Building Longevity into the Design of a Historical Geographic Information System:
The Atlas of Historical County Boundaries

John H. Long

Social Science History Association Meeting, Chicago, 16 November 2001, and
Pacific Rim Neighborhood Interim Conference, Guadalajara, 4 December 2001

When we planned the initial products of the County Boundary Project—printed volumes of maps and text detailing all changes in the locations, sizes, shapes, names, and organization of every county in the United States from the seventeenth century to the twenty-first—we could confidently assure our sponsors and funding agencies that the fruits of our efforts and their generosity would serve users for several generations. The shelf lives of books constructed with chemically neutral papers and inks and kept in a safe, benign environment is measured in centuries, and we expected no shorter lifespan for our work.

Book publication is now a low priority, and digital files distributed via The Newberry Library's web site (and possibly some other medium) have taken the place of books as our primary product. An electronic version of our data is very different from a book, but we need not give up the goal of long life. Although we cannot be sure of digital data's longevity, we should aim for at least 100 years.

Books and databases are not the same things, of course, even when their content is identical. Probably nowhere is the difference between those two forms of information transmission and storage more evident than in their projected lifespans and the factors that affect their effective lives. Just look at the media and the environments, for example. The predicted useful life of storage media like Zip disks and compact disks (CDs) is measured in decades. More serious, history indicates that the life expectancy of any computer operating system, software, or format is probably only a decade or so. That acid-free book can sit on a shelf for a long, long time with no ill effects, but most computer-experts warn that storing digital records for an extended period will require more active maintenance, probably a measure of personal attention from the conservator.[1]

As we convert from a methodology that required a lot of manual drawing to a fully computerized system, and as we produce our first all-digital product, we are trying to minimize and to put off for as long as possible the problems that will confront the future stewards of our data. How do we expect to do that? For ease of analysis and presentation, let's break the answer into three parts. First is the matter of form, particularly the nature of electronic systems and media. Second is documentation. Third is relations with other organizations and individuals engaged in similar or related work, now and in the future.

Form

Let us return to the media and environments in which digital data will be saved and stored. Because of the relatively brief lives of software—chiefly programs, and operating systems—one measure we will take to insure the long-term preservation of the data will be to print hard copies of the maps and text and to store them in libraries like the Newberry and the Library of Congress. Obviously, such printouts will be much like the printed books we produced in the 1990s. Because they will emerge from the database, however, they will be different in detail from the volumes (e.g., one map per screen or page versus many instances of several maps per page in the books). They could be consulted just as the books can be, but their primary purpose will be to preserve the data. Should someone discover that the compilation for a state suffers from an error or omission, an addition or correction can be written directly on the paper copies, and a brief annotation explaining the change could be added to the documentation. Such changes would also be made to the electronic version, of course, but updating the paper versions of the data means the essential data would continue to be there to recreate the digital files, should they be lost or somehow rendered useless.

Another important step is utilizing a format not dependent on particular hardware, software, or operating system. It has been pointed out to me that one format has proved remarkably stable and durable: ASCII. Therefore, we will try to capitalize on that and make a copy of each set of data in ASCII.

Within the universe of digital media, we shall try to store the data in the form that promises the greatest longevity, as CDs do today. We cannot predict future operating systems, so we will work within the most widely used one available. At this time, that appears to be Microsoft Windows. The choice of application software also must include allowance for potential longevity. We presume that programs that are flexible, widely used, and backed by a solid company are most likely to last a long time. Our choices are ArcView, a geographic information system (GIS) from Environmental Systems Research Institute or ESRI, for maps and Oracle or Microsoft's Access for non-map data.

