

Geographic Data for Academic Research: Assessing Access Policies

Bastiaan van Loenen and Harlan J. Onsrud¹

ABSTRACT: Data availability is a key issue affecting the collective well-being of society. Economic and legal scholars have argued that the current, relatively open, access-to-data environment in the United States is beneficial to advancing knowledge and the economy. However, little empirical evidence exists to validate the extent to which various access policy environments do or do not contribute to the productivity of academic researchers. Our research aimed to evidence support or lack of support of various data policies in the context of access to, and use of, geographic data within the university research environment. We synthesized a set of twenty-three recommended access-to-data principles from recommendations set forth in the literature. An online questionnaire strove to gain sufficient information to determine whether recommended principles were adhered to in the acquisition of each specific data set and whether scientists were productive in their use of such data sets. Productivity was assessed in terms of five measures. We hypothesized that data-sharing relationships would be more productive for science if the data policies confronted by scientists in their use of digital geographic data conformed with the recommended policies advocated in the literature. The data indicated relatively clear statistical significance in testing the principles of “adherence to pricing at marginal cost or less” and “provision for availability of metadata.” Correlated with the productivity of scientists, the collected survey data evidenced non-support of the first principle and support of the second. The latter finding suggests that government, private sector, and academic suppliers of geographic data should give high priority to the documentation of metadata in order to stimulate the more widespread use of available spatial data. This article describes the survey and statistical methods employed in researching this problem and presents the results of testing the two recommended principles. The implications of the findings are discussed.

KEYWORDS: Geographic data, GIS, data access principles, productivity, measures of success, metadata, open access, cost recovery, t-test, chi square test

Introduction

Data availability is a key issue affecting the collective well-being of society. Data and information are the raw materials for the production of useful knowledge. The possibilities for discovering new insights about the natural world, which have both commercial and public interest value, are extraordinary (NRC 1999a, p. 34). The academic community has taken advantage of inexpensive and efficient opportunities to share data and knowledge across digital networks with relatively few legal, policy, or technological encumbrances. The characteristics of digital data (data sets) and collections of data (databases) that make them easy to share and help advance science. Yet, these same characteristics provide disincentives for the provisioning of data from across broad segments of society; "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, p. 85). This begs the counter question: If access to data is overly constrained by legal or technological methods, how can we realistically use the data in advancing the well-being of society?

Some foresee that the “open-access-to-data” environment in academe will expand ultimately because “information wants to be free” (Stewart Brand’s slogan cited in Barlow 1994, p. 89). Others contend that the real future of the information age lies “in metering every drop of knowledge and charging for every sip” (Okerson 1996, p. 80). Most suggest models that balance between the two extremes (see e.g., Varian 1995, p. 201; Reichman and Samuelson 1997; Maurer et al. 2001). Pressure by the private sector to shift the legal balance by increasing the protection for databases through legislation (see Reichman and Uhler 1999; HR 354 1999; HR1858 1999; NRC 2000; Maurer et al. 2001) 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 (NRC 1997, p. 6; Reichman and Samuelson 1997, p. 68), private funding of academic research (Nelkin 1984, p. 97; NRC 1997, pp. 111, 132), and pressure by university administrators to generate royalties from

¹ Bastiaan van Loenen and Harlan J. Onsrud, *Geographic Data for Academic Research: Assessing Access Policies*, *Cartography and Geographic Information Science*, Vol. 31, No. 1, 2004, pp. 3-17

the products of faculty (Reichman and Samuelson 1997, p. 68) are other developments decreasing or threatening to decrease access to data by academics using geographic, scientific and technical data.

However, empirical data about the effects of competing policy approaches on academic access to digital data for research are 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 relatively open “access-to-data” environment in the United States is beneficial to advancing knowledge and the economy (Weiss and Backlund 1997; Pira 2000; Pluijmers and Weiss 2002). Lopez (1998, p. 169), for example, found evidence that “U.S. academic sector players significantly benefit from the [open] dissemination policy of the U.S. federal government.” The main objective of our research was to evidence support, or lack of support, for various recommended data-access principles, in the context of access to and use of geographic data for knowledge advancement purposes within the university research environment.

Research Method

The general steps used in the research are summarized in Figure 1. The research began with literature reviews on both access-to-data policy issues and on appropriate research methods for pursuing knowledge advancements within the domain.

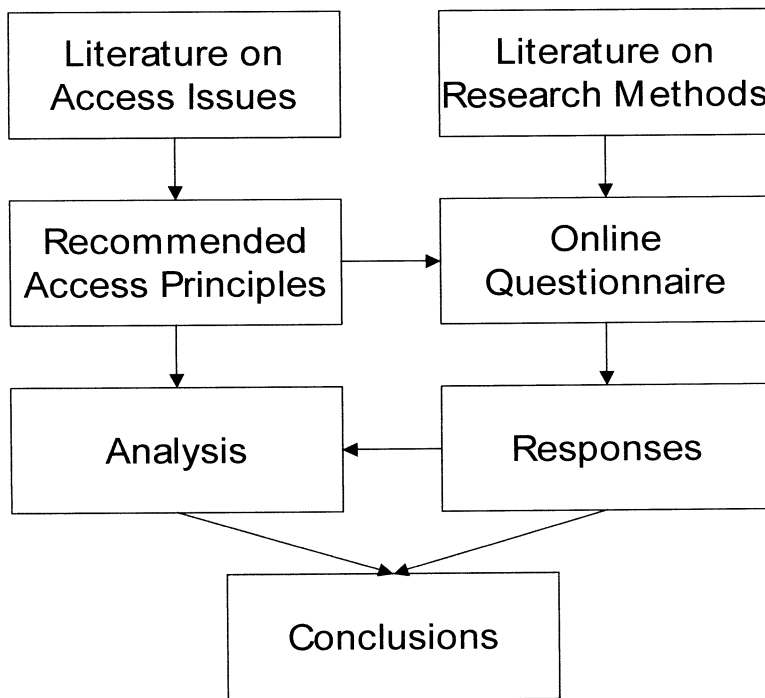


Figure 1. Research process for testing validity of access principles.

