BARBARA P. BUTTENFIELD

Geographic Information Systems and Digital Libraries: Issues of Size and Scalability

The term "scalability" has specific connotations in Geographic Information Systems (GIS) that conventionally relate to monitoring and predicting growth of geographic phenomena. A family of computational models has been developed to predict changes in structure associated with changes in size. These models have been applied in physical science, social science, and cartographic science to study growth and may assist in monitoring the growth of digital libraries as well. As the size of the digital library increases, challenges for data organization and collection maintenance tasks will also increase. However, the rate of increase may not be in linear proportion to library size. At some critical scales, existing procedures will fail and new procedures must be implemented to accommodate further growth. Allometric principles may be applied to estimate these critical scales. Three aspects (data volume, indexing, and metadata recordation) will be discussed in the context of implementing and maintaining a digital library containing spatial data archives distributed across local or global electronic networks.

THE SCALE OF GEOGRAPHIC INFORMATION

The nature of spatial data is that phenomena and processes take on varying appearances with variations in resolution at which they are measured and observed. For example, a high altitude, 79-meter resolution satellite image of a large city may display land-water distinctions and areas of vegetative cover distinguished from paved or built-up areas, while a lower altitude, 10-meter aerial photograph can be used to identify urban land use patterns and large landmarks. Geographic processes that are evident at small scales include geologic processes such as continental plate tectonics or global migration patterns. At larger scales, processes such as erosion and urban zoning constraints become evident. Environmental and atmospheric scientists are sensitive to the scales at which a phenomenon or process can be identified in a map or satellite image and choose their data sources accordingly.

To meet the needs of the environmental and atmospheric scientists, cartographers must preserve details that are necessary to identify geographic phenomena and processes when they change the scale of map data. Many think that changing map scale involves either enlarging or eliminating detail. This view is incorrect. The cartographic challenge is that new details are needed to preserve realism at larger scales. As one "zooms" in or out, geographic details may appear, then disappear, and sometimes reappear. Some feature domains (hydrography) change rapidly with changes in map scale, while others (transportation) stabilize at a particular level of detail. The scales at which changes occur differ from one domain to another and are not straightforward to predict due to variations in terrain, soil type, and other mediating factors. Much of the focus in current cartographic research centers on formalizing knowledge about what geographic details can be identified at particular data scales, and on developing computational and graphic procedures to meet user requirements for information whose appearance may change with scale. The challenge to our discipline is to create and maintain digital cartographic data sets that embody multiple representations of features and attributes, and to identify the ranges of scale for which data processing algorithms are effective. At the limits of these ranges, operating procedures must be changed to preserve geographic and visual logic in the database.

MEASURING THE SIZE AND SCALE OF LIBRARIES

It is intriguing for a cartographer to discover that library scientists face a similar challenge. The challenges in creating and maintaining a digital library include issues of interface design (Siegel, 1991), adaptation of existing procedures (Cohn et al., 1992; Weibel, 1992), and the new roles for library and information scientists arising with emerging technologies (Smith & Dalrymple, 1992). The challenge relating to size and scale is that as the library expands and matures, much of the collection maintenance must be modified to meet information demands and user expectations. In some cases, the larger size will require altering the contents of existing archives. It is likely that as more information becomes available, it becomes more difficult to access, retrieve, and catalog. This paradox becomes especially evident for digital libraries that are electronically networked to other distributed archives. The effectiveness of digital libraries will be based in part upon the ability of library scientists to "scale up" operating procedures. Scaling up cannot be accomplished by mimicry of existing procedures and often involves evolutionary or revolutionary change. For example, conversion from the Dewey Decimal cataloging scheme to Library of Congress Cuttering scheme in response to the increase in library holdings at sites across the country also introduced new methods of formal specification describing the contents of archived items (Molz, 1984).

This discussion will expand upon concepts of scale and growth as they are applied in a variety of natural and social science disciplines and present models by which scale progressions and their consequences may be anticipated. The objective of such study in other disciplines can facilitate information gathering or data analysis, or estimate consequences of predicted growth on large or complex phenomena, or improve management by foreseeing points of growth at which operating procedures must change. The scaling concepts are referred to in other disciplines as allometry, which is the study of changes in size that are accompanied by changes in form or structure. Allometry has been applied to biological evolution (Gould, 1966), architectural engineering (Bon, 1973), industrial management (Haire, 1973), urban planning (Woldenberg, 1971, 1973), and in many other disciplines. There is a long history of allometric modeling in GIS which began with early cartographic studies by Richardson (1961) and Mandelbrot (1967). Their work demonstrated that the length of coastlines and other map features tends to increase without apparent limit with finer units of measurement (see Figure 1). A recent survey by Lam and DeCola (1993) demonstrates the breadth of recent applications in GIS and geography.

