the digital environment as well. Thus the assumption would be that one could transfer a copy of a purchased dataset to an archive for use by others rather than being forced to dispose of the data. Several technical solutions are available for ensuring that only a single user is using a purchased dataset at any one time.

A proposal to legislate a right comparable to the "first sale doctrine" in digital environments may have value since the information economy is moving towards an environment of licensing rather than selling, making the first sale doctrine obsolete (see section licensing of this chapter and e.g. Reichman and Uhlir 1999, 809).

2.2.1.4 Unfair Competition Law (in State Common Law)

In the most general sense, unfair competition law protects a business from a competitor gaining a free ride on the goodwill of the first business (Perritt 1999A, 436). Unfair competition law protects some general types of misappropriation resembling copyright infringement (Perritt 1999A, 438). The main question one should ask in deciding about unfair competition is: when does fair use cross the line of free-riding? Guidelines to determine this are described under unfair competition law.

The classic American case in this tradition is *International News Service, Inc. v. Associated Press* ("*INS*") (248 U.S. 215 (1918)). In this case, the Associated Press ("AP") successfully challenged the practice of International News Service ("INS") agents

who bought early editions of newspapers affiliated with AP and read the war news these papers contained via telephones to INS agents in California. The latter then published this news in competition with AP-affiliated newspapers. If the INS case would be applied to (digital) information goods Gordon thinks the producer's efforts should be defended by the courts if all of the following criteria are met:

(1) the costs of developing an information product are high;

(2) the costs of copying are low;

(3) copying yields a substantially identical product;

(4) which a copyist can price cheaply, not having substantial research and development costs to recoup; and

(5) when consumers, believing the two products are substantially identical, decide to purchase the cheaper one, thereby inducing market failure because the first comer is unable to recoup its expenses; and

(6) when such a market failure could have been averted by a period of protection that would allow the first comer to recoup its expenses and justify its investment in developing the information product

(Wendy J. Gordon cited in: Reichman and Samuelson 1997, 140).

In the unfair competition model most scientific community uses would not be viewed as competing with the commercial interests of the current rights holder in the data, although specific instances give rise to the issue.

However, like the fair use doctrine, the unfair competition doctrine does not provide affirmative rights of access to data. It only enforces use of the data once data is reported.

Private parties, universities, not-for-profit organizations and state and local government may be free to negotiate conditions with potential data-users. Organizations, libraries among them, have in the first sale doctrine a means by which to be legal subsequent users without the need to gain permission of the originator of the data.

2.2.2 Self-Help Methods to Protect Data

2.2.2.1 Contracts

A contract is an exchange of promises or other things of value between two or more people. Contracts generally determine limitations on duplications, resale, and derivative products. They also allow data suppliers to receive economic gain at privately negotiated prices (Goldstein, 1977). An online contract can include the right to access a database, services or resources. Contracts provide data suppliers with means to protect the content of factual datasets. This is not possible under federal copyright law alone. Contracts provide some but not comprehensive protection to a vendor for the actions of a third party. The vendor relies on copyright or other laws to restrict use of the data in the copy.

Traditionally, contracts are used to settle a sale. Sales involve a complete transfer of ownership rights, in particular copies from the vendor to the purchaser, following which the purchaser could largely do whatever he or she wished. In digital environments licenses, a special form of contract, are popular for protecting the interests of the vendor.

2.2.2.2 Licensing

A license is a contract imposing express limits on the use of the data (Dreyfuss 1999, 203). One can generally redistribute a licensed copy only if especially contracted for the right to do this (Samuelson 1998, 17). License agreements in the digital era are of two types: bargained agreements for custom software, and unbargained "shrink-wrap licenses" imposed on mass-market purchasers (Lemley 1995, 1239). A shrink-wrap license is a license agreement for a software or data product not accessible to the user until the box has been opened. Click-wrap licenses may be the digital equivalent of a shrink-wrap license or may additionally require that you affirmatively respond that you have read the terms supplied on the screen and that you agree to the terms by pressing the "I agree" button.

A landmark case about the enforceability of the terms of click-wrap licenses is *Pro CD v. Zeidenberg*[2]. The district court held that because the defendant was not able to examine the terms of the license prior to his purchase, those terms could not be enforced against him. The small-print reference stating that use of the software was subject to the terms of the enclosed licensing agreement was not held adequate (Loundry 1996, 5).

Although this confirmed the general assumption of the legal status of the shrink-wrap license (unenforceable), the federal Court of Appeals upheld the shrink-wrap license in *Pro CD v Zeidenberg*[3]. Where a vendor has clearly stated that detailed terms apply and where the purchaser has the opportunity once the detailed terms are available, to back out of the deal and obtain his or her money back, the Court of Appeals was unwilling to

---

[2] *ProCD, Inc. v. Zeidenberg*, 908 F. Supp. 640, 659 (W.D. Wis. 1996)

[3] *ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir.1996)

conclude that a contract could not be formed on the vendor's stated terms (Hutcheson 1996). Conflicting law exists in other federal circuits. For example, in *Vault Corp. v. Quaid Software Ltd.*[4] the district court held that the shrink-wrap license was an unenforceable contract of adhesion, and that Louisiana statute that would have authorized shrink-wrap licenses was preempted by the federal copyright act. In *Step - Saver Data Systems, Inc. v. Wyse Technology* [5] the court's ruling established that a contract comes into existence when the purchaser submitted a purchase order, and the licensor shipped the software. The court treated the terms on the box of the software as new or additional terms of the agreement. The court ruled that the license on the box was not enforceable because the licensor had not clearly expressed its unwillingness to proceed with the transactions in the absence of the box-top license, and because the license contained additional terms that would "substantially alter the distribution of risk between the parties". Therefor the additional terms were held unenforceable.

Streff Jr. and Norman (1997) summarize the usability of click-wrap licenses as follows: "The enforceability of shrink-wrap licenses is a fact-specific determination- one that depends heavily on the rules selected by the court in its analysis. A court treating post sale terms as new or additional terms to an already formed contract may not enforce the license agreement. However, a court treating the sale as conditioned on assent to the license agreement is likely to enforce the agreement, especially if it contains a right of refund if the purchaser opts to reject it".

---

[4] *Vault Corp. v. Quaid Software Ltd*. 847 F.2d 255 (5th Cir. 1988)

[5] *Step- Saver Data Systems, Inc., v. Wyse Technology,* 939 F.2d 91 (3rd Cir. 1991)

Partly because of this uncertainty the American Law Institute (ALI) and the National Conference of Commissioners on Uniform State Laws (NCCUSL), representing the major software developers, introduced UCC Article 2B to create more clarity in the enforceability of click-wrap licenses (see § 2.2.3.2 and § 2.2.3.3).

2.2.2.3 Technical Controls over Data(sets)

Technical methodologies consists of technologies inside the software that help the originator of the data enforce his or her license conditions. Programming the software to self-destruct if the license engages in a particular kind of abuse (like copying the data) or embedding a block of code in the program capable of disabling its operation are examples of technical self-help constructions (Samuelson 1997, 13). Other means that may be used to control access are: encryption of data, watermarking, limitations in downloading data, database access control, and trusted systems (see in more detail NRC 2000, 68).

Technical control gives originators of databases a technical lead-time to recover their investments. The con of it is that "one man's self-help, may be another man's virus of worm" (Samuelson 1997, 13). If a lessee's (e.g. academic researcher) existence depends completely on the data of the licensor (e.g. commercial vendor) after a certain period, the licensor has the power to enforce conditions, which may be unfair to the lessee. Moreover, if a lessee accidentally uses the dataset in violation with the terms in the license, the technical self-help construction may terminate the program/ dataset without any warning.

Perritt (1999A, 458-459) provides examples of several economic or technical alternatives, among them: content encryption, planned obsolescence and system access controls.

Whether State law should recognize technical self-help remedies is one of the issues of the current draft of UCC Article 2B (see below under Proposed Legislation UCC Article 2B) and of its predecessor the Uniform Computer Information Transaction Act (UCITA).

2.2.3 Proposed Control over Data(sets)

2.2.3.1 Collection of Information Anti-piracy Act (HR 354)

Collections of information (databases) are very expensive to create, compile, verify, update and to format. However, once created the distribution or dissemination of the collection is very cheap and so is its reproduction. Someone acquiring a dataset or database, can distribute it now to others cheaply, and thus go into competition with the owner of the dataset or database.

The European Union has responded to these theoretic threats for database originators with the Directive on Databases (96/9 EG March 11 1996), which the European member states must convert into domestic legislation. In short (see Reichman and Samuelson 1997, 84-94 for a discussion in detail) the directive imposes strong economic and legal restrictions on the conditions of availability and use of factual data in databases. It has barely taken into account the interests of competitors, intermediaries and end-users

(Pluijmers 1998A, 378). The academic community is one of the groups facing more restrictive access to data(bases) than before the enactment of the directive.

Similar *sui generis* (type specific) legislation for databases is pending in (106th) U.S. Congress (Collection of Information Antipiracy Act HR 354; formerly HR 2652 rejected in Senate in 1999). Although the *sui generis* legislation is only a proposal, it or similar bills are expected to continually arise in Congress with strong support from the information industry. If so, it may be the biggest threat to the availability of collections of data for academia.

The main concern is that collections of facts and data would now be protected. This implies that scientific activities that were previously permissible would become infringing acts under the new law. The draft states that the user of all or a *substantial* part, measured either *quantitatively* or *qualitatively*, of the "Collection of Information" causing harm to the *actual or potential* market of the originator of the database is liable to a civil action (paragraph 1402 HR 354). The following example illustrates the impact of paragraph 1402 for the scientific and technical community.

> An academic researcher publishes the results of tests that investigate the reliability of a car navigation system. A second researcher rechecks the reliability of the same car navigation system and the system directs him into the ocean, through houses and, when he wants to go to the nearest hospital, to Walmart. He publishes his findings and reproduces in his article the results of the first researcher in order to challenge those results. The second researcher would likely be held liable to civil action for infringement of the proposed database legislation, despite the fair use provision in HR 354.

Although the number of acts of reasonableness (fair use) incorporated into the legislation have increased as the process has gone forward, use of data for academic research purposes has not been excluded from paragraph 1402. Publishers view the academic community as a major market and would like to expand the number of academic users paying for access to both intellectual works and datasets published in conjunction with research results.

Many scholars and others have expressed their concerns about the impact of the proposed *sui generis* legislation on the practice of university researchers (see for example Reichman and Samuelson 1997, Samuelson 1996A, 1999B, Ginsburg 1997, Reichman and Uhlir 1999, Lederberg 1999). The critics focus and have focused on: the scope of the proposal, the duration of protection (currently 15 years), the use of vague terminology, and the use of the insufficient fair use doctrine instead of the more favorable unfair competition doctrine for academia.

HR 354 does not apply to collections of data gathered, organized, or maintained by or for a government entity, whether Federal, State or Local (paragraph 1404). However, this provision does not affect data collected and created in public and private partnerships (PPPs) (see for example *Delorme* in section FOIA, see also Neal 1999).

If Congress passes the current draft of HR 354 or an equivalent, it would give collections of data, including collections of facts, more protection than is available for copyrighted works.

2.2.3.2 Uniform Commercial Code Article 2B (UCC 2B)

The sales of goods is regulated in UCC article 2. In the information age, however, most goods (e.g. software) are licensed and not sold or leased. Article 2 does not apply to licensed transactions. The National Conference of Commissioners on Uniform State Laws (NCCUSL) and the American Law Institute (ALI) attempted to address this "gap" in article 2 by drafting UCC 2B.

UCC article 2B aspires to provide a standard set of rules that will regulate online and mass-market transactions (Ginsburg 1998, 945). It intends to clarify the current uncertainty about the enforceability of click-wrap licenses and it may permit the use of technical self-help measures. The draft also includes a broad range of methods for electronic contract formation. For example, a record replaces the traditional writings requirements; authentication replaces the traditional signature requirements; and a contract may be formed by a programmed electronic agent even though there is no actual review by the parties of the terms of their agreement (Streff 1997).

If the current UCC Article 2B is enacted, it will influence the way academics access data of others. Data will be available and accessible online as set forth by the terms of the data supplier and technical self-help measures will "control" the use of the data.

Many legal scholars have reviewed and discussed Article 2B. The main concern is that the draft meets only the interests of the major software companies (the sponsors of the draft) (see Nimmer 1999, 70, McManis 1999, 173, Dreyfuss 1999, 198, Lessig 1998, Onsrud 1999, Reichman and Franklin 1999, Reichman and Uhlir 1999, 798, Streff 1997). For example it would validate licenses that override public interest exceptions that favored users, including the scientific and technical community.

Other issues of discussion are its relationship and interaction with federal copyright law (Litman 1998, Lemley 1999, 170), its scope (Samuelson 1999, 23) its use of unclear and inconsistent terminology (Litman 1998, 939, Dreyfuss 1999 206-209, 220, Ginsburg 1998, 949) and its need (Samuelson 1999, 3, Lessig 1998).

The commission charged with UCC2B failed consensus and efforts to move it forward are currently dormant. However, the proposed Article 2B issues have resurfaced under the guise of the Uniform Computer Information Transaction Act (UCITA).

2.2.3.3 Predecessor to UCC 2B: UCITA

UCITA is a draft state law for contracts relating to software and other forms of computer information. The NCCUSL drafted this model law. UCITA, some call it "Lex Microsoft", mirrors the UCC Article 2B initiative in most respects. According to opposing parties, including FTC, ALI, ACM, IEEE, American Association of Law Libraries, and the American Library Association, it is essentially the same bill (Lousin, 1999, 276, Sandburg, 1999) and "dangerously out of balance in favor of large software companies" (Huggins, 2000). The broadness of the proposals made one of the founders of UCC Article 2B, the ALI, withdraw its support “because it would give licensors power to restrict use of information more narrowly than current patent and copyright law” (J. Hazard director ALI). Not surprisingly: the act is solely supported by the US software industry.

Although the proposal is expected to have severe problems to be accepted in many states (Sandburg 1999), it has been enacted into law in Maryland, and introduced in Iowa and New Jersey.

2.3 Data Access and Dissemination Opportunities in the Academic Community

2.3.1 Data Flows in an Academic Research Environment

Data flows in the academic research environment potentially flow into two directions from the perspective of the researcher: data for ones own research and disseminating data for the use of others. The academic researcher often both collects data from others and distributes data to others (see Figure 2.1).

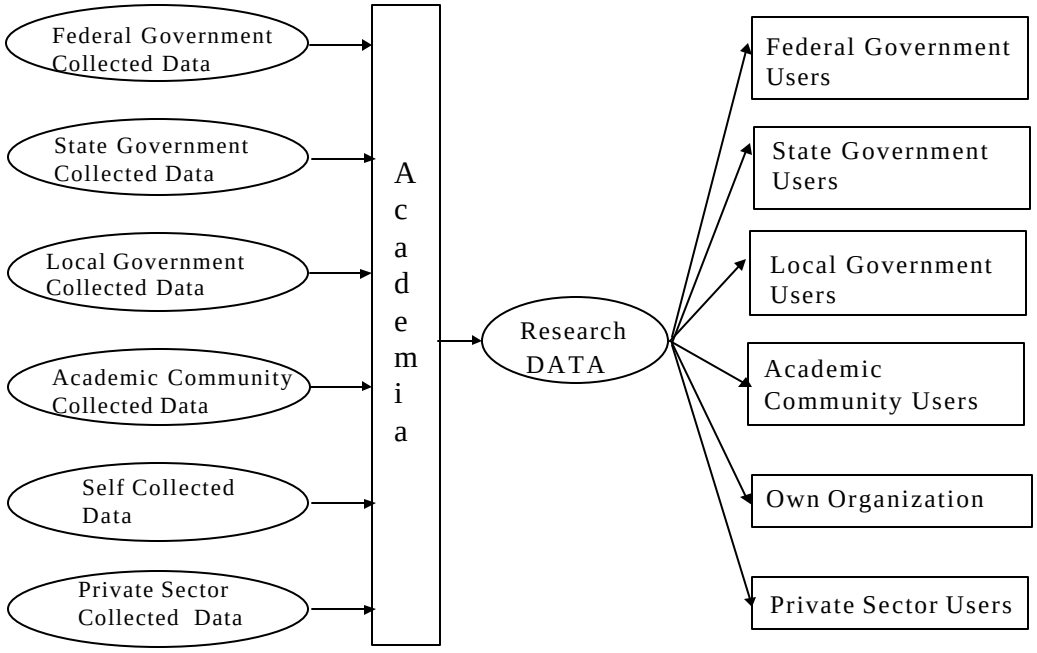


Figure 2.1: Data Flows in an Academic Research Environment

The researcher typically needs data to accomplish his or her goals. After or during the research the researcher disseminates the results of the research including developed or derived datasets to the world. Four primary parties are identified as playing significant roles in both making data available to and obtaining data from the researcher in a U.S. context: (1) federal, state and local government, (2) the academic community, (3)

additional non-profit entities, (4) the private sector. The researcher might typically accomplish substantial independent data collection as well.

2.3.2 Parties Involved in the Access to Data Model

As figure 2.1 shows, academia depend on five sources to access their data. Typically each source has a business model (or non-business model) that reflects its mandate and environment. The types of data and services it provides, the restrictions it imposes on users, the quality and standards for the data all reflect this business model.

Here we discuss the legal means to control data and the obligation of the different parties to the public on a source basis. The first source we discuss is the government.

2.3.3 Use of Data Collected by the U.S. Government

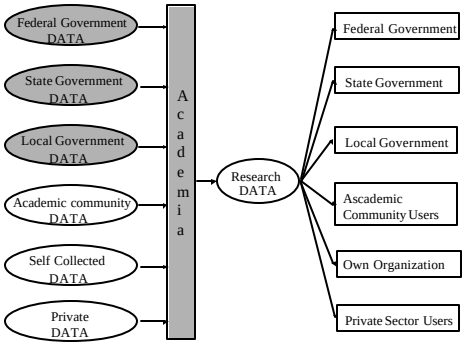


Figure 2.2 Government Data and Academia

Many (federal, state and local) government agencies have a public duty for collecting, archiving, or distributing information. Branscomb (1994) distinguishes at least four different types of government information: "(1) that which is necessary for citizens acting in their roles as voters engaging responsibility in the electoral process; (2) that which is necessary for law- abiding residents in order to comply with the legislative enactments and judicial decisions that are the law of a land; (3) that which is mandated by the purpose for which the agency is established; (4) that upon which the very essence of the deliberative process rests, and which cannot be collected reliably and accurately in the private sector. Such information assets fall within the public domain." (Branscomb 1994, 164-165).

Data collected or created by the government[6] is subject to several legal restrictions and obligations. Three Acts mark the legal framework for access to governmental data: the Copyright Act (17 USC 1976 last amended 1997), Paperwork Reduction Act 1995 and Freedom of Information Act (1966 last amended 1996).

2.3.3.1 Copyright Act

For copyright a distinction must be made between data collected by the federal government, state government and local (county) government. Unlike most European countries (e.g. The Netherlands, Great Britain, France), but like most other developed countries the United States does not recognize Crown Copyright: "Copyright protection is not available for any work of the United States Government, but the United States Government is not precluded from receiving and holding copyrights transferred to it by assignment, bequest, or otherwise" (17 USC 105 1988). However, this does not extend to state or local government agencies. They can claim copyright in their datasets.

2.3.3.2 Paperwork Reduction Act 1995

The Federal government is held to the Paperwork Reduction Act 1995 (44 USC 3506 (d) Paperwork Reduction Act 1995). It rules about the dissemination of federal government data.

---

[6] See for an overview of free spatial data provided by the U.S. government:

http://www.cast.uark.edu/local/hunt/ (website 7)

"With respect to information dissemination, each agency shall –

(1) ensure that the public has timely and equitable access to the agency's public information, including ensuring such access through –

(A) encouraging a diversity of public and private sources for information based on government public information;

(B) in cases in which the agency provides public information maintained in electronic format, providing timely and equitable access to the underlying data (in whole or in part); and

(C) agency dissemination of public information in an efficient, effective, and economical manner;

(2) regularly solicit and consider public input on the agency's information dissemination activities;

(3) provide adequate notice when initiating, substantially modifying, or terminating significant information dissemination products; and

(4) not, except where specifically authorized by statute -

(A) establish an exclusive, restricted, or other distribution arrangement that interferes with timely and equitable availability of public information to the public;

(B) restrict or regulate the use, resale, or redissemination of public information by the public;

(C) charge fees or royalties for resale or redissemination of public information; or

(D) establish user fees for public information that exceed the cost of dissemination."

2.3.3.3 Freedom of Information Act

"A popular government, without popular information or the means of acquiring it, is but a Prologue to a Farce or a Tragedy or perhaps both. Knowledge will forever govern ignorance, and a people who mean to be their Governors, must arm themselves with the power knowledge gives" (James Madison Letter to W.T. Barry, Aug. 4, 1822 cited in Branscomb 1994, 164).

Since the enactment of the Freedom Of Information Act (FOIA, 5 USC 552) in 1966, records of the federal government are subject to the federal Freedom of Information Act. It offers judicially enforceable procedures for compelling government agencies to release information to the public (Branscomb 1994, 167). States and local governments records are subject to State Freedom of Information Acts. These records are not identical to FOIA, nor is state court interpretation of similar language in such state statutes necessarily the same as federal court interpretation of FOIA (Perritt 1999A, 479).

The federal FOIA provides that agencies shall act actively in disseminating certain public information to the public (5 USC 552 (a) (1) and 552 (a) (2)). Moreover, it provides that any person has the right to request access to federal agency records or information (5 USC 552 (a) (3) (A)). This right of access is enforceable in court (5 USC 552 (a)(4)(B)). In making any record available to a person, an agency shall provide the record in any form or format requested by the person if the record is readily reproducible by the agency

in that from or format (5 USC 552 (a)(3)(B)). Although FOIA does not specifically identify datasets as a governmental record, the federal courts have consistently held that computer records are public records for the purposes of FOIA (Onsrud and Lopez 1997, 160, Perritt 1994, 13).

All agencies of the United States federal government are required to disclose records upon receiving a written request for them, except for those records that are protected from disclosure by the nine exemptions and three exclusions found in the FOIA. Those include documents concerning "national security," trade secrets, and information relating to an individual's privacy. It also allows a federal agency to withhold materials if the materials are exempt from disclosure by statute other than the FOIA, as *Delorme*[7] confirmed. *Delorme* ruled that the agency must possess and control the dataset in order to be able to disseminate the data on the terms in FOIA. The plaintiff, an electronic map publisher, sought disclosure of digital nautical charts from the defendant under the FOIA. The defendant used the Federal Technology Transfer Act (FTTA) to justify its refusal to disclose the material. The FTTA (and the judge) allowed the agency to withhold the materials for five years because it produced the material together with a private company (extracted from Perritt 1999B, 232).

The federal FOIA also does not provide a right of access to records held by Congress, the courts, or by private businesses or individuals.

---

[7] *Delorme Publishing Co. v. National Oceanic & Atmospheric Administration of United States Department of Commerce*, 917 F. Supp. 867 (D. Me. 1996)

2.3.3.4 Cost-recovery Under FOIA

Agencies are able to recover their costs of dissemination in accordance with the guidelines of the Office of Management and Budget. It shall provide that "fees shall be limited to reasonable standard charges for document duplication when records are not sought for commercial use and the request is made by an educational or noncommercial scientific institution, whose purpose is scholarly or scientific research" (5 USC 552 (a) (4)(A)(ii)(II)). The most recent version of the guidelines recommends that Federal information resources be disseminated at the marginal costs of dissemination in order to encourage access and use through a diversity of channels (OMB Circular A-130 1992). Marginal pricing allocates the smallest nonzero cost to users and thus is consistent with the principle of full and open exchange of data.

2.3.3.5 Technical Limitations in Accessing the Data

New technology is significant in that it creates an opportunity for people to access information previously unavailable. However, one needs to use the technology efficiently and effectively in order to take advantage of the opportunity. In order to "disseminate public information in an efficient, effective, and economical manner" (PRA 1995 (1) (C)) sufficient and appropriate hard- and software programs, standards to communicate between agencies and between agencies and requesters of data, and adequate documentation (metadata) to guarantee the quality of the dataset are required. Affirmative programs by government that anticipate records and data in greatest demand by the public and that actively release such records and data in electronic environments appear to be the most sufficient means for overcoming technical limitations.

2.3.3.6 Extension of Federal Principles to State and Local Government Agencies

Most state and local governments believe they have the option of asserting copyright in their public records if they choose to do so (NRC 1999, 57). However, some legal scholars argue the economic validity of the argument (Epstein 1990). Others argue the legal validity that federal government, but also state and local governments restrict access to public data to ease budget pressures (Perritt 1995, 450). To realize the potential of geographic information systems, federal, state and local government must honor two policies: (1) make electronic formats available, and (2) allow and promote a diversity of channels and sources of public information (Perritt 1995, 455). This is only possible if governments "resist the temptation" of selling of data to generate revenues and thus asserting copyright in their public records.

In this study we follow Perritt's reasoning and will treat federal, state and local government data alike. The recommended principles of "Access to Government Scientific and Technical Data" apply to federal, state and local government.

2.3.3.7 Access to Government Scientific and Technical Data: Recommended Principles

As the data above shows, *Federal* United States public information policies are based upon an attempt to guarantee broad access to information as a precondition to economic and political opportunity (Onsrud and Lopez 1997, 160). In a nationwide and international comparison between governments in different jurisdictions, Lopez found that "open access approaches were more conducive to contributing to access and commercialization of geographic data than those information policies that attempted to restrict access and protect the revenues of a government franchise" (Lopez 1996, 208).

Furthermore his study found evidence that "US Academic and private sector players significantly benefit from the dissemination policy of the US Federal government" (Lopez 1996, 210, see also Onsrud, Johnsson and Winnecki 1996, Matsunaga and Dangermond 1994, Litman 1994, Lederberg 1999). Lopez' findings suggest that current federal public data laws and policies (for geographic data) should be adhered to by all government agencies, including federal government and state and local government.

The principles the federal government adheres to are translated into access to governmental data principles below. They are extracted from the literature (NRC 1995A and B, NRC 1997, ICSU 1998, NRC 1999A, Perritt 1999A) and current legislation applying to policies of the US federal government.

1. Government agencies should ensure that electronic data, information and value-added features developed with public funds are available to the public. (see also Perritt 1999A, 499, PRA 1995 (1), FOIA (a) (1) and (2))
2. Government agencies should adopt affirmative programs of electronic public information dissemination so that scientists do not need to resort to Freedom of Information requests in order to gain access to government records (see also Perritt 1999A, 499, FOIA (a) (1) and (2)).
3. Government agencies should anticipate requests by the general public (including the scientific community) for electronic information and should build features into their electronic information systems so that information most likely to be requested by the public may be actively released (such as publishing datasets on web servers or CDs along with appropriate retrieval software) (see also Perritt 1999A, 499, PRA 1995 (2), FOIA (a) (3) (B), Lopez and Onsrud 1997).