Recommended Access Principles

We synthesized a set of recommended access-to-data principles drawn from recommendations set forth in various study reports issued by the National Research Council or recommended in academic literature, which relate to policies for providing access to scientific and technical data. The recommended principles relate to the practices and policies of government agencies, the academic community, and the private sector. While a discussion of each principle may be found in Van Loenen (2001), the twenty-three principles found in literature and considered in the research are summarized below.

Access Principles for Data Provided by Government

- Government agencies should ensure that electronic data, information, and value-added features developed with public funds are available to the public.
- 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.
- 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 data sets on web servers or CDs along with appropriate retrieval software).
- 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.
- Scientific and technical data collected or maintained by or under authority of a government agency should be documented adequately with metadata.
- 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.
- 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.
- 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.
- 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.
- 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.

The above principles were derived or synthesized primarily from the Copyright Act, FOIA, ICSU/CODATA (1998), Onsrud and Lopez (1998), NRC (1995a; 1995b; 1997; 1999a; 1999b), Paper Reduction Act (1995), and Perritt (1999).

Access Principles for Data Provided by Academia

- 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.
- Scientific and technical data sets created by university and other not-for-profit researchers or their employing institutions, which 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 the publication of the results of the research.
- When publishing research articles, scientists should concurrently publish or otherwise make available electronically the data sets upon which their research depends or from which it is derived.
- 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.
- Scientific and technical data sets created or collected in conjunction with research or educational projects by university and other not-for-profit researchers or their employing institutions, which may be of current or future use to the scientific community should be retained and placed in a publicly accessible archive.
- Scientific and technical data sets made available in a publicly accessible archive should be documented adequately with metadata.
- For research and scholarly work partially or entirely financed with government funds or public university funds, university and other not-for-profit researchers who create data sets 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 data sets.

- 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.
- Scientific and technical data sets 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.

The above principles were derived or synthesized primarily from ICSU/CODATA (1997; 1998), NRC (1995b; 1997; 1999a; 1999b).

Access Principles for Data Provided by the Private Sector

- Commercial derivative products should be required to identify the government sources used.
- Scientific and technical data sets 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.
- 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.
- All mass market contracts, access contracts, and contracts imposing restrictions on uses of computerized information goods should be made on fair and reasonable terms, with due regard for the public interest in education, science, research, technological innovation, freedom of speech, and the preservation of competition.

The above principles were derived or synthesized primarily from ICSU/CODATA (1998), Reichman and Franklin (1999), and NRC (1999a).

Questionnaire Process

Specific questions were developed for each principle in an attempt to acquire enough information from respondents to determine whether a principle was being followed in a specific geographic data use or acquisition instance. This approach was ultimately problematic for testing some of the principles. For example, for some tested principles, almost all respondents were “following the principle,” such that a population “not following the principle” which would be sufficient for comparative testing failed to exist (or vice versa). Comparisons between populations adhering and not adhering to the principle relative to their productivity were thus not always possible. However, for those principles we were able to test, we found the results interesting.

Questionnaire Content

The questionnaire consisted of three sections. The first section, General Information, asked for general background information (e.g., name of the researcher, use of geographic data). The questions in this section made it possible to separate the user of geographic data 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, was intended to elicit more specific background information: name of the research project, field of research, sources of funding and data sets used for the research. The third section, Dataset Specifics, asked for specific information about a dataset: its price, physical means of acquisition, licensing approach encountered, and quality of the documentation, among others. Every question in the third section was based on one or more of the principles presented in this paper’s section “Recommended Access Principles”. While the complete questionnaire may be found online (http://www.geo.tudelft.nl/gigb/personen/loenenvan/Access_to_Data_Survey.htm), the Appendix to this article provides the questionnaire items germane to the principles discussed in the remainder of this article.

Measures of Success

We hypothesized that geographic data sharing relationships would be more productive for science and achieving research objectives if principles recommended in the literature were followed. Therefore, in addition to determining through sets of questions whether a principle was followed, the questionnaire explored whether scientists were successful in their use of each data set. Measures of success or productivity were gleaned from responses to questions in the following areas (see questions E, F, G, H, and I in the Appendix):

Measure 1: Factors allowing successful use of the data set;

Measure 2: Factors impeding use of the data set;

Measure 3: Task accomplishment using the data set;

Measure 4: Satisfaction with the data set; and

Measure 5: Contribution of the data set to the accomplishment of the overall research objective.

Responses were evaluated statistically. We tested whether data sets adhering to a recommended access principle contributed to more productive research than did those not adhering to the principle. For the productivity measures 3, 4, and 5 we asked the respondents to assess their productivity with the data set as excellent, good, fair, poor, or non-existent. We used a t-test (Figure 2) to test for statistical significance in these three measures of productivity. A 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, p. 504). Responses were grouped into those where academics used geographic data adhering to the selected principle and those where data not adhering to the selected principle were used. We assumed equal variances in the samples.

$$T = \frac{Y_1 - Y_2}{s_p (1/N_1 + 1/N_2)^{0.5}}$$

where

$$s_p^2 = \frac{(N_1 - 1) s_1^2 + (N_2 - 1) s_2^2}{N_1 + N_2 - 2}$$

Y_1 = mean sample 1
 Y_2 = mean sample 2
 N_1 = size sample 1
 N_2 = size sample 2
 s_1^2 = variance sample 1
 s_2^2 = variance sample 2
df = $N_1 + N_2 - 2$

Figure 2. T-test statistic.

Furthermore, an assessment was made in terms of success factors and impediments in the use of the data set. We used the chi-square test (Figure 3) to evaluate statistically these two productivity measures. As with the t-test, two groups were tested: one adhering to a principle and one not. This allowed us to test for differences in two groups' distribution across categories (Zikmund 1991, p. 500). The chi-square test requires that for a 3 x 2 matrix, no expected values can be zero, and a maximum of 20 percent of the expected values are between 1 and 5 (Sirkin 1995, p. 363).

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

$$R_i = \sum_{j=1}^c n_{ij} \quad C_j = \sum_{i=1}^r n_{ij} \quad E_{ij} = \frac{R_i C_j}{N}$$

E_{ij} = Expected frequency
 n_{ij} = Observed frequency
 R_i = total frequency in i-th row
 C_j = total frequency in j-th column
 N = total number of cases
 r = number of classifications (rows)
 c = number of groups (columns)
 df = $(r-1)(c-1)$

Figure 3. Chi-square statistic.