Allometric study should also prove useful for management of digital libraries, particularly as it becomes clear that electronic information depositories will continue to come online and to grow even though it is currently difficult to predict just how large and intertwined these may become. Kemeny's (1962) projections of exponential library growth accept that holdings may be distributed in multiple branches, but his assumption that the library items take up physical space have been surpassed with the advent of electronic archival dissemination. A real challenge for those who monitor the growth of digital libraries lies in determining how to measure the size of holdings at any point in time.

CONCEPTS OF SCALE AND SCALE CHANGE

For the purpose of this presentation and the discussion that it may generate, let the reader accept the label "scalability" to encompass the range of issues associated with changing relative or absolute scale. Scalability takes the same linguistic root as the word scale. In cartography, scale is the ratio between distance on a map and distance on the earth, and customarily reported as a Representative Fraction (RF). To belabor the point, an RF value of 1:24,000 describes a map where one map unit equals 24,000 earth units or 1 inch to 2,000 feet. A more general definition implying increases of scale denotes "a succession or progression of steps or degrees; a graduated series (the scale of taxation, a social scale) or a point on such a scale" (*Random House Unabridged Dictionary of the English Language*). The last clause of the definition ("a point on such a scale") leads in many cases to considering scale and size interchangeably. A variety of

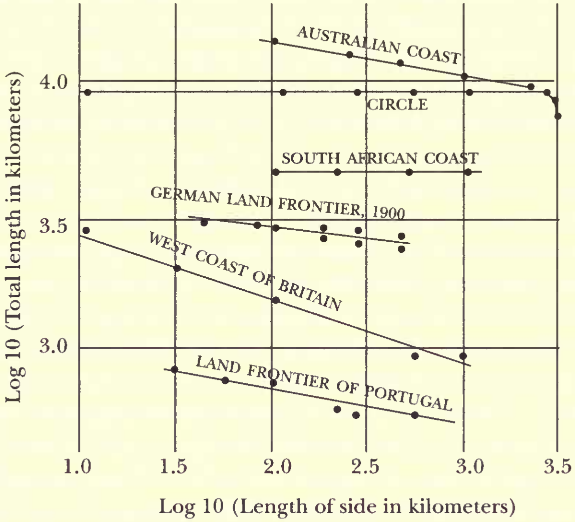

Figure 1. Increase in the length of cartographic lines measured at finer and finer resolution. (Redrawn from Richardson, 1961)

attributes are applied to quantify scale progressions. It is possible to create unusual cartographic transformations by intermingling scale units, as for example in scaling the size of countries in proportion to energy generation or consumption.

Scale measures may be applied to understand the size and growth of libraries and of digital libraries in particular. Numerous units of measurement might be formulated, for example, the number of items (books, journal series, map series) acquired per year, the number of items contained in these items (number of chapters, journal articles, or map sheets, respectively), the number of patrons, the number of branches if the library is distributed, etc. RF values might be developed to compare the acquisitions per library branch, the average number of items requested per patron per year, or the number of library staff in relation to the number of items cataloged. Clearly it is not the place for a cartographer to tell a librarian how to determine what measures of library size are appropriate. Likewise, telling the library community that digital library archives will tend to grow without apparent limit is a form of "preaching to the choir." The intention of this discussion is to outline quantifiable methods which may assist in estimating rates of digital library expansion, and adjusting operational procedures to accommodate increasing acquisitions and item retrieval.

A very powerful approach has been used in other scientific disciplines to model changes in scale. As it turns out, changes in gross structure (for example overall size) are in proportion with changes in substructures. To take a simple example, while the perimeter of a circle ($2\pi r$) increases linearly, its area (πr^2) increases geometrically. For very large circles, as the great and small circles of latitude circling the globe, the difference between the coefficients π and 2π becomes inconsequential with respect to the difference between r and r^2, and thus scalability is a factor of the radius alone. This knowledge enables cartographers to estimate linear distances on the sphere which vary in direct proportion to the cosine of latitude. What is important here is not the specific formula, but that these estimations can be made, and made much more simply than actually going to a place and surveying distances. In biology, the focus of attention is more often placed upon animal weight and animal height (or length). So, for example, a fish may be seen to double its weight in growing from four to five inches long, and knowing this a biologist may predict the weight of a very large fish simply by measuring its length and applying the scale ratio $W=kL^3$ (Thompson, 1977, p. 16). In this way, potential fishing yields can be estimated by measuring the length of schooling fish on air photographs.