4. Scientific and technical data collected or maintained by or under authority of a government agency which may be of current or future use to the scientific community should carry with it the obligation to retain the data collected and to place the data in a publicly accessible archive. (see also NRC 1995B, 32 and NRC 1997,11, PRA 1995 (1) (C) and (1) (A)).
5. Scientific and technical data collected or maintained by or under authority of a government agency should be documented adequately with metadata (NRC, 1995B, 36, PRA 1995 (1) (B)).
6. Scientific and technical data collected or maintained by or under authority of a government agency should be made available to all requesters at the marginal cost of dissemination or less. (see also Perritt 1999A, 499, PRA 1995 (4) (D) FOIA 4A ii II, NRC 1999, 6 and ICSU 1998)
7. Scientific and technical data collected or maintained by or under authority of a government agency should be made available for exploitation by both not-for-profit and commercial entities alike on a non-exclusive basis. (see also Perritt 1999A, 498, PRA 1995 (1), NRC 1999, 6 and ICSU 1998))
8. Government agencies should not hold copyrights in scientific and technical data collected or maintained by or under their authority (see also Perritt 1999A, 499, 17 USC 105) and federal agencies should not establish or maintain exclusive arrangements for access to scientific and technical data (see also Perritt 1999A, 499, PRA 1995 (4) (A)).

9. Government agencies should ensure that electronic data, information and value-added features developed with public funds are available without restrictions on subsequent uses of the materials. (see also Perritt 1999A, 499, PRA (4) B and C)
10. Scientific and technical data collected or maintained by or under authority of a federal, state or local government agency that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means (Onsrud personal correspondence)

Appendices C and D show how each of the principles are addressed by which specific question(s) in the online questionnaire. In the appendices C and D one may see that we initially addressed principle 4, "level of accessibleness by archive", through an analysis of Questions 1, 2, 3, 7, 9, 10, 12, and 13. However, the options for responses to these questions proved to be inappropriate for testing principle 4. In short this principle asks respondents to comment on conditions not yet prevalent in the GIS scientific community and therefore testing could only be based on speculations by the respondents. Thus testing of this principle ultimately was not achieved through this thesis work

2.3.4 Use of Data Collected by the Academic Community

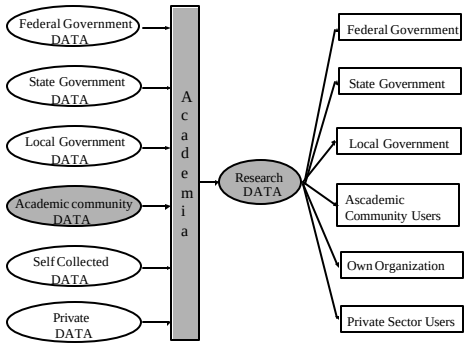


Figure 2.3: Not-for-profit Data and Academia

Traditionally, the main purpose of universities is to seek "truth" in an independent and objective way (Richman 1974, 119). Academic science is and has been a public resource, a repository for ideas and a source of relatively unbiased information (Nelkin 1984, 29). Scientists serve as advisors to policymakers, consultants to government and private enterprises, expert witnesses in the courts, technical administrators and bureaucrats, social critics, popularizers, advocates for public interest groups and above all, educators (Nelkin 1984, 94). Their incentives are the creation of new knowledge, the thrill of discovery, and the enhancement of professional status (Lederberg 1999).

Until recently, data collected and created by universities for research purposes typically were not subject to legal obligations to provide actively or passively access to the data. Similarly, few regulations exist to restrict the use by university researchers of such devices as intellectual property rights, contracts, (click-wrap) licenses or technical means in the use by others of their research data.

However, on February 4 1999, the Office of Management and Budget (OMB) released a proposed revision to OMB Circular A-110 (Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-profit Organizations). It responded to a provision included in the Omnibus Appropriations Bill (the Shelby amendment) that directed OMB to require federal awarding agencies to ensure that data produced under a federal award be made publicly available via the Freedom of Information Act. Many in the higher education, research, and scientific community have expressed serious reservations with the original legislation mandating the OMB revision. It would allow businesses to harass scientists working in controversial fields, such as air quality and tobacco (Zitner, 1999). An opposing bill (HR 88) that would repeal the Shelby Amendment did not pass and on November 8, 1999, a revised version of OMB Circular A-110 became effective. An important finding by OMB is that the statute construed is not requiring scientists to make research data publicly available while research is still going on. The Chamber of Commerce is trying to overturn the White House of OMB's ruling, which it said has illegally narrowed the scope of the law.

The process is still going on. An update on the ruling may be found at the OMB Watch Website (website 4).

Yet the tradition and underlying philosophy of universities has been to commit to full and open sharing of knowledge (Boonin 1987, 260) emphasizing a not-for-profit policy and adhering to an open access to data environment.

The policy framework for the model of access to data collected and created by universities, whether public or private, revolves around three primary issues: rights to data, technical accessibility to data, and the time between concluding research and the dissemination of results.

2.3.4.1 Rights to Data

The issue of who owns data created by a university researcher is not a new topic. Although most researchers haven't thought about the ownership of data resulting from their research work, university administrators and private companies funding the research often do. Nowadays, the creation of new knowledge, regardless of how ambiguous the term might be, depends heavily on research funding. The more money research generates the more research may be done is the line of reasoning of increasing numbers of university administrators. Although the government supports the bulk of academic research, the influence of the private sector is increasing. Thus, more data will likely be removed from the public domain in the form of income producing products (Reichman and Uhlir 1999, 819).

This raises the question of who should control the data created with the help of public money and what dissemination policy should researchers or university administrators follow to promote the progress of science and to satisfy both the academic community and the funding agencies? Here we discuss recommended principles if research is entirely or partly funded with public funds.

The basic principle of exclusivity of public funded research is found in David:

“A critical feature [of public subsidizing the production of knowledge] is that producers are denied exclusive rights to the output of their R&D activity: once it is produced, the knowledge is made freely available to all who care to use it” (David 1995, 32). “The conduct of science is a public good” (NRC 1997, 111). Accordingly, researchers should guarantee the full and open exchange of research data. Full and open availability means that "data and information derived from publicly funded research are made available with as few restrictions as possible, on a non-discriminatory basis, for no more than the cost of reproduction and distribution" (NRC 1997, 15). In the information age, this cost can be very close to zero.

Although not everyone agrees with the continuance of this principle (e.g. D’Andrea Tyson 1997, 17), the open policy of the federal government has (had) a positive influence on the advancement of science (see e.g. Lopez 1996).

A distinction should be made between research entirely funded with public money and research partly funded with public money.

Research solely funded with public money is in principle nothing different from data collected and created by the government. This data should be in the public domain. With the data in the public domain, intellectual property right, license and contract issues disappear.

Applying this principle to research not substantially (<50%) supported with public funds is ambiguous. In order to accomplish the research other financial sources are acquired, e.g. funds from private corporations. These funding parties mostly fund to benefit from the research. If the research data and results are subject to principles similar to the federal

FOIA, private corporations may decide that supporting academic research is not as beneficial anymore and stop the funding. This may put academic research with substantial private funding at risk.

However, commercial sector control over the products created by public entities like universities tends to lead towards more restrictive access to scientific and technical data environments (Samuelson and Reichman 1997, 151). Examples of this situation are found in the publishing sector where publishers like Thomson and Reed/ Elsevier require exclusive rights over the published paper (see e.g. Samuelson 1994, 21 or website 6). This already makes researchers pay for the use of their own research results (Elsevier Science charges start at approximately U.S.$7 per paper).

Proposed Open Access Policies

In order to guarantee the full and open exchange of data, the researcher should always maintain at least full and non-exclusive rights in the created data. There are several options to accomplish this principle. Copyright law or *sui generis* legislation could force the creator of data or works created through public funding to maintain an exclusive right to sell copies (Masson 1997) or ban the creator from transferring exclusive rights (Onsrud 1999). The political process accompanying the introduction in law can be time consuming and in the end not satisfactory due to political concessions. Another, more practical, option is that public agencies who fund research and public universities should require the researcher to keep full but non-exclusive rights in the data as a condition for accessing their funds or resources. Private or government entities funding research should be allowed to obtain the same full but non-exclusive rights in the research results

(Guernsey 1998). That is, they receive more in the way of ownership rights than the general public but their interests are not exclusive. They are shared with the creator or author. In this way a balanced "access to academic data partly funded with public funds" principle is found. The dissemination policy of a researcher is likely to be to disseminate as much as possible to obtain (academic) recognition for the achievement (NRC 1997, 49). The private entity can use the data for its own purposes. Finally the public agency that funded the research may publish the data in a publicly accessible archive (see below). Adherence to this principle, offers the public a variety of potential sources with similar or the same data. Access to and use of the dataset should be guaranteed.

Recently, a new provision in the OMB Circular A-110 requires federal awarding agencies to ensure that data produced under a federal award be made publicly available via the Freedom of Information Act (see 2.3.4 for a discussion). Although the discussion is still on-going, access to federal funded research data may now legally be guaranteed.

2.3.4.2 Technical Accessibility of the Data

Science builds on science. New knowledge best advances when the data and results from previous work is available. It is important that data is stored adequately for the use of others later. A successful archive (database) is one which is affordable, durable, extensible, evolvable, and readily accessible (NRC 1995B, 50). To meet these requirements effectively and efficiently, data should be maintained in a publicly accessible archive with adequate documentation.

Publicly Accessible Archive

Once the data is collected, created and published the academic researcher moves on to the next project. Keeping the data accessible for the use of others and increasing the awareness of the existence of the research does not anymore have the first, if any, priority. This implies the danger of loosing valuable (digital) data for the use of others.

The novel data(set) should be retained in a publicly accessible archive for the use of others. One could think of a public depository or library in a traditional meaning or in a more modern sense of an archival website.

The researcher should be required by the funding agency to archive the research in such an archive and to allow others to freely read the data at a minimum. The burden of maintaining the new data in the archive, or integration into other databases should be borne by an entity other than the researcher, such as the government or a library system supported by government. The researcher was funded to do the research and not to maintain an independent archive over time.

The Documentation of the Data(set)

Adequate explanatory documentation or metadata can eliminate a great barrier to use of scientific data. One way of guaranteeing this all is to require and fund metadata creation and appropriate archiving of research datasets in public depositories or libraries as standard conditions of grants.

Standards in the geographic discipline, are of significant interest because of the potential for increased access and sharing of geographic data, reduced data loss in the data

exchange, reduced duplication of data acquisition, and increased quality and integrity of geographic data (Brewer 1999, 221). One useful standard already mandatory in the federal government is the Metadata Standard of the Federal Geographic Data Committee (FGDC) which adheres to Federal Information Processing standard (FIPS) 173, and which are being proposed in whole or in part as an ISO standard.

2.3.4.3 Delaying Research Data Publication

"The right to search for truth implies also a duty: one must not conceal any part of what one has recognized to be true" (Albert Einstein).

Scientists employ secrecy to support their positions in disputes, to protect their work from plagiarism, to divert competition, to avoid external interference, and to ensure the accuracy of results before disclosure (Nelkin 1984, 97, NRC 1997, 50). Some of these are valid reasons for not releasing data and others are not. In certain situations, complete secrecy in science is justifiable: for example, for national security reasons, the protection of endangered species, and to protect the personal privacy of data subjects.

However, David makes clear that, in theory, society at large does not benefit from secrecy or delayed dissemination of new data: "Wider distribution and timely inexpensive access to new findings reduces wasteful duplication in effort in research. By putting research data into the hands of a more diverse population of researchers, these conditions tend to increase the probability of useful new products and processes arising from novel and unanticipated combinations" (David 1995, 22).

The time between a discovery of new information and the dissemination to society can be very important. Consider the classic spatial analysis case of Mr. Snow in London where

he discovered that the distribution of people with cholera was positively correlated with the location of poisoned wells (Snow 1855). If he had not disclosed this information the epidemic would have been far more severe and many more residents of London would have died.

The academic responsibility of open communication inevitably conflicts with the commercial responsibility to maintain proprietary secrecy (Nelkin 1984, 25, for empirical evidence in life science see Blumenthal et al 1997). The pressure on the researcher rises when the amount of privately funded research is increasing. “The imposition of secrecy on scientific research for any reason, threatens both science and the public interest” (Nelkin 1984, 101).

But life-threatening situations are not daily occurrences in academic professions using spatial data. Secrecy is mainly held in data for a period of time in order to guarantee the publication of the research; the main incentive to do the research (NRC 1997, 49). The researcher should be allowed to keep a reasonable time period of proprietary use in the data to allow publication of the results of the research.

The National Institute of Health (NIH) considers 60 days a reasonable time to allow for publication (website 5, Blumenthal, 1997). However, in many disciplines, the process of publication of research takes more than a year after a paper is submitted, due to a wide variety of reasons (Egenhofer, personal correspondence).

2.3.4.4 Access to Data Principles for the Academic Community

Thinking along the lines mentioned and discussed above the following access to data principles should apply to data collected by universities and not-for-profit organizations:

1. The not-for-profit scientific and technical community should continue to promote and adhere to the policy of full and open exchange of data at both the national and international levels (NRC 1999, 94, ICSU 1998).
2. Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions that have been collected for projects entirely or primarily financed with public funds should be treated by the creators from a science policy perspective as being in the public domain, after a reasonable time period to allow for publication of the results of the research (ICSU 1997, 9).
3. When publishing research articles, scientists should concurrently publish or otherwise make available electronically the datasets upon which their research depends or from which it is derived (ICSU 1998).
4. Public agency grant conditions and university policies should establish that all scientists conducting publicly funded research should make their data available immediately, or following a reasonable period of time for proprietary use. The maximum length of any proprietary period should be expressly established by the particular scientific communities (NRC 1997, 9), and compliance should be monitored subsequently by the public funding agency (NRC 1997, 11).
5. Scientific and technical datasets created or collected in conjunction with research or educational projects by university and other not-for-profit researchers or their employing institutions that may be of current or future use to the scientific community should be retained and placed in a publicly accessible archive (Similar to NRC 1995B, 32 and NRC 1997, 11).

6. Scientific and technical datasets made available in a publicly accessible archive should be documented adequately with metadata (NRC 1995B, 36).
7. For research and scholarly work partially or entirely financed with government funds or public university funds, university and other not-for-profit researchers that create datasets should be required by the granting agency or their employing institutions to not grant or otherwise transfer exclusive rights in the works. The recipient of public funds should retain at least full but non-exclusive rights to such databases when submitting them for publication, for incorporation into other databases, or when entering into any other contractual relations regarding the datasets (similar to NRC 1999A, 90).
8. Scientific and technical data collected or maintained by or under authority of an academic institution that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means (Onsrud personal correspondence)
9. Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions should be made available to all requesters at the marginal cost of dissemination or less (NRC 1997, 7).

Appendices E and F show by which specific question(s) in the online questionnaire the principles are addressed. We found that principle 5, "Level of Accessibleness by Archive", could not be addressed sufficiently in this research. This principle asks

respondents to comment on conditions not yet prevalent in the GIS scientific community and therefore testing could only be based on speculations by the respondents.

We also did not test principle 7 for academic data. The goal of the thesis was to address primarily the use of datasets and the problems in acquiring them. As such dissemination practices are not directly addressed. For instance, we did not ask whether respondents retained at least full but non-exclusive rights to their works when submitting them for publication, for incorporation into other databases, or when entering into any contractual relations regarding the datasets. Further we did not fully explore appropriate measures of success for this principle. Possible measures of success may be the number of times a datasets is used by others or the number of times a dataset has been downloaded. Qualitative research methods, such as in-depth case studies, may be more appropriate to address the relevance of principle 7 in supporting access to scientific and technical data.

2.3.5 Use of Data Collected by the Private Sector

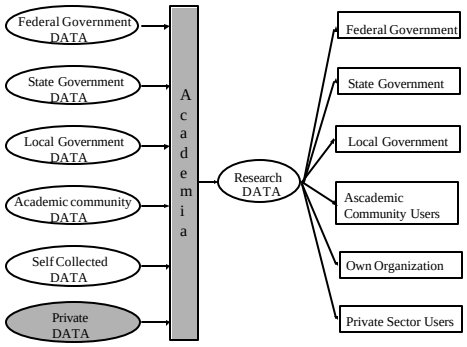


Figure 2.4: Private Data and Academia

The private sector collects and creates new data(sets) to support profit generation activities. The private corporate sector supports internal and external research for the same reason. In order to guarantee their existence in the future they must make a profit and policies in regard to the distribution of research data are expected to comport with that overall goal. If profit making suggests that expected research results or data should be kept secret, such research is often accomplished internally. If profit making suggests open dissemination of data and results will still benefit the company, that form of research may be accomplished in cooperation with external parties, such as universities.

Assuming internal development of data, a private corporation typically controls subsequent uses of the data through contract language, negotiated or otherwise, and through technical methods of protection.

Assuming external development of data, a private entity will seek ways to promote access to the data. One way may be to place the data in a publicly accessible library with the right to read the data for all patrons.

The new electronic world provides alternatives to the traditional way of doing business. Barlow (1994 84, 85-87) states that information economics will be based more on relationships than possession. Dyson sees a future in controlling the relationship with customers through subscriptions defined by contract or licensing language, memberships in the ancillary market and data metering (Dyson 1995).

Here we focus on the traditional contractual relationship. In order to encourage the use of negotiated contracts that respond to licensees' actual needs, as revealed in the emerging information economy, without unduly impeding licensors from resorting to standardized terms and conditions that do not threaten competition or undermine present or future public interest uses of information goods, all mass-market contracts, (non-negotiable) access contracts, and contracts imposing (non-negotiable) restrictions on uses of computerized information goods should be made on fair and reasonable terms and conditions, with due regard for the public interest in education, science, research, technological innovation, freedom of speech and the preservation of competition (Reichman and Franklin 1999, 930).

The fairness of "take it or leave it" contracts will need to be decided on a case -by- case basis in the courts. The level of judicial scrutiny may appropriately vary with such factors as the market power of the licensor, and the potential harm to public interest uses of information likely to ensue from widespread adherence to the terms or practices in question (Reichman and Franklin 1999, 930).

Furthermore, scientists need data that is fit for the purposes intended by their research. Documentation of data that may be used by others is important so that the accuracy and usability of the data is evident. Lack of adequate documentation (metadata) makes a

dataset less likely to be useful to others. Part of the documentation is notification of the source of the data. It is a well-known complaint of governmental agencies that their data is used in commercial datasets but not mentioned as such.

Privately Funded Research

Private universities, supported by private entities, accomplish research with making a profit as one of the main goals, while keeping secrecy in their data. A way to establish the principles of the governmental model in the privately funded research is to rely on the ethics of science. The university should only accept support for their research under conditions ensuring open access to the data for academia. The central theme here is: "Researcher stay in control of your own research". Ethics of science must assure that no researcher gives away any intellectual property right.

Publicly Funded Research

One exception to for-profit distribution are datasets created solely or primarily with public funding. As mentioned before, those datasets are in principle not very different from datasets collected and created by the government itself. Thus the principles and their testing would be similar to geographic data use by the academic sector that was obtained from government.

Access to Private Sector Scientific and Technical Data: Recommended Principles

Thinking along the lines set forth above we think that private entities should adhere to the following principles in order to advance science:

1. All mass-market contracts, access contracts, and contracts imposing restrictions on uses of computerized information goods should be made on fair and reasonable terms and condition, with due regard for the public interest in education, science, research, technological innovation, freedom of speech and the preservation of competition (Reichman and Franklin 1999, 930, ICSU 1998)
2. Scientific and technical datasets created by private universities and other for-profit organizations that have been collected for projects entirely or primarily financed with public funds should be treated by the creators from a science policy perspective as being in the public domain, after a reasonable time period to allow for publication of the results of the research.
3. Scientific and technical data collected or maintained by or under authority of a private entity that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means (Onsrud, personal communication).
4. The commercial derivative product should be required to identify the government source(s) used (NRC 1999, 7).

Appendices G and H show which questions in the questionnaire address the principles. Principle 1 and 3 are not addressed in the questionnaire. The terminology of principle 1 is difficult to interpret. What would one consider fair and reasonable contracts?

Although we could have guessed what one would consider fair and reasonable provisions in contracts, this would have been ambiguous. Reichman and Franklin (1999, 930) advocate that this should be left to the courts on a case by case basis. In this research we did not test this principle.

Principle 2 was addressed properly in the questionnaire. However, of the 21 datasets coming from a private entity only 5 were created with public R&D funds and 3 with public funds only. One dataset was originally created with both public funds and public R&D funds. Thus the total group to be tested would consist of 9 datasets. We consisted this group too small to be useful for a statistical test.

Also principle 3 was addressed properly in the questionnaire. However, none of the respondents acquired datasets from the private sector through a publicly accessible library. We were unable to test principle 3.

Principle 4 is addressed in the questionnaire by questions 17d and 18 d. We wanted to test whether datasets identifying the source allow more successful use of the datasets than datasets that do not identify the source. However, we did not ask for data about the documentation of the source of the data. This makes it difficult to make an appropriate division of subgroups. The only data we had about source identification is whether the documentation of the source is a success or an impediment to the use of the dataset. We did not test this principle.

2.3.6 Use of Data Collected by Not-for-profit Organizations

Not-for-profit organizations groups (i.e. research laboratories, conservation groups, professional associations, private universities) fall between and have characteristics of

both academic sector and private sector. They typically respond to one or more public interest means. Therefor they also tend to support open exchange of knowledge but tend to be more restrictive due to the need to ensure that all expenses are paid for by income.

Many not-for-profit organizations consider the advancement of knowledge as an intrinsic good and exploiting data for financial gain is subordinate to fulfilling public-interest objectives (NRC 1999A, 41). But an increasing number of not-for-profit organizations seeks to maximize the revenues from their databases, subject to the constraints of their tax-exempt status, to finance future R&D and database development in order to remain at the forefront of their respective fields (see NRC 1999A, 31, 41). They are exploring means to recover their costs of production and distribution, or to generate revenue streams to support their expensive data activities, thereby making them function in a manner similar to private enterprises (NRC 1999, 31).

Most not-for-profit organizations, however, fall somewhere in the middle in trying to reconcile their public interest mission, but need to generate sufficient revenues to accomplish this mission (NRC 1999, 41).

Chapter 3 Questionnaire

3.1 Introduction

The first objective of this research was "to gather information on the policies and processes confronted by university researchers using geographic data(sets) in gaining access to data for their research".

There are generally four ways of collecting this data: a self-administrating questionnaire, a mail questionnaire, telephone survey and a personal survey (Zimmerman 1995, 123). Given time and financial constraints we chose an online self-administrating questionnaire. This method allows us to question a large group in a relatively short period of time, in an inexpensive way. It also enables us to generalize the data obtained from the questionnaire to a larger population (Zimmerman 1995, 123). Furthermore it is a way to ask questions with long or complex answers, asking batteries of similar questions, and the respondent does not have to share answers with an interviewer (Fowler 1993, 66).

A self-administered questionnaire has potential drawbacks. Careful questionnaire design is needed, open questions are often not useful for detailed comparative analysis, and quality control is not exercised due to the absence of an interviewer (Fowler 1993, 66).

An online questionnaire has the advantages of economy and speed over a paper based questionnaire. The interviewer does not have to print out the questionnaire, put it in a envelope, address each envelope, place the questionnaire instructions in a self addressed envelope inside each envelope, pay for stamps, and deliver the mailings to the post office.

Similarly the respondent does not have to go through a physical handling and mailing process. Once an online questionnaire is filled out and submitted, the interviewer receives it immediately versus several days of delay due to the mail for the paper-based questionnaire. Another advantage of an online questionnaire is the ability to make the questionnaire extra attractive and customized by using interactive elements such as motion, links, background colors, and the addition of extra information about the topic after the questionnaire is submitted. A major advantage of an online questionnaire is that responses are already in digital form which greatly facilitates the ability to process the data.

At the current time, many potential respondents may be unfamiliar with web technology (what is a hyperlink?). Does one need an advanced web user to take advantage of the new features, or does the user interface allow anyone to fill it out in the most convenient way? This concern is lessened by avoidance of jargon and providing basic instructions to novice web form users. The limited overview (the size of the screen) of the questionnaire may be another disadvantage. Questions like "where am I" and "how many more questions are there?" are difficult for the respondent to assess. Furthermore there are computer related problems; the host server may be "down" or "busy" when a participant submits the questionnaire, and the compatibility of the program that runs the questionnaire (MS Frontpage) and the web server (UNIX) can be non-existent. Further problems may exist in the precision and accuracy of the data processing.

3.2 Sample group

The research explores current research environments of researchers in universities in the U.S. using geographic scientific and technical data. Our total sampling group consists of 705 people.

The significance of the outcome of the survey depends on the size of the sample group. Fowler states that:

> "The first prerequisite for determining a sample size is an analysis plan. Usually the key component of an analysis plan is an outline of the subgroups within the total population for which separate estimates are required, together with some estimates of the fraction of the population that will fall into those subgroups. Most sample sizes are concentrated on the minimum sample sizes that can be tolerated for the smallest subgroup of importance" (Fowler 1993, 35).
>
> He continues:
>
> " Like most decisions relating to research design, there is seldom a definitive answer about how large a sample should be for any given study. There are many ways to increase the reliability of survey estimates. Increasing sample size is one of them. …..three approaches to deciding on sample size are inadequate. Specifying a fraction of the population to be included in the sample is never the right way to decide on sample size. Saying that a particular sample size is the usual or typical approach to studying a population also is virtually always the wrong answer. Finally, it is very rare that calculating a desired confidence interval for one variable for an entire population is the best way to decide how big a sample should be" (Fowler 1993, 35).

Our means of dealing with sample size was to attempt to identify as many researchers as possible known to be using GIS or digital geographic data in their research work. Thus, we attempted to contact the entire population of researchers at major universities in the U.S. connected with active interdisciplinary or campus wide geographic information science research groups or those that had received funding from the National Science Foundation (NSF) for scientific research with geographical information. We inevitably missed mailings to some of the population but we believe that most visible and active researchers using GIS received requests.

Several measures were taken to increase the response rate of the survey. Confidentiality of the answers was guaranteed and emphasized on the letter accompanying the questionnaire and included in the questionnaire itself. This is also a way to increase the accuracy of the answers and to decrease "socially desired or correct answers" (Fowler 1995, 28, 30-31, Zimmerman 1995, 121). The confidentiality was secured in the analysis by not having any link (no respondent addresses on questionnaire form, no questionnaire numbers, etc.) between the research sample and their responses.