Some of you may be wondering about that reference to two different database programs, not just the geographic information system. The central idea is that the design of the database also can contribute to longevity. We have chosen the model created by Larry Crissman for the China Historical Geographic Information System.[2] As he has pointed out, various GISs will accept Access data tables, and it is easier to add or integrate new, variant, or alternative data sets in Access, than in ArcView or other GIS. Crissman uses an off-the-shelf GIS package like ArcView for the specialized tasks of creating and managing maps or polygons, but does not ask it to do tasks for which it is not so well suited. First, ArcView simply cannot match a generalized database like Oracle or Access for dealing with dates, names, source citations, etc. Second, Access is easier to use, more readily available, and less expensive than ArcView or any other GIS, so people without the special training in

ArcView can operate it. In addition, if it becomes desirable to switch from one GIS to another, having the non-map data in an independent format will minimize problems.

Another feature of our design will be separate fields for the year, month, and day of each date, rather than a single date field that would hold all three elements of a date together. At first glance, the separate fields may appear to be unnecessarily complicated and space consuming. In practice, however, it is relatively easy to enter data into the several fields. Most important, that arrangement makes it easy to use our data with any number of GISs, regardless of what date format they require. Moreover, the addition of a fourth field will make it possible to provide modifiers of dates (e.g., the word circa or the abbreviation c. in front of an approximate date) without making it difficult for the program to sort the dates.

Another application of the idea that the better the design, the longer the product will last is in the design of the web site from which the data will be disseminated. Ease of access, clear directions and navigational tools, and accurate anticipation of what readers will want, both in content and in form, will be keys to success or failure. We expect to adopt some of the features in our books, such as the list of counties that actually functions as a county index that includes name changes, extinct counties, and unsuccessful proposals for new counties. Software that takes advantage of common facilities like point-and-click affect the convenience of the user and, therefore, also will contribute to longevity.

Documentation

Beyond the basic information about the names, dates, and maps of historical counties, there is additional content that I believe will increase the chance of our GIS enjoying a long life. Documentation is an obvious feature that will have great influence. In the books we detailed our methods, provided a long bibliography for each state, and cited the original authorities for each county creation, change, and extinction. That documentation will continue in the online database.

A second type of documentation must be added, of course. That is technical metadata, information about out data and the GIS. We will follow the standards of the Electronic Cultural Atlas Initiative and the metadata standards of the Federal Geographic Data Committee.[3] Other standards that must be met include those of our host institution, The Newberry Library, which will post the data on its web site.

In November 2001, at the annual meeting of the Social Science History Association in Chicago, a strong case was made for also documenting the design of a database, particularly its purpose and anticipated applications.

It is worth emphasizing that nothing undercuts the value and authority of a collection of data like the lack of adequate documentation. Once the original compilers have moved on and there is no one to vouch personally for the quality of a

dataset, the documentation is the only authoritative reference available to operators and users. Nearly everyone has heard a horror story about valuable data assembled through a long, difficult, and expensive compilation and then left undocumented. When the time eventually came that there was no person in the organization who could testify at first hand about the quality of the dataset, the lack of documentation doomed it. The data had to be judged untrustworthy and were discarded. Worse, in some of these horror stories the punch line is that the same data had to be compiled all over again. We intend to avoid that sort of disaster.

<center>Relationships with Others</center>

Any new geographic information system or dataset will be able to claim a more or less large audience for a relatively long time only to the extent that it is compatible with existing data that is employed by a wide range of researchers. The most obvious example is setting up a GIS that will accept digitized population statistics from past U.S. federal censuses. Compatibility with those data is the minimum standard of utility for a compilation of historical counties, and the key to that compatibility is the way the counties are identified.

Since the 1930s the U.S. Census Bureau and other agencies in both government and the private sector have employed FIPS codes to identify states and counties. FIPS stands for Federal Information Processing Standard, and the codes are five-digit numbers. The first two digits stand for the state, the next three for the county. For example, the FIPS code for Cook County, Illinois, is 17031. The initial 17 represents Illinois; within Illinois, Cook County's number is 031. Notice that the FIPS code implicitly refers to the county as a fixed institution, not as a geographic entity that could change size, shape, or location. The FIPS system has no provision for identifying the different geographic versions of the county. [4]