The accuracy of such estimation is based of course on the assumption that the shape of the globe, or the shape of the fish at both sizes is equivalent or isometric. This is called the Principle of Similitude and can be applied to estimate multiscaled processes such as cohesion, chemical, electrical, and gravitational attraction at molecular and astronomical scales. The principle sounds pretty simple, and for geometric objects (like circles) it is. However, isometric relationships occurring in nature tend to hold true only within finite ranges of scale. At certain critical scales, the isometric model fails to generate an accurate estimate. Growth beyond these critical scales is associated with changes in form and proportion that are called allometric. Thompson (1977) demonstrates the principle for an engineering example:

> the strength of an iron girder obviously varies with the cross-section of its members, and each cross-section varies as the square of a linear dimension; but the weight of the whole structure varies as the cube of its linear dimensions. It follows at once that, if we build two bridges geometrically similar, the larger is the weaker of the two, and is so in the ratio of their linear dimensions. (p. 18)

And later in the passage, Thompson (1977) refers to Galileo's writings in the fifteenth century, and a comment that when building things at increasing scale, eventually

> beams and bolts would cease to hold together; nor can Nature grow a tree nor construct an animal beyond a certain size, while retaining the proportions and employing the materials which suffice in the case of a smaller structure. The thing will fall to pieces of its own weight unless we either change its relative proportions, which will cause it to become clumsy, monstrous or inefficient, or else we must find new material, harder and stronger than was used before. (p. 19)

Allometric models apply numeric power laws relating internal factors (growth, weight, or mass) acting within an organism to the external forces (environmental, gravitational, etc.) acting upon it. The approach focuses on apparently paradoxical changes in structure that accompany growth, evolution, and maturation. The premise in allometry is that as critical size thresholds are passed at certain points in the growth process, internal physiological changes take place to accommodate the increased external forces acting on the organism. Allometry describes changes in shape and form that accommodate necessary changes in physiology that occur with scale change. Inversely, by identifying the critical points where changes in form occur, one can predict when, in the growth process, the internal or physiological changes ought to take place.

In physics, chemistry, architecture, an so on, the "materials" referred to in the quotes above are physical materials such as gypsum, wood, and steel. In social and informational sciences (cartography, communication, library science), the "material" may be considered as a metaphorical reference to information content or structure. New structure is generated by reshaping the existing structure, whether through changes in organizational operating procedures, or selection of different computational algorithms. The parameters guiding the algorithms modify the form of the organization or the structure of the information.

The scale at which size change becomes allometric can be identified where the ratio changes between the subcomponent parts and the whole. If the rate of change exceeds the proportion for isometric conditions, one speaks of positive allometry. Otherwise, the allometric relationship is said to be negative. Allometries may be quantified: when plotted on a graph, isometric relations will display values in proportion one would expect given the scale ratio. Signed allometries will have slopes that are greater than or less than this expected value respectively (see Figure 2). In the Euclidean cross-tabulation mentioned above, each cell in the matrix is characterized by allometric relations of a fixed magnitude. Comparisons between volume allow extensions from consideration of a single organism to consider the growth of organizations. For example, Haire (1973) studied industrial firms, comparing the number of employees who dealt primarily with activities inside or outside the firm. The internal staff include personnel officers, for example, while the external staff are in marketing and purchasing. Haire (1973) argued:

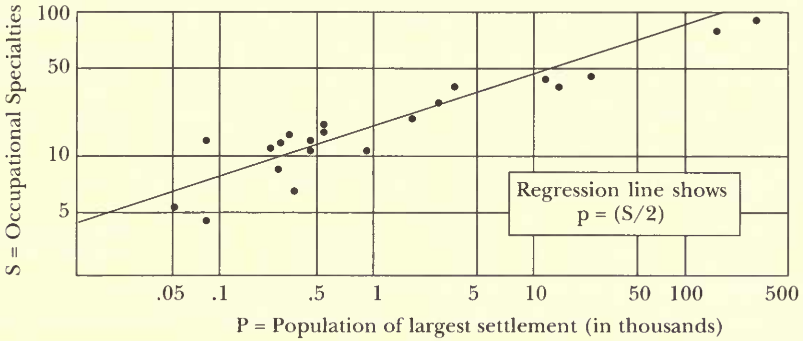

Figure 2. Relation between total number of occupational specialties and population of largest tribal settlement. (Redrawn from Naroll and Bertalanffy, 1956)