Furthermore, the email was directed to the participants personally. This is preferred over sending the questionnaire to a group of people. This also responds to privacy concerns more properly.

Finally, the "new" way of approaching people and guiding them to an attractive online questionnaire to participate in a survey may have increased the response rate.

3.3 Design of Questions

The goal of the survey was to gather information on the policies and processes confronted by university researchers using geographic data in gaining access to data for their research (objective 1 of the thesis). With this information we are able to assess to what extent the current access environments meet or violate the presented principles of access to spatial scientific and technical data (objective 3 of the thesis).

The questionnaire is presented in Appendix A of this work. It consists of four sections. Section 1, *General Information*, asks for general background information (e.g. name of the researcher, use of geographic data). It makes it possible to separate the geographic data user from the non user and to direct the latter very quickly to the end of the questionnaire.

The second section, *Most Recent (Current) Research Project Dealing With Geographic Data*, deals with more specific background information: name of the research project, field of research, sources of funding and datasets used for the research.

The third section, *Dataset Specifics*, addresses the third objective of the thesis. Every question in the third section is linked or based upon one or more of the principles presented in chapter 2. Whether a dataset adheres to each principle is determined from information provided in this section. Appendices C - H show the correlation between the principles being tested and the questions constructed to test each principle. That is each principle is listed followed by the explicit questions on the questionnaire that were used to gain information about whether the principle was met or not. After determining whether a principle was met or not for a specific dataset, a measure of productivity was made of the researcher's use of the data. Productivity has five different measures: (1)

factors of successful use, (2) impediments in the use, (3) task accomplishment, (4) satisfaction about the dataset, and (5) overall objective accomplishment.

The fourth section, *Desired Datasets*, measures indirectly whether the principles not addressed in section 3 are adhered to or not.

3.4 Design of the Questionnaire

The design of the questionnaire follows the guidelines for questionnaires provided in the literature (Fowler 1993, 100). The questions in the questionnaire are simple, using clear terminology. We added concise explanation in places where confusion about terminology was likely to arise. We restricted the questionnaire to closed answers: no open answers were allowed, except for the "other" category.

The guidelines state that the questionnaire should be self-explanatory. This means that in order to fill out the questionnaire properly no instructions should be needed. This appeared to be impossible for the online questionnaire in this stage of the information age. We assumed that a significant part of the sampling group had never filled out an online questionnaire previously. The pretest of the questionnaire confirmed this. It showed that especially the hyperlink feature was not understood by everyone.

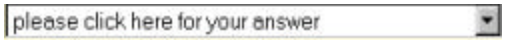
In order to present a clear and uncluttered questionnaire we used as few question and answer forms as possible. Questions are direct and use active language whenever possible.

For several answers we asked the participants to rate the answer. Consistent with recommendations in the literature, we used, two different five-point scales (Dole 1988, 264-265) depending on the circumstances. The first scale consist of: almost never, some,

about half, most, and almost always The second consists of: non-existent, poor, fair, good, and excellent.

3.5 Online Questionnaire Design

The online questionnaire was created in Frontpage 2000, a Microsoft product. Frontpage 2000 is designed to create webpages. It also provides a tool to make online forms. The form tool has several useful features. It is possible to link the participant in the survey to a confirmation page after the form is submitted (see Appendix B for the one we used). Responses can be sent to the interviewers personal email address, to a Frontpage file or directly to a database. Due to limitations in the available server we could not use the database option for the research.

Four interactive answer features were used: dropdown menus, radio buttons, check boxes and text boxes. A dropdown () menu can be described as a menu with all the possible answers predefined available but showing only one option directly on the screen. One click on this option opens the menu. In the menu you can chose the answer you wish. It is possible to choose more than one answer but this requires an advanced user (press CTRL and click). We categorized our sampling group as not advanced. Thus, we considered the multi answer possibility of this feature as one with too many drawbacks. Another disadvantage is that the dropdown menu does not allow a participant to explain the “other” choice.

Radio buttons (μ) are features that allow only one answer. Selecting a new choice cancels the previous selected choice automatically. The interface of the radio button is similar to

the interface of the check box. This is one of the reasons we chose in section 3 and 4 radio buttons instead of dropdown menus.

Check boxes (θ) find their look-a-likes in the paper-based questionnaire. Any answer provided may be checked. Check boxes allow multiple answers in a non-advanced way. Text boxes (☐) are used for the "open" questions about, for example, the used datasets and were used to give the participant the opportunity to explain their "other" answer.

3.6 Pre-testing the Online Questionnaire

The questionnaire was pre-tested by eight University of Maine students and professors. The suggested pre-test approach by Fowler was used to test the survey (Fowler 1993, 102-103).

First, the respondent filled out the questionnaire as if he or she was part of the survey; the interviewer kept track of the time and wrote down possible problems, hesitation, questions that took longer and noticed the use of the new features. Then the questionnaire was discussed with the test-person. Here we addressed the clearness of the instructions and questions, missing answers, the flow of the sections and questions, the length of the sections, and the total length of the questionnaire. From the test we drew the following conclusions:

- Answering the same questions for more than three datasets was too burdensome
- The hyperlink feature was not clear for everyone
- Some instructions were not clear

- The dropdown menu, radio buttons, check boxes, and the text boxes were clear and used properly
- The flow of the questions and the clarity of the questions were sufficient

3.7 Analysis of the Survey Responses

We used a combination of MS Excel97 and MS Access97 to apply statistical testing in respond to the results. That analysis is presented in chapter 5 of this work.

3.8 Mailing Process

The electronic mailing process consisted of three rounds. In the first letter we sent to participants we explained the purpose of the survey and introduced ourselves (see Appendix I). The importance of participation of the researcher was emphasized. This email was sent out on October 19, 1999.

The second electronic letter reminded people of the questionnaire and offered help if problems were encountered with filling out the questionnaire (see Appendix J). This email was sent on October 26, 1999.

The third e-mail explained again the purpose of the survey, emphasized the need of the participants help, and offered personal help (see Appendix K). This email was sent on November 3, 1999.

As of the self-imposed deadline date of November 8 1999 we had received 300 (42.5%) responses. Five questionnaires were received after this date and not considered in the analysis. 148 responses (21% of the total) were found to be useful for this study. Those

that were not useful typically were from researchers who indicated that they were not actively using geographic data, were not using a geographic information system in their research work, did not have time to fill out the form, were not doing academic research, or had privacy concerns about filling out such questionnaires. The distribution of useful responses across the original lists is as follows: 134 UCGIS members (22% of total UCGIS), 11 researchers part of the URISA "group" (33% of total URISA), and 3 researchers with NSF funding (6% of NSF total) filled out the questionnaire.

Table 3- 1: Overview of Responses

Responses total	*305*
Useful responses	148 (49%)
Not useful responses	157 (51%)
Reasons for not useful response:	
Respondent did not perform research with geographic information or GIS	74 (24%)
Respondent did not have time to fill out the questionnaire	21 (7%)
Respondent did not accomplish academic research	11 (4%)
Respondent did not fill out for privacy reasons	2 (1%)
Response was received after closing date	5 (2%)
Other reasons for not filling out the questionnaire	44 (14%)

Chapter 4 Survey Results

4.1 Introduction

This chapter presents and discusses the results of the survey. The paragraphs in this chapter correspond with the sections in the questionnaire. Paragraph 4.2 corresponds with section 1 in the questionnaire, paragraph 4.3 corresponds with section 2, paragraph 4.4 with section 3, and paragraph 4.5 with section 4. Within the sections we present, on a question by question basis, the answers the participants provided. A specification is made for the categories of data providers identified in chapter two. The database, Microsoft Access, was used to select the appropriate fields, and to count.

4.2 Section 1 General Information

Section 1 of the questionnaire deals with the selection of the appropriate participants. Out of 305 respondents, 148 indicated that they use geographic information or a Geographic Information System in accomplishing academic research. Thus our sample group consists of 148 academics.

4.3 Section 2 Most Recent (Current) Research Project Dealing With Geographic Data

Section 2 orients the participants by asking them some simple background questions and focuses their attention on one specific research project. We asked for the title of the research project for which GIS was used, the status of the researcher in this project, the discipline he or she associates most closely to this research project, and the data sources used for this project. The section ends with a question about the datasets used in the project, and the name of the agency that provided this dataset. The counts for the questions 5 - 8 are presented on a question by question basis.

QUESTION 5: W*hat is your status in the project?*

The status of the participants in the project is important, since the level of project involvement may result in the inability to answer some questions on some issues as set forth in section 3 of the questionnaire. For example, a principal investigator will typically know the details of the conditions and constraints under which a dataset has been acquired whereas a graduate student working on the project may have little or no knowledge of these contractual constraints.

In the “other” category, participants mentioned their status as GIS consultants, advisor of graduate student, research faculty, visiting professor, and data coordinator. The majority of the respondents were principal investigators or co-investigators, as shown in Table 4-1.

Table 4-1: Status of Participants in the Project

Status in the project	*Total*	*Percentage*
Principle investigator	70	47%
Co- investigator	40	27%
Graduate Student	13	9%
Staff	11	7%
Other	13	9%
Unanswered	1	1%
Total	148	100%

QUESTION 6: *With which disciplinary field do you most closely associate this project?*

The returns of the survey cover a wide variety of disciplines. The majority of participants (33%) associated themselves with the classic spatial profession: GIS/ surveying/ photogrammetry/ remote sensing and geography. A fair number of responses came from people in ecological research, earth sciences, and planning. A summary of the respondent disciplines is contained in Table 4-2. A summary of the distribution of the subject matter of the research projects is provided in Table 4-3.

Table 4-2: Disciplines of Participants

Discipline	*Total*	*Percentage*
GIS/ surveying/ photogrammetry/ remote sensing	27	18%
Geography	22	15%
Ecological research	15	10%
Earth sciences	13	9%
Planning	14	9%
Natural resources/ environmental	10	7%
Engineering/ architecture/ construction	8	5%
Forestry	7	5%
Social sciences	6	4%
Education	5	3%
Medical/ health	4	3%
Agriculture/ farming	3	2%
Economics	2	1%
Emergency services	2	1%
Business/ banking/ finance/ insurance	1	1%
Legislative/ policy making	1	1%
Meteorology/ air quality	1	1%
Utilities	1	1%
Wildlife management	1	1%
Unanswered	5	3%
Total	148	100%

Table 4-3: Subject Matter of Research Projects

Subject Matter Project	*Counts*	*Percentage*
Spatial Analysis	47	16%
Building a database/ mapping	39	13%
Tools for GIS	23	8%
Other	37	13%
Not filled out	144	50%
Total	290	100%

QUESTION 7: *From which of the following sources did you acquire data for use in this specific research project?*

Table 4-4 shows the number of respondents who indicated their use of datasets from each listed data source. The "total" column on the right indicates that 112 respondents used federal government data in the research project for which they responded, 71 used state government data and so on for the other providers. The column on the far right indicates the number of datasets from each source for which the questionnaire was completed. That is, some of the 112 respondents using federal government datasets filled out the form for more than one federal government dataset. Although 71 respondents indicated use of state datasets, fewer than that number answered the questionnaire for those datasets.

The results indicate heavy and multiple use of federal government geographic information datasets as compared to the other sources.

Table 4-4: Source of Project Data

Source	*Total*	*Number of datasets addressed in questionnaire*
Federal government agency(S) (U.S.)	112	133
State government agency(s) (U.S.)	71	60
Local government agency(s) (U.S. county or municipality)	47	24
Not-for-profit organization or foundation	48	30
Private commercial firm	42	21
Other sources – please specify	34	22

In the “other” category respondents specified primarily foreign government and self collected data.

QUESTION 8: *Please provide the explicit name(s) of one or two agencies or organizations in each of the indicated categories from which you acquired data and name a specific dataset that you acquired or accessed from that organization.*

This question was used to focus the researcher’s attention on specific datasets for further consideration. However, it is interesting to note those agencies or organizations most frequently mentioned as sources for data by this responding group of academic researchers. Table 4-5 shows the sources most often mentioned.

Table 4-5: Name of Data Provider

Name of Agency	*Counts*	*Percentage*
USGS	42	14%
US Bureau of the Census	15	5%
USDA	14	5%
NASA	11	4%
NOAA	6	2%
Other	202	70%
Total	290	100%

4.4 Section 3 Dataset Specifics

To aid assessment of responses to questions in this section of the questionnaire, total responses to each question are accompanied by a breakdown in accordance with the classes set forth above in Table 4.4. That is, the original creator of the dataset was designated as follows:

F = federal government agency(s) (U.S.)

S = state government agency(s) (U.S.)

L = local government agency(s) (U.S. county or municipality)

N = not-for-profit organization or foundation

P = private commercial firm

O = other sources

QUESTION 1: *From whom did you directly acquire this dataset?*

As shown in Table 4-6, the majority of datasets are acquired from the creator of the dataset or an intermediate non-commercial entity. Only a few datasets were acquired from intermediate commercial data providers.

As shown in Table 4-6, 78 federal government datasets are obtained directly from the originating federal agencies, 10 federal government datasets are obtained from intermediate commercial entities, 105 are obtained from intermediate non-commercial entities and so on. While the number of researchers using datasets from private commercial firms is small, approximately half of those using private datasets in their research acquired access to them through a library or some other non-commercial organization (i.e. 10 out of 21).

Table 4-6: From Whom Data Acquired

Source	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
The creator of the dataset	159	78	36	16	14	5	10
An intermediate non-commercial entity, not being the primary creator of the dataset	105	44	22	7	15	10	7
An intermediate commercial entity, not being the primary creator of the dataset	19	10	1	1	0	5	2
Do not know	6	1	1	0	1	1	2
Unanswered	1	0	0	0	0	0	1
Total	290	133	60	24	30	21	22

QUESTION 2: *How did you find out about the availability of this specific dataset?*

Datasets used in academic settings are found primarily through either personal inquiries or common knowledge (see Table 4-7). The Internet as a search device is not (yet) commonly used to find specific datasets but already this method is used more often than traditional means of finding information, such as through the library or print literature. Of special interest is that 27% of the datasets provided by not-for-profit organizations (i.e. 8 out of 30) were found through the Internet, more than for any other category, 15% of federal datasets were found in this manner while for all other categories fewer than 10% of the datasets used by the respondents were found through the Internet.

Table 4-7: Finding Out About Datasets

Find out through:	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Personal inquiries (by phone, email, personal contact)	150	45	44	17	17	11	16
Existence commonly known in the discipline	101	60	14	7	2	9	9
General Internet search	35	20	6	0	8	0	1
Print literature (including supplier catalogs)	16	9	0	0	2	3	2
Search of a specific database	9	5	0	0	1	2	1
Library catalog search (on-line or otherwise)	7	2	1	0	1	0	3
Advertisements (print or on-line)	1	0	0	0	1	0	0
Other	26	8	5	2	5	3	3

The “other” class includes the following answers: given by client, specialist meeting, self created/ generated, clearinghouse, came with the software, contracted to have it created

QUESTION 3: *What was the physical means by which you acquired this (digital) dataset?*

As shown in Table 4-8, most datasets are acquired on a digital portable medium (47%, that is 137/290) or are downloaded across the Internet (38%; 109/290). Paper based acquisition with conversion to digital is rarely used (13%; 39/290). Digital portable media may be favored over Internet accessible datasets because of the large size of geographic datasets that use a great deal of memory and may take very long times to download. Portable media also allow more reliable storage of datasets and downloading datasets may be problematic (server may be down, computer not available during download process, etc.).

Table 4-8: Physical Means of Acquiring Datasets

Physical means	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Acquired on a digital portable medium (e.g. CD-ROM or disk)	137	57	25	19	7	17	12
Downloaded across the Internet	109	67	20	3	15	3	1
Acquired on paper and converted	39	17	5	3	4	1	9
Shipped by e-mail (ftp, LAN)	20	10	7	0	2	1	0
Self-collected	16	2	3	2	4	1	4
Other	5	2	2	1	0	0	0

The “other” class included BPI tape, 8mm tape, and printed air photos.

QUESTION 4: *Did you need to make a specific request to an agency or organization in order to obtain a copy or access to this dataset?*

As shown in Table 4-9, data acquired from federal government agencies, and not-for-profit organizations was accessed in at least 50% (i.e., 67/133, and 16/30) of the cases without a specific request. Most datasets acquired from local government (79%; 19/24), state government (58%; 35/60) and datasets categorized as "other" (53%; 12/22) were acquired after making specific requests.

Table 4-9: Specific Request Made

Specific Request	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes	144	56	35	19	9	13	12
No	126	67	23	5	16	8	7
Do not know	20	10	2	0	5	0	3
Total	290	133	60	24	30	21	22

QUESTION 5: *Were you required to identify yourself prior to being allowed to access the dataset?*

As shown in Table 4-10, relatively fewer individuals were required to identify themselves when accessing federal (35%; 47/133), not-for-profit (40%; 12/30) and "other" datasets (45%; 10/22) than were required to identify themselves when accessing state (57%; 34/60), local government (71%; 17/24), or private datasets (52%; 11/21).

Table 4-10: Identification Required Before Access

Identify	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes	131	47	34	17	12	11	10
No	122	69	21	3	13	8	8
Do not know	33	14	5	4	5	2	3
Unanswered	4	3	0	0	0	0	1
Total	290	133	60	24	30	21	22

QUESTION 6: *Were you required to explain your intended use of the dataset prior to being allowed to access the dataset?*

Table 4-11 shows that federal agencies (68%; 91/133), and not-for-profit organizations (43%; 13/30) asked the least about the intended use of the dataset. Local government agencies (67%; 16/24) asked the most about the intended use.

Table 4-11: Intended Use Requirement

Intended use	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes	97	30	26	16	9	8	8
No	154	91	27	5	13	7	11
Do not know	35	10	7	3	7	6	2
Unanswered	4	2	0	0	1	0	1
Total	290	133	60	24	30	21	22

QUESTION 7: *Was all or a substantial portion of this dataset or database originally developed by a government agency using exclusively or primarily public funds?*

As expected, most respondents believe that the federal, state, and local government datasets they used in their research were funded exclusively or primarily from public funds. Further, most believe that the private datasets they used were not originally developed by a government agency using exclusively or primarily public funds. However, it is noteworthy that almost a fifth (4 out of 21) of the private datasets were stated to be funded with public money.

Table 4-12: Substantial Government Contribution of Database Using Public Funds

Public Funds	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes	227	118	55	23	14	4	13
No	34	2	1	0	10	14	7
Do not know	25	12	3	1	6	2	1
Unanswered	4	1	1	0	0	1	1
Total	290	133	60	24	30	21	22

QUESTION 8: *Was all or a substantial portion of this dataset or database originally developed by a university or private firm (profit or not-for-profit) using exclusively or primarily publicly financed research and development funds?* (e.g. government research grant to a public or private university or to a private company)

Table 4-13 shows that respondents believe that most of the data acquired by them from not-for-profit organizations was developed using public research and development funds (57%; 17/30). Perhaps it is surprising that 29% (6 out of 21) of the datasets acquired from private entities were believed by respondents to have been originally developed using exclusively or primarily publicly financed research and development funds.

Table 4-13: Substantial University or Private Sector Contribution to Creation of Database Using Public Funds

Public R&D funds	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes	48	9	9	3	17	6	4
No	185	97	40	17	6	10	15
Do not know	48	22	8	4	6	5	3
Unanswered	9	5	3	0	1	0	0
Total	290	133	60	24	30	21	22

Whether produced by governmental agencies, universities, or the private sector, the responses shown in Table 4-12 and Table 4-13 indicate that academic researchers using geographic data depend substantially on data originally developed using public funds.

QUESTION 9: *What specific contractual or licensing approach, if any, was imposed on your use of this dataset?*

Table 4-14 shows that the preponderance of datasets (81%; (52 + 184)/290) involves no licensing approach or assumes that no licensing approach applies to the used dataset. Federal government agencies (91%; (29 + 92)/290), state government agencies (78%; (8 + 39)/60), local government (67%; (3 + 13)/24), and not-for-profit organizations (87%; (3 + 14)/22) contribute highly to this general conclusion. According to the respondents, private entities do not appear to impose any restrictions on their data in 43% ((3+6)/21) of the cases.

When we compare percentages of approaches between local government and the rest we see that relatively many local government datasets were acquired on a boilerplate license basis; 17% (4/24) of local government datasets versus 7% (18/266) of the remainder of the datasets. Also more local government datasets were acquired after negotiating the license (13% (3/ 24) for local government versus 2% (5/266) for the rest). While the sample is relatively small, initial indications are that the data access policies of local government tend to be as restrictive or more restrictive than the policies of private firms; 19% (4/21) of the private datasets were acquired with a boilerplate license, and a license was negotiated for 14% (3/21) of the private datasets.

Table 4-14 shows that licensing or contract restrictions were imposed on the use of the dataset from the federal government 8% of the time (11/133). The percentage of the time that restrictions were imposed by other sources is as follows: state government 20% (12/60), local government 33% (8/24), not-for-profit organizations 13% (4/30), private entities 53% (11/21), and "other" organizations 23% (5/22).

Table 4-14: Licensing Approach Imposed on the Use of the Dataset

Licensing approach	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
No licensing or purchase contract provisions were involved in our use of this dataset (or in our use of a database from which the data was extracted)	184	92	39	13	20	6	14
We acquired this specific dataset in such a manner that we assumed that no contract or licensing provisions applied to our use of the data	52	29	8	3	6	3	3
"Boilerplate" license or purchase contract provisions were offered on a take-it or leave-it basis in response to our request for a specific or custom produced data set and we were required to sign or otherwise respond affirmatively to those provisions	18	3	3	4	2	4	2
License or purchase contract provisions were placed in writing by the supplier of the dataset or database when supplied but we were not required to sign or otherwise affirmatively assent through a volitional act to the terms	15	3	6	0	2	2	2
"Shrink-wrap" license or purchase contract provisions were offered on a take-it or leave-it basis (e.g. terms were contained in the packaging of a CD)	8	4	0	1	0	2	1
License or purchase contract provisions were negotiated with the supplier of the dataset or database	8	0	2	3	0	3	0
"Click-wrap" license or purchase contract provisions were offered on a take-it or leave-it basis (e.g. terms were stated on our computer screen to which we were required to affirmatively respond prior to downloading a dataset, accessing an on-line database or having a dataset shipped)	2	1	1	0	0	0	0
Unanswered	3	1	1	0	0	1	0
Total	290	133	60	24	30	21	22

QUESTION 10: *What restrictions, if any, were imposed on your use of this dataset or on your use of the computer database from which the data was acquired?*

As shown in Table 4-15, most datasets (65%; 189/290) could be used without any restriction imposed by the data provider. The federal government (78%; 104/133) especially allows access of datasets without restrictions.

We added the total number of restrictions per agency category and divided it by the number of datasets used in this study for the corresponding category in order to make the data of table 4-15 more transparent. The 24 datasets acquired from local government had restrictions imposed whereas 10 did not. Those imposing restrictions averaged 1.5 restrictions per dataset (i.e. 36 restrictions imposed by the 24). For private entities, 5 datasets were acquired without restrictions as 16 datasets had restrictions imposed. Those restrictions averaged 1.1 restrictions per dataset (i.e. 26 restrictions imposed by the 21). The same conclusions as for the licensing question may be drawn: the data access policies of local government tend to be as restrictive or more restrictive than the policies of private firms.

In absolute terms state government imposes restrictions on value-added products more than any of the other data providers. State government imposes restrictions in passing on digital data for 27% of their datasets mentioned in this study.

For 6 federal datasets a monetary payment was required. In one case this was a price based on a minimal statutory fee, and in another the price was based on the cost of dissemination to the user. However, two datasets were acquired at market price and two for a market price less a discount for the university or other not-for-profit user. If the respondents assessment is accurate and a statutory exception does not apply, the pricing

structure imposed does not conform to the marginal cost recovery rules of the federal government.

Table 4-15: Restrictions Imposed on the Use of the Dataset

Restriction	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Not applicable, no explicit or implied restrictions were imposed	189 65%	104 78%	35 58%	10 42%	19 63%	5 24%	16 73%
Provisions stated that we could not pass on the provided digital data to any other parties	48 17%	10 8%	16 27%	7 29%	5 17%	7 33%	3 14%
Provisions stated that our use could be for only academic or research purposes	54 19%	13 10%	15 25%	8 33%	7 23%	9 43%	2 9%
A monetary payment was required	19 7%	6 5%	1 2%	5 21%	0 0%	5 24%	2 9%
Provisions stated that any value-added products that we developed through use of the data (1) required explicit permission of the data supplier prior to dissemination of the value-added products by us, (2) vested an ownership interest in the original data supplier, or (3) required a royalty due to the data supplier	9 3%	1 1%	4 7%	1 4%	1 3%	1 5%	1 5%
Provisions stated that the data supplier would not be liable to us for any losses that we or others might incur due to any errors or other shortcomings in the data supplied	9 3%	2 2%	1 2%	2 8%	1 3%	2 10%	1 5%
Our understanding is that state legislation or other state law does not allow some of the uses we made of the dataset in this research project without first acquiring the permission of the data supplier (We therefore obtained that permission or ignored the law)	2 1%	1 1%	1 1%	0 0%	0 0%	0 0%	0 0%
Provisions stated that we are liable to the supplier of the data for any losses the supplier might incur to a third party through our inappropriate use of the data	1 0%	0 0%	0 0%	1 4%	0 0%	0 0%	0 0%
Our understanding is that federal copyright law does not allow some of the uses we made of the dataset in this research project without first acquiring the permission of the data supplier (We therefore obtained that permission or ignored the law)	1 0%	1 1%	0 0%	0 0%	0 0%	0 0%	0 0%
Other or alternative restrictions were imposed on the data	16 6%	5 4%	6 10%	2 8%	1 3%	2 10%	0 0%

Among other or alternative restrictions we found confidentiality of the data, respecting privacy of individuals, not for for-profit use, only for use of employees of this university, and "for cost recovery reasons they requested we not post data on our free FTP site for a year".

QUESTION 11: *What did you pay for access to or a copy of the dataset?*

As shown in Table 4-16, most datasets were freely accessible for the respondents (76%; 221/290). If we include the counts of price based on the cost of dissemination to the use (16 counts) and the price based on a minimal statutory fee (5 counts) then 83% ((221+16+5)/290) of the indicated datasets were available at a nominal price.

However, 16 times (12%; 16/133) respondents indicated that federal agencies charged the market price or the market price less a discount for the university or other not-for-profit user. This is also true for 2 datasets acquired from state government agencies and 1 dataset from a local agency.

