19

Geoprocessing, Workflows, and Provenance

Jason A. Tullis
University of Arkansas

Jackson D. Cothren
University of Arkansas

David P. Lanter
CDM Smith

Xuan Shi
University of Arkansas

W. Fredrick Limp
University of Arkansas

Rachel F. Linck
University of Arkansas

Sean G. Young
University of Iowa

Tareefa S. Alsumaiti
United Arab Emirates University

Acronyms and Definitions

ACSM	American Congress of Surveying and Mapping
API	Application programming interface
CDAT	Climate data analysis tool
CI	Cyberinfrastructure
DBMS	Database management system
FGDC	Federal Geographic Data Committee
HMM	Hidden Markov model
ISO	International Standards Organization
LIDAR	Light detection and ranging
LULC	Land use/land cover
MODAPS	MODIS adaptive data processing system
NCDCDS	National Committee for Digital Cartographic Data Standards
NSDI	National Spatial Data Infrastructure
OGC	Open Geospatial Consortium
OpenMI	Open modeling interface
PROV-DM	PROV data model
REST	Representational state transfer
SOAP	Simple object access protocol
SOC	Service-oriented computing
SOI	Service-oriented integration
SQL	Structured query language
W3C	World Wide Web Consortium
WPS	Web processing service
WSDL	Web services description language
XML	Extensible markup language
XSEDE	eXtreme Science and Engineering Development Environment

19.1 Introduction

Integrated remote sensing and GIS-assisted problem solving now supports a remarkable array of domains (e.g., food and agricultural security, climate change, forest management, heritage preservation, and urban and regional planning) and is being configured in a great variety of technical means. Given the sheer quantity of innovations reported in journals and books (including the Remote Sensing Handbook), any one expert may be keenly aware of only a fraction of the detailed remote sensing and related geospatial methods available to address a

given problem statement. Regardless of the remote sensing application under study or review, some reliance (whether implied or reported) is always made upon the *geoprocesses* and *workflows* associated with any geospatial artifacts produced. In the context of a specific geospatial decision support artifact (e.g., a map of predicted crop yield in kg/ha), a record of the specific geoprocesses may be termed geospatial *provenance* (or *lineage*; see Section 19.1.1). This chapter explores how remote sensing–assisted geoprocessing and related GIS workflows have been or may be combined with digital provenance information in order to augment scientific reproducibility, comparison, trust, or to otherwise improve remote sensing–assisted decision support.

Increasingly of interest in computer systems, digital provenance has relatively early geospatial origins that date back to at least the 1980s (e.g., Chrisman 1983), with a definite resurgence around 2009 (e.g., Yue et al. 2010a). The early and expanded geospatial interest and connection to provenance are driven in large part by the question of methodological innovation. For remote sensing and GIS integration to best improve the quality of decision making tools across a range of applications and domains, it seems reasonable that, if possible, such innovation must first be machine recognizable. Unfortunately, many geospatial decision support tools lack suitable means to even replicate their findings, and innovation reported is naturally bracketed by complex questions of accuracy, fitness for use, and a variety of other qualitative and quantitative metrics related to reliability and trust. So, while there is broad conceptual agreement that machine-interpretable source and process history records are vital and may even be scientifically transformative in the modern era, questions remain unanswered on how provenance information may simultaneously benefit multiple domains (including the geospatial domain), and what mechanisms for its digital capture and exchange will most successfully convey those benefits.

There are at least two good reasons to believe that even partial success toward machine-interpretable geospatial process history records will be rewarded. First, correct expert interpretation of the full scope of relevant methods, procedures, algorithms, and expert knowledge is subject to entropy and constitutes an increasingly complex, even daunting companion to the twenty-first-century *big [geospatial] data* (Hey et al. 2009). Second, as remote sensing and other geospatial techniques are communicated in the scientific literature, there is a well-known continuing expectation and scholarly requirement that previously published studies are carefully acknowledged for their relevant achievements and/or limitations. Failure to increasingly harness machine power on these two fronts (but to continue interpretations by experts alone) is probably not a viable long-term option. In a related example from computer systems, Buneman (2013) notes that the underappreciated machine-managed provenance in software version control systems has helped prevent a total disaster in software engineering.

It is clear that absent the kinds of methodological analyses enabled in part through exchange of provenance information, an increasingly data-intensive *geo-cyberinfrastructure* (Di et al. 2013a) renders comprehensive remote sensing–assisted geospatial workflow interpretations, comparisons, and knowledge transfers ever more difficult by experts alone. Furthermore, depending on the geospatial laboratory setting and the capabilities of a given research team, the actual digital methods linked to published materials may overlap significantly with previously reported work, may offer similar results using a more or less computationally efficient means of problem solving, and/or may be idiosyncratic to individual skills and experience. In an integrated geoprocessing, workflow, and provenance cycle, expert refinement of remote sensing–assisted decision support knowledge may be augmented by software agents capable of automated exchange and recognition of innovation (Figure 19.1).

Over the past 25 years, various prototype forms of geospatial provenance have been implemented in shared workflow environments, including those specialized for high-performance capabilities. In spite of the potential of these prototypes, single user/workstation geoprocessing and workflow design continue to be a dominant tradition with many active options (e.g., from Hexagon Geospatial, Exelis Visual Information Solutions, and ESRI). There is therefore a discrepancy between futuristic collaborative goals and the actual state of the art of remote sensing–assisted software. There are also variations in how provenance itself is defined, whether specifically in a remote sensing or geospatial-related forum, or more broadly in computer systems. It therefore seems reasonable to report progress in terms of what the actual computational environments entail and which definitions are implied.

19.1.1 Working Definitions

Though commonly understood in a broad remote sensing and geospatial computation parlance, Wade and Sommer (2006) define *geoprocessing* in the context of the many tools available in one software platform (ESRI's ArcGIS) with an emphasis on input GIS datasets, operations performed, and associated outputs. More generically, its root, *process*, implies an instance of a computer program execution, and this is naturally compatible with a geospatial/remote sensor data processing software context. Of course, identical geospatial computer programs operating on identical input datasets may produce different results as a function of additional configuration parameters. For example, raster-based geoprocessing tools in ESRI's ArcGIS 10 platform allow for an *environment setting* called *Snap Raster*. This setting allows the user to specify the spatial grid on which computations are made. In practice, use of this parameter allows pixels in an output raster layer to be exactly aligned with another raster having the same cell size. To a novice, the resulting subpixel geometric shift may seem inconsequential at the overview scale. However, remote sensing experts know that when geoprocessing tools are chained together into a *workflow* (in the present context, a repeatable

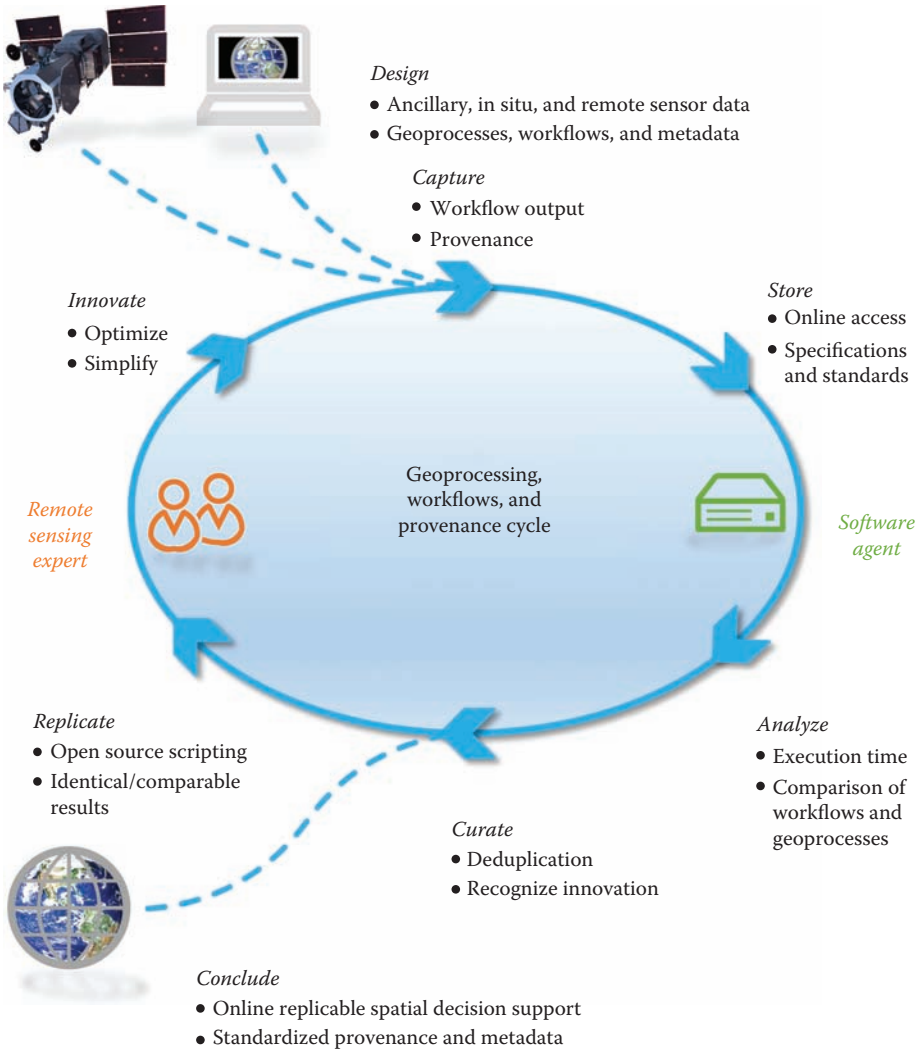

FIg u r e 19.1 Integrated geoprocessing, workflows, and provenance may be conceptualized as a positive developmental cycle used to refine remote sensing knowledge before decision support is communicated. Highlighted aspects of this cycle suggest a capacity of remote sensing experts, in conjunction with software agents, to cooperatively capture, store, analyze, curate, replicate, and innovate remote sensing–assisted decision support methods. (Artist image of WorldView-3, Courtesy of DigitalGlobe, 2014.)

sequence of geoprocesses of interest to a person or group), environment settings like *Snap Raster* can affect the logic of a decision support conclusion.

Provenance traces back to 1294 in Old French as a derivative of the Latin *provenire*, and while Merriam-Webster (2014) emphasizes provenance as a *concept* (e.g., ownership history of a painting), Oxford University Press (2014) highlights the *record* of such provenance (Moreau 2010). In the art domain where the term is very well established, provenance entails an artifact's complete ownership history, but ideally will also include artistic, social, and political influences upon the work from its creation to the present day. There is an established research process for obtaining an artifact's trusted provenance, and the information is highly valued, particularly to authenticate real versus fraudulent works (IFAR 2013; Yeide et al. 2001). As a related term, provenance is now increasingly used in a broad range of fields (e.g., archaeology, computer science, forestry, and geology) with usually overlapping definitions.

Computational definitions of provenance are more numerous than in other domains, largely because of (1) the difference between *concepts* of digital records and *actual* digital records, and (2) the variation in software environment such as a database management system (DBMS) versus file-based processing (Moreau 2010). Understanding provenance *within* DBMS queries requires more computationally detailed observations than understanding provenance at a more generalized workflow level (where one step in the workflow may entail multiple database queries). Various traditions further influence how provenance is viewed, for example, whether it is conflated with *metadata* or *trust*, two closely related but distinct concepts (Gil et al. 2010). Given the infrastructural importance of the web in remote sensing–assisted decision support, the following W3C

Provenance Incubator Group's working definition of provenance (in a web resource context) carries significant weight:

> Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.
>
> **g il et al. (2010)**

It should be noted that while *provenance* and *lineage* are here used interchangeably, one can argue that there are subtle differences in their meanings. Process history seems to fit more easily with the many definitions attributed to provenance, and lineage implies a kind of genealogy or data pedigree record relative to a remote sensing–assisted decision support artifact. While these semantic differences are not a point of the present focus, each word will appear in its historical context (beginning with lineage). Also, a number of surveys have been conducted on provenance including some with a geoprocessing and workflow flavor. For example, Yue and He (2009) provide a review covering various aspects of geospatial provenance. For a broader perspective, Bose and Frew (2005) provide a review covering provenance in geospatial as well as other domains. More recently, Di et al. (2013b) provide an overview of geoscience data provenance.