> As the organization grows, its internal shape must change. Additional functions of coordination, control and communication must be provided and supported by the same kind of force that previously supported an organization without these things....If each increment in size produced one increment of additional supportive function, there would be no limit. However, in the organism, the proportion of skeleton needed to support the mass grows faster than the mass itself and...hence comes to consume a disproportionate amount of the productive capacity of the organization. It becomes important to identify the skeletal support of the firm, the forces it resists, and the rates at which the support must grow. (p. 264)

Haire's data are shown in Figure 3. He plots the cube root of internal staff (assuming that the "mass" of an organization should increase volumetrically) against the square root of the external staff (assuming these to represent the "skeletal" support structure), and finds nearly linear relationships in every case. He remarks: "There is no immediately obvious organizational artifact that imposes this orderly progression on each of the growth patterns. It looks as if a geometry very similar to conventional spatial description can be used in picturing social bodies" (Haire, 1973, p. 265).

"SCALING UP" DIGITAL LIBRARIES

It is clear from the literature that allometric models can be used to formalize estimations of size and growth, and these models can inform organizational planning and policy. How can allometric models be generated that describe the growth of digital libraries? Three areas come to mind where measures of scale might be applied to monitor growth and change: data volume, indexing, and metadata recordation. The three

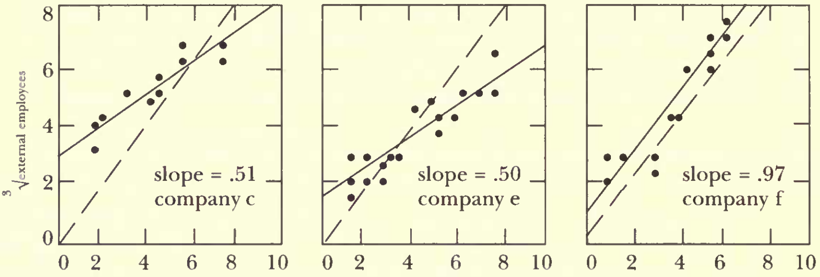

Figure 3. Allometry of Growth in Three Industrial Firms (dashed line indicates one-to-one ratio). Redrawn from Haire, 1973, p. 265.

areas are illuminated in the context of spatial data, which is the information customarily found within a Geographic Information System. GIS data are commonly troublesome for library archives in these three respects at least. To avoid these troublesome issues, most libraries separate regular (text) archives from (map and image) special collections. A truly workable digital library should eliminate distinctions between text and special archives (e.g., graphics, sound, video) or at least make such distinctions transparent to users. As the size of the digital library increases (this is the internal force or mass of the library), these issues are likely to become weak points in the supporting skeletal structure. Without proper adjustment, the sheer "weight" of the digital library will stress the overall library structure to the point where access and retrieval functions are likely to collapse.

Data Volume

As the number of items archived in a digital library increase, one can expect increasing strain on mass storage space, on data structures that organize the archives, and on the length of time required to retrieve any single item. One operational adjustment that would reduce overall storage volume is data compression. Storing maps and imagery in compressed form will reduce storage needs, and may shorten the time required for electronic data delivery. Many types of compression algorithms are designed to compress specific types of data, including static imagery (JPEG compression) or animated sequences (MPEG compression). The efficiency of the compression algorithm (total reduction in file size) is offset by consideration of how much information is lost when the file is reconstructed. Wavelet decomposition provides a hierarchical compression method characterized by little or no loss of original information, and this method should apply well for reducing the volume of GIS archives. The disadvantage of archiving data in compressed form is that, at present, it is not possible to search the content of a compressed file. One must decompress a satellite image file to determine its geographical content. If images could be searched prior to decompression, query response times would be reduced, especially if data archives were distributed across a network.

Indexing

For geographical data in particular, it is advantageous to organize maps or files into series. Journal and monograph series run in linear sequence (volume 1, volume 2, etc.). Map series require two or more dimensions for proper organization (the third dimension is needed to archive duplicate map sheets and/or multiple editions). Organizing maps into spatial files in the digital archive shortens access times for patron requests for information that is contained in abutting map sheets. Digital indexing schemes for efficient searching of a two-dimensional array include Peano curves, Quadtrees, and Morton indexing. Interface designs to facilitate user views on an archived map series are easy to design for single map series. Because patrons do not customarily request a series sheet name or reference number, but instead base their request on geographical place names or features, the user interface design must also display selected classes of geographic content in addition to the series footprints. However, as the digital archives incorporate additional series covering the same region, the user interface can become cluttered and unusable. These design issues must be considered before the library grows, and GIS technology (map overlay, feature buffering and selection) can be applied to assist the user in decomplicating interface designs for map indexing.

