Advancing Geospatial Data Curation*

Rajendra Bose[1], Femke Reitsma[2]

[1] Digital Curation Centre and School of Informatics
University of Edinburgh
rbose@inf.ed.ac.uk
[2] Institute of Geography, School of Geosciences
University of Edinburgh
femke.reitsma@ed.ac.uk

Abstract. *Digital curation* is a new term that encompasses ideas from established disciplines: it defines a set of activities to manage and improve the transfer of the increasing volume of data products from producers of digital scientific and academic data to consumers, both now and in the future. Research topics in this new area are in a formative stage, but a variety of work that can serve to advance the curation of digital geospatial data is reviewed and suggested. Active research regarding geospatial data sets investigates the problems of tracking and reporting the data quality and lineage (provenance) of derived data products in geographic information systems, and managing varied geoprocessing workflow. Improving the descriptive semantics of geospatial operations will assist some of these existing areas of research, in particular lineage retrieval for geoprocessing results. Emerging issues in geospatial curation include the long-term preservation of frequently updated streams of geospatial data, and establishing systematic annotation for spatial data collections.

1 Introduction: Digital Curation

The dictionary definition of *curation* is the supervision of a collection of preserved or exhibited items, by those responsible for the care or charge of the collection [1]. Recent UK Joint Information Systems Committee (JISC) reports including [2], however, focus on the impending "data deluge" associated with ongoing e-science and e-research initiatives, and propose a domain of *digital curation* activities to manage and improve the transfer of this increasing volume of data products from the producers of digital scientific and academic data to consumers.

Digital curation is a new term, but it encompasses ideas from the established digital library and digital preservation communities, as well as other disciplines. Curating digital data collections involves providing potential users with the means to discover and access trustworthy and adequately documented data products. It also involves the traditional archival activities of assessing and selecting particular products and processing details for long-term preservation.

* This paper has been submitted to the International Journal of Digital Curation (www.ijdc.net)

Research topics for digital curation are in a formative stage; to introduce this area we discuss the research agenda for the newly established UK Digital Curation Centre (DCC). The DCC explores issues related to scientific databases and curation (here we use the term *database* generally to refer to collections of structured data stored in relational database management systems (DBMS), XML, or other formats). These issues include:

•*Enabling provenance retrieval to increase the utility of query results and data products:* Buneman *et al.* [3] use the bioinformatics community as a prime example of the activity of extraction and compilation of records from various public and private databases in support of biological research. Tracking the provenance—that is, the sources or origin—of query results from derived databases is crucial to determining the validity of biology research results. This tracking and retrieval of provenance is also notoriously difficult to implement in today's environment of database-driven web applications. Recent related work also investigates methods for tracking custom, satellite-derived geospatial data products [4].

•*Using systematic annotation to extend the research record:* Annotation often refers to affixing structured description or interpretation over an existing body of data. Projects within the Earth system science community have long sought to build systems for collaboration [5, 6], which cross-disciplinary annotation of both data products and processing steps could now help to achieve. While special-purpose annotation systems such as BioDAS [7] are starting to meet the needs of the bioinformatics community, there are few examples of successes with this idea elsewhere in the digital scientific domain.

•*Publishing and integrating scientific databases:* Scientific organizations sometimes use curators who are ultimately responsible for editing and presenting the content of databases. Current work is related to publishing parts of these curated databases that other researchers or scientists may easily bring into or integrate with their own "research databases." This includes developing techniques to extract records from a number of differently structured relational databases and translate them into XML documents that share the same structure.

•*Archiving scientific data:* Current research concerns developing techniques to efficiently archive large databases that continuously change, such as genome databases that grow rapidly as biological research moves forward.

Over a half century ago, Vannevar Bush noted that "a record, if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted" [8]. The digital curation research topics listed above aim to ensure this situation is possible for all types of scientific data collections, including the record of *research computing* in geography and the environmental sciences, which often involves processing of georeferenced, time-varying 2D or 3D spatial data.