Sampling Group

The sample for which we strove 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. Our sample of researchers using geographic information was developed and drawn from three sources.

The first group consisted of 619 academics listed as having interests in GIS on the web site 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 understanding of geographic processes and spatial relationships through improved theory, methods, technology, and data (<http://www.ucgis.org>). The second group consisted of 33 additional academics drawn from an Urban and Regional Information Systems Association (URISA) list of individuals who have indicated an interest in geographic information science. [This sentence seems out of place in a paragraph describing the sampling groups; would delete it.] The Urban and Regional Information Systems Association is a non-profit international association of information professionals with specific emphasis on applications in state and local government (<http://www.urisa.org>). The third group consisted of 53 academic researchers funded by the National Science Foundation (NSF) who indicated they intended to use a GIS in their research work. These individuals were identified through key word searches of the NSF web site (<http://www.nsf.gov>). Only those researchers were selected who had research proposals accepted in 1994 or more recently. The total sampling group consisted of 705 academics using geographic data in their work.

Survey Responses

The invitation to participate in the research attracted 305 responses from 705 people contacted. Of the 305 respondents, 148 (21 percent of 705) provided useful responses for this research. A response was considered useful when for at least one data set the majority of questions were answered. The 157 responses 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.

Disciplines of Respondents

[Too verbose. How about: "Because many disciplines use geographic data in research, the responses we received were indicative of the cross-disciplinary nature of our sample."] Because a broad spectrum of disciplines use

geographic data in scientific research, one would expect that the data provided may be indicative of the responses across many research domains due to the cross disciplinary nature of our sample. The majority of respondents indicated that they did research work in the field of GIS, surveying, remote sensing, or photogrammetry (all together 19 percent). Other major respondent fields were geography (15 percent), ecological research (10 percent), earth sciences (9 percent) and planning (9 percent).

Distribution of Data sets

The on-line questionnaire asked participants to fill out each question for at least one and at the most three specific data sets. The questionnaire was completed for 290 data sets. Figure 4 shows the distribution of the sources of the data sets of the respondents. The majority of the data sets (75 percent) in this research came from a public source. This suggests a current heavy dependence by U.S. researchers on public rather than private sources of geographic data. The explicit reasons for such heavy reliance on public data may be worthy of a future scientific investigation.

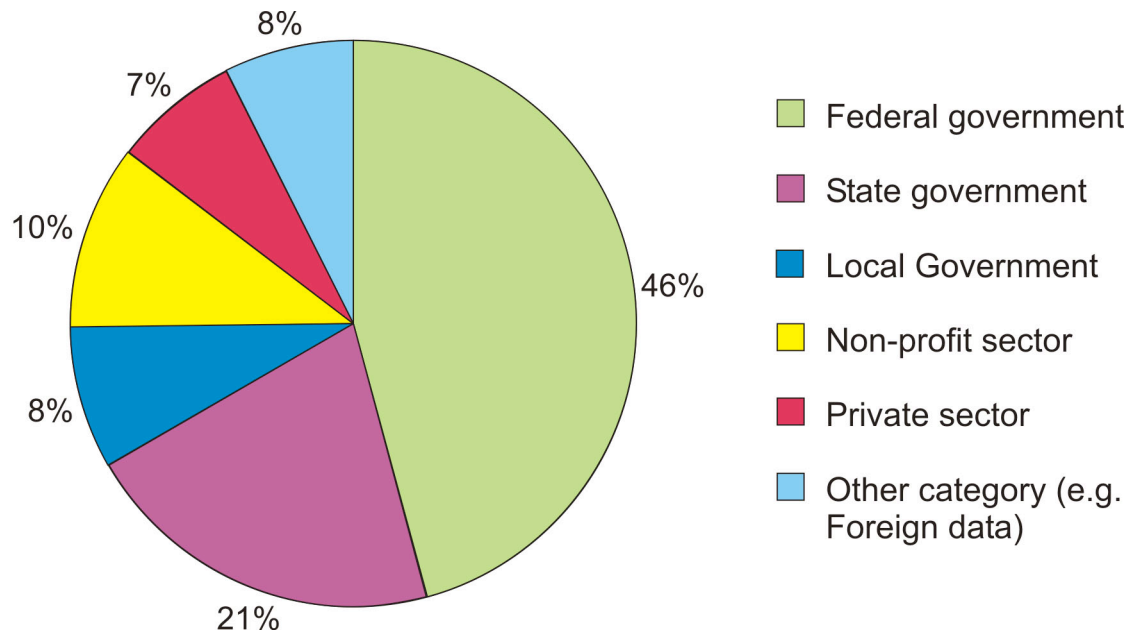


Figure 4. Sources of geographic datasets used in academic research (percentage of total of 290 datasets).

Analysis of Selected Principles and Discussion of Results

In this section we present results in regard to two principles for which the statistical procedures evidenced significant findings. Both principles focus on access to and use of government data by academic researchers. Of the 217 government data sets reported, 133 were federal datasets, 60 were state data sets, and 24 were local government datasets. In order for a data set to be categorized as a government data set, a respondent had to answer in the affirmative also to the question, “Was all or a substantial portion of this data set or database originally developed by a government agency using exclusively or primarily public funds?” (See also question A in the Appendix.) Based on the responses to this question, the total number of government data sets that were used in the statistical tests was 196. We tested the group of data sets “adhering to” against the group of data sets “not adhering to” for each recommended principle for government data sets.

Principle: "Adherence to Pricing at Marginal Cost or Less"

One of the most prevalent access issues discussed in the GIS literature has been the sale of government agency data (Rhind 1992; Litman 1994; Perritt 1996). The value of (geographic) data comes from their use, and the argument often is made that restricting access to government data by asking high (market) prices discourages the effective use of such data within the academic research community. It is assumed that the higher the price of the data, the less they will be used, and the less value the data set will have in respect to advancing knowledge. Throughout the 1990s many

discussions focused on this issue (Onsrud 1992a; 1992b; Weiss and Backlund 1997). The hypothesis tested in this research was:

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.