Metadata Recordation

Metadata are information about data. It includes, but is not limited to, the information customarily included in a MARC record. For digital geographical data, as of this calendar year, federal agencies distributing and exchanging spatial data are mandated to include metadata reports. The content of the metadata report is standardized first to determine what data exist, and second to determine fitness for use. Measures of data quality form the basis for determining fitness for use, which can be determined by empirical or deductive testing. Data quality measures include evaluation of positional and attribute accuracy, data completeness and currentness, and the logical consistency of underlying data structures. Third, as a digital library grows, and as its files increase in size, patrons should be able to browse a file's metadata records to learn in what data format, data structure, and data model a data set will be delivered. In the optimal setting, patrons should understand the consequences of requesting an actual dataset before they download it in the event it is especially large (satellite image sets can run into hundreds of megabytes in size, and these archives grow larger every year) or complex. Fourth, metadata should report a chronology of data processing steps, that is, a lineage of the data since its original distribution. For GIS data in particular, lineage information provides a real challenge to record data changes including data filtering, changes in projection, category aggregation or reselection, and so on. Should every processing change be recorded, one might anticipate metadata reports containing unlimited record lengths—plainly this is unrealistic, especially for data delivery over a network. The question of how to transport and exchange metadata must be resolved quickly since metadata records (and particularly lineage information) will grow with data use, regardless of the growth of the digital library as a whole. Patrons' metadata browsing behaviors are not well understood, which provides another area that needs research. What may become necessary is the implementation of query mechanisms for searches that have not been anticipated by system designers.

SUMMARY

This discussion presents issues of scale as a cartographer views them and reviews how a class of models geared toward the description and analysis of scale change have been applied in several disciplines, including natural and social science. Classes of allometric relations can be defined taxonomically in a matrix cross-tabulating one- , two- , and three-dimensional phenomena in Euclidean space. This presentation takes examples from a few of the possible combinations. One class describes the growth and size of linear phenomena (such as the length of coastlines) as often applied in cartography. A second class of models identifies a square-cube relationship in which external growth of an organism or an organization will falter due to weakened internal support structures that must be modified if external growth is to continue.

Allometric models appear to have relevance for adoption of digital library technology, particularly as it becomes clear that electronic information repositories will continue to come online and to grow even though it is currently difficult to predict just how large and intertwined they may become. Quantifiable examples are more appropriately formulated by the library and information science community, since these are the individuals who currently measure and monitor library size and growth rates. The discussion has covered three areas of growth for which a digital library of GIS information will become especially challenged during intensive growth phases. These are presented as examples to guide discussion and are not intended as an exhaustive list. It is up to the library community to select quantitative parameters for the models and to interpret them.

In closing, it is helpful to consider the perspectives from geography and library and information science alike.

> Allometric concepts and the analysis of relative growth may help to fill our current vacuum of ignorance concerning relevant norms of societal growth and change. As humanity takes conscious control of the planet which shaped the species, the analysis of relative growth can indicate what changes are possible, which are most likely, and to some degree, which may be desirable. By attending to changes in shape of our social organism, we may become more competent in shaping change. (Dutton, 1973, p. 306)

"Embedded in the public library movement is the belief that people have a fundamental right to know. No matter how rich you are, how old you are, where you come from, or where you call home, you have a right to both information and knowledge" (Bremmer, 1994, p. 1). In addressing issues of digital library growth, there is an implicit requirement to attend to the needs of both library patrons and of library staff, meaning that access to archived items must serve multiple purposes, and information delivery must be flexible with respect to both content and presentation. One can assume that user needs will change with changes in the scale of the library. In this regard, the target "user community" and the set of information requirements is somewhat more complex than for the cartographic situation posed at the beginning of this discussion.

There are other issues germane to scalability of digital libraries, including issues of copyright and intellectual freedom, issues of privacy, issues of equality of access, and economic factors. Then, too, there is the role of the librarian in the digital library. In some circles, the advent of digital libraries is seen as a threat to the job security of library staff. Nothing could be further from the truth—the role of library and information science has never been so important as it is now.