Maintaining and extending the scientific record for digital products that result from combining and transforming georeferenced spatial data presents unique challenges for many scientific communities. Unfortunately, anecdotes of lost or undecipherable research computing results and data sets are legion in academic and government laboratories. But large interdisciplinary environmental research projects such as the U.S. National Science Foundation-supported Long-Term Ecological Research (LTER) activity begun in the 1980s, and the newer National Ecological Observatory Network

(NEON) and the Geosciences Network (GEON) visions of collaboration, however, depend on the principles of comprehensive digital curation to construct a persistent record of research.

In the following sections we review active research areas related to curating geospatial data, and we outline previous work related to a vital topic within this domain: classifying geospatial data operations. We then discuss the additional topics of preserving and annotating data sets used within geographic information systems (GIS).

2 Active Research Areas in Geospatial Curation

Geographic information has long been attributed "special" status because of factors including, for example, the distinctive data structures, indexing systems and algorithms required for its processing. More recently this status has been de-emphasized by Longley *et al*. [9] as they recognize that software now increasingly hides from the user the special structures necessary for manipulating geographic data. Although this may be true, data curators are not absolved of the responsibility for capturing and communicating the vital aspects of geographic information that Longley *et al.* do identify: spatial dependence; spatial heterogeneity; and the possible underlying correlations among different "layers" of data (spatial autocorrelation).

In addition to the geodetic reference system and coordinate system used to specify location, other critical metadata for a georeferenced data product includes measures of data quality, discussed in Section 2.1, which will be partly dependent on its provenance and processing history. With data quality, other curation-related research topics active for the past decade or more include tracking the lineage or provenance of derived GIS layers, and managing workflow for geospatial data processing (geoprocessing).

2.1 Digital Geospatial Data Quality

The *quality* of finished cartographic products, a notion which encompasses imprecise concepts like accuracy and completeness, has been of concern to mapmakers and map users for centuries.

In the last several decades, computing environments for geospatial data processing have introduced the ability to effortlessly combine and transform varied data sources, creating products that range across a continuum from "rough" to "final." Derived data products are often assumed to be error-free by their consumers, however. Measures of accuracy or error are not always included with derived products, or if they are, these measures are limited in scope. For example, current metadata standards only facilitate the documentation of an overall accuracy measure over the entire map area, providing no means for documenting the spatial heterogeneity of uncertainty over a field or individual objects. This difficulty is compounded by GISs that typically do not provide the means for automatically updating metadata following the modification of spatial data (Zhu, 2005).

Research concerning these issues is flourishing: in a review of [10], Heuvelink [11] mentions nine edited books and special journal issues published in the period from 1989-2003 that cover the topic of spatial data quality. He also calls for a text to establish an improved foundation for understanding this complex area, however.

Veregin [12] frames the scope of error for map overlay operations, which also serves as a general aid for understanding issues of error in GIS. Identifying potential sources of error occupies the first level of the framework, with detection and measurement of error within a data product following from this. Attempting to model the propagation of error through operations is next, while the highest levels of the framework concern strategies for error management and reduction.

Recently, the discussion about data quality has been recast with the focus on data uncertainty. In their comprehensive treatise, Zhang and Goodchild [13] conclude that conveying uncertainty to spatial data consumers in an effective manner, for example, through visualizations of the stochastic simulation of uncertainty, represents one important area of further research. The University Consortium for Geographic Information Science (UCGIS) seems to concur in spirit, suggesting in their recently published *Research Agenda for Geographic Information Science* that there is a need to continue to explore tracking and reporting the uncertainty of spatial information operations [14].

2.2 Lineage Tracking for GIS Data Layers

U.S. metadata standards for geospatial data have included specifications for lineage as a component of data quality information since the Spatial Data Transfer Standard (SDTS) [15] became a Federal Information Processing Standard (FIPS 173) in 1992. The SDTS, developed for transferring georeferenced spatial data between dissimilar applications or computer systems, includes a text description of lineage as part of a data quality report. This report is required to accompany the data in a standard transfer, but also must be obtainable separately from the actual data. No mechanism is stipulated for how this is to be achieved.

In the early 1990s, coincident with the data quality report requirements of the SDTS, Lanter investigated tracking the lineage of new coverages created within Arc/Info GIS [16]. In his Lineage Information Program (LIP) prototype, command-line GIS operations were intercepted by a programming shell and inserted into a meta-database external to the GIS itself. This lineage meta-database could then accept lineage queries, for example: what other GIS layers (coverages) contributed to the creation of layer X? What GIS layers are children of layer Y?