T-test

A measure of “adherence to pricing at marginal costs or less” was established through an analysis of the question: “What did you pay for access to or a copy of the dataset?” (See also question B in the Appendix). The distribution of responses to this question is shown in Figure 5. The highest ranking of adherence to the principle was assigned to those data sets receiving the following responses: the data set was free, the price was based on a minimal statutory fee, and the price was based on the cost of dissemination to the user. The lowest ranking of adherence to the principle was assigned to data sets receiving the following responses: the price was based on partial or full cost recovery, market price less a discount for the not-for-profit user, and market price. Of the total of 196 government data sets, 171 qualified as adhering to the principle while 23 qualified as not adhering to the principle. For two data sets, the question concerning the price of the data set was not filled out. Thus, a total of 194 data sets were used for the statistical analyses of this principle.

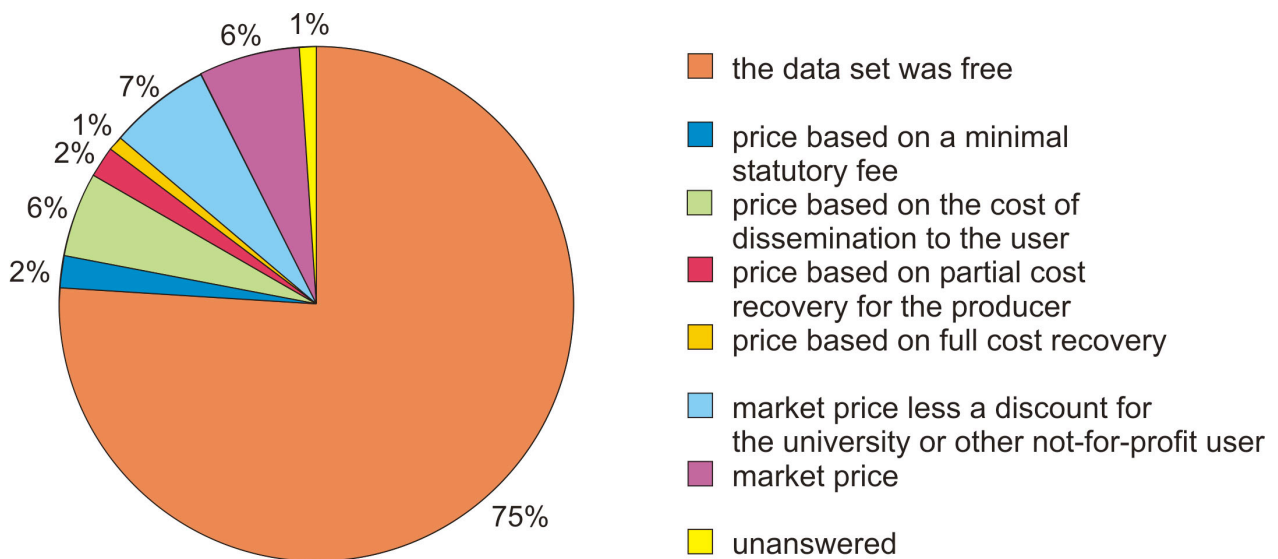


Figure 5. Price respondents paid for data.

Note: Figure includes all sources, i.e. government sources, private sources, non-profit sources and other sources.

We statistically tested whether data sets adhering to the principle contributed to a more productive research environment than did data sets not adhering to the principle. The t-test provided conflicting results for the different measures of productivity (see Table 1). Respondents who acquired data sets at high costs were able to perform significantly (at a level of significance of 0.10) more tasks with the data set (Measure 3) than did respondents who accessed their data sets for marginal costs or less (t-value of 1.647). However, respondents who acquired data sets at marginal costs or less were significantly (at a 0.20 level of significance) more satisfied (Measure 4, t-value of 1.432) and accomplished significantly (at a level of significance of 0.10) more overall objectives (Measure 5, t-value of 1.863).

Productivity Measure	Task Accomplishment		Satisfaction		Overall Objective Accomplishment	
	Yes	No	Yes	No	Yes	No
Data set acquired at marginal costs or less?						
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

Table 1. T-test for adherence to marginal cost or less.

These mixed results suggest a range of possible explanations. While respondents using expensive data were able to perform more tasks, those using free or low cost data appear to have been encumbered with far fewer limitations in pursuing their ultimate goals. Another explanation may be in the expectations of the recipients of the data sets. The (potential) user may have very low expectations for inexpensive data sets and, therefore, almost any contribution to the research may satisfy the researcher. In contrast, the expectation for more expensive data sets to contribute to research objectives may be higher. If no contribution is evident, or is less than expected, satisfaction with the more expensive data set may diminish. Finally, the expensive data sets may be data sets that were made fit-for-use on request by a particular researcher. The data sets that were acquired free or at low cost may be data sets that were not modified. The specific reasons behind these differences could be best determined through case study research.

Chi-square Test

We also performed a chi-square test. We asked respondents to indicate which factors contributed to the successful use of their data set (Measure 1) and which impeded the use of their data set (Measure 2). (Questions E and F in the Appendix provide the specific questions and the response options.) For both groups, the category “Cost of the data set was a significant impediment in the use of the data set” only included respectively two and one score. As a result, the cell frequency requirement (minimum of 5 scores per category) was not satisfied. We combined the categories “Cost of the data set was a significant impediment in the use of the data set” and “Cost of the data set was not a significant success factor or a significant impediment in the use of the data set” to satisfy the requirement. Table 2 gives the test results. The two groups were not uniform (at the very low 0.90 level of significance) in responding to whether “the cost of the data set” contributed to or impeded their success in its use. We concluded therefore that the chi-square test did not provide sufficient evidence that the two groups are significantly not uniform.

Data set acquired at marginal costs or less?	Yes	No	Total
Cost of the data set was a significant factor in allowing the successful use of the data set	77	12	89
Cost was no success or a significant impediment in the use of the data set	94	11	105
Total	171	23	194
Chi square value df=1	0.42		

The critical value at one degree of freedom is: 0.0158 at the 0.90 level of significance.

Table 2. Chi-square test for adherence to marginal cost or less.