Private entities charged the market price or the market price less a discount for the university or other not-for-profit user 57% of the time ((5+7)/21). For the nine private datasets used for free, one is led to wonder whether the dataset was paid for by another party and perhaps borrowed from, for instance, a library, whether the private company allowed the free use as an incentive or marketing technique for sale of its own products, or whether other dynamics were at work. It is for example known that private entities market their products with free demo versions, maps or other free material. It is however doubtful that this free data will be of any use for the academic researcher other than to assess the relevance of the dataset for a specific research project.

Table 4-16: Price of the Dataset

Price of the dataset	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Not applicable, the dataset was free	221	105	51	15	27	9	14
Market price less a discount for the university or other not-for-profit user	19	8	1	1	0	7	2
Market price	18	8	1	0	1	5	3
Price based on the cost of dissemination to the user	16	5	3	4	1	0	3
Price based on partial cost recovery for the producer	5	1	2	2	0	0	0
Price based on a minimal statutory fee	5	3	0	1	1	0	0
Price based on full cost recovery	3	2	0	1	0	0	0
Unanswered	3	1	2	0	0	0	0
Total	290	133	60	24	30	21	22

QUESTION 12: *How good was the documentation regarding the dataset?*

As shown in Table 4-17, approximately 53% (155/290) of the datasets used by academics were considered documented good or excellent. The majority of local government datasets (71% fair documentation or less; (4+4+9)/24), and datasets provided by not-for-profit agencies (53% fair documentation or less; (7+4+5)/30) are considered not well to be documented.

Table 4-17: Quality of the Documentation

Documentation	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Excellent	6	3	1	0	1	0	1
Good	149	74	33	7	13	11	11
Fair	61	28	10	4	7	7	5
Poor	16	4	2	4	4	1	1
Non-existent	53	22	11	9	5	2	4
Unanswered	5	2	3	0	0	0	0
Total	290	133	60	24	30	21	22

QUESTION 13: Which of the following did the documentation of the dataset (digital catalogue files or metadata) help you accomplish?

Table 4-18 shows some conflicting results with Table 4-17. In question 12, 22 federal government datasets, and 11 state government datasets were categorized as datasets without documentation. Question 13, in contrast, indicates that 33 federal government datasets, and 20 state government datasets lacked documentation helpful in accomplishing the tasks listed.

Table 4-18 shows that the documentation of a dataset is used extensively in the assessment of the usability of datasets for academic purposes. The documentation of approximately one out of three datasets was used to determine the relevance (121/290), technical suitability (117/290), quality or accuracy (100/290) and/ or timeliness (74/290).

Table 4-18: Accomplishment through Documentation

Accomplishment:	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Allowed us to assess the relevance of the dataset for our research project (e.g. data type, description entities)	121	60	21	10	16	9	5
Allowed us to assess the technical suitability of the dataset (e.g. data structure)	117	58	21	9	12	8	9
Allowed us to assess the quality or accuracy of the dataset	100	52	19	3	9	10	7
Not applicable, no documentation or metadata was available	81	33	20	9	5	5	9
Allowed us to assess the timeliness of the dataset for our purposes	74	37	16	1	11	8	1
Allowed us to find the dataset through a computer search	34	25	5	1	2	1	0
Allowed us to assess contractual or other legal constraints on the use of the dataset	20	9	7	0	2	2	0

QUESTION 14: *Was access to this dataset or database made available to you within a reasonable period of time of requesting access?*

Table 4-19 shows that datasets were made available to the researcher immediately or within a reasonable period of time 92% of the time ((136+130)/290). Most datasets disseminated by a federal government agency (55%; 74/133) and not-for-profit organization (57%; 17/30) were accessed immediately. This may be because these agencies allow access to their datasets through the Internet to a greater extent than others (see Table 4-8).

In only a few instances was the time for availability unreasonable (6%; 18/290). Datasets acquired from local government were acquired within an unreasonable period of time 17% of the time.

Table 4-19: Timeliness of Accessing the Dataset

Time of requesting access	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Yes, access was immediate	136	74	25	4	17	6	10
Yes, the time between the request and obtaining the data was reasonable	130	50	29	16	12	14	9
No, the time between the request and obtaining the data was unreasonable	18	7	4	4	1	1	1
Unanswered	6	2	2	0	0	0	2
Total	290	133	60	24	30	21	22

QUESTION 15: *If you acquired access to this dataset through a database service to which your university library subscribes or participates in supporting, how was this database made available to you?*

Table 4-20 shows that there was only a marginal role for the library in accessing geographic datasets for academic use by the respondents. Only 10 datasets (3%) were accessed through the library and those were primarily federal government datasets, probably distributed on CD's to libraries as part of the government documents library depository program.

Table 4-20: Access Dataset through a Library

Access through a Library Service	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Not applicable to this dataset	266	115	58	24	29	20	20
We paid a per use fee, the library paid a per use fee, or we acquired special permission that might not be granted to all library patrons	0	0	0	0	0	0	0
We acquired access through an open access policy applied to all library patrons; no per use fee was charged nor was special permission required	10	7	0	0	1	0	2
Unanswered	14	11	2	0	0	1	0
Total	290	133	60	24	30	21	22

QUESTION 16: *Is it possible to access the same or similar dataset meeting your needs from another source?*

Table 4-21 shows that most of the datasets (55%; 160/290) were the only dataset the respondents could realistically use. For example, most data from local government are, according to the respondents, only accessible through local government (83%; 20/24). Also state government (65%; 39/60), not-for-profit organizations (57%; 17/30), and data from other sources (82%; 18/22) are major sole resources for datasets used by the respondents.

Federal government was the sole realistic data provider only in 43% of cases (identical to private datasets). This may be a result of the federal government's open access policy,

stimulating other entities to use free data, and to make improved versions of federal datasets accessible to others.

Furthermore, the expenses of other data were barriers in the accessibility of existing alternatives. For this reason, federal government data was preferred twelve times (9%), and private data five times (24%) over alternatives.

Table 4-21: Existence of Alternative Datasets

Alternatives	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
No, this was the only realistic source for the dataset	160	57	39	20	17	9	18
Yes, but access through this source was more convenient	100	59	15	3	11	8	4
Yes, but the expense of other sources was not as responsive to our needs	24	12	2	1	3	5	1
Yes, but the quality of the dataset from other sources was not as responsive to our needs	15	9	2	0	1	3	0
Yes, but the restrictions imposed by other sources were not as responsive to our needs	7	3	2	1	0	1	0

QUESTION 17: *Which of the following, if any, were significant factors in allowing you to successfully use this dataset?*

Table 4-22 shows the counts for the factors allowing successful use of the dataset. The major factors of allowing successful use in the dataset are: sufficient quality or accuracy (62%; 181/290), physical means for gaining access (56%; 163/290), suitable format or compatibility with the software or hardware used (52%; 152/290), timeliness (46%; 132/290), cost (42%; 121/290), and personal or institutional willingness to giving access to the dataset (38%; 111/290).

Table 4-22: Factors Allowing Successful Use of the Dataset

Success factors in use of the dataset	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Sufficient quality or accuracy of this dataset for our purposes	181	78	40	15	19	16	17
The physical means for gaining access to this dataset	163	84	30	12	18	9	10
Suitable format or compatibility with the software or hardware we used	152	78	28	10	18	10	12
Timeliness of this dataset for our purposes	132	57	21	11	16	15	12
Cost of this dataset	121	64	22	9	12	4	10
Personal or institutional willingness to giving us access within the organization that created the dataset	111	41	31	15	16	4	4
Adequate documentation or metadata for this dataset	96	54	13	5	10	7	7
Sufficient identification of the sources used to create this dataset	82	39	16	6	13	5	7
Lack of application of copyright law to our uses of this dataset	72	43	13	6	4	1	5
Lack of application of specific data protection legislation to our uses of this dataset (e.g. local ordinance, state statute, federal statute)	49	30	8	4	2	0	5
Availability of a search capability allowing the ability to find this dataset or database	34	26	2	0	4	1	1
Contractual provisions facilitating our uses of this dataset	18	8	3	3	1	3	0
Contractual provisions regarding further dissemination of this dataset	7	2	1	0	1	3	0
Contractual provisions regarding liability	2	1	0	0	1	0	0
Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset	1	0	0	0	1	0	0
Other	8	2	1	1	3	1	0

Other factors that were mentioned in allowing successful use of the data were: met our spatial needs, prior access to this site, hard drive space, identifiable agreement to disclaimer terms not required, and only appropriate dataset.

QUESTION 18: *Which of the following, if any, were significant impediments to your use of this dataset?*

As shown in Table 4-23 important factors of concern to the use of datasets are: documentation (21%; 62/290), physical means for gaining access (12%; 35/290), quality or accuracy (10%; 30/290), lack of alternative datasets (10%; 28/290), timeliness (8%; 23/290), and lack of identification of the sources to create the dataset (7%; 21/290).

Table 4-23: Factors Significant Impediments to Use of the Dataset

Impediment	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Inadequate documentation or metadata for this dataset	62	22	14	8	7	7	4
The physical means for gaining access to this dataset	35	16	4	3	4	3	5
Inadequate quality or accuracy of this dataset for our purposes	30	15	7	1	2	3	2
Lack of alternative datasets meeting our needs	28	13	5	4	0	3	3
Timeliness of this dataset for our purposes	23	8	6	4	4	0	1
Lack of identification of the sources used to create this dataset	21	6	4	2	2	3	4
Lack of a search capability allowing the ability to find this dataset or database	20	7	4	5	4	0	0
Lack of suitable format or compatibility with the software or hardware we used	19	5	5	3	3	1	2
Personal or institutional resistance to giving us access within the organization that created the dataset	13	1	5	4	1	1	1
Cost of this dataset	6	1	0	2	0	3	0
Restrictions imposed on our use of the dataset by specific data protection legislation (e.g. local ordinance, state statute, federal statute)	4	1	1	1	0	1	0
Contractual restrictions imposed on our uses of this dataset	4	0	1	1	1	1	0
Contractual restrictions regarding further dissemination of this dataset	4	0	1	1	1	0	1
Restrictions imposed on our use of the dataset by copyright law	2	0	0	1	0	1	0
Contractual provisions regarding liability	1	0	0	1	0	0	0
Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset	0	0	0	0	0	0	0
Other	11	4	3	2	0	0	2

The other class included the following impediments: elements in data gathering process, lack of conversion software, changing projections, and learn how to use.

QUESTION 19: *Even though contractual, legal, technical and other impediments may have constrained your use of the specific dataset, to what degree were you able to accomplish research tasks that were dependent upon use of this dataset?*

Table 4-24 shows that most datasets allowed the accomplishment of almost all or most research tasks dependent on the datasets. This is logical, since the dataset was used in the project.

However, if *almost all* and *most* research tasks are considered one subgroup of options, interesting differences between data provider categories appear. Data provided by the federal and local government scores in this subgroup for 80% ((80+26)/133, (9+10)/24) of the counts. Data provided by a state government agency (92%; (40+15)/60), private firms (90%; (12+7)/21), and data providers in the other category (91%; (17+3)/22) seem to allow a more productive use of their datasets, although this difference may not be statistically significant.

Table 4-24: Tasks Accomplishment of the Dataset

Task accomplishment	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Almost all research tasks dependent on this dataset were accomplished	175	80	40	9	17	12	17
Most research tasks dependent on this dataset were accomplished	66	26	15	10	5	7	3
About half of the research tasks dependent on this dataset were accomplished	16	4	3	3	3	1	2
Some of the research tasks dependent on this dataset were accomplished	15	10	1	2	2	0	0
Almost none of the research tasks dependent on this dataset were accomplished	2	1	0	0	1	0	0
Unanswered	16	12	1	0	2	1	0
Total	290	133	60	24	30	21	22

QUESTION 20: *How would you rate your satisfaction with your use of this specific dataset or database?*

Satisfaction is more uniformly distributed over the data provider categories (see Table 4-25. Respondents expressed their satisfaction with use of the dataset as excellent or good 82% of the time (102+137)/290). This overall percentage is similar for each separate category. Only datasets provided by local government agencies score more than ten percent lower (71%; (5+12)/24) than the overall score.

Table 4-25: Satisfaction with the Dataset

Satisfaction	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Excellent	102	53	21	5	11	6	6
Good	137	60	31	12	12	11	11
Fair	45	17	8	6	6	3	5
Poor	3	1	0	1	1	0	0
Non-existent	0	0	0	0	0	0	0
Unanswered	3	2	0	0	0	1	0
Total	290	133	60	24	30	21	22

QUESTION 21: *Use of this specific dataset was important in accomplishing the overall objectives of the research project*

The answers to the statement of question 21 are presented in Table 4-26. Most respondents strongly agreed or agreed with the statement above. No substantial differences between the groups are evident.

Table 4-26: Importance of Dataset for Accomplishment of Overall Research Objectives

Dataset was important in accomplishing overall research objectives	*Total*	*F*	*S*	*L*	*N*	*P*	*O*
Strongly agree	198	95	38	16	15	15	19
Agree	82	35	19	8	13	5	2
Disagree	3	1	1	0	1	0	0
Strongly disagree	0	0	0	0	0	0	0
Do not know/ no opinion	0	0	0	0	0	0	0
Unanswered	7	2	2	0	1	1	1
Total	290	133	60	24	30	21	22

4.5 Section 4 Desired Datasets

Section 4 of the questionnaire allows respondents to fill out questions about a dataset that they would have liked to use for their project but failed to acquire.

Out of the 148 useful responses 20 participants (14%) indicated that they desired to use other datasets. Due to this small percentage of responses and the high variability in the reasons why the preferred datasets were not acquired we chose not to use statistical analysis in evaluating these responses. Instead, counts are presented on a question by question basis.

QUESTION 1: *Why did you want this particular dataset?*

Mostly technical reasons were mentioned in favor of the desired dataset.

Table 4-27: Why Dataset was Desired

Why did you want this particular dataset?	*Count*
The dataset consists of more accurate or reliable data	7
The dataset is more comprehensive or complete	7
The dataset has higher quality data	6
The dataset is more up-to-date	4
The dataset is better documented	1
The dataset is more flexible	1
The dataset is more user friendly	0
Other. Please specify	9

Other reasons why respondents wanted to access a desired dataset were: more relevant to the ultimate objectives of the project, would be useful, new research area, advisory committee priority, compliments data set we used, had needed variables, needed to complete regional coverage, dataset contains desirable or otherwise useful data not currently used for our project, and needed flow data.

QUESTION 2: *Why didn't you acquire access to this particular dataset?*

Table 4-28 shows that six respondents (30%) stated that the expense of the other dataset was the reason for not acquiring it. Incompatibility with software, or hardware was mentioned 3 times (15%). In the "other" category the answer "the data provider did not respond to our request" was filled out 4 times (20%). Also the following responses appeared in the other box: difficult to create a reasonable GIS layer, project time constraints and expense with respect to data quality, insufficient coverage of project area, legally protected confidentiality of data, limited utility programs to convert data, not available, too many companies to contact, potentially too many formatting problems.

Table 4-28: Reasons for Not Acquiring Desired Dataset

Why didn't you acquire access to this particular dataset?	*Count*
The dataset was too expensive	6
The dataset was incompatible with our software or hardware limitations	3
Until (very) recently the existence of this dataset was unknown to us	2
The restrictions imposed on this dataset were not responsive to our needs	1
The dataset was no longer available in digital format	0
The documentation of the dataset was inadequate or not responsive to our needs	0
Exclusive rights were given to another organization	0
Other reason(s), please specify:	8

QUESTION 3: *From whom could you directly acquire this dataset?*

Table 4-29 shows that datasets available from the creator of the dataset, and intermediate commercial entities were not available to the respondents. Most of these datasets are created with the support of public funds (see Table 4-30).

Table 4-29: Acquire Desired Dataset From

From whom could you directly acquire this dataset?	*Count*
The creator of the dataset	12
An intermediate commercial entity, not being the primary creator of the dataset	4
An intermediate non-commercial entity, not being the primary creator of the dataset (e.g. public library, university, government agency, etc.)	1
Do not know	3

Table 4-30: Public Funds Used to Create Desired Dataset

Public Funds?	*Count*
Yes	15
No	2
Do not know	3

QUESTION 5: *Was all or a substantial portion of this dataset or database originally developed by a university or private firm (profit or not-for-profit) using exclusively or primarily publicly financed research and development funds?* (e.g. government research grant to a public or private university or to a private company)

One respondent believed that his desired dataset was created with the help of public R&D funds, as 15 did not.

Table 4-31: Public R&D Funds Used to Create Desired Dataset

Public R&D Funds?	*Count*
Yes	1
No	15
Do not know	4

Chapter 5 Support or Nonsupport of Access to Scientific and Technical Data Principles

5.1 Introduction

This chapter presents the results of the survey analysis. The responses provide opportunities to test the data in many ways. For example to test the most productive answer per question. However, in this thesis we only assess whether the data access principles introduced in chapter 2 are adhered to or not. We present the analysis on a principle by principle basis.

We use three different measures of productivity for the assessment. These are:

(1) task accomplishment of the dataset,

(2) satisfaction with the dataset, and

(3) overall accomplishment of objectives of the research project.

The t-test is used to test for statistical significance.

Furthermore, an assessment is made in terms of success or impediments in the use of the dataset. We use the chi-square (χ^2) test to address this statistically.

The results presented in this chapter are based only on the returned questionnaires assessed as being useful for this project. Participants who indicated that they did not accomplish academic research with either geographic data or GISs are not included in the analysis.

Due to a variety of reasons we were unable to test use of the academic sector of datasets acquired from the private sector. The primary limitation was often lack of sufficient

sample size. For the other categories, geographic data acquired from government and other academics, we present in this chapter the results of the analyses.

The spreadsheet package Microsoft Excel is used to perform the statistical t-test and chi-square test on the selected fields from the database Microsoft Access.

5.2 Statistical Justification

In this research we asked for different types of data. On the one hand respondents provided us with interval data. They had to indicate how productive the dataset was to them on a scale varying from very low (score 1) to very productive (score 5) (see Questions 19, 20, 21 Section 3). On the other hand we used counts of successful use (Question 17 Section3) or impediments in the use (Question 18 Section 3). The different types of data are tested with different statistical tests: the t-test for the interval data and the chi-square test for counts of success. These tests are described in this paragraph but first we explain the level of significance and the degrees of freedom, applying to both tests.

5.2.1 Level of Significance

In both tests we test the data on a certain level of statistical significance. The level of significance indicates how great the risk is of rejecting the null-hypothesis. If the level of significance is 5% (0.05), the probability of falsely rejecting the null-hypothesis is 5% (Mark Shirkin 1995, 189).

In order to make a decision it is custom and tradition to choose a level of significance of 5% (Mark Shirkin 1995, 195). But one is free to choose a higher or lower level if one chooses to do so.

In this research we only indicate at what level of significance the hypotheses are accepted or not. One should decide whether this level is acceptable in order to decide on it.

5.2.2 Degrees of Freedom

Tests of significance, like the χ^2 and the t-test, use a critical value to decide on significance. Critical values vary from one test to another depending on the degrees of freedom (df). We need to find the degrees of freedom in order to find the critical value. The degrees of freedom refer to the number of unknowns in an equation that are free to vary. For example, the equation a + b + c =10 has 2 degrees of freedom: two of the unknowns are free to vary, the third is fixed.

The degrees of freedom for the t-test we used are calculated by:

Df = (total number of datasets included in the test) – 2

The degrees of freedom for the χ^2 –test are calculated as follows:

Df = (number of rows –1) * (number of columns –1)

5.2.3 The Statistical T-test

The t-test may be used to test a hypothesis stating that the mean scores on some variable will be significantly different for two independent samples of groups (Zikmund 1991, 504). We used the two-sample t-test to test for differences of means in the productivity

measures (task accomplishment, satisfaction, and overall objective accomplishment). To use the t-test we assume a normal distribution for all the samples, equal variances between the samples, and we assume interval data.

In this research we divide the responses into two groups: one with datasets adhering to the proposed principles and one with datasets ignoring or violating the principles. The null hypotheses state that the group adhering to the proposed principles is more productive than the group ignoring or violating the principles. Exceeding the critical t-value means statistical significance in differences of means. In other words we can say that the subgroup with the most productive responses should be favored over the other subgroup.

Table 5-1 shows an example of how we present the results of the t-test in this chapter. We tested datasets acquired at marginal costs against datasets acquired for more than marginal costs. 242 datasets were acquired for marginal costs as 45 datasets were acquired for more than the marginal costs. Satisfaction was measured on a scale varying from 1 (non existent) to 5 (very satisfied). The mean is 4.204 for datasets available at marginal costs.

Table 5-1: Example T-test for Costs of Datasets

Productivity Measure	*Satisfaction*	
Costs marginal?	*Yes*	*No*
Counts	242	45
Mean	4.204	4.023
Variance	0.515	0.534
Df	285	
T-value	2.159	

The critical t-value for 285 degrees of freedom at a 0.05 level of significance is 1.960. The t-value in our test is 2.159 , exceeding the critical t-value. Thus, we conclude, at

a level of significance of 5%, that respondents using datasets acquired at marginal costs, are significantly more satisfied with their dataset than respondents using datasets acquired for more than the marginal costs (4.204 v. 4.023).

5.2.4 Chi-square (c^2) Test

The chi-square distribution provides a means for testing the statistical significance of contingency tables. This allows us to test for differences in two groups' distribution across categories (Zikmund 1991, 500).

The chi-square test may be used for a "goodness of fit" test. This test compares the observed distribution with the expected distribution. We expect the proposed access principle to be the most "ideal" situation. Observations identical to the expected value would be deemed "most successful and productive for academic researchers".

Respondents could indicate success factors and impediments for the dataset they use(d). Respondents with a dataset adhering to the data access principles was expected to choose the success option. We did not expect datasets adhering to the principles to have any impediments mentioned. This implies that, in Table 5-2, the expected values in the rows no success and impediments will be zero. However, the chi-square test requires that for a 3x2 matrix no expected values can be zero and maximum of 20% of the expected values are between 1 and 5 (Mark Sirkin, 1995, 363). Due to these requirements we were unable to use the chi-square test in this manner (see Table 5-2 for an example).

Table 5-2: Example Chi-square Test "Goodness of Fit" for Costs of Datasets

	No Costs Observed value	*No Costs Expected value*	*Total*
Success	50	100	150
No success	40	0 (ERROR)	40
Impediment	10	0 (ERROR)	10
Total	100	100	200

Instead, we used the chi-square test to compare the distribution of two different groups: one group adhering to the principles and one group ignoring the principles. In this way we circumvent the requirement of the expected cell frequencies. The test describes uniformity of the distributions. If the distributions are significantly not uniform (do not belong to the same sample), we can conclude that one distribution allows more successful use of a dataset than the other. In order to measure this, we compared the group percentages of successful use and impeded use. The group with the highest percentage successful use and the lowest percentage impediments is preferred over the group with the lower percentage successful use and higher percentage impediments in the use of the dataset.

In the analysis, all null hypotheses state that the distributions are uniform; the distribution of the two groups are not significantly different. Exceeding the critical value makes us reject the null hypothesis. If so, we accept the alternative hypothesis stating that the distributions are not uniform and decide which group is more successful in the use.

We use an example to demonstrate the theory. In Table 5-3 we see that from datasets that were acquired for free, respondents indicated 50 times that the cost of the dataset was a success factor in the use of the dataset. Similarly respondents mentioned for the datasets not acquired for free that in 50 cases they found the costs of the dataset an impediment.

Table 5-3: Example of Chi-square Test: Cost of Datasets

Cost of the dataset a success?	*Dataset is free Observed values*	*Dataset is not free Observed values*	*Total*
Success	50	0	50
No success	40	10	50
Impediment	0	50	50
Total	90	60	150
Chi-square value (df=2)	117		

We test the null hypothesis stating that the distributions of the two observed groups are uniform. In this case we find a Chi square value of 117. The chi-square value from the observation exceeds the critical chi-square value (13.82) at a level of significance of 0.001 (df =2). Thus, the distributions of the groups in terms of allowing successful use are significantly not uniform. If we compare the percentages of success we see that the cost issue was considered a success in 50/90 (56%) times for the free datasets and an impediment in 0/90 (0%) of these cases. The issue was never considered a success for the "not free" datasets (0/60; 0%) but in 50/60 (83%) of these cases an impediment in the use. We conclude that the free datasets allow more successful use than the datasets acquired at costs.

5.3 Principles for Data Provided by the U.S. Government

A decision on whether principles for data provided by the U.S. government was adhered to in a specific instance, may be established by first determining whether a respondent used datasets produced or provided by a federal, state or local agency (Question 7 section 2). Then we determined whether the respondent filled out the questionnaire for such a dataset (First Question, Section 3). The responses of Question 7 Section 2 and the First Question of Section 3 only indicate whether a government dataset was used and do not

necessarily correspond with the dataset for which the questionnaire was filled out. Thus, we only used the responses of the First Question of Section 3 to decide on government datasets.

Of the datasets reported 133 are federal datasets, 60 are state datasets and 24 are local government datasets (a total of 217 as reported in chapter 4). Further the datasets had to receive a yes response to Question 7 of Section 3 (substantial portion of the dataset developed with public funds). This resulted in a total of 196 datasets used for the tests in this paragraph. We tested these datasets in a group adhering to against a group of datasets ignoring the recommended principles for government datasets, as set forth in chapter 2. Where applicable, we created more than two levels of adherence to the proposed principles.

5.3.1 Principle 1: "Level of Availability"

> Government agencies should ensure that electronic data, information and value-added features developed with public funds are available to the public.

A measure of availability was established through an analysis of the Questions 3, 4, 5, 6, 9, 10, 11, 12 and 14 in Section 3. The highest ranking would have had the following responses to these questions: Question 3: any answer except acquired on paper or self-collected, Question 4: no, Question 5: no, Question 6: no, Question 9: no licensing or we assumed no contract or licensing provisions, Question 10: not applicable, Question 11: not applicable, cost of dissemination, or minimal statutory fee, Question 12: good or

excellent, Question 14: immediate or reasonable. Responses for 30 datasets adhered to this highest ranking.

The lowest ranking would have had the following responses to these questions:

Question 3: any answer but acquired on paper or self-collected, Question 4: yes

Question 5: yes, Question 6: yes, Question 9: any answer but "no licensing" or "we assumed no contract or licensing provisions", Question 10: any answer but "not applicable", Question 11: market price, market price less a discount, price based on full or partial cost recovery, Question 12: non-existent, poor or fair, Question 14: unreasonable. None of the datasets adhered to all the qualifications of the lowest ranking. Instead we used another low level of availability: Question 3: any answer, Question 4: yes, Question 5: yes, Question 6: yes, Question 9: any answer but "no licensing" or "we assumed no contract or licensing provisions", Question 10: any answer but "not applicable", Question 11: any answer, Question 12: any answer, Question 14: any answer. 13 datasets were categorized in this low level group.