Vert *et al.* [17] list what are essentially curation challenges related to processing GIS data: "the need to manage a wide range of differing applications' data formats, group files into meaningful organizations that are driven by geospatial concerns, provide lineage and version support, provide documentation of the nature of the data and other descriptive information, and keep track of relations among data files and groups of data files." Their prototype "GIS Workbench" is limited in scope, but it provides a preliminary proof of concept for a platform-neutral architecture to resolve some of these issues. The prototype is noteworthy because it is a rare example of system architecture in the GIS-related literature where tracking the lineage or provenance of

versioned data sets is an inherent property. Other examples of such systems include the previously mentioned work by Lanter, the design for a system to track the changes of cadastre boundaries over time [18], and the Geo-Opera prototype system [19], discussed in the next section.

2.3 Managing Geoprocessing Workflow

To some extent, trends in geospatial data processing mirror the evolution of general computing environments. Writing and executing programs and scripts at a workstation command line may still be the dominant paradigm, but visual programming environments, web services and Grid concepts are impacting the conduct of geoprocessing. Consider the widely used ArcGIS platform (ESRI, Inc.): the most recent version (9.1) bundles data processing operations as "tools," which may be executed via dialog boxes or on a command line. Additionally a workflow of tools may be assembled in a visual editor, or tools may be called in Component Object Model (COM)-compliant scripting languages or with the custom Arc Macro Language (AML).

True workflow management systems (WFMS), with a broader scope than the ArcGIS ModelBuilder facility, are usually built on top of a DBMS and require additional administration, but are meant to allow greater control of potentially complex data processing applications. Prototypes that feature WFMS for geoprocessing include the WASA (Workflow-based Architecture to support Scientific Applications) and Geo-Opera systems, described below.

Weske et al. suggest a design for WASA, a client/server WFMS that is more tailored to scientific workflows than commercial systems, in the context of a geoprocessing application [20]. The Geo-Opera extension of the OPERA (Open Process Engine for Reliable Activities) kernel provides a management system for distributed geoprocessing that incorporates elements of workflow management, transaction processing, and lineage tracking for an Earth science example of hydrologic modeling [19, 21]. Geo-Opera data objects include a set of system-maintained attributes supporting automated versioning, change propagation, and lineage recording. These systems are discussed in more detail in [22].

The interest in combining and manipulating geospatial data via web services introduces a different context for managing geoprocessing workflows. Web service standards associated with streaming geospatial data introduced by the Open Geospatial Consortium (OGC) (www.opengeospatial.org) continue to evolve, but adoption of these standards are presenting profound challenges with regard to strategies for long-term data preservation (personal communication, Steven Morris, Head of Digital Library Initiatives, North Carolina State University Libraries, May 2005). The use of complex combinations of web services for geoprocessing may also suggest the need to manage workflows of web services. Related research efforts in bioinformatics involving WFMS and web services [23] may offer some guidance.

3 Advancing Geospatial Curation Through Improved Semantics

Expressing geospatial data quality, geoprocessing workflow, and the lineage of geoprocessing results requires adequate descriptive semantics. This involves recording the semantics of the data regardless of its representation, and recording the changes that result from modifying the data.

3.1 Recording the Semantics of Geospatial Data Processing

Geospatial data is exhaustively treated by a number of standards that specify the syntax of description: the federally mandated U.S. Content Standard for Digital Geospatial Metadata; the ISO 19115 Geographic Information Metadata standard; and markup languages such as GML 3.0 (Geographic Markup Language). Yet there are no standards for expressing the formal semantics of geospatial data or the metadata describing it [24]. Such formal semantics would define the relationships between, or the content within, those syntactic atoms, such as the metadata tags of "Title", "Supplementary Information", or "Use Constraints", or spatial data tags of "MultiPolygon", "surfaceProperty", or "attribute". This would allow services for geographic data to utilize curated data in an automated fashion.