> I believe it is time we take much more seriously the important responsibility we hold in adopting the technologies now rolling out of Silicon Valley workshops. We need to evaluate them carefully before we buy [into] them. We need to make others aware of potential problems we see before others buy them. We urgently need "environmental impact studies" for new information technologies, so as to protect those good parts of our world information environment–like scholarly journals and neighborhood newspapers–that are on the "endangered species" list. Above all, we need to learn more about economics, and learn fast. (Nielson, 1981, p. 112)

ACKNOWLEDGMENTS

This paper forms a portion of the Alexandria Project, funded jointly by NSF, ARPA, and NASA (NSF grant IRI-94-11330). The research is also related to

Research Initiative 8 "Formalizing Cartographic Knowledge," a part of the National Center for Geographic Information and Analysis (NCGIA) (NSF grant SBE-88-10917). Support from these funding agencies is acknowledged. Comments on early versions by my colleagues Geoff Dutton and Michael Woldenberg have improved the logic and content considerably, and I am grateful for their advice and debate.

REFERENCES

Bon, R. (1973). Allometry in the topologic structure of architectural spatial systems. *Ekistics, 36*(215), 270-276.

Bremmer, S. W. (1994). *Long range planning* (How-to-do-it manual for librarians series, Number 40). New York: Neal-Schuman Publishers.

Cohn, J. M.; Kelsey, A. L.; & Fiels, K. M. (1992). *Planning for automation* (How-to-Do-it Manual for Librarians Series, Number 25). New York: Neal-Schuman Publishers.

Dutton, G. (1973). Criteria of growth in urban systems. *Ekistics, 36*(215), 298-306.

Gould, S. J. (1966). Allometry and size in ontogeny and phylogeny. *Biological Reviews, 41*(4), 587-640.

Haire, M. (1973). Biological models and empirical histories of the growth of organizations. *Ekistics, 36*(215), 263-269.

Kemeny, J. G. (1962). A library for 2000 A.D. In M. Greenberger (Ed.), *Computers and the world of the future* (pp. 135-162). Cambridge, MA: MIT Press.

Lam, N. S., & DeCola, L. (Eds.). (1993). *Fractals in geography.* Englewood Cliffs, NJ: Prentice-Hall.

Mandelbrot, B. B. (1967). How long is the coast of Britain? Statistical self-similarity and fractal dimension. *Science, 156*(May 5), 636-638.

Molz, R. K. (1984). *National planning for library service, 1935-1975.* Chicago, IL: American Library Association.

Naroll, R. S., & Bertalanffy, L. V. (1956). The principle of allometry in biology and the social sciences. *General Systems Yearbook,* 1, 76-89.

Nielson, B. (1981). Technological change and professional identity. In L. C. Smith (Ed.), *New information technologies – new opportunities* (Papers presented at the 1981 Clinic on Library Applications of Data Processing, Urbana-Champaign, 26-29 April 1981, pp. 101-113). Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Richardson, L. F. (1961). The problem of contiguity; an appendix to the statistics of deadly quarrels. *General Systems Yearbook, 6,* 139-187.

Siegel, M. A. (Ed.). (1991). *Design and evaluation of computer/human interfaces: Issues for librarians and information scientists* (Papers presented at the 1988 Clinic on Library Applications of Data Processing, Urbana-Champaign, 5-7 April 1988). Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Smith, L. C., & Dalrymple, P. W. (Eds.). (1992). *Designing information: New roles for librarians* (Papers presented at the 1992 Clinic on Library Applications of Data Processing, Urbana-Champaign, 5-7 April 1992). Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Thompson, D. W. (1977). *On growth and form* (Abridged edition). London, England: Cambridge University Press.

Weibel, S. (1992). Automated cataloging: Implications for libraries and patrons. In F. W. Lancaster & L. C. Smith (Eds.), *Artificial intelligence and expert systems: Will they change the library?* (Papers presented at the 1990 Clinic on Library Applications of Data Processing, Urbana-Champaign, 25-27 March 1990, pp. 67-80). Urbana-Champaign, IL: Graduate School of Library and Information Science, University of Illinois.

Woldenberg, M. J. (1971). *Allometric growth in social sciences* (Harvard Papers in Theoretical Geography, Technical Report No. 6). Cambridge, MA: Harvard University Graduate School of Design.

Woldenberg, M. J. (1973). An allometric analysis of urban land use in the United States. *Ekistics, 36*(215), 282-290.