We used the t-test to test the differences in productivity of datasets with the highest ranking with datasets ranked as a low level of availability. Questions 19, 20 and 21 of section 3 were used as measures of productivity. The results are presented in Table 5-4.

Table 5-4: T-test for Level of Availability

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall objective Accomplishment*	
Open access?	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Count	30	13	30	13	30	13
Mean	4.621	3.923	4.517	4.077	4.724	4.769
Variance	0.530	1.577	0.330	0.744	0.207	0.192
Df	41		41		41	
T-value	2.183		2.713		-0.670	

The critical t-value for 41 degrees of freedom at a 0.01 level of significance is: 2.702, and at a 0.05 level of significance 2.020.

Here we see that datasets adhering to principle 1 are more satisfying at a level of 0.01 statistical significance (4.517 v. 4.077) and researchers using these datasets accomplish, at a level of significance of 0.05, significantly more tasks (4.621 v. 3.923) than researchers with datasets acquired through a less open environment.

We also tested the two levels of availability in a Chi square test. As measures of success we used the following answers to Questions 17 and 18 of Section 3: physical means for gaining access, adequate documentation, timeliness of the dataset, personal or institutional willingness to giving us access, lack of application of copyright, lack of application of specific data protection legislation, cost, lack of the other mentioned legal restrictions. We counted the corresponding impediments in Question 18. This resulted in the following Table 5-5.

Table 5-5: Chi-square Test for Level of Availability

Measure:	*Open Access Highest (30 datasets)*			*Open Access Low (13 datasets)*		
	S	*NS*	*I*	*S*	*NS*	*I*
Timeliness	15	13	2	6	4	3
Personal willingness	12	18	0	6	4	3
Lack/ application of copyright	13	17	0	2	11	0
Lack/ application of data protection legislation	10	20	0	2	11	0
Cost	19	11	0	6	6	1
Physical means	25	4	1	9	1	3
Documentation	20	8	2	4	4	5
Contractual provisions	2	27	1	5	6	2
Total	116	118	6	40	47	17
X2 value df=23	40.4					

S=Success, NS = No Success, I = Impediment

Critical chi-square value at a level of significance of 0.05 is 35.17

The chi-square value we found exceeds the critical value at a level of significance of 0.05. Thus, the distributions are not significantly uniform at this level of significance. If we count how many times a success and impediment were filled out for both groups and divide this by the number of datasets in the corresponding group we may get an indication of the datasets of most successful use to the researcher. The highest group scores 389% (116/30) for success and 20% (6/30) for impediments. The lowest level group scores 307% (40/13) for success and 131% (17/13) for impediments.

We conclude that the highest level group allows more successful use of datasets than the lowest level group at a 0.05 level of significance.

The numeric results of both the t-test and the chi-square test in assessing the relation of success with conformance to principle 1 are included in Table 5-40. Similarly this table shows the results of the further tests of principles discussed throughout the remainder of this section.

5.3.2 Principle 2: "Level of Affirmativeness in Dissemination"

> Government agencies should adopt affirmative programs of electronic public information dissemination so that scientists do not need to resort to Freedom of Information requests in order to gain access to government records.

A measure of availability was established through an analysis of the Questions 1, 2, 3, and 4 of Section 3 of the questionnaire. A highest ranking would have had the following answers to these questions: Question 1: creator, Question 2: anything but personal inquiries, Question 3: Internet, digital portable medium or e-mail, Question 4: No.

29 responses were categorized as highest level.

A lowest ranking would have had the following answers: Question 1: creator, Question 2: anything but personal inquiries, Question 3: paper or other analogue medium, Question 4: yes. Only 2 datasets were ranked as lowest. We were unable to test the highest ranked datasets against the lowest ranked datasets.

An alternative level of affirmativeness may be found when Question 1, 2, 3 and 4 are analyzed differently. Here, a dataset will be included in the test group if the answer to Question 1 is creator, Question 2 is anything but personal inquiries, and the answer to Question 3 is Internet, digital portable medium or e-mail. Question 4 Section 3: Specific

request decides on the level of affirmativeness. Datasets acquired with a specific request score on this level low as datasets accessed without a specific request score high. 29 datasets scored high and 25 datasets scored low. Table 5-6 shows the results of the t-test.

Table 5-6: T-test for Level of Affirmativeness in Dissemination

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Specific Request?	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Count	25	29	25	29	25	29
Mean	4.318	4.538	4.417	4.310	4.625	4.690
Variance	1.180	0.978	0.514	0.579	0.332	0.222
Df	52		52		52	
T-value	-0.750		0.708		-0.853	

At no level of significance were the differences in the mean of the two groups significantly different. We conclude that it does not make a difference in productivity whether one obtains his dataset with or without a specific request.

We also tested the two levels of affirmativeness in dissemination with a Chi square test. As measures of success we used the following answers to Questions 17 and 18 of Section 3: Physical means for gaining access to this dataset, personal or institutional willingness to giving us access within the organization that created the dataset, personal or institutional resistance to giving us access within the organization that created the dataset. Table 5-7 shows the results.

Table 5-7: Chi-square Test for Level of Affirmativeness in Dissemination

	Highest level (total of 29 datasets)	*Low level (total of 25 datasets)*
Success physical means	19	19
No success physical means	6	6
Impediment physical means	4	0
Success personal willingness	8	6
No success personal willingness	13	19
Impediment personal resistance	2	0
Total	58	50
Chi Square (df=5)	6.00	

The critical chi-square value at a 0.10 level of significance is 9.24 (5 df).

The chi-square value we found, is 6.00. Thus we assume that the two groups belong to the same group. We follow the conclusion of the t-test: the issue of a specific request does not significantly influence the successful use of the datasets.

This conclusion may make sense when one realizes that the datasets we asked for were already in the possession of the researcher. Thus for these datasets the researcher had a positive experience with the specific request issue. This may have influenced the results of the tests. In theory however, a positive response of a data producer to a specific request should highly satisfy a researcher when this specific request resulted in tailor made datasets. In this respect the results of the analysis provide some evidence that datasets adhering to the proposed principle are considered as good as datasets for which a specific request was made and accepted.

5.3.3 Principle 3: "Level of Activity in Release"

> Government agencies should anticipate requests by the general public (including the scientific community) for electronic information and should build features into their electronic information systems so that information most likely to be requested by the public may be actively released (such as publishing datasets on web servers or CDs along with appropriate retrieval software)

A measure of activity was established through an analysis of the Questions 2, 3 and 13 of Section 3. A highest ranking would have had the following answers: Question 2: Internet, Specific Database or Library, Question 3: Internet, digital portable medium or e-mail, Question 13: find through documentation. We found 11 datasets that adhered to the highest level.

The second level of activity would have had the following answers: Question 2: any answer but through personal inquiries, Question 3: Internet, digital portable medium or e-mail, Question 13: find through documentation. 28 datasets qualified for the second level.

A low ranking would have had the following answers, Question 2: any answer but through personal inquiries, Question 3: paper or other analogue medium, Question 13 any answer but found through documentation. 22 Datasets qualified for this rank.

The lowest ranking would have had the following answers: Question 2: find through paper or other analogue medium, Question 3: paper or other analogue medium, Question 13: any answer but found through documentation. Only 2 datasets qualified for this group.

We tested the highest level of activity against the low (not the lowest!) ranked datasets in a t-test. The results are presented in Table 5-8.

Table 5-8: T-test for Level of Activity in Release

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall objective Accomplishment*	
	Highest level	*Low level*	*Highest level*	*Low level*	*Highest level*	*Low level*
Count	11	22	11	22	11	22
Mean	4.545	4.368	4.455	4.048	4.727	4.810
Variance	0.873	1.135	0.673	0.448	0.218	0.162
Df	31		31		31	
T-value	0.453		2.076		-1.224	

Datasets in the highest level group satisfy researchers significantly more than datasets in the low ranked group at a level of significance of 0.05 (critical value 2.043).

We also tested the second level of activity against the low ranked datasets in a t-test. Table 5-9 shows the results.

Table 5-9: T-test for Level of Activity in Release II

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall objective Accomplishment*	
	Second level	*Low level*	*Second level*	*Low level*	*Second level*	*Low level*
Count	28	22	28	22	28	22
Mean	4.556	4.368	4.464	4.048	4.821	4.810
Variance	0.487	1.135	0.480	0.448	0.152	0.162
Df	48		48		48	
T-value	0.787		3.137		0.267	

The critical value at a level of significance of 0.01 is 2.660. Thus, Table 5-9 above provides evidence that researchers are more satisfied with datasets acquired from an environment adhering to the principle than datasets ignoring the principle.

We also tested the different groups with a Chi square test. As measures of success we used the following answers to Questions 17 and 18 of Section 3: Physical means for gaining access to this dataset, availability of a search capability allowing the ability to find this dataset or database, lack of a search capability allowing the ability to find this dataset or database. The results are presented in Table 5-10.

Table 5-10: Chi-square Test for Level of Activity in Release

	Highest level	*Low level*	*Total*
Success physical means	8	12	20
No success physical means	3	3	13
Impediment physical means	0	7	7
Success search capability	4	2	6
No success search capability	7	19	26
Impediment search capability	0	1	1
Total	22	44	66
Chi square value df=5	8.63		

The critical value at five degrees of freedom is: 9.24 at the 0.10 level of significance. The chi-square value does not exceed the critical value so no significant differences exist between the two groups. We also tested the second highest ranked group with the low group. The results are presented in Table 5-11.

Table 5-11: Chi-square Test for Level of Activity in Release II

	Second highest level (28 datasets)	*Low level (22 datasets)*	*Total*
Success physical means	19	12	31
No success physical means	6	3	9
Impediment physical means	3	7	10
Success search capability	11	2	13
No success search capability	16	19	35
Impediment search capability	1	1	2
Total	56	44	100
Chi square value df=5	9.36		

The critical value at five degrees of freedom is: 9.24 at the 0.10 level of significance.

Now we see that the distribution within the groups are significantly not uniform at a 0.10 level of significance. The second highest group scores for successful use 107% (30/28) as the low group scores 64% (14/22). The second highest group also scores better for the impediment 14% ((3+1)/28) v. 36% ((7+1)/22).

We conclude that at a 0.01 level of significance the group adhering to the principle allows more successful use of the dataset than the group ignoring the principle.

5.3.4 Principle 5: "Level of Metadata Availability"

> Scientific and technical data collected or maintained by or under authority of a government agency should be documented adequately with metadata.

A measure of availability was established through an analysis of the Question 12 of Section 3. A highest ranking would have had the following answers: Question 12: good or excellent documentation. 109 datasets qualified for the highest level of adherence.

A lowest ranking would have had the following answers: Question 12: fair, poor or non-existent documentation. 84 datasets qualified for this lowest level of adherence.

3 respondents did not fill out this question. So a total number of responses of 193 was analyzed. The results are presented in Table 5-12.

Table 5-12: T-test for Level of Metadata Availability

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Adequate Documentation?	Yes	No	Yes	No	Yes	No
Counts	109	84	109	84	109	84
Mean	4.657	4.146	4.398	3.904	4.726	4.614
Variance	0.371	1.287	0.391	0.576	0.201	0.289
Df	191		191		191	
T-value	3.935		7.090		3.176	

The critical T-value for 191 degrees of freedom at a 0.001 level of significance is: 3.291.
The critical T-value for 191 degrees of freedom at a 0.01 level of significance is: 2.576.

The test shows that datasets with adequate documentation are at a 0.001 level of significance for two measures of productivity (task accomplishment and satisfaction) more productive than datasets with inadequate documentation. Datasets with adequate documentation also allow, at a level of significance of 0.01, significantly more overall objectives to be accomplished than datasets with inadequate documentation. We conclude that people who indicated that they had datasets with adequate documentation are more productive in their research than researchers working with datasets with inadequate documentation.

We also tested the two levels of Availability of Documentation in a Chi square test. As measures of success we used the following answers to Questions 17 and 18 of Section 3: adequate documentation or metadata for this dataset and inadequate documentation or metadata for this dataset. Table 5-13 shows the results.

Table 5-13: Chi-square Test for Level of Metadata Availability

Adequate Documentation?	*Yes*	*No*	*Total*
Success	56	6	62
No success	43	51	94
Impediment	10	27	37
Total	109	84	193
Chi-square value (df=2)	46.2		

The critical chi-square value for 2 degrees of freedom at a 0.001 level of significance is: 13.82.

At a 0.001 level of significance the two groups are significantly not uniform. The group with adequate documentation scores 51% (56/109) for the success measure as only 7% (6/84) of the datasets did in the other group. The group with datasets with adequate documentation also scored better on the impediments measure: 9% (10/109) versus 32% (27/84). The chi-square test confirms that the availability of adequate documentation allows significantly more successful use of a dataset than datasets lacking adequate documentation.

One may wonder what adequate documentation is. The responses to Question 13 Section 3 provide us with background information on the documentation of a dataset. Question 13 provides 6 metadata features. Documentation may be considered adequate when a certain number of metadata qualities of a dataset allows significantly more productive use than datasets with less than this number of metadata qualities. The conclusion should be consistent with the results of the t-test provided above. Thus when datasets with at least 4 metadata qualities allow more productive use than datasets with only 2 metadata qualities, adequate documentation would be at least 4 features of metadata.

We t-tested the responses to question 13 to determine on what one may consider adequate documentation. The tables 5-15, 5-16, 5-17 and 5-18 provide the result of these tests.

Table 5-14: T-test of Determination of Adequate Documentation I

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
4 or more metadata qualities?	Yes	No	Yes	No	Yes	No
Counts	33	160	33	160	33	160
Mean	4.645	4.380	4.485	4.120	4.818	4.647
Variance	0.303	0.949	0.383	0.540	0.153	0.256
Df	191		191		191	
T-value	1.586		3.691		3.699	

Table 5-15: T-test of Determination of Adequate Documentation II

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
3 or more metadata qualities?	Yes	No	Yes	No	Yes	No
Counts	56	137	56	137	56	137
Mean	4.647	4.338	4.436	4.081	4.786	4.632
Variance	0.353	1.016	0.362	0.564	0.171	0.265
Df	191		191		191	
T-value	2.215		4.362		4.022	

The critical T-value for 191 degrees of freedom at a 0.001 level of significance is: 3.291.
The critical T-value for 191 degrees of freedom at a 0.01 level of significance is: 2.576.
The critical T-value for 191 degrees of freedom at a 0.05 level of significance is: 1.960.
The critical T-value for 191 degrees of freedom at a 0.20 level of significance is: 1.282.

Table 5-16: T-test of Determination of Adequate Documentation III

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
2 or 1 metadata qualities?	None	Yes	None	Yes	None	Yes
Counts	55	64	55	64	55	64
Mean	4.314	4.344	4.130	4.032	4.698	4.548
Variance	1.300	0.896	0.756	0.386	0.215	0.285
Df	117		117		117	
T-value	-0.151		0.907		3.197	

Table 5-17: T-test of Determination of Adequate Documentation IV

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
# of metadata qualities	One	None	One	None	One	None
Counts	36	55	36	55	36	55
Mean	4.314	4.314	4.000	4.130	4.583	4.698
Variance	0.869	1.300	0.400	0.756	0.307	0.215
Df	89		89		89	
T-value	0.002		-0.944		-2.098	

The tables presented above suggest that adequate documentation would be datasets with at least three of the features of Question 13 Section 3 documented.

We also tested the impact of one option of Question13 Section 3: the documentation allowed us to assess the relevance of this dataset for our research project.

Table 5-18: Relevance Assessment Through Documentation

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Relevance through documentation?	Yes	No	Yes	No	Yes	No
Counts	79	114	79	114	79	114
Mean	4.493	4.377	4.308	4.097	4.692	4.667
Variance	0.686	0.961	0.398	0.607	0.242	0.242
Df	191		191		191	
T-value	0.922		2.705		0.723	

Here we see that when researchers are able to judge the relevance of the dataset for their research, this leads to significantly more satisfied researchers than when they are unable to check the relevance.

We also tested all the metadata qualities of Question 13 section 3 individually with the chi-square test (see Table 5-19).

Table 5-19: Chi-square Test for Each Individual Metadata Quality

	Yes				*No*				χ^2 *value (df= 2)*
Success measure:	*S*	*NS*	*I*	*T*	*S*	*NS*	*I*	*T*	
Documentation Adequate?	56	43	10	109	6	51	27	84	46.4
Technical suitability assessed through documentation?	57	18	3	78	47	62	9	118	20.9
Quality/ Accuracy assessed through documentation?	51	12	6	69	70	43	14	127	7.1
Timeliness assessed through documentation?	27	15	7	49	43	94	10	147	16.6
Relevance assessed through documentation?	44	28	8	80	18	68	30	116	34.9
Contractual restrictions assessed through documentation?	1	15	0	16	14	164	2	180	0.23
Documentation not available	2	34	19	55	60	62	19	141	30.6

S = Success, NS = No success, I = Impediment, T = Total

The critical chi-square value for 2 degrees of freedom at a 0.05 level of significance is: 3.84
The critical chi-square value for 2 degrees of freedom at 0.01 level of significance is: 9.21
The critical chi-square value for 2 degrees of freedom at a 0.001 level of significance is: 13.82

The results of the chi-square test show that the two groups significantly differ in ability to assess the technical suitability, and ability to assess the relevance of the dataset for the research project at a level of significance of 0.001. For the group of datasets enabling the assessment of the technical suitability contributed 73% (57/78) of the times to a successful use, as 4% (3/78) was mentioned as an impediment. In the group without the possibility to assess the technical suitability of the dataset 40% (47/118) mentioned this as a success and 8% (9/118) an impediment. We conclude that datasets that allow the

assessment of technical suitability allow more successful use of the dataset than datasets which do not have this quality.

For the assessment of the relevance of the dataset for a particular research project, we found that datasets enabling the assessment allowed successful use in 55% (44/80) of the cases and 10% (8/80) found this an impediment. Datasets not providing attributes for the relevance assessment scored respectively 16% (18/116) for success and 26% (30/116) for impediments. Thus, datasets allowing the assessment of relevance through documentation allowed more successful use than datasets lacking this documentation.

At a level of significance of 0.05 the group allowing assessment of the quality of the dataset and the group not allowing assessment of the quality of the dataset are not uniformly distributed. The former scores 74% (51/69) on allowing successful use, and 10% (6/69) on an impediment in the use. The latter scores respectively 55% (70/127) on successful use and 11% (14/127) on impediments. Again, datasets with the metadata quality allow significantly more successful use than the datasets lacking this quality.

5.3.5 Principle 6: "Adherence to Marginal Cost or Less"

> Scientific and technical data collected or maintained by or under authority of a government agency should be made available to all requesters at the marginal cost of dissemination or less.

A measure of adherence to marginal costs was established through an analysis of the Question 11 Section 3. The highest ranking of adherence to marginal cost or less would be one with the following responses: no costs, cost of dissemination, or a statutory fee.

The lowest ranking would be the responses market price, market price less a discount, full or partial cost recovery.

171 of the government datasets qualified for the highest level of adherence as 23 did to the lowest. Respondents for 2 datasets did not fill out Question 11 Section 3. We tested the two groups with the t-test and the chi-square test. The results of the t-test are presented in Table 5-20.

Table 5-20: T-test for Adherence to Marginal Cost or Less

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Marginal costs or less?	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*
Counts	171	23	171	23	171	23
Mean	4.410	4.714	4.212	4.045	4.690	4.591
Variance	0.881	0.214	0.523	0.522	0.239	0.253
Df	192		192		192	
T-value	-1.647		1.432		1.863	

The critical T-value for 192 degrees of freedom at a 0.05 level of significance is: 1.960.
The critical T-value for 192 degrees of freedom at a 0.10 level of significance is: 1.645.
The critical T-value for 192 degrees of freedom at a 0.20 level of significance is: 1.282.

We see conflicting results for the different measures of productivity. Respondents who acquired datasets at costs were able to perform significantly more tasks with the dataset than respondents who accessed their datasets for marginal costs or less (at a level of significance of 0.10).

However, respondents using "inexpensive" datasets were significantly more satisfied (at a 0.20 level of significance) and accomplished significantly more overall objectives (at a level of significance of 0.10). Maybe these respondents could use the funds initially meant for the acquisition of datasets for other elements important for the research project.

We also performed a chi square test. As measures of success we used the following answers to Questions 17 and 18 of Section 3: cost of the dataset.

Table 5-21: Chi-square Test for Adherence to Marginal Cost or Less

Marginal costs or less?	*No*	*Yes*	*Total*
Success	12	77	89
No success	9	93	102
Impediment	2	1	3
Total	23	171	194
Chi square value df=2	9.75		

The critical value at two degrees of freedom is: 9.210 at the 0.01 level of significance. The chi square value exceeds the critical value at a 0.01 level of significance: our two groups are significantly not uniform. The group with datasets available for "free" scores 45% (77/171) on the successful use and 0% for the impediment score (1/171). The group with the datasets available at cost score 52% (12/23) for successful use and 9% (2/23) for impediments. Thus, not one group is preferred over the other or allows more successful use of datasets. The measure of success in this test focused on successful use of the dataset. The issue of money may not influence the use of the dataset since one first acquires and then uses the data.

5.3.6 Principle 7: "Adherence to Non-exclusivity Availability"

> Scientific and technical data collected or maintained by or under authority of a government agency should be made available for exploitation by both not-for-profit and commercial entities alike on a non-exclusive basis.

A measure of availability was established through an analysis of the Questions 5 and 6 of Section 3. A highest ranking would have had the following answers: Question 5: No, Question 6: No. 80 datasets were ranked as highest.

A middle ranking would have had an yes on one of the two question (27 datasets) and a lowest ranking would have been a confirming answer to both questions (64 datasets). 12 respondents were unable to indicate whether they had to identify themselves or not and did not know whether they had to explain their intended use. One respondent did not fill out both questions. One respondent filled out "No" for the intended use question and did not fill out the other question. Ten respondents answered "No" to one of the two questions and do not know to the other. Finally one respondent did not fill out one question and did not know the answer to the other (Check: 80 + 27 + 64 + 12 +10 + 1 + 1 + 1 = 196).

The results of the different tests are provided in the tables Table 5-22, Table 5-23, and Table 5-24. We see in the tables that datasets with the highest ranking are not necessarily more productive than datasets with the lowest ranking. An explanation may be that the way respondents acquire their datasets does not impede the way they perform the research. Respondents who were required to identify themselves and to explain their intended use with the dataset accomplished significantly more overall objectives than datasets adhering to the highest (!) level of this principle (see Table 5-22, and Table 5-23). This suggests that the dataset provider is more likely to help an academic researcher than strangers working on an unknown project.

Table 5-22: T-test for Adherence to Non-exclusivity Availability I

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
	No ID and no intended use	*ID and intended use*	*No ID and no intended use*	*ID and intended use*	*No ID and no intended use*	*ID and intended use*
Counts	80	64	80	64	80	64
Mean	4.480	4.267	4.241	4.156	4.628	4.750
Variance	0.739	1.216	0.467	0.610	0.263	0.222
Df	142		142		142	
T-value	1.298		0.939		-2.958	

The critical value at 0.01 level of significance at 142 degrees of freedom is: 2.576
The critical value at 0.20 level of significance at 142 degrees of freedom is: 1.282

Table 5-23: T-test for Adherence to Non-exclusivity Availability II

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
	No ID and no intended use	*ID or intended use*	*No ID and no intended use*	*ID or intended use*	*No ID and no intended use*	*ID or intended use*
Counts	80	27	80	27	80	27
Mean	4.480	4.593	4.241	4.185	4.628	4.741
Variance	0.739	0.328	0.467	0.541	0.263	0.199
Df	105		105		105	
T-value	-0.764		0.511		-2.035	

The critical value at 0.05 level of significance at 89 degrees of freedom is: 1.985

Table 5-24: T-test for Adherence to Non-exclusivity Availability III

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
ID AND Intended use v. ID OR Intended use?	*AND*	*OR*	*AND*	*OR*	*AND*	*OR*
Counts	64	27	64	27	64	27
Mean	4.267	4.593	4.156	4.185	4.750	4.741
Variance	1.216	0.328	0.610	0.541	0.222	0.199
Df	89		89		89	
T-value	-1.368		-0.213		0.187	

The critical value at 0.20 level of significance at 89 degrees of freedom is: 1.293

We may conclude that datasets ignoring the principle allow a higher productivity than datasets adhering to the proposed principle.

For the chi-square test we used the same criteria for the highest and lowest ranking as for the t-test. The measures of success for the chi-square test are the personal or institutional willingness (Question 17) or resistance (Question 18).

Table 5-25: Chi-square Test Identification before Access

Productivity Measure	*Highest level*	*Lowest level*	*Total*
Success personal willingness	27	34	61
No success	51	24	75
Impediment personal resistance	2	6	8
Total	80	64	144
Chi square value df=2	10.88		144

The critical value at two degrees of freedom is: 9.210 at the 0.01 level of significance.

The distribution of the two groups is at a level of 0.01 of significance significantly not uniform. Respondents who acquired datasets without the need to identify themselves, mentioned personal or institutional willingness to giving access to the dataset 34% (27/80) of the times, as respondents who acquired datasets with identification did 53 % (34/64) of the times. Respondents of the highest level group mentioned in 3% (2/80) of the responses that personal or institutional resistance to giving access to the dataset was considered an impediment as 9% (6/64) did for the lowest level group.

Thus we may conclude that the lowest level group is more successful in the use than the highest level group. On the contrary we may not conclude this since the percentage impediments of the more successful group is higher than the percentage impediments of the higher level group.

5.3.7 Principle 8: "Adherence to No Exclusive Partner Arrangements"

> Government agencies should not hold copyrights in scientific and technical data collected or maintained by or under their authority and federal agencies should not establish or maintain exclusive arrangements for access to scientific and technical data.

A measure of availability was established through an analysis of Question 16 Section 3. A dataset with the highest ranking would have had the following answer: Question 16: any of the answers but no, indicating that the same dataset could have been accessed elsewhere. A dataset with the lowest ranking would have had a no answer to Question 16. One respondent did not know the answer to Question 1 and was excluded from the analysis. We tested for significance by using the t-test. Table 5-26 shows the results of this test. The test does not provide any evidence for a preference for either one of the two groups.