19.2 Historical Context

The earliest work in geospatial lineage was spurred in the United States through the formation of the National Committee for Digital Cartographic Data Standards (NCDCDS) by the American Congress of Surveying and Mapping in 1982 (Bossler et al. 2010). In 1988, chaired by Dr. Harold Moellering from Ohio State University, the NCDCDS proposed five fundamental components of a geospatial data quality report, including (1) lineage, (2) positional accuracy, (3) attribute accuracy, (4) logical consistency, and (5) completeness. The NCDCDS described lineage in detail, which they presented as the *first* quality component. Less than a third of their description for lineage follows (Moellering et al. 1988, p. 132):

> The lineage section of a quality report shall include a description of the source material from which the data were derived, and the methods of derivation, including all transformations involved in producing the final digital files. The description shall include the dates of the source material…

As geospatial workflows began to transition from analog to digital environments, it became clear that lineage-implied geoprocesses would need to be tracked from their origins, through revisions to the data, and finally to the output (Moore 1983). Chrisman (1983) noted that unfortunately over its lifetime, lineage information in quality records would be subject to entropy or fragmentation as a result of continuous GIS maintenance. He described *reliability diagrams* (for intelligence and other reliability-sensitive applications) embedded with lineage-related geometry and attributes (e.g., polygons identifying specific aerial photographic sources) and recommended them to be incorporated in typical GIS design. While not typically portrayed as lineage or provenance today, this type of lineage-related geodata, such as DigitalGlobe image collection footprints accessible in Google Earth, is extremely useful for visualization purposes and may resist digital entropy due to established geodata interoperability.

Beyond the challenges presented by digital records of lineages for multiple geodata versions, Langran and Chrisman's (1988) emphasis on multitemporal GIS highlighted additional record complexity that would be required. Nyerges's (1987) discussion on geodata exchange implied that quality metadata (including lineage information) could eventually facilitate geoprocessing design (workflows) with the two being mutually dependent. Others including Grady (1988) reasoned that lineage need not only support records of data quality but could in turn be used to record societal mandates (e.g., legislative drivers of geodata development) in the lineage information. While the existence of these additional complexities and potential requirements for geospatial lineage/provenance did not thwart attempts to forge ahead with possible software solutions, they pointed to significant challenges.

19.2.1 Digital Provenance in Remote Sensing and Geospatial Workflows

Over the last few decades and especially in the last 5 years, there has been significant attention given to understanding lineage/provenance in computer systems, and a variety of formalisms have been developed to understand their role in scientific workflows (e.g., Bose and Frew 2005; Buneman and Davidson 2010; Hey et al. 2009; Simmhan et al. 2005). In the following text, we highlight pioneering digital advances with geospatial lineage (circa 1990s) and more recent geo-cyberinfrastructure advances in provenance (circa 2000s to present).

19.2.1.1 Pioneering Work in Geospatial Lineage

As Chrisman (1986) suggested, "evaluation and judgment of fitness of use must be the responsibility of the user, not the producer. To carry out this responsibility, the user must be presented with much more information to permit an informed decision" (p. 352). Moellering et al. (1988) later emphasized producers' obligation to first document and update the lineage of their data in order to trace all the work (whether analog or digital) from original source materials through the intermediate processes to final digital output. It became obvious that both GIS software and international standards would be needed to facilitate the development and the maintenance of such records.

An early version of ESRI's ARC/INFO Geographic Information System featured a LIBRARIAN module capable of capturing

and querying some aspects of geospatial lineage. Using the module's CATALOG command, a database administrator could retrieve information on map production status as well as review time stamps and coordinates of recent map updates (Aronson and Morehouse 1983). In the mid-1980s, the U.S. Geological Survey (USGS) began development of a GIS-linked automated cartographic workflow system called Mark II with partial lineage capabilities. An important part of Mark II's design was its capacity to track the location (e.g., network address) of datasets and their progress from curated archive toward final map products (Anderson and Callahan 1990; Guptill 1987). While it was envisioned this system would play a key role in fulfilling the National Mapping Program's mission through 2000, the agency focus transitioned by the mid-1990s toward GIS data development including the National Map. The first reported development of a system to specifically and directly address geospatial lineage was David Lanter's *Geolineus* project commenced in the late 1980s as part of his doctoral research at the University of South Carolina's Department of Geography (Lanter 1989). As the prototype pioneering work in geospatial lineage/provenance, this is reviewed in detail with added explanation.

Lanter invented a method and means to capture, structure, and process geospatial lineage to determine and communicate the meaning and integrity of the contents of a GIS database (Lanter 1993a). His metadata and processing algorithms track and document remotely sensed and other geodata sources and analytic transformations applied to them to derive new datasets. In addition to differentiating between source and derived datasets, Lanter further distinguished intermediate and product-derived datasets. More concisely, let

$$\text{Datasets} = \{\text{Dataset}_i : i = \text{source}, \text{derived}\},$$

$$\text{Dataset}_{\text{derived}} = \{\text{Dataset}_{\text{derived}.k} : k = \text{intermediate}, \text{product}\}.$$

Source datasets can be the results of in situ sampling and data collection, remote sensing, or ancillary data (e.g., digitization of maps, or thematic data resulting from digital processing of remotely sensed data). Initially, only source datasets are available for geoprocessing and transformation into a derived dataset (Figure 19.2; $n \geq 1$, $m = 0$). Later, new datasets can be generated exclusively from derived datasets ($n = 0$, $m \geq 1$) using spatial analysis transformations such as reclassification, distance measurement (buffering), connectivity, neighborhood characterization, and summary calculations. Alternatively, new datasets can be derived from inputs that include sources, derived, or both ($n + m > 1$) using multi-input transformations such as arithmetic, statistical, and logical overlays, as well as drainage network and viewshed determinations.

Lanter classified datasets into source, intermediate, and product types (Figure 19.2), and related them to one another as inputs and outputs of each data processing step of an analytical application. He gave input datasets *parent* links pointing to output datasets they were used to create (Who am I the parent of?)

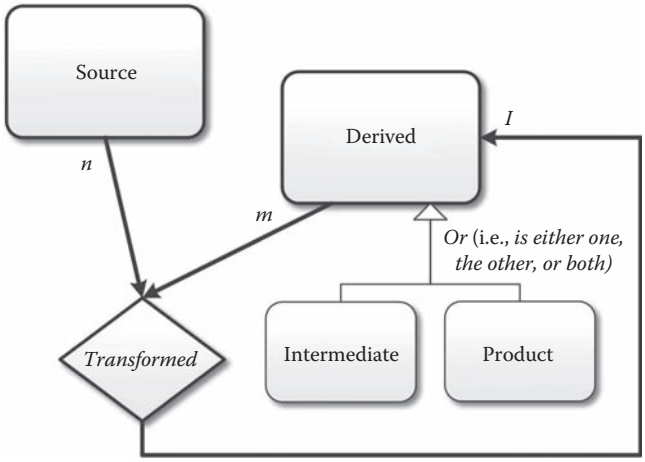

FIg u r e 19.2 Relationship among source and derived datasets, where each instance of the latter may be either an intermediate or product dataset, or both.

and provided output datasets *child* links connecting them back to their input datasets (Who am I the child of?). Each parent-and-child relationship was defined as an ordered pair of input and output datasets. Lanter's parent relationship identified the derived output given a source or derived input dataset, while his child relationship would identify a derived or source dataset when given an output dataset.

Child links connecting output datasets to their inputs enabled automatic deduction of which datasets within an analytic database are sources and which are derived (Lanter 1993b). Derived datasets are connected to their inputs by child links, while sources lack such links. Lanter defined his *child* operator to take a derived dataset, access its child links, and identify inputs used to create it. His *Ancestors* algorithm applied the *child* operator and by a recursive function traced the child links to identify datasets used to create a derived dataset, including any sources in the geoprocessing application. Lanter defined the *parent* operator to take a source or derived dataset as input, and access and traverse its parent links to identify all the outputs derived from it. His *Descendants* function recursively traced parent links and identified all datasets derived from a source or other derived input dataset used within a geoprocessing application.

Classification of datasets into source, intermediate, and product paved the way to structuring additional lineage metadata attributes. Lanter used the artificial intelligence *frame* data structure to organize knowledge about the metadata properties of source, intermediate, and product dataset types. Each source dataset was provided a frame for storing source properties such as its name, feature type(s), date(s), responsible agency, scale, projection, and accuracy attributes. He provided each derived dataset with a frame for storing detailed metadata elements about where it is physically stored, the command applied to derive it, the command's parameters, who derived it, and other aspects of its derivation. Lanter saw products as derived datasets that were provided an additional frame for metadata detailing

the analysis goal the dataset was intended to meet, intended audience/users of the dataset, when it was released, etc.

More formally, each Dataset_i (i.e., source, intermediate, or product) was provided an ordered list of metadata properties, A_j, such that $A_j = \{A_{j1}, A_{j2}, \ldots, A_{jk=f(i)}\}$. Specifically,

$$\text{Dataset}_{\text{source}} A_{\text{source}} = \{\text{Name}, \text{Features}, \text{Data}, \text{Scale}, \text{Projection}, \text{Agency}, \text{Accuracy}, \ldots\},$$

$$\text{Dataset}_{\text{intermediate}} A_{\text{intermediate}} = \{\text{Name}, \text{Command}, \text{Parameters}, \text{User}, \text{Date}, \ldots\},$$

and

$$\text{Dataset}_{\text{product}} A_{\text{product}} = \{\text{Goal}, \text{Audience}, \text{Release Date}, \text{Intended Use}, \ldots\}.$$

Given $w \in \{source, intermediate, product\}$, m a metadata property of w, and a_{wm} a value of A_{wm} then a $\text{dataset}_w = (a_{w1}, a_{w2}, \ldots, a_{wl})$.

Lanter's lineage metadata structure represented datasets as nodes coupled with source, command, and product properties, and connected them with parent and child links (Figure 19.3). Lanter adapted the *Ancestors* function to respond to lineage queries and to report on data sources and the sequence of processing (i.e., data lineage) applied to sources and intermediates to derive a target dataset (Lanter 1991). He integrated the *Ancestors* function with a rule-based processor that checked the inputs of each user-entered GIS command, determined their related sources, and evaluated their metadata to detect and warn users when they were entering commands that would otherwise combine datasets of incompatible properties such as projections, scales, and dates (Lanter 1989). Lanter subsequently modified the *Descendants* function to automatically generate and run GIS scripts and propagate new source data to update dependent intermediates and products (Lanter 1992a).

19.2.1.1.1 Geolineus

Lanter and Essinger designed the *lineage diagram*, an icon-based flowchart graphical user interface (GUI), to enable users direct interaction with lineage metadata to understand, modify, and maintain their analytical applications and ESRI's ARC/INFO's spatial data contents (Essinger and Lanter 1992; Lanter and Essinger 1991), and implemented it in *Geolineus* (Lanter 1992b)—the first lineage-enabled geospatial workflow system.

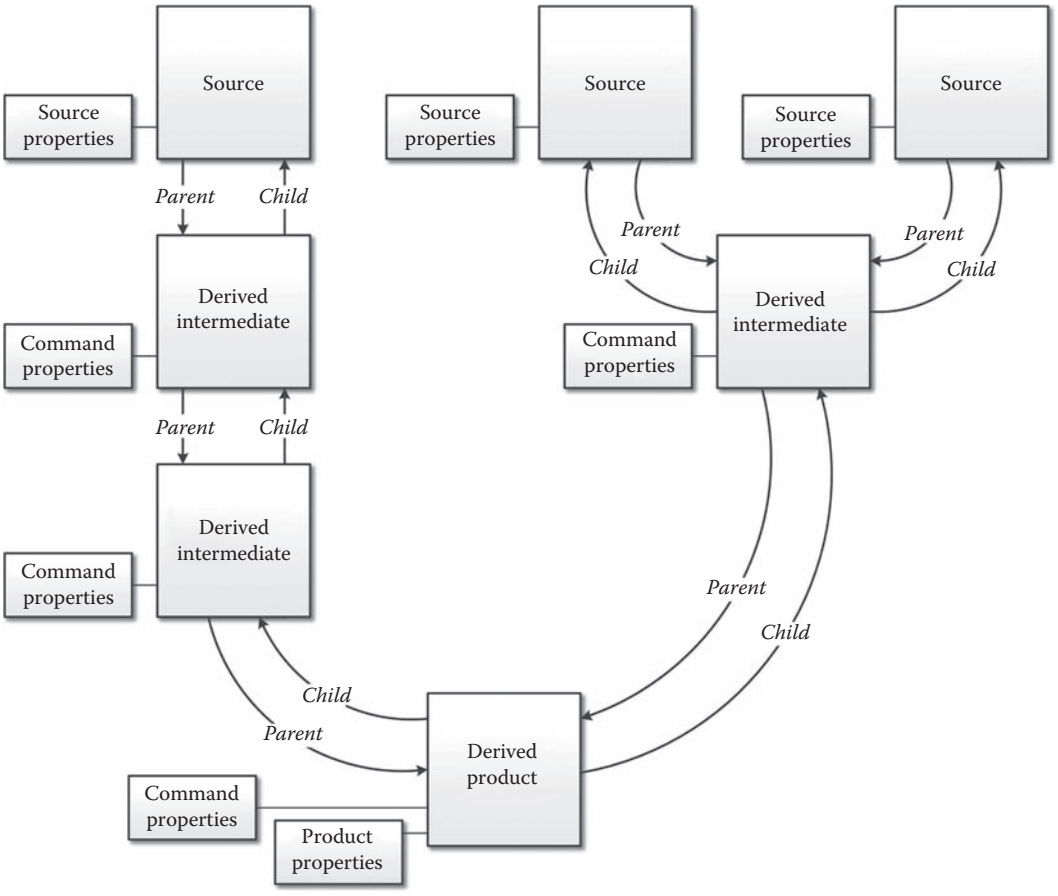

FIgure 19.3 Lineage represented as structured metadata consisting of parent and child links connecting source, intermediate, and product datasets. While each source possesses a frame containing metadata properties, frames for derived datasets detail the GIS command used in its creation. In addition, derived product datasets possess a frame describing analytic goals, release date, and users.