The chi square test suggests that the price of a data set does not necessarily impact the productivity of the academic researcher. Yet, the measure of success in the chi-square test focused on successful “use” of the data set. The scientists who responded that they were able to use geographic data productively had, of necessity, to be able to first overcome the hurdle of acquiring the data. Cost may thus not influence substantially the successful use of a data set, assuming that the user is able to overcome the economic hurdle in acquiring the data. For instance, researchers who had included the cost of data in their research proposal budgets and received funding for the proposal likely would not view the high cost of data as an impediment.

Discussion of Results

The conflicting statistical information and the lack of background information that may have allowed us to explain the results forces us to conclude that although this research suggests that the price of geographic data may not affect the ultimate productivity of an academic researcher, additional research would be required to understand the limits of such a conclusion. Alternative research methods such as interviews and case studies would appear to be eminently suitable for this type of research.

Most of the respondents in our study accessed geographic data at low cost (see Figure 5). One of the impediments mentioned in the introduction—revenue generation from the sales of data by local governments—currently does not seem to cause a substantial problem for academic researchers in the U.S. Figure 6 shows that 87 percent of the data sets coming from local government were obtained at marginal cost or less. A substantial number of geographic data sets generated by the private sector also appear to have been made available to academic researchers at the cost of dissemination or less (40 percent)—perhaps as a contribution to academic research by the private sector.

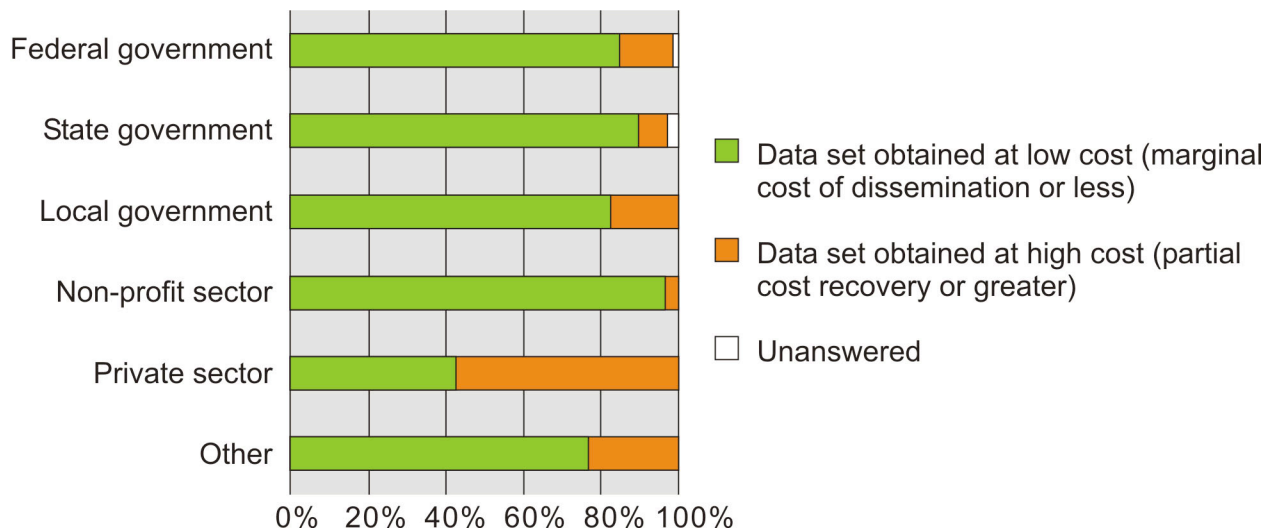


Figure 6. Cost of the dataset per source (in percentages).

Principle: "Adherence to Metadata Availability"

Metadata are data about data, telling the user where the data are located, how the data were collected and maintained and by whom, how the data can be accessed, and what are their characteristics (McLaughlin and Nichols 1994, p. 71). The major uses of metadata are to:

- Help organize and maintain an organization's internal investment in spatial data;
- Provide information about an organization's data holdings to data catalogues, clearinghouses, and brokerages; and
- Provide information to process and interpret data received through a transfer from an external source (FGDC 1997).

Metadata can eliminate barriers in the use of scientific and technical data. They are one of the key components in the Federal Geographic Data Committee (FGDC) strategy to develop the U.S. National Spatial Data Infrastructure, as evidenced by a statement made in a 2000 report: "If you think the cost of metadata production is too high, you haven't compiled the costs of not creating metadata—loss of information with staff changes, data redundancy, data conflicts, liability, misapplications, and decisions based upon poorly documented data" (FGDC 2000).

The recent Infrastructure for Spatial Information in Europe (INSPIRE) initiative also considers the documentation of metadata and the access to metadata as one of the critical components in developing the European Spatial Data Infrastructure (INSPIRE 2003). However, information system managers may regard the additional costs of cleaning up and documenting the data they collect so that they can be shared with others as outweighing the benefits to be obtained by gaining access to other data sets (Masser and Ottens 1999, p. 37). Harvey (2001, p. 37), for example,

found that local government suppliers of geographic data do not always recognize the documentation of metadata as being of significant importance. The hypothesis we tested was:

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

Although this hypothesis was primarily applied to government data sets, we also obtained indications of the quality of data set documentation in all the respondent sectors.

The responses to the question, “Which of the following did the documentation of the data set (digital catalogue files or metadata) help you accomplish?,” provided us with background information on the documentation of a data set (see also question C in the Appendix). Positive responses that at least three of the following features were addressed in the documentation of the data provided a measure of the sufficiency of the metadata:

1. Technical suitability of the data set;
2. Quality/accuracy of the data set;
3. Timeliness of the data;
4. Relevance of the data set;
5. Contractual restrictions or other legal constraints to the use of the data sets; or
6. Allows users to find the data set through a computer search.

The sufficiency level was determined through statistical testing. We found that for datasets with three or more positive metadata responses, the results were similar or better than those determined by the overall t-test of adequate versus inadequate metadata documentation.