Ontologies have been developed for geographic features, such as the work by Manov *et al.* [25], who have developed an ontology for spatial entities as the basis for geographic information extraction. This work extends the flat structure of gazetteers by utilizing the transitivity of parthood relations, such as their *subRegionOf* relation, and classes such as *country* and *city* and their relations. Yet this does not allow for the semantics of the spatial data itself to be expressed; as noted by Tomai and Kavouras [26], spatial aspects of geographic ontologies are often underspecified or totally absent from extant ontologies. Ontologies have also been conceptualized and prototyped as a basis for a GIS in an attempt to seamlessly integrate geographic information based on its semantic content regardless of its representation [27, 28].

Although spatial data infrastructures (SDIs) at the national level are proposed as one way of encoding and unifying the semantics of geospatial data processing, there is a worldwide trend of declining use, management, and content of national clearinghouses for spatial data [29]. We believe the means to clearly communicate the semantics of spatial data processing needs to be introduced at the level of the individual organization or research group. This requires building the ability to record data processing semantics directly into the software and systems used for processing. There is, however, no complete reference currently available for classifying operations on geospatial data.

3.2 Classifying Geospatial Data Operations

Extant classification schemes for geospatial data transformations are not comprehensive: they either focus on a particular data structure or data model (such as the raster data model to which Map Algebra [30] applies) or otherwise serve a special case. Likewise, no standard has been developed for describing the complete set of opera-

tions available in GIS [31]. Several reviews of existing classifications for describing spatial data handling have been presented [32, 33], followed by proposals for new classifications. Existing approaches include the following five ways of organizing spatial operations:

By Data Processing Flow. This type of classification scheme is structured around the typical flow or "pipeline" of GIS processing steps; for example, by data input, data management, data analysis, and data output. Stefanakis and Sellis [31] describe an example of a taxonomy defined by [30], which divides GIS operations into the following classes: programming operations, data preparation operations, data presentation operations, and data interpretation operations.

From a User Perspective. Operations can be organized from the perspective of a user's cognition of GIS processing tasks. Albrecht [32], for example, presents twenty universal GIS operations, which are independent of data structure, task-oriented, and which aim to cover the full range of possible analytical capabilities.

From an Implementation Perspective. This scheme involves classifying operations based on whether they implement low-level procedures that operate at the data structure level, such as the conversion between raster and vector data, or higher-level file or data set management tasks that are independent of the data model used, such as the transfer of data from one software environment to another. Voser and Jung [34] define such a framework, with levels of abstraction for analytical GIS-Operators. These are: *management*, which are high level operators that are independent of data structure; *controlling*, which specify mid-level operations such as analysis operations; and *processing*, which are low level operations that might involve the processing of an algorithm.

By Spatial or Attribute Relationships. Space can provide the guiding construct for defining a taxonomy of GIS operations. For example, Map Algebra [30] classifies operations according to their spatial relationships, namely *local*, *focal*, *zonal*, and *global* neighborhoods. The modification of attributes can also serve as a basis for classification. Goodchild [35] specifies six classes of GIS operations based on the relationships between the classes of pairs of "GIS objects" within an ideal data model, and whether the attribute or locational information for the object classes is required.

By Spatial Data Transformation. GIS operations can be classified by the way they transform spatial data. This approach highlights the impact of the operation on the spatial data and recognizes the import of knowing whether the procedure can be reversed in order to reconstruct the original data. Chrisman [33] presents a classification framework for GIS operations based on the transformational view of cartography, where the change in spatial data and the existence of invertible operations provides the basis of classification. This approach differs from the others in that it seeks explicit semantics for GIS operations.

The five different classification approaches discussed above neglect to communicate the impact of data transformations to future consumers of a derived geospatial data product. The transformational approaches come closest to this purpose, but they do not comprehensively cover all data models and operations.

Creating an improved and inclusive classification scheme for operations on geospatial data that delivers unambiguous processing history is crucial to providing useful provenance retrieval. The work on composing and conveying provenance for custom satellite-derived geospatial data products in [36] focuses on recording processing steps within scripting environments favored by the earth sciences community such as MATLAB (The Mathworks, Inc.) or IDL (Research Systems, Inc.). We plan to extend this work with a new classification scheme for geospatial operations, and apply it to scripted workflows within commercial and open source GIS.

4 Emerging Geospatial Data Curation Issues

Further potential research topics in the area of geospatial data curation include the long-term preservation of spatial data collections and their annotation.