Today, computer memory and storage are plentiful and inexpensive, so we do not need to use compact, abstract coding systems, and we could simply identify each state and county clearly by name or abbreviation and each version of the county map by a number. Cook County, Illinois, for example, could be ILCook, and the map of Cook's second configuration could be ILCook2. For anyone working with the counties, this is easier to recognize and remember than the FIPS codes and it demonstrates how we probably will identify the maps or polygons. But we cannot ignore the FIPS codes because they are built into so many datasets. Therefore, we will construct a special table in Oracle or other relational database. That table will match the unique code for a polygon, such as ILCook2, to a combination of (1) a FIPS code for the county and (2) a pair of dates that demarcate the period when that particular county institution had that particular configuration and location. A researcher who wants to map data tagged by a county FIPS code simply must specify the FIPS code and the date of the data. The relational database program will find the county map/polygon whose FIPS code matches and whose beginning and ending dates embrace the date of the information. With equivalence established between our unique identifiers and the FIPS codes with dates, many researchers will be able

to utilize our historical counties for the analysis and display of those data. Of course, a separate field can be added to this table for any other coding system to coordinate it with our polygon/map identifiers in the same fashion.[5]

Clearly, we believe that as long as a many people and organizations use your datasets, those datasets will persist and enjoy long life. Another way in which we can pursue that goal is to establish cooperative relationships with other data providers and distributors. The Electronic Cultural Atlas Initiative (ECAI) has invited us to participate in its programs and has laid the foundation for possible publication of our data through the California Digital Library. We believe that other online data stores will be interested in using our data or in establishing links to our data. For example, that massive historical reference, the *New Handbook of Texas*, has an article on the history of each Texas county, all 254 of them, but there are no illustrations or maps.[6] The *Handbook* is now available in its entirety online, and it still has no maps. I have talked to at least one trustee about the prospect of establishing links that would make it easy for a reader of the *Handbook* to get to our Texas county maps and chronologies and vice versa, and I think something will come of that. The U.S. Census Bureau frequently receives requests for information about historical counties, and, according to conversations we held a couple of years ago, the Census would be interested in forwarding such inquiries to our web site. Of course, I'd be happy to hear from any other organization that might be interested in a cooperative link.

Conclusion

I have just described the principal ways in which we on the Atlas of Historical County Boundaries Project intend to provide a long life for our data in digital form. If anyone has any additional suggestions to offer us, I'd be pleased to hear them. Thanks for your attention.

---

[1] Jim Coates, "Jumping into Computerized Storage of Records? Not So Fast," *Chicago Tribune* (26 July 1999), sec. 4, p. 1; Media Sciences, Inc., "Do Gold CD-R Discs Have Better Longevity than Green Discs?" at www.msciences.com/faq53.html (19 Oct. 2001).

[2] Lawrence W. Crissman, "Draft Database Design and Geocoding System (Version 011/09/00)", paper presented at the International Workshop on Historical Geographic Information System, Fudan University, Shanghai, 24 august 2001.

[3] Information about the ECAI standard can be found at its web site at www.ecai.org. The "Content Standard for Digital Geospatial Metadata (CSDGM)" can be downloaded from the Federal Geographic Data Committee's web site at www.fdgc.gov.

[4] For a complete listing of the FIPS codes, including codes invented for counties that appear in historical census reports but that became extinct before the FIPS system was instituted, see Richard L. Forstall, comp. and ed., *Population of States and Counties of the United States: 1790 to 1990 from the Twenty-one Decennial Censuses* (Washington, D.C.: U.S. Department of Commerce, Bureau of the Census, March 1996).

[5] Given our goal/standard of absolute comprehensiveness—covering all changes in all counties—this approach is relatively safe. Should a compiler miss a change, a new polygon would be created. Revising and enlarging the relational database to reflect the new data would be a relatively easy task.

[6] *The New Handbook of Texas*, ed. Ron Tyler et al., 6 vols. (Austin: The Texas State Historical Association, 1996).

John H. Long, Editor
Atlas of Historical County Boundaries
The Newberry Library
60 W. Walton St.
Chicago, IL 60610
Tel: 312-255-3602
FAX: 312-255-3696
Email: longj@newberry.org
URL: www.newberry.org/ahcbp