T-test

One of the measures of availability of adequate metadata was obtained through an analysis of the following question: “How good was the documentation regarding the data set?” The response options were: excellent, good, fair, poor, and non-existent (see also question D in the Appendix). The distribution of responses across sectors is shown in Figure 7. Data sets receiving a response of excellent or good (109 data sets) were deemed to be in accordance with the stated principle, while those receiving a response of fair, poor, or non-existent (84 data sets) were deemed to be in non-conformance. For three data sets the question about the documentation was not completed. Thus, a total of 193 data sets were used for the statistical analyses of this principle.

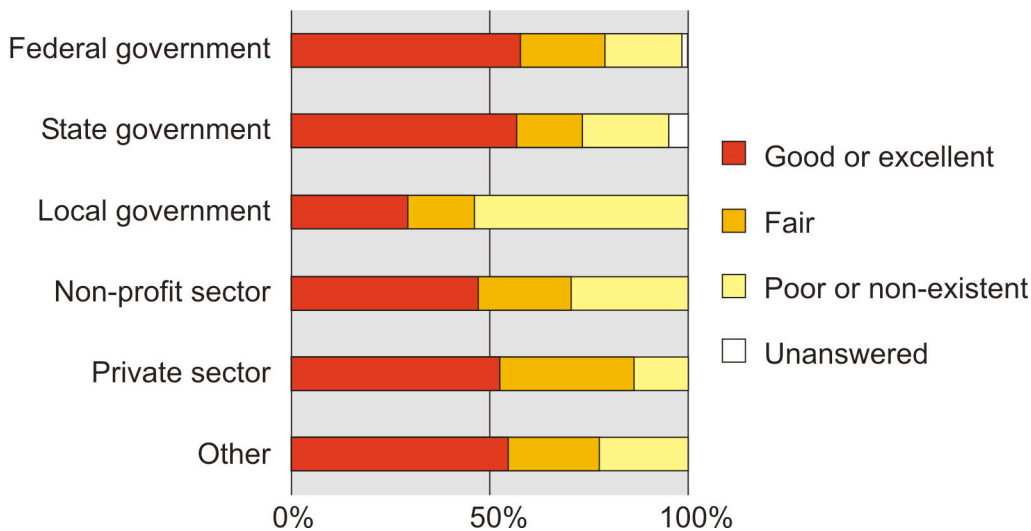


Figure 7. Quality of the documentation per source.

Table 3 provides the key values of the t-test. The t-test showed that academics using data sets with adequate documentation are significantly more productive than academics using data sets with inadequate documentation. The t-test showed that for the measures task accomplishment (t-value of 3.935) and satisfaction (t-value of 7.090) the productivity of academics using data sets with adequate documentation was significantly higher (at a 0.001 level of significance). Further, the t-test showed that data sets with adequate documentation allow significantly more overall

objectives to be accomplished (at a level of significance of 0.01) than do data sets with inadequate documentation (t-value of 3.176). There is thus a strong indication that data sets with adequate documentation are more productive for academic researchers than data sets with inadequate documentation.

Productivity Measure	Task Accomplishment		Satisfaction		Overall Objective Accomplishment	
	Yes	No	Yes	No	Yes	No
Data set with adequate documentation?						
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.

Table 3. T-test for level of metadata availability.

Chi-square Test

The measure of success in the chi-square test was the response “adequate documentation or metadata for this data set,” given to the question: “Which of the following, if any, were significant factors in allowing you to successfully use this data set?” (see also question E in the Appendix). Answers to the question “Which of the following, if any, were significant impediments to your use of this data set?” which included the statement “inadequate documentation or metadata for the data set” were used as the measure (see also question F in the Appendix). Table 4 presents the results of the chi-square test.

Data set with adequate documentation?	Yes	No	Total
Adequate documentation or metadata for this data set was a significant factor in allowing the successful use of the data set	56	6	62
Documentation or metadata was no success factor or impediment in the use of the data set	43	51	94
Inadequate documentation or metadata for this data set was a significant impediment to the use of the data set	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.

Table 4. Chi-square test for level of metadata availability.

The two groups were significantly non-uniform (at a 0.001 level of significance). The group of data sets with adequate documentation or metadata scored 51 percent for the success measure while the group with inadequate documentation scored only 7 percent for the success measure. The data sets with adequate documentation or metadata also scored better on the impediments measure: 9 percent versus 32 percent. The chi-square test confirmed the findings of the t-test suggesting that the availability of adequate documentation or metadata allows significantly more success in the use of a data set.

Summary Assessment of the Principle "Adherence to Metadata Availability"

The tests provide evidence that academic users of spatial data highly value their adequate documentation with metadata. Although the study only included the experiences of academic users of geographic data, the results may be valid for most other users as well. If so, this research suggests that suppliers of geographic data may attract more users generally if their data are documented adequately with metadata. Adequate metadata records, in addition to

benefiting suppliers and users of geographic data, are the “backbone” of catalogues used in establishing spatial data infrastructures and geospatial digital libraries. The documentation of metadata is, however, still costly. The use of existing metadata documentation utilities, such as ESRI’s ArcCatalog or Geodan’s GeoKey, and emerging documentation capabilities may reduce these costs.

Conclusions

This study explored current access policies imposed on researchers in U.S. universities using geographic scientific and technical data. Because a broad spectrum of disciplines use geographic data in scientific research, we suspect that the results obtained in this study may be *indicative* of user responses across many research data domains.

Our research suggests that cost alone is probably not a good indicator of the likelihood of an academic researcher to productively use geographic data. Additional factors likely affect whether cost matters in affecting the productive use of geographic data. Further research is warranted in exploring why most U.S. researchers choose to use low-cost geographic data sets and the additional factors affecting success as correlated with cost levels.

This research provided evidence that academic users of geographic data highly value the existence of metadata. Moreover, the research showed that the productivity of the academic researcher with a particular data set, as measured by task accomplishment with the data set, satisfaction with the data set and overall objective accomplishment with the data set, is positively correlated with the existence of metadata. Statistically significant results support the position held by FGDC that documentation of metadata should be a high priority in advancing spatial data infrastructures. Determining the specific utility of metadata and which constituent components are most critical would require further research.

ACKNOWLEDGMENTS

The authors acknowledge the financial support of the U.S. National Science Foundation (Grant # SBR-9700465) and the VSB-foundation in the Netherlands (<http://www.vsbfonds.nl>).