Geolineus enabled users of ESRI's ARC/INFO and GRID (for image processing) to capture, create, save, exchange, analyze, and reuse lineage metadata to maintain their GIS databases. Geolineus's user interface included a lineage data flow diagram within one panel, coupled with another panel containing its own command line processor in place of the command line processor of ARC/INFO and GRID. As users added source datasets, they were presented with a form to document them, after which they were displayed along the top of the data flow diagram, each with a square icon with a bar at its top. Symbols within the icons would identify if the dataset contained points, lines, polygons, raster grids, and/or value attribute tables. Icons further down the flowchart represent datasets derived with geospatial analysis operations such as CLASSIFY, BUFFER, and INTERSECT. Geolineus would create icons and arrows connecting them to the flowchart automatically as these commands were used. Icons at the bottom of the flowchart signifying products, that is, derived datasets that represent the final step in the geospatial application, each included a bar along its bottom edge (Figure 19.4).

Written in Common LISP, Geolineus used multiprocessing capabilities of UNIX to run the geospatial processing software as a background job while providing its own command line window to the user. As the user would enter a command transforming one or more spatial datasets to derive a new one (e.g., classify, union, and intersect), Geolineus would parse, extract the identities of the input and output datasets and the command and its parameters, and pass the command off to the geospatial processing software running in the background. Geolineus monitored the processing and feedback messages returned from the geospatial processing software and presented them to the user within its own command line window in real time to provide the user with the illusion that they were interacting directly with the geospatial software. The system detected whether the processing successfully completed and, if so, the input/output relationships and command information would be stored within its metadatabase and the data flow diagram dynamically updated with a new icon for the output dataset connected by dotted arrows (labeled with the command) emanating from its data sources. When the final data product was reached, the user could click on its icon and fill in the displayed product form to document the analytical goal it represented (e.g., wells at risk from nearby leaking pipes) and who should be contacted if it was updated or changed.

Geolineus also monitored each dataset in the diagram to determine if it was edited or replaced. If a source or derived dataset was found to be modified, its icon would turn yellow in the diagram. If the dataset needed its topology rebuilt in response to an edit, the polygon or line feature symbol within the icon would turn red. If a derived dataset was potentially out of date because one of the sources it was derived from was edited, its icon would turn orange. Users could click on a source icon to

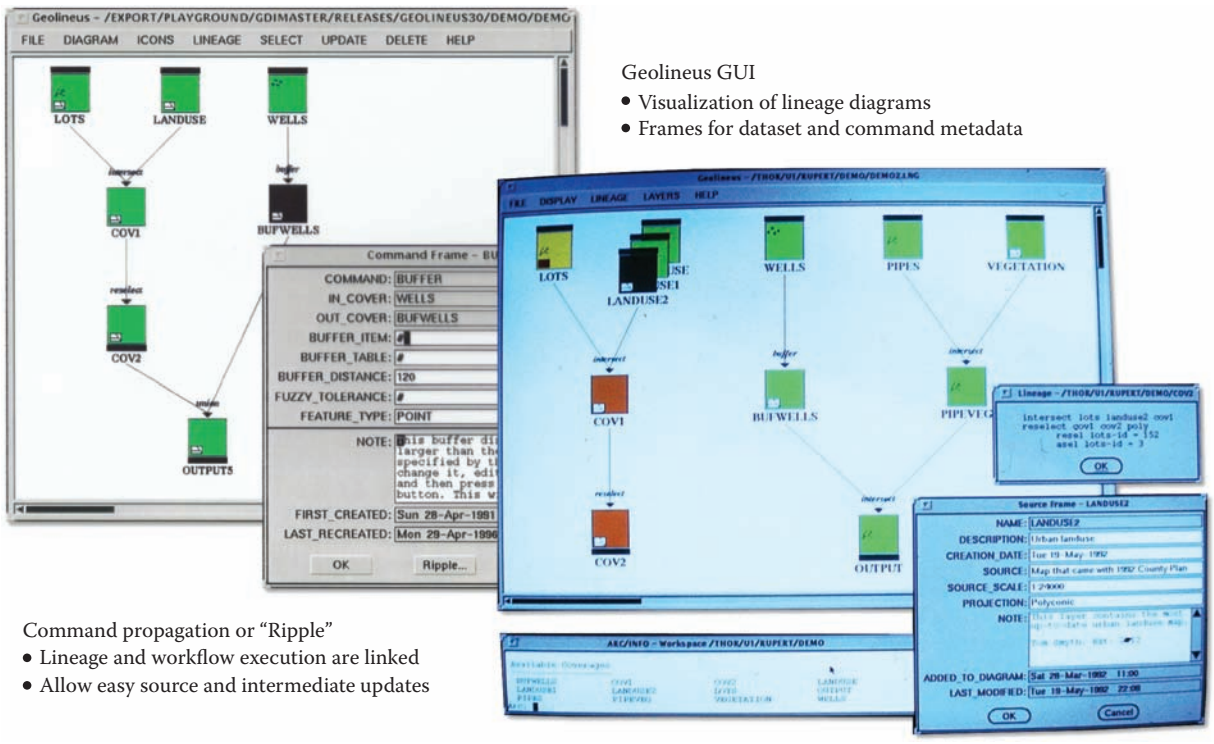

FIg u r e 19.4 Examples of Geolineus' interactive lineage diagram GUI. The left screen shot illustrates linkage of a command frame to the BUFWELLS dataset highlighted in black; clicking on the "Ripple" button at the bottom of the command frame propagated changed buffer command parameters throughout the workflow. The right screen shot illustrates the LANDUSE2 dataset's source frame, and commands applied to that source to derive the COV1 and COV2 datasets.

view metadata about what it represented, when it was created, where it came from, and cause a propagation (*ripple*) of its data through sequences of commands updating intermediate and product datasets originally derived and created from it. Users could also click on a derived dataset to rerun the commands necessary to pull new, updated, and modified source data through the flowchart's processing logic and update the derived geodata (Lanter 1994b). Geolineus enabled users to save, exchange, and import lineage metadata in ASCII file format to meet the Federal Geographic Data Committee's (FGDC's) Content Standard for Digital Geospatial Metadata, document exchanged datasets, accompany source datasets, provide logic for use within other instances of the software to reconstitute a derived geospatial database, and serve as reusable analytic application logic templates to snap to replacement source datasets associated with different study areas.

Lanter and Veregin (1991, 1992) modified the lineage metadata to store error measures and demonstrated new algorithms for mathematically modeling how error measures of data sources are transformed and combined through a sequence of spatial analysis functions to determine the quality of a derived spatial analytic product dataset. They added properties to the source frame for storing user-entered measures of data source error, and properties to the command frame for storing derived error measures for each derived dataset. Geolineus's *Ancestors* and *Descendants* functions were modified, enabling them to access error properties of input datasets, select and apply an appropriate error propagation function to derive, store and present the error measure of the derived geospatial dataset as the user typed in their spatial analysis commands. Lanter (1993b) followed this by modifying the lineage metadata and *Ancestors* and *Descendants* functions to use commercial costs of data storage and central processing time to calculate and compare the relative costs of storing versus using lineage metadata to re-derive intermediate and product datasets when needed. The results enabled Geolineus to determine an optimal spatial database configuration and choose which datasets to delete and re-create when needed. Veregin and Lanter (1995) modified the metadata frames and *Ancestors* and *Descendants* functions to demonstrated lineage metadata-based error propagation techniques for identifying the best data source to improve based on cost value per product quality improvement achieved. Geolineus was programmed to systematically vary the error value of each source, iteratively applying mathematical error propagation functions and determining its effect on product quality. Comparing slopes of lines graphing source error versus resulting product enables determination of relative impact each data source has on data product quality.

To help analysts and auditors understand undocumented preexisting analytically derived GIS datasets, Lanter provided Geolineus with capabilities to extract lineage metadata and create a lineage diagram from ARC/INFO log files. Similar to the history list the UNIX operating system recorded user commands into, the ARC/INFO GIS copied user-entered GIS commands into log files, which it stored and maintained within the operating system file system directories or workspaces. Geolineus's "Create from log" option automatically extracted lineage metadata and created a lineage diagram reflecting the commands contained in the log file of a targeted workspace. While the log files contained the name of the dataset and the file system path indicating where the dataset was stored, they did not include other source metadata (i.e., thematic feature type, date, agency, scale, projection accuracy, etc.) necessary for achieving a clear understanding of contents and qualities of each source. To resolve this, analysts and auditors working with Geolineus clicked on the source icons within the lineage diagram, brought up source frames, and filled in missing source metadata if available.

Lanter (1994a) formulated metadata comparison functions that enabled him to automatically determine if two spatial analytic datasets were equivalent and if two geospatial datasets were similar. These were implemented within Geolineus to identify common and unique geospatial data processing conducted in and among multiple GIS workspaces (Lanter 1994b). His search for datasets common to different lineage metadata representations began with a determination of source equivalence. Source datasets were considered equivalent when their source metadata properties were found to have equivalent values, assuming these properties are sufficient to uniquely identify their contents and qualities. This enabled the detection of equivalent and possibly redundant source datasets that are stored in different file system locations but contain equivalent content. Source data equivalence was implemented in Geolineus's "Merge" function, which enabled users to analyze log files of data processing applications run in different workspaces and produce a single unified lineage diagram illustrating their common and unique data sources (Figure 19.5).

In turn, Lanter considered derived datasets equivalent when (1) their input datasets were equivalent and (2) when transformations applied to compute them from their inputs were found to be equivalent. Derived data equivalence was implemented in Geolineus's "Condense" function. Condense enabled Geolineus's users to detect the lineage representations of redundant processing and resulting copies of derived data stored under different names or in different file system locations, remove the redundant data, and consolidate the transformational logic applied in their derivation within the unified metadata and lineage diagram.

Lanter and Surbey (1994) put Geolineus's capabilities to work in the first enterprise GIS database and geoprocessing quality audit. They systematically evaluated the geospatial data sources, products, and geoprocessing applied to derive 40 GIS data products, developed within 14 projects, for eight departments of a large southwestern electric utility. Lanter and Surbey identified 54 data sources among the 806 raster (GRID) and vector (ARC/INFO) GIS datasets produced for the electric utility's decision makers. They interviewed the department's GIS specialists, filled in as much missing source metadata that could be recalled and confirmed, and noted findings about what was unknown about the source data. In addition to assessing adequacy of source data documentation, Lanter and Surbey analyzed the resulting lineage diagrams they created and measured the complexity of spatial analysis logic employed within the 14 GIS application projects.

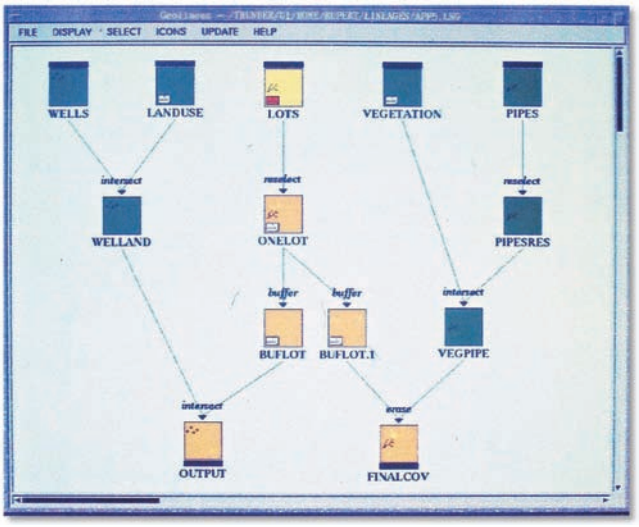

FIg ur e 19.5 Geolineus GUI illustrating the results of the "Merge" function unifying two lineage diagrams at their common source LOTS dataset, and "Condense" function which removed redundant processing and derived datasets unifying them at their common intermediate ONELOT dataset. The red mark on the yellow LOTS dataset indicated an edit and need for polygon topology repair, and the orange color in derived datasets reflected the need for changes to be propagated using the "Ripple" function to update the OUTPUT and FINALCOV products.