Further, we used the responses to Question 1 to test the principle in more depth. It enabled us to create two groups: one with datasets acquired from the creator and one with datasets acquired from a for-profit or not-for-profit intermediate. Table 5-28, Table 5-29, and Table 5-30 show the results of the analysis. Surprising is the appreciation of the intermediate entities here. Respondents are significantly more satisfied when a dataset can be obtained from multiple sources than from intermediaries alone (see Table 5-29) and also significantly more satisfied when a dataset can be obtained from the public creator of the dataset alone (see Table 5-30) than with datasets that can only be obtained from intermediate entities. The tests with the datasets available through multiple sources

and datasets accessible only through the public creator did not provide this information; no significant differences were found.

Table 5-26: T-test for Adherence to No exclusive Partnership Arrangements I

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Access possible through multiple sources?	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Counts	104	91	104	91	104	91
Mean	4.398	4.482	4.173	4.213	4.667	4.697
Variance	0.819	0.872	0.513	0.556	0.244	0.236
Df	193		193		193	
T-value	-0.696		-0.527		-0.867	

Table 5-27: T-test for Adherence to No exclusive Partnership Arrangements II

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Commercial or not for profit intermediate?	*Commercial*	*Not for profit*	*Commercial*	*Not for profit*	*Commercial*	*Not for profit*
Counts	9	65	9	65	9	65
Mean	4.667	4.429	4.111	4.063	4.444	4.710
Variance	0.500	0.829	0.611	0.472	0.278	0.242
Df	72		72		72	
T-value	0.837		0.279		-3.026	

The critical t-value at 0.01 level of significance at 72 degrees of freedom is: 2.651

Table 5-28: T-test for Adherence to No exclusive Partnership Arrangements III

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Access possible through multiple sources?	*Yes*	*No, only through public creator*	*Yes*	*No, only through public creator*	*Yes*	*No, only through public creator*
Counts	91	77	91	77	91	77
Mean	4.482	4.451	4.213	4.260	4.697	4.658
Variance	0.872	0.737	0.556	0.511	0.236	0.255
Df	166		166		166	
T-value	0.251		-0.558		1.021	

Table 5-29: T-test for Adherence to No exclusive Partnership Arrangements IV

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Access possible through multiple sources?	*Yes*	*No, only through intermediate*	*Yes*	*No, only through intermediate*	*Yes*	*No, only through intermediate*
Counts	91	27	91	27	91	27
Mean	4.482	4.259	4.213	3.926	4.697	4.692
Variance	0.872	1.046	0.556	0.456	0.236	0.222
Df	116		116		116	
T-value	1.114		2.451		0.085	

The critical value at 0.02 level of significance at 116 degrees of freedom is: 2.358

Table 5-30: T-test for Adherence to No exclusive Partnership Arrangements V

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Creator alone v. intermediate alone	*Creator*	*Intermediate*	*Creator*	*Intermediate*	*Creator*	*Intermediate*
Counts	77	27	77	27	77	27
Mean	4.451	4.259	4.260	3.926	4.658	4.692
Variance	0.737	1.046	0.511	0.456	0.255	0.222
Df	102		102		102	
T-value	1.036		3.002		-0.624	

The critical value at 0.01 level of significance at 102 degrees of freedom is: 2.638

No chi square test was performed since this principle does not deal with the use of the dataset.

5.3.8 Principle 9: "Adherence to No Restrictions on Subsequent Uses"

Government agencies should ensure that electronic data, information, and value-added features developed with public funds are available without restrictions on subsequent uses of the materials.

A measure of availability was established through an analysis of Question 10 Section 3. Respondents with a dataset with the highest ranking would have had answered no to Question 10. Respondents with a dataset with the lowest ranking would have had the following responses: any of the answers but no.

We tested this with the t-test. Table 5-31 shows the results.

Table 5-31: T-test Adherence to No restrictions on Subsequent Uses I

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Any restrictions?	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Counts	139	57	139	57	139	57
Mean	4.492	4.304	4.219	4.123	4.669	4.696
Variance	0.724	1.088	0.511	0.574	0.238	0.252
Df	194		194		194	
T-value	1.418		1.155		-0.718	

The critical value at a level of significance of 0.20 is: 1.282 (df=194)

We also did a more in depth analysis of the productivity of adherence to this principle. We use the same Question 16 for this test. A highest ranking would have had any of the answers but "provisions stated that we could not pass on the provided digital data to any other parties" or "provisions stated that any value-added products that we developed through use of the data (1) required explicit permission of the data supplier prior to dissemination of the value-added products by us, (2) vested an ownership interest in the

original data supplier, or (3) required a royalty due to the data supplier" or "provisions stated that our use could be for only academic research purposes".

Datasets with a lowest ranking would have had a mark on at least one of the following answers: "provisions stated that we could not pass on the provided digital data to any other parties" or "provisions stated that any value-added products that we developed through use of the data (1) required explicit permission of the data supplier prior to dissemination of the value-added products by us, (2) vested an ownership interest in the original data supplier, or (3) required a royalty due to the data supplier" or "provisions stated that our use could be for only academic research purposes". Table 5-32 provides the results of the t-test.

Table 5-32: T-test Adherence to No restrictions on Subsequent Uses II

Productivity Measure	*Task Accomplishment*		*Satisfaction*		*Overall Objective Accomplishment*	
Use restricted by value/ pass on or academic use only?	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Counts	148	48	148	48	148	48
Mean	4.489	4.277	4.219	4.104	4.683	4.660
Variance	0.708	1.204	0.490	0.648	0.232	0.273
Df	194		194		194	
T-value	1.496		1.300		0.575	

The critical value at a level of significance of 0.20 is: 1282 (df=194)

The data in the table shows that datasets with no restrictions on subsequent use allow more productive research than datasets with a restriction on subsequent uses.

We also tested the principle with the chi-square test. Question 17 and 18 were used to determine about the success or impediments of legal restrictions or lack of restrictions.

Table 5-33: Chi-square Test for Adherence to No Restrictions on Subsequent Use I

	Highest level (total of 139 datasets)			*Lowest level (total of 57 datasets)*		
Success measure:	*S*	*NS*	*I*	*S*	*NS*	*I*
copyright law	49	90	0	7	49	1
specific data protection legislation (e.g. local ordinance, state statute, federal statute)	34	105	0	4	50	3
Contractual restrictions facilitating our uses of this dataset	6	133	0	8	47	2
Contractual restrictions regarding further dissemination of this dataset	1	138	0	2	55	2
Contractual provisions regarding liability	1	138	0	0	57	0
Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset	0	139	0	0	57	0
Chi square value (df=17)	60.4					

S = Success, NS = No success, I = Impediment

The critical chi-square value at 17 degrees of freedom at a level of significance of 0.001 is: 40.79.

The distribution within the two groups is significantly not uniform (at the level of 0.001 of significance). The group ranked as the highest had 65% (91/139) of the times successful use mentioned and 0% (0/139) impediment in the use. The other group had 37% (21/57) successful in the use mentioned and 14% (8/57) impediments. Thus, we conclude that datasets adhering most to the proposed principle allow more successful use than datasets ignoring the principle.

We also tested the principle in more depth with the chi-square test. Question 17 and 18 were used to determine about the success or impediments of legal restrictions or lack of restrictions.

Table 5-34: Chi-square Test for Adherence to No Restrictions on Subsequent Use II

	Highest level (total of 148 datasets)			*Lowest level (total of 48 datasets)*		
Success measure:	*S*	*NS*	*I*	*S*	*NS*	*I*
Contractual restrictions imposed on our uses of this dataset	9	139	0	8	38	2
Contractual restrictions regarding further dissemination of this dataset	2	145	1	2	44	2
Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset	0	148	0	0	48	0
Chi square value df=8	15.4					

S = Success, NS = No success, I = Impediment

The chi square value exceeds the critical value at a level of significance of 0.10 (critical chi-square value is 13.36). At this level the distributions of the two groups are significantly not uniform. When we compare the scores on successful use and impediments, we see that contractual provisions in 7% (11/148) of the datasets in the highest level group allow successful use as 21% (10/48) of the datasets in the lowest level group. For the score on impediments in the use the highest level group scores 1% (1/148) and the lowest level group 8% (4/48). The data provides contradicting data: we cannot prefer either one of the two groups over the other.

5.3.9 Principle 10: "Adherence to Access Through Publicly Accessible Archive"

> Scientific and technical data collected or maintained by or under authority of a federal, state or local government agency that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means

It is possible to test this hypothesis with the type of questions asked in this survey. However, none of the respondents acquired or accessed data through a monetary fee or special permission in the library.

Due to a lack of data, we were unable to test principle 10.

5.4 Principles for Data Provided by the Academic Community

In order to decide whether principles for data provided by the Academic Community was adhered to in a specific instance, may be established by first determining whether a respondent used datasets produced or provided from not-for-profit organization or foundation (Question 7 section 2). Then we determined whether the respondent filled out the questionnaire for such a dataset (First Question, Section 3). The responses of Question 7 Section 2 only indicates whether a 'not-for-profit' dataset was used and does not necessarily correspond with the dataset for which the questionnaire was filled out. Thus, we only used the responses of the First Question of Section 3 to decide on not-for-profit datasets.

Of the datasets reported 30 datasets were identified as datasets originating from not-for-profit organizations or foundations (as reported in chapter 4). We tested these datasets in a group adhering to, against a group of datasets ignoring the recommended principles for Academic Community datasets, as set forth in chapter 2. Where the data allowed us, we created more than two levels of adherence to the proposed principles.

5.4.1 Principle 1: "Level of Full and Open Exchange of Data"

> The not-for-profit scientific and technical community should continue to promote and adhere to the policy of full and open exchange of data at both the national and international levels

A measure of full open and exchange was established through an analysis of the Questions 3, 4, 5, 6, 9, 10, 11, 12 and 14 in Section 3. The highest ranking would have had the following responses to these questions: Question 3: any answer except acquired on paper or self-collected, Question 4: no Question 5: no, Question 6: no, Question 9: no licensing or we assumed no contract or licensing provisions, Question 10: not applicable, Question 11: not applicable, cost of dissemination, or minimal statutory fee, Question 12: good or excellent, Question 14: immediate or reasonable. Responses for 1 dataset adhered to this highest ranking.

The lowest ranking would have had the following responses to these questions:

Question 3: any answer except acquired on paper or self-collected, Question 4: yes

Question 5: yes, Question 6: yes, Question 9: any answer but "no licensing" or "we assumed no contract or licensing provisions", Question 10: any answer but "not

applicable", Question 11: market price, market price less a discount, price based on full or partly cost recovery, Question 12: non-existent, poor or fair, Question 14: unreasonable. None of the datasets adhered to all the qualifications of the lowest ranking. Instead we used another low level of availability: Question 3: any answer, Question 4: yes, Question 5: yes, Question 6: yes, Question 9: any answer but "no licensing" or "we assumed no contract or licensing provisions", Question 10: any answer but "not applicable", Question 11: any answer, Question 12: any answer, Question 14: any answer. 2 datasets were categorized in this low level group.

Due to a lack of datasets adhering to and ignoring the proposed principle we were unable to test the principle statistically with the most comprehensive definition of full and open exchange.

When we deteriorate the meaning of full and open to its most important issues: no restrictions what so ever and no licensing approach, only the marginal costs of the dataset involved, and after reasonable time or immediate access to the dataset then we come to 18 datasets adhering to this high level of adherence.

The lowest level would then be: any licensing approach but no licensing or the assumption of no licensing, at least one restriction involved, cost of the dataset higher than the marginal costs, and able to access the dataset after an unreasonable time after the request was made. No dataset adhered to the adjusted lowest level.

We were unable to test the principle. However, the majority (60%) of the datasets already adheres to the proposed principle.

The numeric results of both the t-test and the chi-square test in assessing the relation of success with conformance to principle 1 are included in Table 5-41. Similarly this table

shows the results of the further tests of principles discussed throughout the remainder of this section.

5.4.2 Principle 2: "Level of Accessibleness"

> Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions that have been collected for projects entirely or primarily financed with public funds should be treated by the creators from a science policy perspective as being in the public domain, after a reasonable time period to allow for publication of the results of the research.

First we determined what datasets were created with the help of public Research and Development funds (a yes on Question 8 Section 3).

A highest level of adherence would have had the following answer to Question 9 Section 3 "no licensing or purchase contract were involved" or "we assumed that no contract or licensing provisions applied to our use of the data", the answer to Question 10 Section 3 was "not applicable", the response to Question 11 Section 3 was "not applicable", “price based on the cost of dissemination” or “the price based on a minimal statutory fee”, and the answer to Question 14 Section 3: “access was immediate” or “the time between a request and obtaining the data was reasonable”. 9 datasets adhered to these characteristics.

The lowest level of accessibleness would have had any of the answers not mentioned for the highest level. Only 1 dataset qualified for this lowest level. Because of this we were

unable to test the principle. This might however be an indication that the principle is agreed upon within the not-for-profit community.

5.4.3 Principle 3: "Level of Dissemination Datasets for Concurrent Publishing"

> When publishing research articles, scientists should concurrently publish or otherwise make available (electronically) the datasets upon which their research depends or from which it is derived.

This principle is not directly addressed by any of the questions in the questionnaire.

The questions assessing whether a dataset is in the public domain or not (Questions 9, and 10) address the principle indirectly.

In this respect, a dataset would qualify for the highest level of adherence when the following responses were filled out: Question 9 Section 3 no licensing or purchase contract were involved or we assumed that no contract or licensing provisions applied to our use of the data, and Question 10 Section 3 not applicable. 19 datasets were ranked as highest level datasets.

A lowest level of electronic availability would have had the following answers: Question 9 Section 3: any answer but the answers of the highest level, Question 10 Section 3: provisions stated that we could not pass on the provided data to any other parties, or provisions stated that our use could be for only academic or research purposes, or provisions stated that any value added products that we developed required permission prior to dissemination. 3 datasets qualified for this lowest level group.

Table 5-35: T-test Level of Availability

Principle 3	*Task Accomplishment*		*Satisfaction*		*Overall Accomplishment*	
	Adhering	Ignoring	Adhering	Ignoring	Adhering	Ignoring
Total # datasets	19	3	19	3	19	3
Mean	4.471	2.667	4.158	3.000	4.667	3.667
Variance	0.890	4.333	0.585	1.000	0.235	0.333
Df	20		20		20	
t value	1.804		2.919		6.521	

The critical t-value for 20 degrees of freedom at a level of significance of 0.05 is 1.725
The critical t-value for 20 degrees of freedom at a level of significance of 0.01 is 2.845
The critical t-value for 20 degrees of freedom at a level of significance of 0.001 is 3.850

Table 5-35 shows that consistently at different levels of significance datasets adhering to the proposed principle are more productive to their academic user than datasets ignoring the principle.

Table 5-36: Chi-square Test Level of Availability

	Highest level (total of 19 datasets)			*Lowest level (total of 3 datasets)*		
Success measure:	*S*	*NS*	*I*	*S*	*NS*	*I*
Timeliness	10	7	2	2	0	1
Copyright	3	16	0	0	3	0
Specific data protection legislation	1	18	0	0	3	0
Contractual restrictions imposed on our uses of this dataset	1	18	0	0	2	1
Contractual restrictions regarding further dissemination of this dataset	1	18	0	0	2	1
Contractual provisions regarding liability	1	18	0	0	3	0
Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset	1	18	0	0	3	0
Chi square value df=6	12.3					

S = Success, NS = No success, I = Impediment

Critical value at a level of significance of 0.10 is: 10.64

At a level of significance of 0.10 our chi-square value exceeds the critical value. This provides some evidence that adherence to the principle allows more successful use than ignoring it.

5.4.4 Principle 4: "Adherence to (at least) Reasonable Time for Proprietary Use Before Dissemination of Dataset "

> Public agency grant conditions and university policies should establish that all scientists conducting publicly funded research should make their data available immediately, or following a reasonable period of time for proprietary use. The maximum length of any proprietary period should be expressly established by the particular scientific communities, and compliance should be monitored subsequently by the public funding agency.

The problem with this principle, in the light of the questionnaire, is that we do not know whether the owner of the dataset disseminates the latest data or disseminates it directly after creation. We only test whether the time between request and access is reasonable. Furthermore, reasonable is according to the researcher's own impression. This is not an absolute term. Further we found it impossible to test whether compliance should be monitored by the public funding agency.

However, the following principle is tested:

> "R&D funded research should be disseminated immediately or following a reasonable period of time for proprietary use after a request for this data is made".

First we selected the datasets of which a substantial portion originally was developed using exclusively or primarily research and development funds (a yes to Question 8 Section 3). 17 datasets qualified. Then we used the answer to Question 14 Section 3 to divide the datasets into two groups. The highest level group consists of datasets that could be accessed immediately or after a reasonable period of time. In the lowest level

group datasets could be accessed after an unreasonable period of time according to respondents.

16 datasets qualified for the highest level group as only 1 qualified for the lowest level. Again we are unable to test the principle. Since almost all datasets adhered to the principle we may characterize the not-for-profit community as a community where a 'sharing datasets' spirit rules.

5.4.5 Principle 6: "Adherence to Adequate Metadata"

> Scientific and technical datasets made available in a publicly accessible archive should be documented adequately with metadata.

This principle is related to principle 5 of federal government data. There we concluded that datasets adhering to this principle allow researchers to be more productive than datasets that are not. For datasets acquired from academic institutions we used the same qualifications for datasets coming from government agencies. These were as follows:

A measure of availability was established through an analysis of the Question 12 of Section 3. A highest ranking would have had the following answers: Question 12: good or excellent documentation. 14 datasets qualified for the highest level of adherence.

A lowest ranking would have had the following answers: Question 12: fair, poor or non-existent documentation. 16 datasets qualified for this lowest level of adherence.

Table 5-37: T-test Adherence to Adequate Metadata

	Task Accomplishment		*Satisfaction*		*Overall Objective Accomplishment*	
Adequate metadata?	Yes	No	Yes	No	Yes	No
Total	14	16	14	16	14	16
Mean	4.500	4.000	4.429	3.813	4.615	4.375
Variance	0.885	1.692	0.571	0.696	0.256	0.383
Df	28		28		28	
t value	0.992		2.626		1.987	

The critical t-value for 28 degrees of freedom at a level of significance of 0.01 is 2.763
The critical t-value for 28 degrees of freedom at a level of significance of 0.05 is 2.048
The critical t-value for 28 degrees of freedom at a level of significance of 0.10 is 1.701

Table 5-38: Chi-square Test Adherence to Adequate Metadata

Adequate metadata?	*Yes*	*No*	*Total*
Success Adequate Documentation	9	1	10
No success	5	7	12
Impediment Adequate Documentation	0	8	8
Total	14	16	30
X2 value df=2	14.67		

The critical chi square value for 2 degrees of freedom at a level of significance of 0.001 is 13.82

The two groups are significantly not uniform. 64% (9/14) of the datasets adhering to the principle allow successful use of the dataset as only 6% of the datasets in the group ignoring the principle do. None of the datasets in the former group do not allow successful use as 50% (8/16) of the datasets in the latter group do.

Again the data in the tables show the importance of documentation for the productivity of academic research. Requiring funding agencies to fund metadata creation and appropriate archiving of research datasets in public depositories or libraries as standard conditions of grants is one way to ensure the quality of the documentation of research data.

5.4.6 Principle 8: " Adherence to Access Through Publicly Accessible Archive"

> Scientific and technical data collected or maintained by or under authority of an academic institution that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means.

The principle is identical to principle 10 of government data. We are not able to test this principle due to a lack of useful data. None of the respondents acquired or accessed data in a library through a monetary fee or special permission.

5.4.7 Principle 9 "Adherence to Marginal Costs or Less"

> Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions should be made available to all requesters at the marginal cost of dissemination or less.

A measure of adherence to marginal costs was established through an analysis of Question 11 Section 3. The highest ranking of adherence to marginal cost or less would be one with the following responses: no costs, cost of dissemination, or a statutory fee. The lowest ranking would be the responses market price, market price less a discount, full or partial cost recovery.

29 datasets qualified for the highest level of adherence as only 1 did to the lowest. We were unable to test the principle. Again there is an indication that a significant part of the not-for-profit community adheres to principles promoting open and full exchange of data.

5.5 Overview of Results of Proposed Principles Tested

Above we provided the results of statistical tests. In this section we provide an overview of the conclusions of the tests.

In order to make the statistical conclusions understandable we attached a letter (mark) to the numeric results of the tests. Based on the marks of all four measures of success (task accomplishment, satisfaction, overall accomplishment, and successful use) we assess the relation of success with conformance to the proposed principles.

The following "relation of success with conformance" measure is presented in Table 5-39. Where necessary we iterated a mark based on the measures in this table.

Further we used the following wording in the analysis:

Inconclusive: the statistical tests indicated that no significant relation with conformance exist

Non testable: the principle was not sufficiently addressed by the questions in the questionnaire

Lack of data: responses were not sufficient to test this principle

Either way: 'relations with conformance to the principle' tests provide conflicting results

Negative (NEG): datasets ignoring the principle appear to contribute significantly more to success than datasets adhering to the principle.

Table 5-39: Overview of Relation of Success and Conformance to Principles

All of the four measures significant	Level of Significance	Relation	Mark
	All 0.001	Evident	A
	All 0.01	Very Strong	A-/ B+
	All 0.05	Strong	B
	All 0.10	Average to Strong	C+/B-
	All 0.20	Average	C

Three of the four measures significant	Level of Significance	Relation	Mark
	All three 0.001	Very Strong	A-/ B+
	All three 0.01	Strong	B
	All three 0.05	Average to Strong	C+/B-
	All three 0.10	Average	C
	All three 0.20	Weak-Average	C-/D+

Only two of the four measures significant	Level of Significance	Relation	Mark
	Both 0.001	Strong	B
	Both 0.01	Average to Strong	C+/B-
	Both 0.05	Average	C
	Both 0.10	Weak-Average	C-/D+
	Both 0.20	Weak	D

Only one of the four measures significant	Level of Significance	Relation	Mark
	0.001	Average to Strong	C+/B-
	0.01	Average	C
	0.05	Weak-Average	C-/D+
	0.10	Weak	D
	0.20	Very Weak	E

Table 5-40 Relation Between Success and Conformance to Proposed Principles Government

Principle/ Test	Task accomplish-ment Test	Satisfaction Test	Overall Objective Accomplish-ment Test	Chi square Test	Overall Conclusion
1 Level of Availability	(0.05)	(0.01)	In	(0.05)	C
2 Level of Affirmativeness in Dissemination	In	In	In	In	In
3 Level of Activity in Release	In	I: (0.05) II: (0.01)	In	I: In II: (0.10)	C
4 Level of Accessibleness by Archive	NT	NT	NT	NT	NT
5 Level of Metadata Availability	(0.001)	(0.001)	(0.001)	(0.001)	A
6 Adherence to Marginal Cost or Less	NEG(0.10)	(0.20)	(0.10)	In	Either way
7 Adherence to Non-Exclusivity Availability	I: (0.20) II: In	In	I: NEG(0.01) II:NEG(0.05)	In	Either way
8 Adherence to No Exclusive Partner Arrangements	In	(0.02)	In	NT	C
9 Adherence to No Restrictions on Subsequent Uses	I: (0.20) II: (0.20)	I: In II: (0.20)	In	I: (0.001) II: In	C
10 Adherence to Access Through Publicly Accessible Archive	L	L	L	L	NT

(0.XX) = level of significance where datasets adhering to the principles are more successful in use than datasets ignoring the principle, In = Inconclusive, NT = Not-tested, L = Lack of data ignoring the principle

Table 5-41 Relation Between Success and Conformance to Proposed Principles Academia

Principle/ Test	Task accomplish-ment Test	Satisfaction Test	Overall Objective Accomplish-ment Test	Chi square Test	Overall Conclusion
1 Level of Full and Open Exchange of Data	No data available ignoring the principle				NT
2 Level of Accessibleness	No data available ignoring the principle				NT
3 Level of Availability	(0.05)	(0.01)	(0.001)	(0.001)	- 3 datasets v 19
4 Adherence to Reasonable Time for Proprietary Use Before Dissemination of Dataset	L	L	L	L	NT
5 Level of Accessibleness by Archive	NT	NT	NT	NT	NT
6 Adherence to Adequate Metadata	In	(0.05)	(0.10)	(0.001)	B
7 Adherence to No Transfer Exclusive Rights in the Work	NT	NT	NT	NT	NT
8 Adherence to Access Through Publicly Accessible Archive	L	L	L	L	NT
9 Adherence to Marginal Cost or Less	L	L	L	L	NT

(0.XX) = level of significance where datasets adhering to the principles are more successful in use than datasets ignoring the principle, In = Inconclusive, NT = Not-tested, L = Lack of data ignoring the principle

Chapter 6 Conclusions & Future Work

6.1 Introduction

This research explored current access policies imposed on researchers in U.S. universities that affect geographic scientific and technical data. Because a broad spectrum of disciplines use geographic data in scientific research, we suspect that the data provided by our sample may be indicative of the responses across many research domains due to the cross disciplinary nature of our sample. Although addressed only in part and for a small subset of scientists, the central question guiding this research has been as follows:

Based on theory and evidenced through empirical testing, which specific access principles appear to best enable scientists that use geographic data to achieve success in advancing knowledge and in meeting their research objectives?

We split the responses to the questionnaire into data obtained from U.S. government sources, data obtained from university sources, and data from private entities. We proposed access principles we thought to be most productive and successful for accomplishing academic research. These principles were drawn from the literature and several questions were drafted relative to each principle in order to determine whether the principle was or was not being followed relative to an academic practitioners use of specific datasets. Adherence or non-adherence to principles were compared with success or lack of success in using the dataset. Through this process and through uses of many datasets across many academic users, the principles were tested statistically. Due to a variety of reasons we were not able to test the

hypothesized principles for academic use of data provided by the private sector. For the other categories, academic use of data provided by government and other academics, we present in this section our conclusions.

6.2 Data Collected by the US Government

For datasets produced by or under authority of a federal, state or local agency we found an evident relation of success with conformance to principle 5: Scientific and technical data collected or maintained by or under authority of a government agency should be documented adequately with metadata. The hypothesis was that datasets acquired from government in adherence to this principle would result in greater success in the use of the data by the academic community. The statistical results evidence support of the truth of the proposition.

For datasets produced by or under authority of a federal, state or local agency we found a positive relation between success and conformance to the following principles:

Principle 1: Government agencies should ensure that electronic data, information and value-added features developed with public funds are available to the public.

Principle 3 (level 2): Government agencies should anticipate requests by the general public (including the scientific community) for electronic information and should build features into their electronic information systems so that information most likely to be requested by the public may be actively released (such as publishing datasets on web servers or CDs along with appropriate retrieval software).

Principle 8: Government agencies should not hold copyrights in scientific and technical data collected or maintained by or under their authority (see also Perritt

1999A, 499, 17 USC 105) and federal agencies should not establish or maintain exclusive arrangements for access to scientific and technical data.