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). *WIRED* 2(3): 84-90, 126-129.
- Copyright Act. 17 U.S.C. (<http://www4.law.cornell.edu/uscode/17/>).
- FGDC (Federal Geographic Data Committee). 1997. Metadata standards development: Content Standard for Digital Geospatial Metadata, Version 1.0. (<http://www.fgdc.gov/publications/documents/metadata/metav1-0.html#1.2>).
- FGDC (Federal Geographic Data Committee). 2000. Ten most common metadata errors. FGDC Metadata Education Program. (<http://www.fgdc.gov/metadata/top10metadataerrors.pdf>).
- FOIA (Freedom of Information Act). 5 U.S.C. [[section]] 552. (<http://www.nara.gov/fedreg/legal/apa/552.html>).
- Harvey, Francis. 2001. U.S. National Spatial Data Infrastructure: The local government perspective. *GIM International* 15(3): 36-9.
- HR 354 U.S. 106th Congress. 1999. To amend title 17, United States Code, to provide protection for certain collections of information (the "Collections of Information Antipiracy Act"), January 19.
- HR 1858 U.S. 106th Congress. 1999. To promote electronic commerce through improved access for consumers to electronic databases, including securities market information databases (Consumer and Investor Access to Information Act of 1999), May 19.
- ICSU/CODATA. 1997. Position paper on access to databases. ICSU/CODATA Group on Data and Information. (http://www.codata.org/data_access/wipo.pdf). 20p.
- ICSU/CODATA. 1998. Responses to WIPO survey on database protection. ICSU/CODATA Group on Data and Information. (http://www.codata.org/data_access/position.pdf). 10p.
- INSPIRE. 2003. Consultation paper on a forthcoming EU Legal Initiative on Spatial Information for Community Policy-making and Implementation. INSPIRE: Infrastructure for Spatial Information in Europe. (<http://inspire.jrc.it/reports/INSPIRE-InternetConsultationPhaseII.pdf>).
- Litman, Jessica. 1994. Rights in government-generated data. In: *Proceedings of the Conference on Law and Information Policy for Spatial Databases*. Tempe, Arizona: NCGIA & Center for the Study of Law, Science and Technology. (<http://www.spatial.maine.edu/tempe/litman.html>).
- Lopez, Xavier R. 1998. *The dissemination of spatial data: A North American–European comparative study on the impact of government information policy*. London, U.K.: Ablex.
- Masser, Ian, and Henk Ottens. 1999. Urban planning and geographic information systems. In: John Stillwell, S. Geertman and S. Openshaw (eds.), *Geographical Information and Planning*. Berlin.: Springer-Verlay. pp. 25-42.

- Maurer, S., P.B. Hugenholtz, and H. Onsrud. 2001. Europe's database experiment. *Science* 294: 789-90.
- McLaughlin, John, and Sue Nichols. 1994. Developing a National Spatial Data Infrastructure. *Journal of Surveying Engineering* 120(2): 64-76.
- Nelkin, Dorothy. 1984. Science as intellectual property: Who controls scientific research? *AAAS series on Issues in Science and Technology*. New York, New York: MacMillan Publishing Company.
- NRC (National Research Council). 1995a. *On the full and open exchange of scientific data*. Committee on Geophysical and Environmental Data. Washington D.C., U.S.: National Academy Press.
- NRC (National Research Council). 1995b. *Preserving the scientific data on our physical universe: A new strategy for archiving the Nation's scientific information resources*. Commission on Physical Sciences, Mathematics and Applications. Washington D.C., U.S.: National Academy Press.
- NRC (National Research Council). 1997. *Bits of power: Issues in global access to scientific data*. Committee on Issues in the Transborder Flow of Scientific Data. U.S. National Committee for CODATA, Commission on Physical Sciences, Mathematics, and Applications. Washington D.C., U.S.: National Academy Press. (<http://books.nap.edu/catalog/5504.html>).
- NRC (National Research Council). 1999a. *A question of balance: Private rights and the public interest in scientific and technical databases*. Commission on Physical Sciences, Mathematics and Applications. Washington D.C., U.S.: National Academy Press. (<http://books.nap.edu/catalog/9692.html>).
- NRC (National Research Council). 1999b. *Distributed geolibraries: Spatial information resources*. Mapping Science Committee. Washington D.C., U.S.: National Academy Press. (<http://books.nap.edu/catalog/9460.html>).
- NRC (National Research Council). 2000. *The digital dilemma: Intellectual property in the information age*. Committee on Intellectual Property Rights and the Emerging Information Infrastructure. Washington D.C., U.S.: National Academy Press. (<http://books.nap.edu/catalog/9601.html>).
- 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* 275(1): 80-85.
- Onsrud, H.J. 1992a. In support of cost recovery for publicly held geographic information. *GIS Law* 1(2): 1-7.
- Onsrud, H.J. 1992b. In support of open access for publicly held geographic information. *GIS Law* 1(1): 3-6.
- Onsrud, H.J., and X. Lopez. 1998. Intellectual property rights in disseminating geographic data, products, and services: Conflicts and commonalities among European Union and United States approaches. In: P. Burrough and I. Masser (eds), *European Geographic Information Infrastructures: Opportunities and Pitfalls*. London, U.K.: Taylor & Francis, pp. 153-167.
- Paperwork Reduction Act. 1995. 44 U.S.C. 3501 et seq. (<http://www4.law.cornell.edu/uscode/44/3501.html>).
- Perritt, Henry H. Jr. 1996. Public policy for basic GIS data. *GIS Law* 3(2): 20-25.
- Perritt, Henry H. Jr. 1999. *Law and the information superhighway*. Gaithersburg and New York, U.S.: Aspen Law and Business.
- Pira International Ltd. 2000. Commercial exploitation of Europe's public sector information. Luxembourg, Luxembourg: Office of Official Publications of the European Communities.
- Pluijmers, Y., and P. Weiss. 2002. Borders in cyberspace: Conflicting public sector information policies and their economic impacts. (http://205.156.54.206/2p/Borders_report.pdf).
- Reichman J. H., and J.H. Franklin. 1999. Privately legislated intellectual property rights: Reconciling freedom of contract with public good uses of information. *University of Pennsylvania Law Review* 147(4): 875-970.
- Reichman, Jerome H., and Pamela Samuelson. 1997. Intellectual property rights in data. *Vanderbilt Law Review* 50(51): 51-166.
- Reichman, J.H., and P. Uhler. 1999. Database protection at the crossroads: Recent developments and their impact on science and technology. *Berkeley Law Journal* 14(2): 793-838.
- Rhind, David. 1992. Data access, charging and copyright and their implications for geographical information systems. *Int. J. Geographical Information Systems* 6(1): 13-30.
- Sirkin, R. Mark. 1995. *Statistics for the social sciences*. Thousand Oaks, California: Sage Publications.
- Van Loenen, B. 2001. Access to scientific and technical data in an academic setting. M.Sc. thesis, University of Maine.
- Varian, Hal R. 1995. The information economy. How much will two bits be worth in the digital market place? *Scientific American* (September): 200-201.
- Weiss, Peter, and Peter Backlund. 1997. International information policy in conflict: Open and unrestricted access versus government commercialization. In: Brian Kahin and Charles Nesson (eds), *Borders in Cyberspace: Information Policy and Global Information Infrastructure*. Boston, Massachusetts: MIT Press.
- Zikmund, William G. 1991. *Business research methods*. 3rd ed. Chicago, Illinois: Dryden Press.