Lanter (1994a) extended his dataset equivalence tests and formulated a set of source and derived data similarity tests in order to detect patterns of data usage and derivations within workflows. He coupled these with a geospatial data taxonomy (e.g., Anderson et al. 1976) and a GIS command language taxonomy (e.g., Giordano et al. 1994) to generalize analytic logic employed within prior applications of GIS and find common data analysis patterns. Lanter presented a suite of lineage-based metadata analysis methods for detecting and communicating commonalities and differences among particularly useful spatial analysis applications, with the intent of improving geographers' basic understandings of spatial analytic reasoning and to provide a method and means to answer fundamental geographic questions including the following:

- Are there a finite number of spatial relationships studied within and among different GIS applications areas? If so, what are they?
- Within particular applications areas, are certain spatial relationships stressed more than others? If so, what are they?
- Are common patterns of analytic logic used to build up certain complex spatial relationships? If so, what are they?
- Are certain spatial relationships consistently sought at different spatial, thematic, and temporal scales?

19.2.1.1.2 Geo-Opera

Also incorporating geospatial lineage into its design in the 1990s, Geo-Opera was developed as a prototype geoprocessing support or geospatial workflow management system that would enable interoperability, data recovery, process history records, and data version monitoring in commercial GIS (Alonso and Hagen 1997). Geo-Opera was based on a modular architecture composed of interface, process, and database modules. It used its own process scripting language and was based on the OPERA distributed operating system that allowed for data distribution and process scheduling within a local area network. In Geo-Opera, geodata first had to be registered before being utilized, thus mitigating the common problem (that persists today) of lack of source metadata.

19.2.1.2 Expansion of Limited Provenance in Commercial and Public Geoprocessing

As commercial and public (i.e., free and open-source) GIS applications rapidly matured and grew in analytical power, it became necessary to provide a way for users to build and track workflows involving interactions among many complex and varied geoprocessing operations. At least two approaches to create and manage workflows have emerged—graphical block programming and integrated database style querying. The first is essentially a visual interface to programming, while the second approach appeals to users trained in database management. While both enable at least some form of provenance, enterprise database systems can provide record-level transaction management, which, at least in detail, is beyond the scope of this chapter.

By far the most common approach, due in large part no doubt to its ease of use and graphic nature, is graphic block programming approach. Commercial GIS and remote sensing applications such as ESRI's ArcGIS, Hexagon Geospatial's ERDAS IMAGINE, and PCI's Geomatica expose their complex processing tools in this way (e.g., Figure 19.6). The free and open-source GRASS GIS also provides a visual programming environment for both vector and raster operations. Boundless (formerly OpenGeo) is developing a visual programming environment for QGIS, an open-source GIS. All of these environments capture and store some degree of provenance including in some cases important environmental settings that can significantly affect geoprocessing results. It is important to note that visual programming interfaces can normally be bypassed by skilled users familiar with application programming interfaces (APIs) or scripting languages integrated with GIS.

A less common approach is incorporated almost exclusively in enterprise databases that have integrated spatial operators and native spatial data objects. With this level of integration, spatial operators become just another type of operation exposed through (often extended) structured query language (SQL) interfaces. At a minimum, the SQL commands used to manipulate spatial data objects are recorded and may be inspected in a variety of graphical environments. As of mid-2014, most spatially enabled databases have extensive vector operators but limited raster or image operators of particular interest in remote sensing workflows. However, technologies such as the Oracle Spatial and Graph option for Oracle Database 12c now enable image algebra in addition to other remote sensing–oriented capabilities such as LIDAR data processing.

FIgure 19.6 Geospatial scientists interact with the ASA Hazard Map (Tullis et al. 2012), a remote sensing–assisted silviculture assessment spatial decision support system, and its five downloadable ArcGIS 10 ModelBuilder workflows using a collaborative multitouch display. Each yellow rectangle represents an ArcGIS tool (e.g., for estimating incoming solar radiation using a LIDAR-derived DEM) and, together with inputs, outputs, and other parameters (colored ovals), constitutes a geoprocess. After execution, geoprocesses are marked with shadows that may be cleared only by resetting or changing geoprocess parameters including geoprocessing environment settings. User interaction with shaded geoprocesses effectively provides access to workflow-level provenance information for the most recent execution and facilitates dependent geoprocess updates after any modifications are made.

19.2.1.3 Interest in Provenance as a Component in Geo-Cyberinfrastructure

Cyberinfrastructure (CI) is a concept that has been extensively used since Atkins et al. (2003) *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. As a common infrastructure for scientific data and computing, a variety of components and topics are involved in CI construction, including hardware, software, network, data, and most importantly people. The development of CI can be traced back to the construction of the TeraGrid infrastructure in the 1990s that was replaced by eXtreme Science and Engineering Development Environment (XSEDE) in 2012. By linking supercomputers through high-speed networks, TeraGrid and XSEDE have provided a powerful computing environment and capability to support petascale to exascale scientific computation.

In a broader and general domain, the internet can be regarded as the CI since all computers can be linked together through the network. When varieties of data and databases can be hosted and connected on the internet, data processing and analytics can be conducted through service-oriented computing (SOC). In early 2000, web service technology was proposed to be the solution for software interoperability. In this vision of interoperable software engineering and integration, a service is an API defined in Web Services Description Language, while communication between the service provider and the service requester is based on the Simple Object Access Protocol (SOAP). Meanwhile, Representational State Transfer (REST) services are based on HTTP protocol using its GET/POST methods for mashup online resources (Fielding 2000). Both SOAP- and REST-based services can be deployed for remote procedure calls. Furthermore, with the advancement of telecommunication infrastructure and technology, wireless networking has been providing another approach for data sharing and network computing, while varieties of sensor networks can be connected through wireless networks.

Today, different computing networks can be linked together. Supercomputers on the XSEDE can be accessed through a web portal, while wireless sensor networks can be accessed on the internet. Such a huge but heterogeneous CI increases the difficulty and complexity for geoprocessing, workflows, and provenance research (Wang et al. 2008). In 2007, the National Science Foundation (NSF) released the DataNet program that would support comprehensive data curation research over the CI, and NSF's Data Infrastructure Building Blocks program "will support development and implementation of technologies addressing a subset of elements of the data preservation and access lifecycle, including acquisition; documentation; security and integrity; storage; access, analysis and dissemination; migration; and deaccession," as well as "cybersecurity challenges and solutions in data acquisition, access, analysis, and sharing, such as data privacy, confidentiality, and protection from loss or corruption" (NSF 2014), which are all topics relevant to the themes in provenance.

19.2.2 Specifications and International Standards for Implementation of Shared Provenance-Aware Remote Sensing Workflows

Since the Moellering et al. (1988) proposal identifying geospatial lineage as the first component in a data quality report, a variety of provenance-related standards have been developed including those at the international level. The most current standard in use

is the International Standards Organization's ISO 19115-2, which has been endorsed by the Federal Geographic Data Committee (FGDC; ISO 2009).

19.2.2.1 Metadata Interchange Standards

In the United States, the FGDC has been coordinating the development of the National Spatial Data Infrastructure by developing policies and standards for sharing geographic data. The Content Standard for Digital Geospatial Metadata defines common geospatial metadata about identification information, spatial reference, status information, metadata reference information, source information, processing history information, distribution information, entity/attribute information, and contact information of the geodata creator.

Partially based on the FGDC's 1994 metadata standards, the ISO Technical Committee (TC) 211 published ISO 19115 Metadata Standard, covering a conceptual framework and implementation approach for geospatial metadata generation. ISO/TC 211 suggests that metadata structure and encoding are implemented based on the Standard Generalized Markup Language that has the same format as the Extensible Markup Language (XML). The XML-based ISO metadata standard has exemplified the advantage in implementation covering a variety of elements in standard definition. ISO 19115 Metadata Standards contain a data provenance component in defining the data quality within the metadata. Unfortunately, while Gil et al. (2010) defined provenance in part as "a form of contextual metadata," their emphasis on the clear distinction between provenance and traditional metadata is not reflected in metadata interchange standards for provenance. For instance, geodata cardinality between a land use land cover (LULC) map and its metadata is one to one; in contrast, geodata cardinality between an LULC map and its provenance is potentially one to many, thus leading to much duplicate information in a "provenance as metadata" paradigm.

19.2.2.2 Provenance-Specific (Non-Metadata) Interchange Standards

Provenance-specific (non-metadata) standards have been developed at different levels and in a variety of domains. ISO 8000 has a series of standards that address data quality. ISO 8000-110 specifies requirements that can be checked by computer for the exchange, between organizations and systems, of master data that consists of characteristic data. It provides requirements for data quality, independent of syntax. ISO 8000-120 specifies requirements for capture and exchange of data provenance information and supplements the requirements of ISO 8000-110. ISO 8000-120 includes a conceptual data model for data provenance where a given "*provenance_event* records the provenance for exactly one *property_value_assignment*," and every "*property_value_assignment* has its provenance recorded by one or many *provenance_event* objects."

In order to trace the changing information and the provenance of data (and by implication geodata) over the web, W3C has recently published a series of documents and recommendations (starting with the term PROV) to guide the provenance interchange on the web. Specifically, the current PROV data model for provenance (PROV-DM; Moreau and Missier 2013) "defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world" (Gil and Miles 2013).

To illustrate PROV-DM in a remote sensing and geoprocessing context, the provenance of a 2001–2006 canopy change layer incorporated in the ASA Hazard Map (Jones et al. 2014; Tullis et al. 2012) can be represented using PROV-DM structures. This may be encoded (Figure 19.7; Table 19.1) as agents (e.g., a specific version of PCI Geomatica as a software agent), entities (e.g., a Landsat image clipped to a forest boundary), activities (e.g., ATCOR2 atmospheric correction based on specific calibration and other parameters), and relationships (e.g., wasInfluencedBy to represent the influence of Wang et al. (2007) on the change detection methodology). It is important to note that PROV-DM is extensible such that subtypes of agents, entities, activities, and relationships can be identified as needed for domain-specific applications (Moreau and Missier 2013).

In the geospatial domain, the efforts of the Open Geospatial Consortium (OGC) initially (late 1990s and early 2000s) focused on the development of specifications that encouraged geospatial data interoperability such as the OGC Simple Features Specification. While not directly related to provenance, this effort has led to common ontologies and semantic structures that are foundational to the integration of geoprocessing, workflows, and provenance. In the 2000s, the OGC's attention shifted to web processing and interoperability of various web services. The OpenGIS Web Processing Service (WPS) specification (Schut 2007) has a *lineage* element in defining the request message to *execute* the spatial operation. In case lineage is defined as *true*, the response message from WPS will contain a copy of input parameter values specified in the service request definition. The OGC also developed the Sensor Web Enablement standard, in which the OpenGIS Sensor Model Language (Botts and Robin 2007) has one specific element that documents the observation lineage to describe how an observation is obtained. Elements of a number of earth observation process specifications, such as the Catalogue Services Standard 2.0 Extension Package for ebRIM Application Profile: Earth Observation Products (Houbie and Bigagli 2010), the Sensor Observation Service Interface Standard (Bröring et al. 2012), and others, increasingly have provenance-related components as key elements. The more recent developments in WaterML and the Open Modeling Interface have increasingly emphasized provenance components.