4.1 Long-term Preservation of GIS Data

The digital preservation community is well-established, but historically the issue of archiving geospatial data has received less attention than other types of data collections such as journals and books. This situation may be changing: in the recently initiated (September 2004) U.S. Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP), two of the eight proposals selected involve geospatial collections. One of these initiatives will collect and preserve digital geospatial data resources from state and local government agencies in North Carolina state. This project identifies several risks to digital geospatial data, and to the record of its extensive modification over time by agencies (personal communication, Steven Morris, Head of Digital Library Initiatives, North Carolina State University Libraries, May 2005). These risks include:

•*Limited archiving efforts:* This has resulted in the absence of archives for time-versioned geospatial data. The incentive to create such archives has diminished because of the continuous data access provided by newer web services.

•*Few preservation guidelines:* Few guidelines exist for assembling canonical "preservation metadata," migrating geospatial data, and maintaining data independence from software. This is especially important in an era of on the order of one hundred existing spatial data formats and limited options for repository software.

•*Additional issues unique to geospatial data:* There are also few guidelines for ensuring the long-term survival of cartographic representation as well as the multiple features such as topology, behavior and annotation embedded with spatial data in the proprietary Geodatabase model (ESRI. Inc.). This is significant because use of the Geodatabase model in agencies is increasing.

(Bleakly, 2002) provides a useful overview of current practices for long-term preservation and archiving of spatial data, and concludes that, rather than a single grand solution for these involved issues, a mixture of strategies based on difficult management and policy decisions will serve the way forward for individual groups.

4.2 Annotating Spatial Data

At its simplest, annotation refers to the process of adding or making notes on or upon something [1]. Such notes can serve a variety of purposes, including explaining, interpreting or describing the thing that has been annotated. This basic definition carries the sense of physically making a mark on an item: writing in the margins of a document, or drawing directly on a picture or map, for example. In GIS, annotation usually refers to the labeling of map features with descriptive text, possibly using attribute values associated with the features. In ArcGIS software , annotation "includes a text string, a position at which it can be displayed, and display characteristics" [37].

A broader view of spatial data annotation, however, recognizes the interest, across many scientific disciplines, in linking databases (in the sense of Section 1) of regions of two or more dimensions with databases of descriptive or interpretative text. To provide one simple example, researchers would like to query repositories of brain scans or other medical images using standard anatomical terminology and retrieve relevant images as results [38-40]. For this to be possible, subregions of the images must first be annotated with the appropriate terms. Generally, annotation of databases implies that descriptive or interpretative data is added over an existing structure, which possesses some coordinate system that can serve as the basis of annotation attachment.

Database annotation is a fairly new area of research, with quintessential examples provided by the bioinformatics community. While genomic and protein sequence annotation in biology relies on a one dimensional coordinate system (base pair location within a DNA sequence), annotation of geospatial data is based on 2D or 3D geographic coordinate systems, with the possible inclusion of a temporal dimension.

The role of annotation in scientific work is integral: for many communities, including geography and the environmental sciences, and the large number of policy- and decision-making bodies that use their computational results, the focus is on description or interpretation from trusted sources. Understanding contemporary forms of annotation and spatial databases, and the interplay between the two, is an emerging and important area of research.

5 Conclusion

Curation is an umbrella term, with the aim of this set of activities improving how digital resources are managed, now and in the future. These resources include the results of interactions between databases of all types--curated, public domain, research and personal databases--and the derived results of data processing in various environments. Geospatial data curation presents significant challenges for many commu-

nities, and encompasses existing threads of research such as data quality, lineage tracking, and geoprocessing workflow. Advances will be accomplished through continuing work in these areas, as well as attending to the emerging issues of improved semantics for spatial operations, and the preservation and annotation of GIS and other spatial data.

Acknowledgements

We would like to acknowledge the support of Peter Buneman, Peter Burnhill and the other partners and participants in the Digital Curation Centre (DCC) (http://www.dcc.ac.uk). We would also like to acknowledge the contributions of Steven Morris, Head of Digital Library Intiatives, North Carolina State University Libraries during his May 2005 visit to the DCC, and Guy McGarva of the EDINA National Data Centre and the DCC. The DCC is funded by JISC and the eScience core programme.