Principle 9: Government agencies should ensure that electronic data, information and value-added features developed with public funds are available without restrictions on subsequent uses of the materials.

While conformance with these principles correlated with success in the academic use of datasets, the correlations were not strong statistically. Therefore further study, probably through alternative and complementary research methods, would be advisable in order to further evidence the truth of the propositions.

For principle 2 (Government agencies should adopt affirmative programs of electronic public information dissemination so that scientists do not need to resort to Freedom of Information requests in order to gain access to government records) we did not find a relation of success with conformance to the principle. In fact virtually no scientists used FOIA requests to gain access to the data they use and therefore the principle could not be adequately tested. For principle 10 (Scientific and technical data collected or maintained by or under authority of a state or local government agency that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means) we only had datasets adhering to the proposed principle. Thus, again it is difficult to test a principle when an insufficient sample or no sample exists for comparative work.

Finally, we found conflicting results in the tests of principle 6, “Scientific and technical data collected or maintained by or under authority of a government agency should be made available to all requesters at the marginal cost of dissemination or less", and 7, “Scientific and technical data collected or maintained by or under

authority of a government agency should be made available for exploitation by both not-for-profit and commercial entities alike on a non-exclusive basis". For some measures of success datasets adhering to the principle scored significantly better and for other measures of success those datasets not adhering to the principle scored better. Accounting for these mixed indications, datasets produced by or under authority of a federal, state or local agency evidenced a very weak relation of success with conformance to principle 6. For principle 7, and its mixed indicators, we found a weak negative relation of success with conformance to the proposed principle.

The inconsistencies and minimal statistical significance in arriving at both of these conclusions make them highly questionable. Further research, probably through alternative research methods, is needed to explore the propositions further.

6.3 Data Collected by the Academic Community

For datasets used by academic users that were acquired from another academic source we found a strong relation of success with conformance to principles 3 and 6; principle 3: When publishing research articles, scientists should concurrently publish or otherwise make available electronically the datasets upon which their research depends or from which it is derived, principle 6: Scientific and technical datasets made available in a publicly accessible archive should be documented adequately with metadata. Thus, note that adequate metadata score as a factor in the successful use of data by academic researchers for both data acquired by government and for data acquired from other academics.

For principles 1, 2, and 9 we did not obtain sufficient datasets lacking adherence to the proposed principles. Therefore no comparisons could be made between the results for those adhering and those not adhering to the principle. For convenience, these

principles are restated as follows: Principle 1: The not-for-profit scientific and technical community should continue to promote and adhere to the policy of full and open exchange of data at both the national and international levels. Principle 2: Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions that have been collected for projects entirely or primarily financed with public funds should be treated by the creators from a science policy perspective as being in the public domain, after a reasonable time period to allow for publication of the results of the research. Principle 9: Scientific and technical datasets created by university and other not-for-profit researchers or their employing institutions should be made available to all requesters at the marginal cost of dissemination or less.

Due to a lack of datasets acquired from other academic sources, we were unable to test principles 4 and 8. Principle 4: Public agency grant conditions and university policies should establish that all scientists conducting publicly funded research should make their data available immediately, or following a reasonable period of time for proprietary use. The maximum length of any proprietary period should be expressly established by the particular scientific communities, and compliance should be monitored subsequently by the public funding agency. Principle 8: Scientific and technical data collected or maintained by or under authority of an academic institution that have been legally placed in a publicly accessible library and all databases accessible through public and university libraries should carry with them the right to read the data or databases by all patrons by any means.

6.4 Recommendations

Although this specific study has arrived at inconclusive results or only weak correlations in regard to several factors, the study suggests that in order to advance the progress of science, government agencies and academic suppliers of geographic data should document their data adequately with metadata. We used as a test for the sufficiency of metadata a positive response that at least three of the following features were addressed in the documentation of the data: (1) technical suitability of the dataset, (2) quality/ accuracy of the dataset, (3) timeliness of the data, (4) relevance of the dataset, (5) contractual restrictions or other legal constraints to the use of the datasets, and (6) allows users to find the dataset through a computer search. While metadata documentation generally had a positive correlation with success of academic use of geographic data, determining the specific utility of metadata and which constituent components are most critical would require further investigation.

This research also evidenced as least minimal statistical support for the following propositions. Government agencies should adhere to open access policies, allowing access through digital media. They should not go into exclusive partnership arrangements that would disallow the widespread availability of government data. Nor should they restrict the subsequent uses of their datasets. Further, similar to the results for geographic data supplied by government, datasets created by academia should be documented adequately and academia should continue to adhere to open access policies in order to best ensure success of use by other academics.

6.5 Future Work

This thesis focuses on the "access to data environment of academia" in the U.S. Academic researchers primarily use geographic data produced and disseminated by

U.S. government agencies, and other academic institutions. Research addressing the same principles regarding access to geographic data in the other parts of the world might provide insights on whether the truth of some propositions might be generalizable to other legal systems and cultures. Such studies should enable us to judge the effectiveness of current data policies in national jurisdictions and ultimately provide insights for advancing scientific research globally.

6.6 Suggestions for Accomplishing Future Work

In this research we used an online questionnaire to obtain empirical evidence of success and non-success of geographic use in academic research environments. In the questionnaire we tried to address a comprehensive list of principles drawn from the literature. The results and our experiences suggest several alternative paths for further research.

At the outset we made the decision to test the entire set of derived principles rather than to test a smaller number of principles. Due to the large number of principles and due to the need to keep the questionnaire a reasonable length, only a small number of questions could be asked about each principle. This limited our depth of understanding in knowing whether an access principle was being fully adhered to or not. Testing a shorter list of principles would have allowed more detailed questions about each principle but the overall scope would have been more limited relative to the substantive issues addressed. Another approach would have involved breaking down a small number of principles into sub-principles and testing each sub-principle with a single question or two. Each approach has its advantages and shortcomings. However, it is likely that both of these alternative approaches using multiple questions for each principle being tested would have been arrived at mixed, conflicting, and

inconclusive results. Thus, while those approaches might provide further insights, the results would likely still be insufficient in testing some of the hypotheses.

Further approaches involve abandoning questionnaire and quantitative approaches in favor of qualitative research approaches. The most in-depth treatment and the one most likely to arrive at the most productive insights would be to utilize a research approach taking advantage of complementary quantitative and qualitative methods.

References

Paper Based References

Barlow, John Perry, (1994), The Economy of Ideas, A framework for patents and copyrights in the digital age. (Everything you know about intellectual property is wrong.) pp.84-90, 126-129,WIRED 2.03 March 1994.

Blumenthal, David e.a., (1997), Withholding Research Results in Academic Life Science, JAMA: The Journal of the American Medical Association, April 16.

Boonin, Leonard G., (1987), The University, Scientific Research, and the Ownership of Knowledge, Chapter 14 in: Owning Scientific and Technical Information; Value and Ethical Issues, Vivian Weil and John W. Snapper (eds.), 1987, Rutgers University Press New Brunswick. ISBN 0-8135-1455-X

Boyle, James, (1997), A Sense of Belonging, Times Literacy Supplement, July 4th. http://www.wcl.american.edu/pub/faculty/boyle/tlscopy.htm.

Branscomb, Anne Wells, (1994), Who Owns Information? From Privacy to Public Access. Basic Books.

Brewer, C.A. and R.B. McMaster, (1999), The State of Academic Cartography, CGIS, Vol. 26 No 3 July, pp.215-234.

Cornes, Richard, and Todd Sandler, (1986), The Theory of Externalities, Public Goods, and club Goods, Cambridge University Press, ISBN: 0-521-30184/0-521-31774 (pbk).

Couclelis, Helen, (1998), Worlds of Information: The Geographic Metaphor in the Visualization of Complex Information, Cartography and Geographic Information Systems, October 1998, Vol. 25, No. 4, pp. 209-220.

Crawford, Walt, and Michael Gorman, (1995), Future Libraries, Dreams, Madness & Reality, American Library Association, ISBN: 0-8389-0647-8

D'Andrea Tyson, Laura and Edward F. Sherry, (1997), Statutory Protection for Databases: Economic & Public Policy Issues, The Information Industry Association's Public Policy & Government Relations Council. IIA's Public Policy & Government Relations Council Online Document Library, Published: in Hearing on H.R. 2652 Before the Subcomm. On Courts and Intellectual Property of the House on the Judiciary, 105th Congress.

David, Paul A., Dominique Foray, (1995), Accessing and Expanding the Science and Technology Knowledge Base, 16 Sci. Tech. & Ind. Rev. (OECD ed., 1995), Vol. 16, No. 3, 1995, pp. 13-68, Organisation for Economic Co-operation and Development.

Dreyfuss, Rochelle Cooper, (1999), Do You Want to Know a Trade Secret? How Article 2B Will Make Licensing Trade Secrets Easier (But Innovation More Difficult), California Law Review Vol. 87 January, No. 1, pp. 193-267.

Dyson, Esther, (1995), Intellectual Value, A radical new way of looking at compensation for owners and creators in the Net based economy, WIRED Magazine 3.07, July.

Epstein, Earl F., (1990), In my opinion, URISA Journal, 3 (1), pp. 2-4.

Fowler, Floyd J. Jr., (1993), Survey Research Methods, Second Edition, Applied Social Research Methods Series, Volume 1, SAGE Publications, ISBN: 0-8039-5048-9.

Fowler, Floyd J. Jr., (1995), Improving Survey Questions, Applied Social Research Methods Series Volume 38, SAGE publications, ISBN: 0-8039-4582-5.

Ginsburg, Jane C., (1997), Copyright, Common Law, and Sui Generis Protection of Databases in the United States and Abroad, U. of Cin. L. Rev. vol 66 pp.151-176.

Ginsburg, Jane C., (1998), Authors as " Licensors" of "Informational Rights" Under U.C.C. Article 2B, Berkeley Technology Law Journal, Vol13., No. 3, pp.945-975.

Goldstein, Paul, (1977), Preempted state doctrines, Involuntary transfers and compulsory licenses: Testing the limits of copyright, UCLA Law Review, 24, 1128.

Goldstein, Paul, (1994), Copyright's Highway: The Law and Lore of Copyright from Gutenberg to the Celestial Jukebox. Hill and Wang, New York.

Guernsey, Lisa, (1998), A Provost Challenges His Faculty to Retain Copyright in Articles, Chronicle of Higher Education, September 18, 1998, http://chronicle.com.

Hugenholtz, P. Bernt, (1998), Electronic Rights and Wrongs in Germany and the Netherlands, Winter, Columbia VLA Journal of Law & The Arts, Volume 22 number 2, pp.151-159.

Huggins, James, (2000), UCITA: Uniform Computer Information Transactions Act, http://www.jameshuggins.com/h/tek1/ucita.htm

Hutcheson, J.C, (1996), "ProCD and the Continuing Shrink-Wrap Saga", paper prepared as a White paper for the Software Manufacturing Association; http://www.forthnoxescrow.com/swrapa31.htm.

ICSU/ CODATA Group on Data and Information, (1997), Position Paper on Data Access, September, 20p.

ICSU/ CODATA Group on Data and Information, (1998), Responses to WIPO Survey on Database Protection, April, 10p.

Kern, David G., (1999), A recent Case Study, presentation colloquium at MIT March 29.

Lederberg, Statement of, (1999), Hearing on the COIAA, 18 March, http://www.house.gov/judiciary/106-lede.htm.

Lemley, Mark A., (1995), Intellectual Property and Shrinkwrap Licenses. Southern California Law Review, Vol. 68 July, pp.1239-1294.

Lemley, Mark A., (1999), Beyond Preemption: The Law and Policy of Intellectual Property Licensing, California Law Review Vol. 87 January 1999, No. 1, pp.111-172.

Lessig, Lawrence, (1998), Sign it and Weep, The Industry Standard, November 20, http://thestandard.com/articles/article_print/0,1454,2583,00.html.

Library of Congress, (1997), U.S. Copyright Office, Report on Legal Protection for Databases. August, Washington D.C.

Litman, Jessica, (1994), Rights in Government-Generated Data, Proceedings of the Conference on Law and Information Policy for Spatial Databases, Tempe AZ, October, NCGIA & Center for the Study of Law, Science and Technology, http://www.spatial.maine.edu/tempe/litman.html.

Litman, Jessica, (1998), The Tales That Article 2B Tells, 13 Berkeley Technology Law Journal pp. 935-949.
http://www.law.wayne.edu/litman/papers/UCCtales.pdf.

Lopez, Xavier, (1996), The Impact of Government Information Policy on the Dissemination of Spatial Data, A Thesis Submitted in the Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (in Spatial Information Science and Engineering), University of Maine, August.

Loundry, D., (1996), 'Shrink-wrap' licenses don't shrink access to data, Chicago Daily Law Bulletin, February 8.

Lousin, Ann, Uniform Computer Information Transaction Act, (1999), The John Marshall Journal of Computer & Information Law, Volume XVIII, Winter 1999, Number 2, pp. 275-278.

Masson, Douglas J., (1997), Fixation on Fixation, Why Imposing Old Copyright Law on New Technology Will Not Work, 6p.

Matsunaga Keene and Jack Dangermond, (1994), Promoting a Free Access or Minimal Cost of Dissemination Arrangement for Government-Held Geographic Information Systems Data, Proceedings of the Conference on Law and Information Policy for Spatial Databases, Tempe AZ, October, NCGIA & Center for the Study of Law, Science and Technology.

McManis, Charles R., (1999), The Privatization (or “Shrink-Wrapping”) of American Copyright Law, California Law Review Vol. 87 January, No. 1, 173-190.

NRC, National Research Council, (1995A), On the Full and Open Exchange of Scientific Data, Committee on Geophysical and Environmental Data, National Research Council Washington D.C.

NRC, National Research Council, (1995B), Preserving the Scientific Data on Our Physical Universe, A new Strategy for Archiving the Nation's scientific information Resources, National Academy Press, National Research Council, Commission on Physical Sciences, Mathematics and Applications, Washington D.C.

NRC, National Research Council, (1997), Committee on Issues in the Transborder Flow of Scientific Data, U.S. National Committee for CODATA, Commission on Physical Sciences, Mathematics, and Applications, National Research Council, Bits of Power: Issues in Global Access to Scientific Data, National Academy Press, Washington, D.C. http://www.nap.edu/readingroom/books/BitsOfPower/index.html.

NRC, National Research Council, (1999A), Commission on Physical Sciences, Mathematics, and Applications, Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest, A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases, National Academy Press, Washington D.C.

NRC, National Research Council, (1999B), Mapping Science Committee, Distributed Geolibraries, Spatial Information Resources, National Academy of Sciences, ISBN: 0-309-06540-2.

NRC, National Research Council, (2000), Commission on Physical Sciences, Mathematics, and Applications, The Digital Dilemma: Intellectual Property in the Information Age, National Academy Press, Washington D.C.

Neal, James G., (1999), Statement of, before the Subcommittee on Courts and Intellectual Property, Committee on Judiciary Hearing on "Collections of Information Anti-piracy Act" of March 18.

Nelkin, Dorothy, (1984), Science as Intellectual Property: Who Controls Scientific Research? AAAS series on Issues in Science and Technology. MacMillan Publishing Company, New York.

Nimmer, David, Elliot Brown and Gary N. Frischling, (1999), The Metamorphosis of Contract into Expand, California Law Review Vol. 87 January, No. 1, pp. 17-78.

Okerson, Ann, (1996), Who Owns Digital Works? Computer Networks Challenge Copyright Law, But Some Proposed Cures May be as Bad as the Disease, Scientific American, July, pp.80-84.

Onsrud H. and G. Rushton, (eds.), (1995), Sharing geographic information, New Brunswick NJ: Centre for Urban Policy Research.

Onsrud, H.J., J.P. Johnson, and J. Winnecki, (1996), GIS Dissemination Policy: Two Surveys and a Suggested Approach, Journal of Urban and Regional Information Systems, 8(2).

Onsrud, H.J. and X.R. Lopez, (1997), Intellectual Property Rights in Disseminating Geographic Data, Products, and Services: Conflicts and Commonalities Among European Union and United States Approaches, I. Masser and F. Salgé (eds.), European Geographic Information Infrastructures. GISDATA 5,London: Taylor & Francis, pp. 153-167, ISBN: 0-7484-0756-1 (pbk).

Onsrud, H.J., (1998), The Tragedy of the Information Commons. Policy Issues in Modern Cartography (Elsevier Science) 1998, 1st edition, Chapter 9.

Onsrud, H.J., (1999), Legal Access to Geographic Information: Measuring the Losses or Developing Responses? Paper based on Specialist meeting NCGIA 1998.

Pelman, Harry, (1998), Letter to Hatch and Leahy, September 4.

Perritt, Henry H., (1994), Commercialization of Government Information: Comparisons Between the European Union and the United States, Internet Research, Vol. 4 No. 2, Summer, pp. 7-23.

Perritt, Henry H. Jr., (1995), Should local governments sell local spatial databases through state monopolies? Jurimetrics, Journal of Law, Science and Technology, Volume 35, Number 4, pp. 449-469.

Perritt, Henry H. Jr., (1999A), Law and the Information Superhighway, Aspen Law & Business, Gaithersburg and New York. ISBN 0-471-12624-1.

Perritt, Henry H., Jr., (1999B), Law and the Information Superhighway, 1999 Cumulative Supplement, Aspen Law & Business, Gaithersburg and New York.

Pluijmers, Yvette, (1998A), Juridische bescherming van commerciele geo-informatie, Een vergelijking tussen Nederland en de Verenigde Staten, Geodesia, No 9 september, pp. 373-379.

Pluijmers Yvette, (1998B), Protecting Intellectual Property in Private Sector Spatial Datasets. Thesis leading to the Master of Science degree at University of Maine, May.

Reichman, Jerome H. and Pamela Samuelson, (1997), Intellectual Property Rights in Data, Vanderbilt Law Review Vol. 50:51, pp.51-166.

Reichman J.H. and Jonathan A. Franklin, (1999), Privately Legislated Intellectual Property Rights: Reconciling Freedom of Contract with Public Good Uses of Information, UPenn Law Review, Volume 147, No. 4, April, pp.875-970.

Reichman, J.H. and Paul F. Uhlir, (1999), Database Protection at the Crossroads: Recent Developments and Their Impact on Science and Technology, Berkeley Technology Law Journal, Vol. 14, No. 2, Spring, pp.793-838.

Richman, Barry M. and Richard N. Farmer, (1974), Leadership, Goals, and Power in Higher Education, Jossey- Bass Publishers, San Francisco, First Edition.

Samuelson, Pamela, (1994), Legally Speaking: Self-plagiarism or Fair Use? Communications of the ACM, Vol. 37, No. 8, August, p. 21.

Samuelson, Pamela, (1996A), Legally Speaking: Legal Protection for Database Contents, Communications of the ACM, Vol. 39, No. 12, December, pp. 17-23.

Samuelson, Pamela, (1996B), The Copyright Grab, Wired 4.01, January, pp. 134-138, 188, 190-193.

Samuelson, Pamela, (1997), Legally Speaking: Embedding Technical Self-Help in Licensed Software, Communications of the ACM, Vol. 40, No. 10, October, pp. 13-17.

Samuelson, Pamela, (1998), Legally Speaking: Does Information Really Have to be Licensed? Communications of the ACM, Vol. 41, No. 9, September, pp. 15-20.

Samuelson, Pamela, (1999), Legally Speaking: Good News and Bad News on the Intellectual Property Front, Communications of the ACM, Vol. 42, No. 3 March, pp. 19-24.

Sandburg, Brenda, (1999), UCC2B Is Dead -- Long Live UCITA, Supporters face a big battle to pass a uniform information licensing law for the Internet, The Recorder (access through: http://www.callaw.com/), May 27.

Schmidtz, David, (1991), The Limits of Government, An Essay on the Public Goods Argument, Westview Press Inc., ISBN: 0-8133-0870-4/0-8133-0871-2(pbk).

Snow, John, (1855), On the Mode of Communication of Cholera. London: John Churchill, New Burlington Street, England. See
http://www.ph.ucla.edu/epi/snow/snowbook2.html,
for the Snow map see: http://www.ph.ucla.edu/epi/snow/1859map/map1859.html.

Streff, W.A. Jr. and J.S. Norman, (1997), Courts, UCC Tackle Shrink-Wrap Licenses, The New York Law Journal, October 14.

Varian, Hal R., (1995), The Information Economy. How much will two bits be worth in the digital market place? Scientific American, September, pp. 200-201.

Zimmerman, Donald E., and Michel Lynn Muraski, (1995), The Elements of Information Gathering, A Guide for Technical Communicators, Scientists, and Engineers, ORYX Press 1995, Phoenix, AZ. ISBN 0-89774-800-X.

Zitner, Aaron, (1999), Health study data demanded by US Chamber, suit likely over research done with federal funds, Boston Globe, December 10th, page D01.

Web Based References

Website 1: http://www.ucgis.org
Website 2: http://www.urisa.org
Website 3: http://www.nsf.gov/verity/srchawd.htm
Website 4: http://www.ombwatch.org/info/govhome.html
Website 5: http://www.ama-assn.org/sci-pubs/journals/archive/jama/vol_277/no_15/oc6d11.htm
Website 6: http://www.nih.gov/welcome/director/ebiomed/com0613.htm
Website 7: http://www.cast.uark.edu/local/hunt/

Appendix A Questionnaire on Access to Scientific and Technical Data

---

Introduction

The goal of this questionnaire is to gather information on the policies and processes confronted by university researchers in gaining access to data for their research. The information gathered should also indicate the extent to which current policies and processes for acquiring access to data meet the desires and needs of the university researcher.
The results ultimately will be used to supply evidence of academic community support or lack of support for a range of legal options for protecting databases, some of which are currently being considered by Congress. In order to obtain relatively unbiased answers we will provide informational links about these legal options after you submitted this questionnaire.

****

Instructions

This questionnaire consists of 4 sections: General Information, Most Recent (Current) Research Project Dealing with Geographic Data, Datasets Specific, and Desired Datasets.

Please complete this questionnaire as directed in each section or question. You may skip any question that you do not want to answer or that is not applicable to your situation.

Others who have completed this survey took less than 30 minutes to do so. When you have completed all the questions, please save the complete questionnaire and send it as an attachment to bvanloen@spatial.maine.edu.
Sending the questionnaire implies consent to participate.

Your responses to this questionnaire are confidential and will not be released individually. Your personal information (name and email-address) will be separated from your response. Thus, the survey is anonymous. There is no more risk in participating than in everyday living.

Thank you very much for your time and cooperation.

Bastiaan van Loenen
Graduate Student
Department of Spatial Information Science and Engineering
National Center of Geographical Information and Analysis
University of Maine
Email: bvanloen@spatial.maine.edu

****

Section 1 General Information

Please provide us with your name and e-mail address.
(Providing this information will allow us to remove your name from future e-mail requesting you to complete the form. Again, your responses will be kept confidential.)

Name:

E-mail address:

1. Have you used, created, updated, integrated or distributed geographic data in accomplishing academic research?

☐ Yes If you answered Yes, please click here to proceed with question #2

☐ No If you answered No, please click here to proceed with question #3

2. A geographic information system (GIS) may be defined as a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information. A GIS by this definition includes systems called "geographic information systems" but also includes computerized systems for mapping, urban modeling, environmental modeling, routing, facilities management, direct marketing and similar tasks involving geographically referenced data.

Are you using or have you used a geographic information system (GIS) in any of your research projects (funded or unfunded) within the past five years?

☐ Yes Please click here to complete the remainder of this questionnaire by skipping to Section 2

☐ No Please click here to continue with question #3

3. Please provide us with the name and e-mail address of one or more other researchers in your department or college that may use digital geographic data or a GIS in one of their research projects.

Name:

E-mail address:

Thank you for your cooperation.
Please save the questionnaire and send it as an attachment to bvanloen@spatial.maine.edu. Warning: it may take some time to save this document!

****

Section 2
Most Recent (Current) Research Project Dealing With Geographic Data

In responding to the remaining questions, please refer to the most recent research project or scholarly study in which you have used a geographic information system. (If you have more than one current project using geographic information technologies, select the project using the greatest amount and variety of digital data)

4. Please provide the title for this research project or scholarly study.

5. With which disciplinary field do you most closely associate this project?

6. From which of the following sources did you acquire data for use in this specific research project?
(Note: Please place a check by sources of all types of digital data you used (not just geographic data) and include data accessed or acquired for free, by grant, purchase, license or any other means)

- ☐ federal government agency(s) (U.S.)
- ☐ state government agency(s) (U.S.)
- ☐ local government agency(s) (U.S. county or municipality)
- ☐ not-for-profit organization or foundation (includes domestic or foreign organizations and includes public universities including your own if you acquired a dataset from it)
- ☐ private commercial firm (includes domestic or foreign mass consumer datasets, custom datasets for specific clients, datasets from utilities and datasets from private universities including your own university)
- ☐ other sources - please specify: -

7. Please provide the explicit name(s) of one or two agencies or organizations in each of the indicated categories from which you acquired data and name a specific dataset that you acquired or accessed from that organization.
(Note: Please list **no more than three** of your primary data sources for the project even though you may have used many more sources of data.)

Dataset 1: Agency/ Organization/ Firm Name:
Name (or brief description) of #1 dataset acquired or database used:
Dataset 2: Agency/ Organization/ Firm Name:
Name (or brief description) of #2 dataset acquired or database used:
Dataset 3: Agency/ Organization/ Firm Name:
Name (or brief description) of #3 dataset acquired or database used:

Please continue to Section 3 by clicking here. Use the following two links **only** if you are ***returning*** from section 3 to provide responses in a later portion of section 3
Go to dataset 2 in section 3 Go to dataset 3 in section 3

****

Section 3 Dataset Specifics

The questions in this section need to be answered for each of the datasets listed above under Question 7 in Section 2. Simply "click" on the most appropriate answer or answers for the dataset under consideration or fill in the "text box". Square "check boxes" indicate that all appropriate responses should be marked whereas circular "check boxes" indicate that only one best answer should be marked.