> The purpose of the Open Modeling Interface (OpenMI) is to enable the runtime exchange of data between process simulation models and also between models and other modeling tools such as databases and analytical and visualization applications. Its creation has been driven by the need to understand how processes interact and to

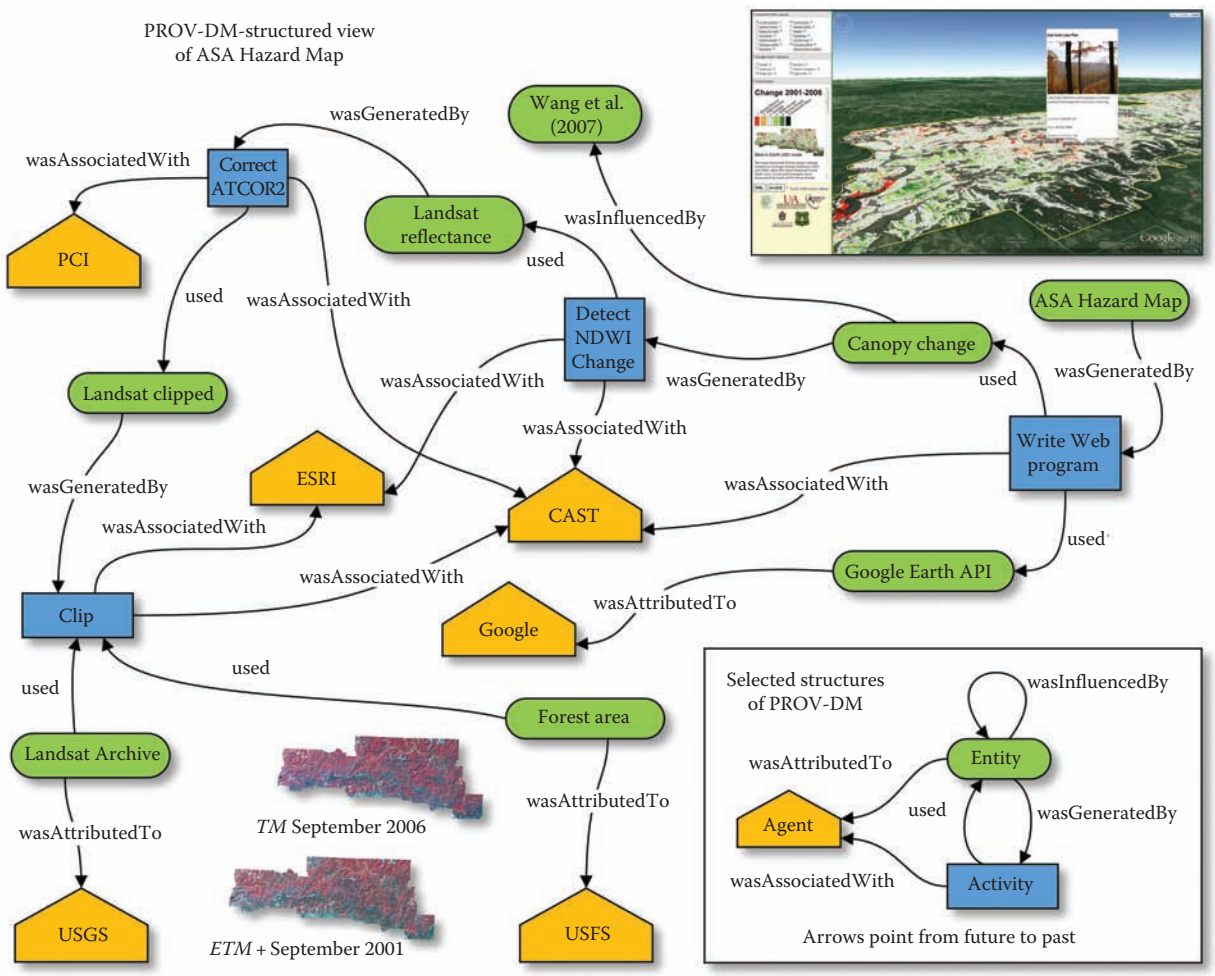

FIg u r e 19.7 Selected provenance of the ASA Hazard Map (Jones et al. 2014; Tullis et al. 2012) structured according to W3C's PROV Data Model (PROV-DM; Gil and Miles 2013; Moreau and Missier 2013; Table 19.1). Arrows (relationships) point from future to past, first from the online ASA Hazard Map to its 2001–2006 canopy change layer, then to various agents, entities, and activities involved in the canopy change layer's creation. Some entities (e.g., "Landsat Archive") represent PROV-DM collections of entities (e.g., individual Landsat images available from USGS), and many potential PROV-DM details are not shown.

> predict the likely outcomes of those interactions under given conditions. A key design aim has been to bring about interoperability between independently developed modeling components, where those components may originate from any discipline or supplier. The ultimate aim is to transform integrated modeling into an operational tool accessible to all and so open up the potential opportunities created by integrated modeling for *innovation* and wealth creation.
>
> **Vanecek and Moore (2014, p. ix, emphasis added)**

It is likely that future OGC efforts will increasingly focus on provenance. The OGC is a major participant in EarthCube (2014). In 2011, NSF's Cyberinfrastructure and Geosciences Divisions established the EarthCube community to promote geosciences data discovery and interoperability. The OGC plays a major role in this community, which, as of 2014, has several NSF-funded research and implementation grants pertaining to provenance records in geoprocessing.

19.3 Why Provenance in Remote Sensing Workflows

As Buneman (2013) argues, a "change of attitude" is in order regarding the role for provenance across a range of computer system-supported domains and activities, including (by implication), remote sensing workflows. He makes the comparison between scientific activities where it is considered obvious that such information should be recorded and other domains where there is little or no awareness of process history or its value. He concludes that "we should worry less about what provenance is and concentrate more on what we can do with it once we have it" (p. 11).

TABLe 19.1 Characteristics of PROV-DM Structures Including Core Types and Selected Relationships (Moreau and Missier 2013), Each with an Example Provided from the Provenance of the ASA Hazard Map (Jones et al. 2014; Tullis et al. 2012; Figure 19.7)

PROV-DM Structure	Interpretive Highlights	Example from ASA Hazard Map Provenance
Core types		
Agent	Need not be a person but could also represent an organization or even a specific software process	Center for Advanced Spatial Technologies (CAST) *agent* (organization) at University of Arkansas
Entity	May be physical, digital, or conceptual	Landsat 5 TM *entity* (satellite image) collected on September 15, 2006, over Ozark National Forest
Activity	Involves entities and requires some time to complete	ESRI ArcGIS 10 for Desktop "Extract by Mask" *activity* (software tool) used to clip the Landsat 5 TM imagery to the bounds of the study area), together with environment settings (e.g., "Snap Raster")
Selected relationships		
wasGeneratedBy	Can represent creation of only new entities (that did not already exist)	Clipped Landsat 5 TM image that has been corrected for atmospheric attenuation *was generated by* running the ATCOR2 algorithm
Used	Only implies that usage has begun (but not that it is completed)	A GIS model for detecting oak-hickory forest decline or growth *used* a clipped and atmospherically corrected Landsat ETM+ image collected September 25, 2001
wasAttributedTo	Links an entity to an agent without any understanding of activities involved	The Google Earth API (used to write a web program to generate the ASA Hazard Map) *was attributed to* Google
wasAssociatedWith	Links an activity to an agent	The ATCOR2 algorithm used to correct Landsat TM and ETM+ imagery for atmospheric attenuation *was associated with* PCI Geomatics through their Geomatica 10 platform
wasInfluencedBy	At a minimum, suggests some form of influence between entities, activities, and/or agents; however, highly specific influence may be captured	The 2001–2006 oak-hickory forest canopy change data produced for the ASA Hazard Map *was influenced by* Wang et al. (2007), who used statistical thresholds of change in Landsat-derived normalized difference water index (NDWI) over time to detect oak canopy changes in the Mark Twain National Forest

19.3.1 Remote Sensing Questions That Only Provenance Can Answer

For volumes that contain primarily raw or unprocessed geodata (e.g., imagery telemetered directly from a satellite sensor), provenance (as used in the present context) may not offer much over traditional metadata. However, when looking at geodata products resulting from complex geoprocessing workflows, there is much valuable information that metadata is ill-equipped to capture and store.

There is sometimes confusion concerning what provenance offers in terms of valuable information to an end user over the far more common and better supported (in terms of software integration) metadata. One way to structure such a discussion is to look at some of the questions data users might ask that can be only reasonably answered using (at least in part) detailed provenance information. For instance, one might ask the following regarding a remote sensing–derived product:

1. What was the processing time necessary to create this product, and what system configuration was implemented (including disk, processor, and RAM information)?
2. In what exact order were processing steps taken, and what precise parameters were used during each intermediate step? Was the process completely automated, or were manual steps (such as onscreen digitization) included in the workflow?
3. What datasets, both source and derived, were used to create this product, and how did each contribute to the product?
4. How were errors expressed and propagated during the product's creation? Is the result statistically significant?

In addition to these, several questions could be asked trying to identify the source of errors or anomalies in the data. For example, one might wonder at what point in the geoprocessing did a specific region get assigned null values and why? Using provenance data, it should be possible to analyze two similar data products and compare their processing history to see how and why they differ (Bose and Frew 2005; Lanter 1994a). The opportunity to better understand and manage the complexity of spatial scale in remote sensing–assisted workflows is a further justification for provenance-enabled geoprocessing (Tullis and Defibaugh y Chávez 2009). Finally, provenance-aware systems could be used to enable and support temporal GIS analyses, which require detailed history of a dataset's change over time to properly function (Langran 1988).

The value of provenance tracking and visualization was demonstrated in a study conducted at the Regional Geospatial Service Center at the University of Texas, El Paso (Del Rio and da Silva 2007). In this study, conducted as part of NSF's GEON Cyberinfrastructure project, web services were built to perform geoprocessing tasks (filtering, gridding, and contouring) required to create a contoured gravity map from a raw gravity dataset. Del Rio and da Silva generated multiple contour maps with incorrect parameters (e.g., a grid size parameter larger than important anomalies in the gravity field), and participants in the study were asked to evaluate each contour map with and without provenance information. Without provenance information, subject matter experts were able to detect errors in only 50% of the cases and to explain cause in only 25% of the cases. Nonsubject matter experts fared much worse (11% and 11%). However, when

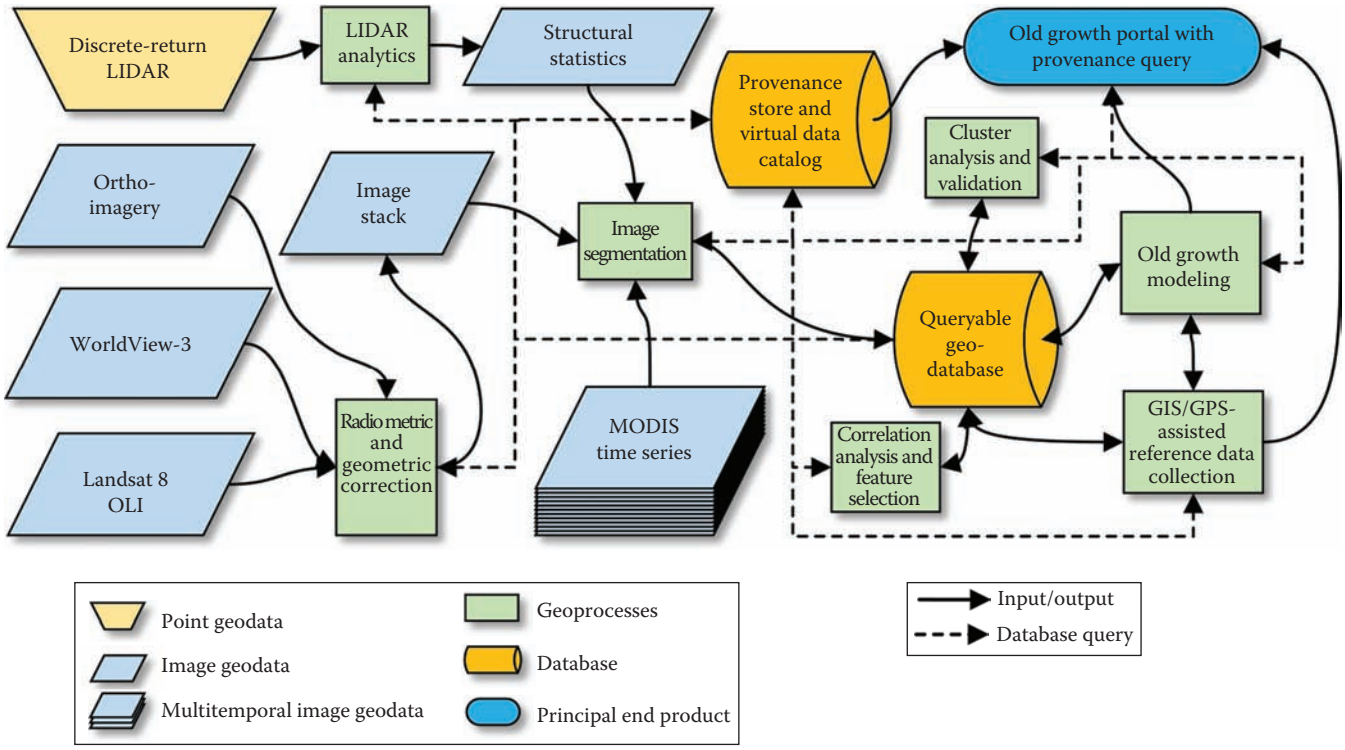

FIg u r e 19.8 A provenance-enabled workflow for extracting old-growth bald cypress (*Taxodium distichum*) forest quality and biophysical parameters from airborne LIDAR and orbital multispectral imagery. Such a framework could enable (a) detailed methodological transparency related to old-growth inventory maps published in the web portal, (b) detailed transparency of the accuracy assessment, and (c) autogenerated scripts for the replication of the detailed workflow despite a complicated blend of commercial and open-source workstation-level software, including cyber-enabled high-performance geoprocessing tools.

provenance information was provided and visualized using Del Rio and da Silva's (2007) ProbeIt! (a provenance visualization tool), the subject matter experts were able to detect and explain all the errors. A more impressive result, though, is that provenance improved the ability of the nonsubject matter experts to detect and explain errors by a factor of 7 (78%).