References

1. Oxford University Press, Oxford English Dictionary (OED Online), 2005, Oxford University Press, <http://dictionary.oed.com > (Last access date: 31 Jan 2005).
2. P. Lord and A. Macdonald, "Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision," The Digital Archiving Consultancy Limited, Twickenham, UK. <http://www.jisc.ac.uk/index.cfm?name=project_escience >
3. P. Buneman, S. Khanna, and W. C. Tan, "Data Provenance: Some Basic Issues," in *Proceedings of the Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2000),* New Delhi, India, 2000, pp. 87-93.
4. R. Bose and J. Frew, "Composing Lineage Metadata with XML for Custom Satellite-Derived Data Products," in *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004),* Santorini, Greece, 2004, pp. 275-284.
5. M. Stonebraker, "An Overview of the Sequoia 2000 Project," Sequoia Technical Report S2K-94-58, Berkeley, CA, 1991. <http://s2k-ftp.cs.berkeley.edu:8000/sequoia/tech-reports/s2k-94-58/ >
6. T. R. Smith, J. Su, D. Agrawal, and A. El Abbadi, "Database and Modeling Systems for the Earth Sciences," *IEEE Bulletin of the Technical Committee on Data Engineering*, vol. 16, no. 1, 1993, pp. 33-37.
7. L. D. Stein, S. Eddy, and R. Dowell, "Distributed Sequence Annotation System (DAS) Specification." <http://www.biodas.org/documents/spec.html >
8. V. Bush, "As We May Think," *The Atlantic Monthly*, 1945.
9. P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, "Introduction," in *Geographical Information Systems*, vol. 1, P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, Eds., 2nd ed: Wiley, 1999, pp. 1-20.
10. W. Shi, P. F. Fisher, and M. F. Goodchild, "Spatial data quality," Taylor & Francis, 2002.
11. G. B. M. Heuvelink, "Book Review: Spatial Data Quality," *International Journal of Geographical Information Science*, vol. 17, no. 8, 2003, pp. 816-818.
12. H. Veregin, "Error modeling for the map overlay operation," in *Accuracy of Spatial Databases*, M. F. Goodchild and S. Gopal, Eds.: Taylor & Francis, 1989, pp. 3-18.