Appendix: Relevant Questions from the Questionnaire

This appendix contains the questions from the online questionnaire that were used for the analysis of the principles discussed in the article. The entire questionnaire may be found at http://www.geo.tudelft.nl/gigb/personen/loenenvan/Access_to_Data_Survey.htm.

A. Was all or a substantial portion of this data set or database originally developed by a government agency using exclusively or primarily public funds?

Answer option in questionnaire:	Classification we used in the research
Yes	Government data
No	No government data
Do not know	No government data

B. What did you pay for access to or a copy of the data set?

Answer option in questionnaire:	Classification of answer with regard to "Adhering to the principle of data available at marginal cost or less"
Not applicable, the data set was free	Yes
Market price	No
Market price less a discount for the university or other not-for-profit user	No
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)	No
Price based on partial cost recovery for the producer	No
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)	Yes
Price based on a minimal statutory fee	Yes

C. Which of the following did the documentation of the data set (digital catalogue files or metadata) help you accomplish? (Mark all that apply.)

Answer option in questionnaire:

- Allowed us to find the data set through a computer search
- Allowed us to assess the relevance of the data set for our research project (e.g. data type, description entities)
- Allowed us to assess the technical suitability of the data set (e.g. data structure)
- Allowed us to assess the quality or accuracy of the data set
- Allowed us to assess the timeliness of the data set for our purposes
- Allowed us to assess contractual or other legal constraints on the use of the data set
- Not applicable, no documentation or metadata was available

D. How good was the documentation regarding the data set?

Answer option in questionnaire:	Classification of answer with regard to "Adhering to principle of adequate documentation"
Excellent	Yes
Good	Yes
Fair	No
Poor	No
Non-existent	No

E. Which of the following, if any, were significant factors in allowing you to successfully use this data set? (Mark all that apply.)

Answer option in questionnaire:

- The physical means for gaining access to this data set
- Availability of a search capability allowing the ability to find this data set or database
- Adequate documentation or metadata for this data set
- Sufficient identification of the sources used to create this data set
- Suitable format or compatibility with the software or hardware we used
- Sufficient quality or accuracy of this data set for our purposes
- Timeliness of this data set for our purposes
- Personal or institutional willingness to giving us access within the organization that created the data set
- Lack of application of copyright law to our uses of this data set
- Lack of application of specific data protection legislation to our uses of this data set (e.g. local ordinance, state statute, federal statute)
- Cost of this data set
- Contractual provisions facilitating our uses of this data set
- Contractual provisions regarding further dissemination of this data set
- Contractual provisions regarding liability
- Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of this data set
- Other, please specify

F. Which of the following, if any, were significant impediments to your use of this data set? (mark all that apply)

Answer option in questionnaire:

- The physical means for gaining access to the data set
- Lack of a search capability allowing the ability to find the data set or database
- Inadequate documentation or metadata for the data set
- Lack of identification of the sources used to create this data set
- Lack of suitable format or compatibility with the software or hardware we used
- Inadequate quality or accuracy of the data set for our purposes
- Timeliness of the data set for our purposes
- Personal or institutional resistance to giving us access within the organization that created the data set
- Restrictions imposed on our use of the data set by copyright law
- Restrictions imposed on our use of the data set by specific data protection legislation (e.g. local ordinance, state statute, federal statute)
- Lack of alternative data sets meeting our needs
- Cost of the data set
- Contractual restrictions imposed on our uses of the data set
- Contractual restrictions regarding further dissemination of the data set
- Contractual provisions regarding liability
- Contractual provisions granting the data supplier certain rights in information, products, or intellectual works arising through our use of the data set
- Other , please specify

G. Even though contractual, legal, technical and other impediments may have constrained your use of the specific data set, to what degree were you able to accomplish research tasks that were dependent upon use of this data set?

Answer option in questionnaire:	Score used in statistical t-test
Almost all research tasks dependent on this data set were accomplished	5
Most research tasks dependent on this data set were accomplished	4
About half of the research tasks dependent on this data set were accomplished	3
Some of the research tasks dependent on this data set were accomplished	2
Almost none of the research tasks dependent on this data set were accomplished	1

H. How would you rate your satisfaction with your use of this specific data set or database?

Answer option in questionnaire:	Score used in statistical t-test
Excellent	5
Good	4
Fair	3
Poor	2
Non-existent	1

I. Use of this specific data set was important in accomplishing the overall objectives of the research project

Answer option in questionnaire:	Score used in statistical t-test
Strongly agree	5
Agree	4
Disagree	3
Strongly disagree	2
Do not know / no opinion	1

Bastiaan van Loenen is a Ph.D. student at Delft University of Technology, the Netherlands. E-mail: <b.v.loenen@otb.tudelft.nl>. **Harlan Onsrud** is Professor of Spatial Information Science and Engineering at the University of Maine. E-mail: <onsrud@spatial.maine.edu>.