It can be argued that extending a traditional remote sensing workflow to include provenance information offers a number of immediate advantages. For example, a web portal displaying a detailed map of old growth forest in the southeastern United States could include provenance query that enables (1) detailed methodological transparency, (2) detailed transparency of the accuracy assessment, and (3) autogenerated scripts for replication of the detailed workflow even though it includes a complicated blend of commercial and open-source software, including cyber-enabled high-performance geoprocessing tools (Figure 19.8).

19.3.2 Provenance and Trust in the Remote Sensing Process

When provenance of sources, intermediates, and products is captured and maintained throughout the digital remote sensing process life cycle, data quality improves, an audit trail is available for reviewers and users, replication is straightforward, attribution is streamlined, and the interpretability of geodata products is enhanced (Simmhan et al. 2005). However, any one of these advantages by itself is not as critical as establishing trust in the remote sensing process, which is clearly not immune to being oversold (Jensen 2007). Generally speaking, statistical validity is a key to establishing trust in scientific processes and has been extensively developed in remote sensing workflows (Congalton 2010). Simply maintaining an audit trail that demonstrates sound choices in statistical processing parameters and methods is one way to help ensure that trust is not erroneously called into question.

In addition to maintaining an audit trail, statistical techniques can be applied to a given workflow to determine trust, and recent work has focused on assigning a measure of trust by using a workflow's provenance. While not demonstrated in a remote sensing context, a nonstationary hidden Markov model can be used to provide a measure of this trust (Naseri and Ludwig 2013). Provenance can also be useful in helping to determine workflow trust for data on the web. When handling linked data provenance, data authentication may serve as an estimation of data trustworthiness. Uniform resource identifiers and digital signatures can be used as a measure of authentication (Hart and Dolbar 2013).

Traditional indicators of trust may also be used in conjunction with provenance. Recommendation, authority, believability,

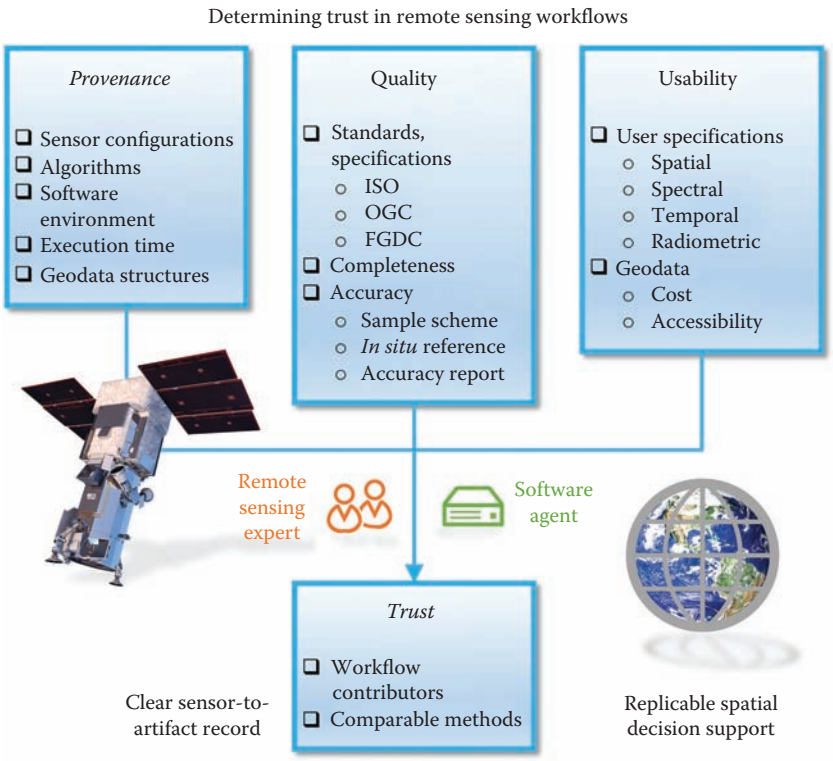

FIg u r e 19.9 Provenance, quality, and usability can be used by remote sensing experts to make a subjective decision on workflow trust (Gamble and Goble 2011; Jensen 2005; Malaverri et al. 2012); a sample of their characteristics is shown. As geoprocessing, workflows, and provenance are integrated, software agents can objectively influence replicable spatial decision support. (Artist image of WorldView-3, Courtesy of DigitalGlobe, 2014.)

reputation, and objectivity can all serve as indicators of trust (Gamble and Goble 2011). In addition to provenance, a workflow's quality and usability should also be evaluated when determining trust. As geoprocessing, workflows, and provenance are integrated, software agents can objectively influence replicable spatial decision support (Figure 19.9). Until more quantitative techniques are developed for measuring trust of geographic workflows using provenance, measures of quality and usability used in conjunction with subjective trust indicators should be examined before making a decision to trust a workflow and its lineage.

19.4 Selected Recent and Proposed Provenance-Aware Systems

Many provenance-aware systems have largely been concerned with provenance capture, and this capability is critical for synergistic geoprocessing, workflows, and provenance of interest in remote sensing applications (Figure 19.1). The characteristics of the captured provenance information can greatly influence how it may benefit the remote sensing process. Of particular significance to geospatial applications is the level of provenance detail or granularity. Fine-grain provenance is obtained at a data level and can even deal with individual pixels (Woodruff and Stonebraker 1997), whereas coarse-grain provenance represents the workflow level (Tan 2007) and can thus be used to facilitate scientific reproducibility. As Henzen et al. (2013) point out, the quality of provenance communication is also very important even when presented in a text format. A number of recent and proposed provenance-aware approaches and systems related to remote sensing (Table 19.2) have addressed these and other issues.

19.4.1 General Approaches

19.4.1.1 Inversion

Inversion was developed for fine-grain data lineage and provenance in database transaction and transformation (Buneman et al. 2001; Cui et al. 2000; Woodruff and Stonebraker 1997). Database queries or processing functions that generate a view, table, or new data product can be registered in a database system (or provenance store). Registered database transformations can be inverted so as to trace the lineage between the data product and the sources that derive the product. For example, when a view is created or updated, the inversion method can help identify the source tables from which the view is generated. Although inversion can be applied in data provenance for geospatial data, not all functions are invertible. However, a weak or general inversion could be substituted to approximate the provenance by returning a fraction (or a projection) of the desired provenance. Examples of inversion can be found in some

TABLe 19.2 Characteristics of Selected Provenance-Aware Systems Reported in Remote Sensing and Other Geodata Applications

Geodata Application	Successes	Limitations or Future Work	References
Earth System Science Workbench and ES3			
Track processing of a laboratory's raw satellite imagery (e.g., AVHRR) into higher level products	a. Automates geodata provenance capture from running processes b. Stores provenance in both XML documents and in a searchable online store	Predates recent geodata interoperability standards and specifications	Frew (2004), Frew and Bose (2001), Frew and Slaughter (2008)
MODAPS and OMIDAPS			
Manage MODIS and other NASA satellite imagery and its provenance	a. Automates version tracking of geodata processing algorithms b. Reduces geodata storage via on-demand processing based on a virtual archive	Identifies science community's lack of appreciation for provenance information	Tilmes and Fleig (2008)
Karma			
Capture provenance for Japan's AMSR-E passive microwave radiometer on Aqua	a. Modularizes architecture to facilitate web service interoperability b. Is compatible with open provenance model (OPM) and ISO 19115-2 metadata standards	Requires additional geodata interoperability standards to facilitate geodata (scientific) reproducibility	Conover et al. (2013), Plale et al. (2011), Simmhan et al. (2008)
Data Quality Provenance System			
Assess quality of agricultural mapping based on SPOT satellite imagery	a. Assigns geodata quality index based on provenance information b. Is compatible with OPM and FGDC metadata standards	Needs to address geodata quality dependencies on provenance granularity	Malaverri et al. (2012)
VisTrails			
Model habitat suitability using WorldView-3 and LIDAR-derived forest structure	a. Provides Python-based open-source provenance and workflow management b. Allows key focus on provenance in rapidly changing workflows (e.g., during remote sensing process development)	Designed to be domain generic, VisTrails may have a steep learning curve	Freire et al. (2012), Talbert (2012)
UV-CDAT			
Analyze large-scale remote sensing–derived climate data	a. Built on top of VisTrails with an extensible modularized architecture that supports high-performance workflows b. First end-to-end provenance-enabled tool for large-scale climate research	Future work could adapt UV-CDAT successes in climate change for other geodata application areas	Santos et al. (2012)
GeoPWProv			
Visualize and navigate city planning geodata (e.g., LIDAR-derived elevation data) provenance via a map	a. Allows geodata provenance to be visualized and explored in a map environment b. Provides for several geospatial levels of provenance query (e.g., via a single polygon versus a larger dataset)	Future work could support geoprocessing replication	Sun et al. (2013)

spatial databases including Oracle Database, Microsoft SQL Server, IBM DB2, and Boundless PostGIS. One early approach was implemented for vector operations in Intergraph's (now Hexagon's) Geomedia product.

19.4.1.2 Service Chaining

In a vision of SOC and service-oriented integration (SOI), different web services can be found, composed and invoked to accomplish certain tasks. The sequence of service discovery, composition, and execution looks like a chain, while alternatively, the composition processes can be constructed through different approaches, such as service orchestration or choreography that could be applied in enabling business processes and transactions on the web. Service composition and chaining could represent a workflow in which scientific computation can be implemented through the SOC/SOI approach. Capturing the provenance information within service-oriented workflows has been explored in geospatial applications (Del Rio and da Silva 2007; Yue et al. 2010a,b, 2011), though feasible and convincing approaches for provenance in SOC/SOI need further exploration and investigation.

19.4.1.3 Virtual Data Catalog Service

A virtual data catalog (VDC) is a provenance approach to trace the derivation route of data product and virtual data generated in the workflow in order to enable scientists to reproduce a data product and validate the quality of the workflow and related simulations. The intermediate data may be generated within a workflow but may not exist physically in a database or computer system (e.g., due to storage limitations). For this reason, such data are called virtual data because it is "the representation and manipulation of data that does not exist, being defined only by computational procedures" (Foster et al. 2002, 2003).

Within virtual data systems, such as Chimera, which is a virtual data grid managing the derivation and analysis of data objects, Virtual Data Language (VDL) is developed to define the workflow, while VDC is a service in the virtual data systems. The latter is defined and implemented based on the virtual data schema (VDS). The VDS defines the relevant data objects and relationships, and VDC can be queried by VDL to construct data derivation procedures from which derived data and output can be recomputed (Foster et al. 2002, 2003; Glavic and Dittrich 2007; Simmhan et al. 2005).

19.4.2 Earth System Science Workbench and ES3

The Earth Systems Science Workbench (ESSW) was an early attempt at automated provenance management and storage. It was a nonintrusive system that made use of Perl scripting techniques and Java to store data provenance as XML documents (Frew and Bose 2001). It contained a registry for provenance and a server for making the information searchable on the web. ESSW was followed up by the Earth System Science Server (ES3), which allowed for more flexibility in client-side implementation, but used essentially the same structure as the ESSW (Frew 2004). ES3, unlike many other systems, automatically captures provenance from running processes. It can also create provenance graphs in XML that can then be visualized using third-party tools like yEd (Frew and Slaughter 2008).