13. J. Zhang and M. F. Goodchild, *Uncertainty in geographic information*: Taylor & Francis, 2002.
14. R. McMaster and E. L. Usery, "A Research Agenda for Geographic Information Science," CRC Press, 2005.
15. U.S. Geological Survey, "Spatial Data Transfer Standard (SDTS)," NCITS 320-1998, American National Standards Institute (ANSI), Reston, VA, June 9, 1998. <http://mcmcweb.er.usgs.gov/sdts/SDTS_standard_nov97/part1b12.html >
16. D. P. Lanter, "Design of a Lineage-Based Meta-Data Base for GIS," *Cartography and Geographic Information Systems*, vol. 18, no. 4, 1991, pp. 255-261.
17. G. Vert, M. Stock, P. Jankowski, and P. Gessler, "An Architecture for the Management of GIS Data Files," *Transactions in GIS*, vol. 6, 2002, pp. 259-275.
18. L. Spery, C. Claramunt, and T. Libourel, "A lineage metadata model for the temporal management of a cadastre application," in *Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA '99),* Florence, Italy, 1999, pp. 466-474.
19. G. Alonso and C. Hagen, "Geo-Opera: Workflow Concepts for Spatial Processes," in *Proceedings of the 5th International Symposium on Spatial Databases (SSD '97),* Berlin, Germany, 1997, pp. 238-258.
20. M. Weske, G. Vossen, C. B. Medeiros, and F. Pires, "Workflow Management in Geoprocessing Applications," in *Proceedings of the ACM 6th International Symposium on Advances in Geographic Information Systems,* Washington DC, 1998, pp. 88-93.
21. G. Alonso, C. Hagen, H.-J. Schek, and M. Tresch, "Towards a Platform for Distributed Application Development," in *Workflow Management Systems and Interoperability*, vol. 164, *NATO ASI Series*, A. Dogac, L. Kalinichenko, M. T. Ozsu, and A. Sheth, Eds. Berlin: Springer, 1998, pp. 195-221.
22. R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," *ACM Computing Surveys*, vol. 37, no. 1, 2005, pp. 1-28.
23. J. Zhao, C. Goble, M. Greenwood, C. Wroe, and R. Stevens, "Annotating, linking and browsing provenance logs for e-Science," in *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (at ISWC 2003),* Sanibel Island, Florida, 2003.
24. L. Bernard, I. Kanellopoulos, A. Annoni, and P. Smits, "The European Geoportal - one step towards the establishment of a European Spatial Data Infrastructure," *Computers, Environment and Urban Systems*, vol. 29, 2005, pp. 15-31.
25. D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard, and H. Cunningham, "Experiments with Geographic Knowledge for Information Extraction," in *Proceedings of the Workshop on Analysis of Geographic References,* Edmonton, Canada, 2003.
26. E. Tomai and M. Kavouras, "Pivotal Issues in Designing Geographic Ontologies," in *Proceedings of the Workshop on Fundamental Issues in Spatial and Geographic Ontologies, COSIT'03,* Ittingen, Switzerland, 2003.
27. F. Fonseca, M. Egenhofer, P. Agouris, and G. Câmara, "Using Ontologies for Integrated Geographic Information Systems," *Transactions in GIS*, vol. 6, 2002, pp. 231-257.
28. F. Fonseca, M. Egenhofer, and C. B. Medeiros, "Ontology-Driven Geographic Information Systems," in *Proceedings of the 7th ACM Symposium on Advances in Geographic Information Systems,* Kansas City, MO, 1999.
29. J. Crompvoets, A. Bregt, A. Rajabifard, and I. Williamson, "Assessing the Worldwide Developments of National Spatial Data Clearinghouses," *International Journal of Geographical Information Science*, vol. 18, 2004, pp. 665-689.
30. D. Tomlin, *Geographic Information Systems and Cartographic Modelling*. Englewood Cliffs, New Jersey: Prentice Hall, 1990.
31. E. Stefanakis and T. Sellis, "Enhancing Operations with Spatial Access Methods in a DBMS for GIS," *Cartography and Geographic Information Systems*, vol. 25, 1998, pp. 16-32.

32. J. Albrecht, "Universal GIS Operations for Environmental Modeling," in *Proceedings of the 3rd International Conference/Workshop on Integrating GIS and Environmental Modeling*, Santa Fe, NM, 1996.
33. N. Chrisman, "A Transformational Approach to GIS Operations," *International Journal of Geographical Information Science*, vol. 13, 1999, pp. 617-637.
34. S. A. Voser and S. Jung, "Towards Hybrid Analysis - Specification of High Level Analytical GIS Operators," in *Proceedings of the First AGILE-Conference (Association of Geographic Information Laboratories in Europe),* ITC, Enschede, The Netherlands, 1998.
35. M. F. Goodchild, "A Spatial Analytical Perspective on Geographical Information Systems," *International Journal of Geographical Information Systems*, vol. 1, 1987, pp. 327-334.
36. R. Bose, "Composing and Conveying Lineage Metadata for Environmental Science Research Computing," *Ph.D. Dissertation*. Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, 2004.
37. A. Perencsik, S. Woo, B. Booth, S. Crosier, J. Clark, and A. MacDonald, "ArcGIS 9: Building a Geodatabase," ESRI, Redlands, CA.
38. A. Burger, R. Baldock, Y. Yang, A. Waterhouse, D. Houghton, N. Burton, and D. Davidson, "The Edinburgh Mouse Atlas and Gene-Expression Database: A Spatio-Temporal Database for Biological Research," in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM'02)*, 2002, pp. 239.
39. R. A. Baldock, C. Dubreuil, W. Hill, and D. Davidson, "The Edinburgh Mouse Atlas: Basic Structure and Informatics," in *Bioinformatics: Databases and Systems*, S. I. Letovsky, Ed.: Kluwer Academic Publishers, 1999, pp. 129-140.
40. M. Gertz, K.-U. Sattler, F. Gorin, M. Hogarth, and J. Stone, "Annotating Scientific Images: A Concept-based Approach," in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM 2002),* Edinburgh, Scotland, 2002.