Dataset 1	Same #1 dataset or database as you indicated under Question 7 in Section 2
Name of agency/ organization/ firm that created this dataset (Please repeat the name from Question 7 for confirmation)	-
1. From whom did you directly acquire this dataset?	○ the creator of the dataset ○ an intermediate commercial entity, not being the primary creator of the dataset ○ An intermediate non-commercial entity, not being the primary creator of the dataset (e.g. public library, university, government agency, community organization) ○ do not know
2. How did you find out about the availability of this specific dataset?	☐ personal inquiries (by phone, email, personal contact) ☐ library catalog search (on-line or otherwise) ☐ general Internet search ☐ search of a specific database ☐ print literature (including supplier catalogs) ☐ advertisements (print or on-line) ☐ existence commonly known in the discipline ☐ other, please specify -
3. What was the physical means by which you acquired this (digital) dataset?	☐ downloaded across the Internet ☐ shipped by e-mail ☐ acquired on a digital portable medium (e.g. CD-ROM or disk) ☐ acquired on paper and converted ☐ self-collected ☐ other, please specify -
4. Did you need to make a specific request to an agency or organization in order to obtain a copy or access to this dataset?	○ yes ○ no

5. Were you required to identify yourself prior to being allowed to access the dataset?	☐ yes ☐ no
6. Were you required to explain your intended use of the dataset prior to being allowed to access the dataset?	☐ yes ☐ no
7. Was all or a substantial portion of this dataset or database originally developed by a government agency using exclusively or primarily public funds?	☐ yes ☐ no ☐ do not know
8. Was all or a substantial portion of this dataset or database originally developed by a university or private firm (profit or not-for-profit) using exclusively or primarily publicly-financed research and development funds? (e.g. government research grant to a public or private university or to a private company)	☐ yes ☐ no ☐ do not know
9. What specific contractual or licensing approach, if any was imposed on your use of this dataset? (select only one)	☐ "shrink-wrap" license or purchase contract provisions were offered on a take-it or leave-it basis (e.g. terms were contained in the packaging of a CD) ☐ "click-wrap" license or purchase contract provisions were offered on a take-it or leave-it basis (e.g. terms were stated on our computer screen to which we were required to affirmatively respond prior to downloading a dataset, accessing an on-line database or having a dataset shipped) ☐ "boilerplate" license or purchase contract provisions were offered on a take-it or leave-it basis in response to our request for a specific or custom produced data set and we were required to sign or otherwise respond affirmatively to those provisions ☐ license or purchase contract provisions were negotiated with the supplier of the dataset or database ☐ license or purchase contract provisions were placed in writing by the supplier of the dataset or database when supplied but we were not required to sign or otherwise affirmatively assent through a volitional act to the terms ☐ we acquired this specific data set in such a manner that we assumed that no contract or licensing provisions applied to our use of the data (e.g. acquired through an openly accessible online government database or web site, through an open public library

	with no contract provisions apparent, received from a colleague, etc.) ☐ no licensing or purchase contract provisions were involved in our use of this dataset (or in our use of a database from which the data was extracted)
10. What restrictions, if any, were imposed on your use of this dataset or on your use of the computer database from which the data was acquired? (mark all that apply and mark all restrictions contained in the contract or licensing language even though you might question the enforceability or legality of some of those provisions)	☐ not applicable, no explicit or implied restrictions were imposed ☐ provisions stated that we could not pass on the provided digital data to any other parties ☐ provisions stated that our use could be for only academic or research purposes ☐ a monetary payment was required ☐ provisions stated that the data supplier would not be liable to us for any losses that we or others might incur due to any errors or other shortcomings in the data supplied ☐ provisions stated that we are liable to the supplier of the data for any losses the supplier might incur to a third party through our inappropriate use of the data ☐ provisions stated that any value-added products that we developed through use of the data (1) required explicit permission of the data supplier prior to dissemination of the value-added products by us, (2) vested an ownership interest in the original data supplier, or (3) required a royalty due to the data supplier ☐ our understanding is that federal copyright law does not allow some of the uses we made of the dataset in this research project without first acquiring the permission of the data supplier (We therefore ☐ obtained that permission or ☐ ignored the law) ☐ our understanding is that state legislation or other state law does not allow some of the uses we made of the dataset in this research project without first acquiring the permission of the data supplier (We therefore ☐ obtained that permission or ☐ ignored the law) ☐ other or alternative restrictions were imposed on the data. Please specify: -

11. What did you pay for access to or a copy of the dataset?	☐ not applicable, the dataset was free ☐ market price ☐ market price less a discount for the university or other not-for-profit user ☐ price based on full cost recovery (e.g. the price was set by the producer by predicting the number of expected purchasers and then spreading the cost across those purchasers but with no profit for the producer) ☐ price based on partial cost recovery for the producer ☐ price based on the cost of dissemination to the user (e.g. costs incurred by the agency in order to respond to your specific request such as duplication and delivery expenses) ☐ price based on a minimal statutory fee
12. How good was the documentation regarding the dataset?	☐ excellent ☐ good ☐ fair ☐ poor ☐ non-existent
13. Which of the following did the documentation of the dataset (digital catalogue files or metadata) help you accomplish? (mark all that apply)	☐ allowed us to find the dataset through a computer search ☐ allowed us to assess the relevance of the dataset for our research project (e.g. data type, description entities) ☐ allowed us to assess the technical suitability of the dataset (e.g. data structure) ☐ allowed us to assess the quality or accuracy of the dataset ☐ allowed us to assess the timeliness of the dataset for our purposes ☐ allowed us to assess contractual or other legal constraints on the use of the dataset ☐ not applicable, no documentation or metadata was available

14. Was access to this dataset or database made available to you within a reasonable period of time of requesting access?	☐ yes, access was immediate ☐ yes, the time between the request and obtaining the data was reasonable ☐ no, the time between the request and obtaining the data was unreasonable
15. If you acquired access to this dataset through a database service to which your university library subscribes or participates in supporting, how was this database made available to you?	☐ not applicable to this dataset ☐ we paid a per use fee, the library paid a per use fee, or we acquired special permission that might not be granted to all library patrons ☐ we acquired access through an open access policy applied to all library patrons; no per use fee was charged nor was special permission required
16. Is it possible to access the same or similar dataset meeting your needs from another source? (mark all that apply)	☐ yes, but access through this source was more convenient ☐ yes, but the quality of the dataset from other sources was not as responsive to our needs ☐ yes, but the expense of other sources was not as responsive to our needs ☐ yes, but the restrictions imposed by other sources were not as responsive to our needs ☐ no, this was the only realistic source for the dataset

17. Which of the following, if any, were significant factors in allowing you to successfully use this dataset? (mark all that apply)	☐ the physical means for gaining access to this dataset ☐ availability of a search capability allowing the ability to find this dataset or database ☐ adequate documentation or metadata for this dataset ☐ sufficient identification of the sources used to create this dataset ☐ suitable format or compatibility with the software or hardware we used ☐ sufficient quality or accuracy of this dataset for our purposes ☐ timeliness of this dataset for our purposes ☐ personal or institutional willingness to giving us access within the organization that created the dataset ☐ lack of application of copyright law to our uses of this dataset ☐ lack of application of specific data protection legislation to our uses of this dataset (e.g. local ordinance, state statute, federal statute) ☐ cost of this dataset ☐ contractual provisions facilitating our uses of this dataset ☐ contractual provisions regarding further dissemination of this dataset ☐ contractual provisions regarding liability ☐ contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this dataset ☐ other, please specify -

18. Which of the following, if any, were significant impediments to your use of this dataset? (mark all that apply)	☐ the physical means for gaining access to the dataset ☐ lack of a search capability allowing the ability to find the dataset or database ☐ inadequate documentation or metadata for the dataset ☐ lack of identification of the sources used to create this dataset ☐ lack of suitable format or compatibility with the software or hardware we used ☐ inadequate quality or accuracy of the dataset for our purposes ☐ timeliness of the dataset for our purposes ☐ personal or institutional resistance to giving us access within the organization that created the dataset ☐ restrictions imposed on our use of the dataset by copyright law ☐ restrictions imposed on our use of the dataset by specific data protection legislation (e.g. local ordinance, state statute, federal statute) ☐ lack of alternative datasets meeting our needs ☐ cost of the dataset ☐ contractual restrictions imposed on our uses of the dataset ☐ contractual restrictions regarding further dissemination of the dataset ☐ contractual provisions regarding liability ☐ contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of the dataset ☐ other , please specify -

19. Even though contractual, legal, technical and other impediments may have constrained your use of the specific dataset, to what degree were you able to accomplish research tasks that were dependent upon use of this dataset?	☐ almost all research tasks dependent on this dataset were accomplished ☐ most research tasks dependent on this dataset were accomplished ☐ about half of the research tasks dependent on this dataset were accomplished ☐ some of the research tasks dependent on this dataset were accomplished ☐ almost none of the research tasks dependent on this dataset were accomplished
20. How would you rate your satisfaction with your use of this specific dataset or database?	☐ excellent ☐ good ☐ fair ☐ poor ☐ non-existent
21. Use of this specific dataset was important in accomplishing the overall objectives of the research project	☐ strongly agree ☐ agree ☐ disagree ☐ strongly disagree ☐ do not know/ no opinion

Was this the last of the datasets listed in question 7 of section 2?

Yes, go to section 4
No, continue with the next dataset below

Thank you for filling out the dataset specifics questions. Please continue with the last section: Section 4.

****

Section 4 Desired datasets

This section is about datasets you wanted to use but were not able to use.
Did a dataset exist that you desired for this research project but you did not acquire?

- Yes Please continue with the following questions
- No Please skip to the end

Desired Dataset	
Name of dataset	-
Name of agency/ organization/ Firm that created the dataset	-
Why did you want this particular dataset?	the dataset consists of more accurate or reliable data the dataset is better documented the dataset is more comprehensive or complete the dataset has higher quality data the dataset is more up-to-date the dataset is more user friendlier the dataset is more flexible Other. Please specify -
Why didn't you acquire access to this particular dataset?	the dataset was incompatible with our software or hardware limitations the dataset was too expensive the restrictions imposed on this dataset were not responsive to our needs the dataset was no longer available in digital format the documentation of the dataset was inadequate or not responsive to our needs exclusive rights were given to another organization until (very) recently the existence of this dataset was unknown to us other reason(s), please specify: -

From whom could you directly acquire this dataset?	☐ the creator of the dataset ☐ an intermediate commercial entity, not being the primary creator of the dataset ☐ an intermediate non-commercial entity, not being the primary creator of the dataset (e.g. public library, university, government agency, etc.) ☐ do not know
Was all or a substantial portion of this dataset or database originally developed by a government agency using exclusively or primarily public funds?	☐ yes ☐ no ☐ do not know
Was all or a substantial portion of this dataset or database originally developed by a university or private firm (profit or not-for-profit) using exclusively or primarily publicly-financed research and development funds? (e.g. government research grant to a public or private university or to a private company)	☐ yes ☐ no ☐ do not know

****

Follow Up Interview

In a later phase of my research, I might want to perform some follow up interviews by telephone. Would you be willing to participate in a follow up call?

☐ No

☐ Yes Please provide the following contact information:

Name

Work Phone

Thank you very much for your cooperation.

Appendix B Confirmation Page

Your response has been successfully sent!

Thank you very much for submitting information about your access to data environment. The results of the survey will be sent to you if you have filled out your name in the questionnaire.

Sincerely,
Bastiaan van Loenen

Informational links

Co-principal investigator professor Harlan J. Onsrud

General information about access issues

- A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Debates (1999) National Research Council, Commission on Physical Sciences, Mathematics, and Applications (CPSMA)
- Proceedings of the Workshop on Promoting Access to Scientific and Technical Data for the Public Interest: An Assessment of Policy Options (1999) National Research Council, Commission on Physical Sciences, Mathematics, and Applications (CPSMA)
- Bits of Power: Issues in Global Access to Scientific Data. A 1997 book by the National Research Council
- International Council of Scientific Unions CODATA- Committee on Data for Science and Technology

Legal scholars writing about "threats" to access of Scientific and Technical Data for academia

- Professor Litman
- Professor Jerome Reichman
- Professor Pamela Samuelson

Information about HR 354 "Collections of Information Antipiracy Act"

- Link to the text of the proposal HR 354 "Collections of Information Antipiracy Act" -- March 18, 1999
- Legislative hearing on H.R. 354, the "Collections of Information Antipiracy Act" -- March 18, 1999
- A comprehensive Association of Research Libraries site including the history of proposal HR 354

Information about the proposed Uniform Computer Information Transactions Act

- A guide to this proposal

Appendix C Answers to Questions Used for χ^2 Test of Proposed Principles for Data Provided by Government

Principle/ Question (Section 3 Questionnaire)	1 Level of Availability	2 Level of Affirmativeness in Dissemination	3 Level of Activity in Release	4 Level of Accessibleness by Archive	5 Level of Metadata Availability	6 Adherence to Marginal Cost or Less	7 Adherence to Non-Exclusivity Availability	8 Adherence to No Exclusive Partner Arrangements	9 Adherence to No Restrictions on Subsequent Uses	10 Adherence to Access Through Publicly Accessible Archive
18q										
18p	X								X	
18o	X								X	
18n	X								X	
18m	X								X	
18l	X					X				
18k										
18j	X								X	
18i	X								X	
18h	X	X					X			X
18g	X				X					
18f					X					
18e					X					
18d					X					
18c	X				X					
18b			X							
18a	X	X	X							
17p										
17o	X								X	
17n	X								X	
17m	X								X	
17l	X								X	
17k	X					X				
17j	X								X	
17i	X								X	
17h	X	X					X			X
17g	X				X					
17f					X					
17e					X					
17d					X					
17c	X				X					
17b			X							
17a	X	X	X							

Appendix D Questions Addressing Proposed Principles for Data Provided by Government

Principle/ Question (Section 3 Questionnaire)	1 Level of Availability	2 Level of Affirmativeness in Dissemination	3 Level of Activity in Release	4 Level of Accessibleness by Archive	5 Level of Metadata Availability	6 Adherence to Marginal Cost or Less	7 Adherence to Non-Exclusivity Availability	8 Adherence to No Exclusive Partner Arrangements	9 Adherence to No Restrictions on Subsequent Uses	10 Adherence to Access Through Publicly Accessible Archive
1		X		X				X		
2		X	X	X						
3	X	X	X	X						
4	X	X								
5	X						X			
6	X						X			
7	X	X	X	X	X	X	X	X	X	X
8										
9	X			X						
10	X			X					X	
11	X					X				
12	X			X	X					
13				X	X					
14	X									
15										X
16								X		
17	X	X	X		X	X	X		X	
18	X	X	X		X	X	X		X	

The explicit questions 1 – 18 may be found on pages 170 – 177.

Appendix E Answers to Questions Used for χ^2 Test of Proposed Principles for Data Provided by Academia

Principle/ Question (Section 3 Questionnaire)	1 Level of Full and Open Exchange of Data	2 Level of Accessibleness	3 Level of Availability	4 Adherence to Reasonable Time for Proprietary Use Before Dissemination of Dataset	5 Level of Accessibleness by Archive	6 Adherence to Adequate Metadata	7 Adherence to No Transfer Exclusive Rights in the Work	8 Adherence to Access Through Publicly Accessible Archive	9 Adherence to Marginal Cost or Less
18q									
18p	X		X						
18o	X	X	X						
18n	X	X	X						
18m	X	X	X						
18l	X								X
18k									
18j	X	X	X						
18i	X	X	X						
18h	X							X	
18g	X	X	X	X		X			
18f						X			
18e						X			
18d						X			
18c	X					X			
18b									
18a	X								
17p									
17o	X		X						
17n	X	X	X						
17m	X	X	X						
17l	X	X	X						
17k	X								X
17j	X	X	X						
17i	X	X	X						
17h	X							X	
17g	X	X	X	X		X			
17f						X			
17e						X			
17d						X			
17c	X					X			
17b									
17a	X								

Appendix F Questions Addressing Proposed Principles for Data Provided by Academia

Principle/ Question (Section 3 Questionnaire)	1 Level of Full and Open Exchange of Data	2 Level of Accessibleness	3 Level of Availability	4 Adherence to Reasonable Time for Proprietary Use Before Dissemination of Dataset	5 Level of Accessibleness by Archive	6 Adherence to Adequate Metadata	7 Adherence to No Transfer Exclusive Rights in the Work	8 Adherence to Access Through Publicly Accessible Archive	9 Adherence to Marginal Cost or Less
18	X	X	X	X		X	X		X
17	X	X	X	X		X	X		X
16									
15								X	
14	X	X		X					
13									
12	X					X			
11	X	X							X
10	X	X	X						
9	X	X	X						
8		X		X					
7									
6	X								
5	X								
4	X								
3	X								
2									
1		X							

The explicit questions 1 – 18 may be found on pages 170 - 177.

Appendix G Answers to Questions Used for χ^2 Test of Proposed Principles for Data Provided by Private Entities

Principle/ Question (Section 3 Questionnaire)	1 Level of adherence to Fair Contractual Provisions	2 Level of Adherence to Public Domain Policy	3 Adherence to Access Trough Publicly Accessible Archive	4 Adherence to Identification of Source
18q				
18p				
18o		X		
18n		X		
18m		X		
18l				
18k				
18j		X		
18i		X		
18h				
18g		X		
18f				
18e				
18d				X
18c				
18b				
18a				
17p				
17o				
17n		X		
17m		X		
17l		X		
17k				
17J		X		
17i		X		
17h				
17g		X		
17f				
17e				
17d				X
17c				
17b				
17a				

Appendix H Questions Addressing Proposed Principles for Data Provided by Private Entities

Principle/ Question (Section 3 Questionnaire)	1 Level of adherence to Fair Contractual Provisions	2 Level of Adherence to Public Domain Policy	3 Adherence to Access Trough Publicly Accessible Archive	4 Adherence to Identification of Source
18		X	X	
17		X	X	
16				
15			X	
14		X		
13				
12				
11				
10		X		
9		X		
8		X		
7		X		
6				
5				
4				
3				
2				
1				

The explicit questions 1 – 18 may be found on pages 170- 177.

Appendix I Letter to Interviewees

Subject: Access to Scientific and Technical Data in an Academic Setting

Dear professor/ Dr./ Ms./ Mr. XXXX,

Legislative efforts are currently being pursued in the United States, the European Union and the World Intellectual Property Organization (WIPO) to alter the legal protection provided to databases. The outcomes of those legislative efforts are likely to affect access to and use of scientific and technical databases. In order to inform the political process, this survey attempts to gather information on the present and preferred practices of the scientific community in using geographic data. This work is performed as part of my graduate thesis work, which is being funded by the National Center for Geographic Information and Analysis at the University of Maine.
As an academic researcher using geographic data, your completion of this survey would be greatly appreciated. In order to generate the survey as quickly and accurately as possible the survey may be completed online at the following web-address:
http://www.spatial.maine/~bvanloen/Questionnaire/survey.htm
If you do not have access to the Internet, please send an e-mail to bvanloen@spatial.maine.edu. A word processing document will be sent to you.
Please fill out the survey as soon as possible, preferably within one week but no later than November 1, 1999.

Your response will remain confidential.
I would like to thank you in advance for your help. In return for your assistance I will inform you of the outcome of the survey results as they are completed,- tentatively in December 1999.

Thank you very much for your cooperation,

Sincerely,

Bastiaan van Loenen

Graduate Student
University of Maine
Department of Spatial Information Science and Engineering
National Center for Geographic Information and Analysis

Appendix J Follow Up Letter 1

Subject: Access to Scientific and Technical Data in Academic Settings

Dear professor/ Dr./ Ms./ Mr. XXX,

About a week ago I sent you an email requesting that you fill out a web-based questionnaire concerning your access to and use of geographic data. As an academic researcher using geographic data, your completion of this survey is very important. Due to potential changes in the law, evidence of the present and preferred practices of the scientific community in accessing and using geographic data is needed in order to better inform the political process.

If you already have filled out the web-based questionnaire on the Internet, please accept my sincere thanks. If not, please go to the web site and complete the questionnaire today. The questionnaire may be completed online at:
http://www.spatial.maine.edu/~bvanloen/Questionnaire/survey.htm

Because I am sending this email to a sample of researchers using geographic data, your response is important so that the results are representative. If you do not have access to the Internet, please send an email to bvanloen@spatial.maine.edu. A word processing document will be sent to you.

If you encounter problems with filling out the questionnaire, please contact me at (207) 581-2210 or bvanloen@spatial.maine.edu Thank you for your help.

Sincerely,

Bastiaan van Loenen

Graduate Student
University of Maine
Department of Spatial Information Science and Engineering

Appendix K Follow Up Letter 2

Subject: Access to Scientific and Technical Data in Academic Settings

Dear professor/Dr./ Ms./Mr. XXXX,

About two weeks ago I sent you an email seeking your help in a national study of researchers using geographic data. The research results should help inform scientists and policy makers of the options and approaches by which data policies and practices advantageous to the research and academic communities might be maintained or improved.

I am writing you again because of the significance each questionnaire has to the usefulness of the study. Those who have already completed the questionnaire indicate that for reporting on 3 datasets the questionnaire took about 30 minutes. For fewer datasets the time commitment is, of course, less.

Please fill out the questionnaire on line at the following webpage:
http://www.spatial.maine.edu/~bvanloen/Questionnaire/survey.htm
If you encounter problems with filling out the questionnaire, please contact me at bvanloen@spatial.maine.edu. I can send you a word processing file if you desire that instead.

Your help is greatly appreciated.

Sincerely,

Bastiaan van Loenen

Graduate Student
University of Maine
Department of Spatial information Science and Engineering

Appendix L Reactions to the Invitation to Participate in the Survey

Dear Mr or Ms. Van Loenen:Thank you for the invitation to participate. I went to the web site and answered a few of the questions but did not submit any of the answers. I was not convinced by your promise of confidentiality. E.g., how secure is your Web site? I conducted a Web-based survey earlier this year, and we were able to promise participants a secure Web site. Nor was I pleased to see requests for specific names and email addresses of others. This kind of linking across respondents must be done very carefully, and it too raises a host of confidentiality issues. Then specific names of projects. Given federal funding and a project name, anyone can figure out my identity. Sorry. Please take me off your list.

I am declining your offer to fill in the questionaire because the information requested might be used in an inappropriate fashion. I do not know you! Also, I think it is odd that you suppose everyone has time to fill out such a questionaire.

While follow-up is a good practice in administering a survey instrument, too many requests may offend the recipient and result in a refusal or even biased answers.
I'm sorry I do not have time to work through this interesting survey

I did not respond to your questionnaire because I do not believe it is applicable. As Department chairperson (alas) my time is taken up by administration instead of interesting stuff, like geography. Quite simply, I do not now qualify as "a researcher using geographic data."

I am actually not a member of "the scientific community" and am also, I'm afraid, rather overwhelmed at the moment and for the near future. Consequently, I won't be able to assist you by completing a survey, though I wish you well in what sounds like a valuable project.

I have recently left the University, due to a budget cut. I cannot participate in your study.

I appreciate your interest in my response. However, as editor of a journal, advisor of 10 dissertation students, PI of several federal grants, etc., it's hard to find extra time in my day.

I am intrigued by your web based approach to survey analysis.

I started to fill out the questionnaire but found it too time consuming and detailed - It would have taken me more than an hour

I have been retired for more than ten years. I am not using any information except for my private use, which is minimal. Good luck to your survey!

Appendix M UCGIS Members Asked to Participate

Boston University

California State University System:

- California Polytechnic State University San Luis Obispo
- CSU Dominguez Hills
- CSU Fullerton
- Humboldt State University
- CSU Los Angeles
- CSU Long Beach
- CSU Monterey Bay
- CSU Northridge
- CSU San Bernardino
- San Diego State University
- San Francisco State University
- Sonoma State University

George Mason University

Hunter College, City University of New York

Oak Ridge National Laboratory

Ohio State University

Oregon State University

State University of New York at Buffalo

Texas A&M University-Corpus Christi

University of California, Berkeley

University of California, Santa Barbara

University of Colorado

University of Delaware

University of Georgia

University of Idaho

University of Illinois

University of Iowa

University of Kansas

University of Kentucky

West Virginia University

University of Maine

University of Massachusetts, Amherst

University of Michigan

University of Minnesota

University of Nebraska

University of Oklahoma

University of Oregon

University of Pittsburgh

University of South Carolina

University of Southern California

University of Texas at Dallas

University of Washington

University of Wisconsin-Milwaukee

University of Wyoming

Virginia Commonwealth University

Appendix N Confidentiality Requirement Form UMaine

Summary of the Proposal

The research explores current research environments of researchers in public universities in the United States using geographic scientific and technical data. A questionnaire will be used to gather information about the research environments.

The research environment will be analyzed using a set of "recommended access to scientific and technical data principles". The main objective of the analysis is to determine whether each acquisition or access arrangement adheres to or violates the recommended principles. Critical success factors of the analysis will be the degree of satisfaction of the individual researcher with his or her research environment and the achievement of their specific project objectives for the research.

This research should result in new knowledge that helps scientists with means to overcome possible impediments. The results ultimately will be used to supply evidence of academic community support or lack of support for a range of legal options for protecting databases, some of which are currently being discussed by Congress.

Personnel

Personnel in contact with subjects or with identifiable data include the applicant and my advisor Harlan J. Onsrud.

Subject recruitment

Participants in this study are members of the University Consortium for Geographic Information Science (UCGIS). The UCGIS is a non-profit organization of universities and other research institutions dedicated to advancing our understanding of geographic processes and spatial relationships through improved theory, methods, technology, and data.

The identity of the participants is acquired through the UCGIS website (http://www.ucgis.org).

Informed consent

Participants in this study have in the questionnaire (see very first question of the attached questionnaire) the choice to either fill out or not fill out their name and email address.

Confidentiality

Confidentiality of individual responses is guaranteed to the participants (see under Instructions of attached questionnaire).

The research process guarantees the anonymity and privacy of the participants. From the UCGIS website a "master list" of email-addresses will be made. The questionnaire will be sent to all the email-addresses listed on the master list.

If the participant decides to fill out the very first question (name and email of participant), his/ her name (1) will be deleted from his/her response and (2) will be deleted from the master list. In this way there will be no relation between the names of the participants and their individual responses.

The complete “master list” (including names of respondents and names of non-respondents) will be the only reference to the participants in possession of the applicant.

Risks to the subjects

I estimate the risks to the participants to be minimal

Benefits

(see summary of proposal)

This research should result in new knowledge that helps scientists with means to overcome possible research impediments. The results ultimately will be used to supply evidence of academic community support or lack of support for a range of legal options for protecting databases, some of which are currently being discussed by Congress.

Biography of the Author

Bastiaan van Loenen was born in Hoorn, the Netherlands. He attended Delft University of Technology in 1992 and graduated in 1998 with a M.Sc. degree in Geodetic Engineering. After his study he worked as a researcher at the Department of Geodetic Engineering. In August 1998, he went to the United States to write his thesis with professor Harlan Onsrud and to work for a Master's degree at The University of Maine. Funded by a NCGIA research assistantship and a grant from the VSB Fonds, he entered the graduate program of Spatial Information Science and Engineering, focusing on legal aspects of Geographic Information Systems.

After receiving his degree, Bastiaan will return to the Netherlands to continue to work in the academic sector. Bastiaan is a candidate for the Master of Science degree in Spatial Information Science and Engineering from The University of Maine in May, 2001.