19.4.3 MODAPS and OMIDAPS

The MODIS Adaptive Data Processing System (MODAPS) and OMI Data Processing System (OMIDAPS) were designed for use by NASA to manage satellite imagery and provenance from MODIS sensors on the Terra and Aqua satellites, and the OMI sensor on Aura respectively (Tilmes and Fleig 2008). Both systems are operational and use a scripting process to track changes in versions of geodata processing algorithms. Using this technique, there is no need to store workflow iterations because enough information is retained that previous versions of the data can be re-created. Further, these systems periodically are tasked with reprocessing past data using the most up-to-date algorithms to maintain a consistent and improved series of data products. MODAPS in particular makes use of these features to maintain a *virtual archive* with provenance information that persists after a geodata product is deleted, allowing the system to re-create data products on demand rather than keeping extensive archives. Data re-creation as implemented in these systems is unique and is something that could be useful in other geospatial provenance systems.

19.4.4 Karma

Plale et al. (2011) make use of the Karma system designed by Simmhan et al. (2008) to collect provenance data for the Advanced Microwave Scanning Radiometer–Earth Observing System (AMSR–E) flown on the Aqua satellite. One of the biggest benefits of Karma is its modular architecture, which simplifies interoperability with Java and other web services. Karma's architecture for this application consists of an application layer, web service layer, core service layer, and a database layer (Plale et al. 2011). The inclusion of open provenance model (OPM) specifications and XML makes its interoperability extend further (Moreau et al. 2011). Conover et al. (2013) also made use of Karma to retrofit a legacy system for provenance capture. They chose the NASA Science Investigator–led Processing System (SIPS) for the AMSR-E sensor on the Aqua satellite. Their system uses a two-tiered approach that captures provenance for individual data files as well as collections or series, both automatically and via manual entry using the ISO 19115-2 lineage metadata standard. Query and display are handled with a database-driven web interface called the Provenance Browser.

19.4.5 Data Quality Provenance System

Taking into account a source's trustworthiness and the data's age, Malaverri et al. (2012) created a provenance system that allows a quality index to be assigned. This approach is based on a combination of the OPM and FGDC geospatial metadata standards. Criteria considered in the quality index include granularity, accuracy of attribute descriptions, completeness of the data, a logical measure of the data, and spatial positional accuracy. Although measures of trust can be very subjective in nature, in this case requiring a domain experts' input, this approach is somewhat unique in that it attempts to quantify data quality (Malaverri et al. 2012).

19.4.6 VisTrails

VisTrails is a free and open-source scientific workflow and provenance management system (Freire et al. 2012). Written in Python/Qt and designed to be integrated with existing workflow systems, VisTrails has been used in a number of research applications ranging from climate (including the UV-CDAT described later) to ecology and biomedical research. Talbert (2012) created software based on VisTrails to capture the details of habitat suitability and species distribution modeling. One of the major advantages of VisTrails is that as an open-source project built in Python, it is interoperable, easily customizable, and benefits from a large community of developers contributing code. A key focus of VisTrails is rapidly changing workflows. The information contained in how workflows are developed (and change over time) may provide highly valuable insight into the creative and development aspects of the remote sensing process.

19.4.7 UV-CDAT

Climate Data Analysis Tools (CDATs) are cutting-edge domain-specific tools for the climate research community, but they are ill-equipped to handle very large geodata and provenance information. The UV-CDAT is a relatively new provenance system for

handling large amounts of climate-based data (Santos et al. 2012). The UV-CDAT uses a highly extensible modular design and makes use of a Visualization Control System and Visualization Toolkit (VTK)/ParaView infrastructure, which allows for high-performance parallel-streaming data analysis and visualization. Its loosely coupled modular design allows for integration with third-party tools such as R and MATLAB® for both analysis and visualization. The UV-CDAT is unique in that it is the first end-to-end application for provenance-enabled analysis and visualization for large-scale climate research. It has already been distributed and is widely used by scientists throughout the climate change field.

19.4.8 GeoPWProv

GeoPWProv is a provenance system specializing in displaying geospatial provenance as an easily accessible interactive map layer. GeoPWProv has the capability to capture provenance at the feature, dataset, service, or knowledge level (Sun et al. 2013). Comparisons can be made between entities in each level or between various levels. In addition to displaying provenance as a map layer, GeoPWProv supports displaying provenance in a workflow or in the more traditional text-based format. Implementation on the client side through use of a browser and *Open Layers* allows for ease of use. GeoPWProv's display of provenance in different formats and at different levels allows for a customizable user experience when evaluating a workflow.

19.5 Conclusions and Research Implications

Integrated geoprocessing, workflows, and provenance may be conceptualized as a positive developmental cycle that enables experts and software agents to capture, store, analyze, curate, replicate, and innovate remote sensing methods. Such integration is increasingly understood as a key to high-quality, replicable remote sensing–assisted spatial decision support. In early discussions in the 1980s, it soon became clear that provenance (or lineage) in particular is a fundamental element in understanding earth observation-related and other geodata quality (Moellering et al. 1988). As commercial GIS accelerated during the early 1990s, the Geolineus project (Lanter 1992b) demonstrated how software dedicated to lineage/provenance capture, management, and visualization can enable such gains as replicable geospatial workflows, automated workflow comparison, data quality modeling, data update management, and increased sharing of expert knowledge of geodata creation. Now with increasingly heightened awareness of provenance in computer systems (Bose and Frew 2005; Ikeda and Widom 2009; Simmhan et al. 2005; Yue et al. 2010a), there has been a maturing appreciation of the need to computationally address provenance capture, management, and exchange in an increasingly big data scenario.

While definitions of geodata provenance have varied, it is quite arguably distinct from and offers unique benefits over traditional metadata in large part because it encompasses process *history*. Regardless of definitions, the application of provenance benefits in remote sensing–assisted decision support workflows cannot be realized without development and demonstration of collaborative software architectures including those in a geo-cyberinfrastructure. Provenance has and will be of increasing interest to and a focus of organizations that create and encourage international specifications and standards (e.g., ISO, W3C, and OGC). As these organizations formulate procedures for the specification of provenance, we will see software developers add this capability to their products in a far more complete implementation than is currently the case. Even before the emerging international standards begin to mature, research is critically needed to demonstrate and fully understand *practical* benefits that user-friendly and integrated geoprocessing, workflows, and provenance can offer. With additional research and development, geospatial provenance has a high potential to benefit quality, trust, and innovation related to remote sensing–assisted spatial decision support.

References

Alonso, G. and C. Hagen. 1997. Geo-opera: Workflow concepts for spatial processes. In *Advances in Spatial Databases*, Springer, Berlin, Germany, pp. 238–258.

Anderson, J.R., E.E. Hardy, J.T. Roach, and R.E. Witmer. 1976. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. Geological Survey Professional Paper 964. Washington, DC: U.S. Government Printing Office.

Anderson, K.E. and G.M. Callahan. 1990. The modernization program of the U.S. Geological Survey's National Mapping Division. *Cartography and Geographic Information Systems* 17(3): 243–248.

Aronson, P. and S. Morehouse. 1983. The ARC/INFO map library: A design for a digital geographic database. In *Auto-Carto Six; Proceedings of the Sixth International Symposium on Automated Cartography*, Vol. 1, Ottawa/Hull, Canada, pp. 372–382.

Atkins, D.E., K.K. Droegemeier, S.I. Feldman, H. Garcia-Molina, M.L. Klein, D.G. Messerschmitt, P. Messina, J.P. Ostriker, and M.H. Wright. 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation: Washington, DC. https://arizona.openrepository.com/arizona/handle/10150/106224. Accessed June 29, 2014.

Bose, R. and J. Frew. 2005. Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys* 37(1): 1–28.

Bossler, J.D., J.B. Campbell, R.B. McMaster, and C. Rizos, eds. 2010. *Manual of Geospatial Science and Technology*, 2nd edn. Boca Raton, FL: CRC Press.

Botts, M. and A. Robin, eds. 2007. *OpenGIS Sensor Model Language* (*SensorML*) *Implementation Specification*. Open Geospatial Consortium. http://portal.opengeospatial.org/files/?artifact_id=21273. Accessed June 24, 2014.

Bröring, A., C. Stasch, and J. Echterhoff, eds. 2012. *OGC Sensor Observation Service Interface Standard*. Open Geospatial Consortium.

Buneman, P., S. Khanna, and W.-C. Tan. 2001. Why and where: A characterization of data provenance. In *International Conference on Database Theory* (*ICDT*), pp. 316–330.

Buneman, P. and S.B. Davidson. 2010. Data provenance—The foundation of data quality. Carnegie Mellon University Software Engineering Institute, Pittsburgh, PA. http://www.sei.cmu.edu/measurement/research/upload/Davidson.pdf. Accessed June 30, 2014.

Buneman, P. 2013. "The Providence of Provenance." In Big Data, edited by G. Gottlob, G. Grasso, D. Olteanu, and C. Schallhart, 7968:7–12. Berlin, Germany: Springer. http://link.springer.com/10.1007/978-3-642-39467-6_3.

Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a geographic information system. *Automated Cartography*, 6: 302–312.

Chrisman, N.R. 1986. Obtaining information on quality of digital data. In *AutoCarto Proceedings of the International Symposium on Computer-Assisted Cartography*, Vol. 1. London, U.K.: Cartography and Geographic Information Society, pp. 350–358.

Congalton, R. 2010. Remote sensing: An overview. *GIScience & Remote Sensing* 47(4): 443–459.

Conover, H., R. Ramachandran, B. Beaumont, A. Kulkarni, M. McEniry, K. Regner, and S. Graves. 2013. Introducing provenance capture into a legacy data system. *IEEE Transactions on Geoscience and Remote Sensing* 51(11): 5098–5014.

Cui, Y., J. Widom, and J.L. Wiener. 2000. Tracing the lineage of view data in a warehousing environment. *ACM Transactions on Database Systems* 25(2): 179–227.

Del Rio, N. and P.P. da Silva. 2007. Probe-It! Visualization support for provenance. In *Advances in Visual Computing*. Springer, Berlin, Germany, pp. 732–741. http://link.springer.com/chapter/10.1007/978-3-540-76856-2_72. Accessed June 24, 2014.

Di, L., P. Yue, H.K. Ramapriyan, and R.L. King. 2013b. Geoscience data provenance: An overview. *IEEE Transactions on Geoscience and Remote Sensing* 51(11): 5065–5072.

Di, L., Y. Shao, and L. Kang. 2013a. Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 1911-2 lineage model. *IEEE Transactions on Geoscience and Remote Sensing* 51(11): 5082–5089.

EarthCube. 2014. EarthCube: Transforming geosciences research. http://earthcube.org/. Accessed June 30, 2014.

Essinger, R. and D.P. Lanter. 1992. User-centered software design in GIS: Designing an icon-based flowchart that reveals the structure of ARC/INFO data graphically. In *Proceedings of the 12th Annual ESRI User Conference*, Palm Springs, CA.

Fielding, R.T. 2000. *Architectural Styles and the Design of Network-Based Software Architectures*. Irvine, CA: University of California. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm. Accessed June 29, 2014.

Foster, I., J. Vöckler, M. Wilde, and Y. Zhao. 2002. Chimera: A virtual data system for representing, querying, and automating data derivation. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, Los Alamitos, CA IEEE, pp. 37–46. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1029704.

Foster, I., J. Vöckler, M. Wilde, and Y. Zhao. 2003. The virtual data grid: A new model and architecture for data-intensive collaboration. In *Proceedings of the First Biennial Conference on Innovative Data Systems Research* (*CIDR*), Vol. 3. Asilomar, CA: Citeseer, p. 12.

Freire, J., D. Koop, E. Santos, C. Scheidegger, C. Silva, and H.T. Vo. 2012. VisTrails. In Brown, A. and Wilson, G., eds. *The Architecture of Open Source Applications: Elegance, Evolution, and a Few Fearless Hacks*, Vol. I. aosabook.org. http://aosabook.org/en/vistrails.html.

Frew, J. 2004. Earth system science server (ES3): Local infrastructure for earth science product management. In *Proceedings of the Fourth Earth Science Technology Conference*, Palo Alto, CA. http://esto.gsfc.nasa.gov/conferences/estc2004/papers/a4p3.pdf. Accessed March 1, 2014.

Frew, J. and P. Slaughter. 2008. ES3: A demonstration of transparent provenance for scientific computation. In J. Freire, D. Koop, and L. Moreau, eds., *Provenance and Annotation of Data and Processes*. Lecture Notes in Computer Science, Vol. 5272. Berlin, Germany: Springer, pp. 200–207. http://link.springer.com/chapter/10.1007/978-3-540-89965-5_21. Accessed June 26, 2014.

Frew, J. and R. Bose. 2001. Earth system science workbench: A data management infrastructure for earth science products. In *Proceedings of the International Conference on Scientific and Statistical Database Management*. Los Alamitos, CA: IEEE Computer Society, pp. 180–189.

Gamble, M. and C. Goble. 2011. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *Proceedings of the Third International Web Science Conference*, Vol. 15. ACM. New York, NY. http://dl.acm.org/citation.cfm?id=2527048. Accessed June 24, 2014.

Gil, Y., J. Cheney, P. Groth, O. Hartig, S. Miles, L. Moreau, and P.P. da Silva, eds. 2010. The foundations for provenance on the web. *Foundations and Trends in Web Science* 2(2–3): 99–241.

Gil, Y. and S. Miles, eds. 2013. PROV Model Primer. W3C. http://www.w3.org/TR/prov-primer/.

Giordano, A., H. Veregin, E. Borak, and D.P. Lanter. 1994. A conceptual model of GIS-based spatial analysis. *Cartographica: The International Journal for Geographic Information and Geovisualization* 31(4): 44–57.

Glavic, B. and K.R. Dittrich. 2007. Data provenance: A categorization of existing approaches. In *Proceedings of the 12th GI Conference on Database Systems in Business, Technology, and Web* (*BTW*), Vol. 7, Aachen, Germany, pp. 227–241.

Grady, R.K. 1988. The lineage of data in land and geographic information systems. In *Proceedings of GIS/LIS'88 American Congress on Surveying and Mapping: Data Lineage in Land and Geographic Information Systems*, Vol. 2, San Antonio, TX, pp. 722–730.

Guptill, S.C. 1987. Techniques for managing digital cartographic data. In *Proceedings of the 13th International Cartographic Conference*, Morelia, Mexico, Vol. 4(16), pp. 221–226.

Hart, G. and C. Dolbar. 2013. *Linked Data: A Geographic Perspective*, 1st edn. CRC Press , London, U.K.

Henzen, C., S. Mas, and L. Bernard. 2013. Provenance information in geodata infrastructures. In *Geographic Information Science at the Heart of Europe*, *III*. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, New York, NY, pp. 133–151.

Hey, T., S. Tansley, and K. Tolle, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Houbie, F. and L. Bigagli. 2010. OGC catalogue services standard 2.0 extension package for ebRIM application profile: Earth observation products. http://portal.opengeospatial.org/files/?artifact_id=35528.

IFAR. 2013. *Provenance Guide*. International Foundation for Art Research. New York, NY, http://www.ifar.org/provenance_guide.php. Accessed September 9, 2014.

Ikeda, R. and J. Widom. 2009. Data lineage: A survey. Technical Report. Stanford University InfoLab, Stanford, CA, http://ilpubs.stanford.edu:8090/918/. Accessed September 13, 2013.

ISO 19115-2:2009(E). 2009. Geographic information—Metadata—Part 2: Extensions for imagery and gridded data. International Organization for Standardization, Geneva, Switzerland.

Jensen, J.R. 2005. In K.C. Clarke, ed., *Introductory Digital Image Processing: A Remote Sensing Perspective*, 3rd edn. Prentice Hall Series in Geographic Information Science. Upper Saddle River, NJ: Prentice Hall.

Jensen, J.R. 2007. In K.C. Clarke, eds. *Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd edn. Prentice Hall Series in Geographic Information Science. Upper Saddle River, NJ: Prentice Hall.

Jones, J.S., J.A. Tullis, L.J. Haavik, J.M. Guldin, and F.M. Stephen. 2014. Monitoring oak-hickory forest change during an unprecedented red oak borer outbreak in the Ozark Mountains: 1990 to 2006. *Journal of Applied Remote Sensing* 8(1): 1–13.

Langran, G. and N.R. Chrisman. 1988. A framework for temporal geographic information. *Cartographica* 25(3): 1–14.

Langran, Gail. 1988. "Temporal GIS Design Tradeoffs." In Proceedings of GIS/LIS '88, 890–99. San Antonio, TX: American Congress on Surveying and Mapping.

Lanter, D.P. 1989. *Techniques and Method of Spatial Database Lineage Tracing*. Columbia, SC: University of South Carolina.

Lanter, D.P. 1991. Design of a lineage-based meta-data base for GIS. *Cartography and Geographic Information Systems* 18(4): 255–261.

Lanter, D.P. 1992a. Propagating updates by identifying data dependencies in spatial analytic applications. In *Proceedings of the 12th Annual ESRI User Conference*, Palm Springs, CA.

Lanter, D.P. 1992b. *GEOLINEUS: Data Management and Flowcharting for ARC/INFO*, *92-2*. Santa Barbara, CA: National Center for Geographic Information & Analysis. http://www.ncgia.ucsb.edu/Publications/tech-reports/91/91-6.pdf. Accessed June 24, 2014.

Lanter, D.P. 1993a. Method and means for lineage tracing of a spatial information processing and database system. Patent No. 5,193,185. United States Department of Commerce Patent and Trademark Office.

Lanter, D.P. 1993b. A lineage meta-database approach toward spatial analytic database optimization. *Cartography and Geographic Information Systems* 20(2): 112–121.

Lanter, D.P. 1994a. Comparison of spatial analytic applications of GIS. In W. K. Michener, J. W. Brunt, and S. G. Stafford, eds., *Environmental Information Management and Analysis: Ecosystem to Global Scales*. CRC Press , London, U.K.

Lanter, D.P. 1994b. A lineage metadata approach to removing redundancy and propagating updates in a GIS database. *Cartography and Geographic Information Systems* 21(2): 91–98.

Lanter, D.P. and C. Surbey. 1994. Metadata analysis of GIS data processing: A case study. In T.C. Waugh and R.G. Healey, eds., *Advances in GIS Research: Proceedings of the Sixth International Symposium on Spatial Data Handling*. London, U.K.: Taylor & Francis Ltd., pp. 314–324.

Lanter, D.P. and H. Veregin. 1991. A lineage information program for exploring error propagation in GIS applications. In *Proceedings of the 15th Conference of the International Cartographic Association*, Bournemouth, U.K., pp. 468–472.

Lanter, D.P. and H. Veregin. 1992. A research paradigm for propagating error in layer-based GIS. *Photogrammetic Engineering and Remote Sensing* 58(6): 825–833.

Lanter, D.P. and R. Essinger. 1991. *User-Centered Graphical User Interface Design for GIS*. Santa Barbara, CA: National Center for Geographic Information & Analysis, pp. 91–96. http://www.ncgia.ucsb.edu/Publications/tech-reports/91/91-6.pdf. Accessed June 24, 2014.

Malaverri, J.E.G., C.B. Medeiros, and R.C. Lamparelli. 2012. A provenance approach to assess quality of geospatial data. In *27th Symposium on Applied Computing*. Riva del Garda (Trento), Italy: ACM.

Merriam-Webster. 2014. Merriam-Webster Online. http://www.merriam-webster.com/.

Moellering, H., L. Fritz, D. Franklin, R.W. Marx, J.E. Dobson, D. Edson, J. Dangermond et al. 1988. The proposed standard for digital cartographic data. *The American Cartographer* 15(1): 9–140.

Moore, H. 1983. The impact of computer technology in the mapping environment. In *Proceedings of the Sixth International Symposium on Automated Cartography*, Vol. 1. Ottawa/Hull, Ontario/Quebec, Canada: Cartography and Geographic Information Society, pp. 60–68.

Moreau, L., B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska et al. 2011. The open provenance model core specification (v1.1). *Future Generation Computer Systems* 27(6): 743–756.

Moreau, L. 2010. The foundations for provenance on the web. *Foundations and Trends in Web Science* 2(2–3): 99–241.

Moreau, L. and P. Missier, eds. 2013. PROV-DM: The PROV data model. http://www.w3.org/TR/2013/REC-prov-dm-20130430/#section-example-two.

Naseri, M. and S.A. Ludwig. 2013. Evaluating workflow trust using hidden Markov modeling and provenance data. In Q. Liu, Q. Bai, S. Giugni, D. Williamson, and J. Taylor, eds., *Data Provenance and Data Management in eScience*. Studies in Computational Intelligence, Vol. 426. Berlin, Germany: Springer-Verlag, pp. 35–58.

NSF. 2014. Data Infrastructure Building Blocks (DIBBs). http://www.nsf.gov/pubs/2014/nsf14530/nsf14530.htm. Accessed June 30, 2014.

Nyerges, T. November 1987. GIS research needs identified during a cartographic standards process: Spatial data exchange. *International Geographic Information Systems Symposium: The Research Agenda* 1: 319–330.

Oxford University Press. 2014. Oxford English Dictionary. http://www.oed.com/.

Plale, B., B. Cao, C. Herath, and Y. Sun. 2011. Data provenance for preservation of digital geoscience data. In A.K. Sinha, D. Arctur, I. Jackson, and L.C. Gundersen, eds., *Societal Challenges and Geoinformatics*. Geological Society of America Special Paper 482. Boulder, CO: Geological Society of America, pp. 125–137.

Santos, E., D. Koop, T. Maxwell, C. Doutriaux, T. Ellqvist, G. Potter, J. Freire, D. Williams, and C.T. Silva. 2012. Designing a provenance-based climate data analysis application. In P. Groth and J. Frew, eds., *Provenance and Annotation of Data and Processes*. Lecture Notes in Computer Science, Vol. 7525. Santa Barbara, CA, pp. 214–219.

Schut, P., ed. 2007. *OpenGIS Web Processing Service*. Open Geospatial Consortium.

Simmhan, Y.L., B. Plale, and D. Gannon. 2005. A survey of data provenance in e-science. *SIGMOD Record* 34(3): 31–36.

Simmhan, Y.L., B. Plale, and D. Gannon. 2008. Karma2: Provenance management for data-driven workflows. *International Journal of Web Services Research* 5(2): 1–22.

Sun, Z., P. Yue, L. Hu, J. Gong, L. Zhang, and X. Lu. 2013. GeoPWProv: Interleaving map and faceted metadata for provenance visualization and navigation. *IEEE Transactions on Geoscience and Remote Sensing* 51(11): 5131–5136.

Talbert, C. 2012. *Software for Assisted Habitat Modeling Package for VisTrails* (*SAHM: VisTrails*) *v. 1*. Fort Collins, CO: USGS Fort Collins Science Center. https://www.fort.usgs.gov/products/23403. Accessed August 31, 2014.

Tan, W.-C. 2007. Provenance in databases: Past, current, and future. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 30(4): 3–12.

Tilmes, C. and A.J. Fleig. 2008. Provenance tracking in an earth science data processing system. In J. Freire and D. Koop, eds., *Provenance and Annotation of Data and Processes*. Lecture Notes in Computer Science, Vol. 5272. Berlin, Germany: Springer-Verlag, pp. 221–228. http://ebiquity.umbc.edu/_file_directory_/papers/445.pdf. Accessed September 19, 2014.

Tullis, J.A., F.M. Stephen, J.M. Guldin, J.S. Jones, J. Wilson, P.D. Smith, T. Sexton et al. 2012. Applied silvicultural assessment (ASA) Hazard Map. University of Arkansas Forest Entomology's Applied Silvicultural Assessment, Fayetteville, AR, http://asa.cast.uark.edu/hazmap/. Accessed August 30, 2014.

Tullis, J.A. and J.M. Defibaugh y Chávez. 2009. Scale management and remote sensor synergy in forest monitoring. *Geography Compass* 3(1): 154–170.

Vanecek, S. and R. Moore. 2014. OGC open modelling interface standard, Version 2.0. https://portal.opengeospatial.org/files/?artifact_id=59022. Accessed June 26, 2014.

Veregin, H. and D.P. Lanter. 1995. Data-quality enhancement techniques in layer-based geographic information systems. *Computers Environment and Urban Systems* 19(1): 23–36.

Wade, Tasha, and Shelly Sommer. 2006. A to Z GIS: An Illustrated Dictionary of Geographic Information Systems. Redlands, CA: Esri Press.

Wang, C., Z. Lu, and T.L. Haithcoat. 2007. Using Landsat images to detect oak decline in the Mark Twain National Forest, Ozark Highlands. *Forest Ecology and Management* 240: 70–78.

Wang, S., A. Padmanabhan, J.D. Myers, W. Tang, and Y. Liu. 2008. *Towards Provenance-Aware Geographic Information Systems*. Irvine, CA: ACM. http://acmgis08.cs.umn.edu/papers.html#posterpapers.

Woodruff, A. and M. Stonebraker. 1997. Supporting fine-grained lineage in a database visualization environment. In W.A. Gray and P.-Å. Larson, eds., *Proceedings of the 13th International Conference on Data Engineering*, Birmingham, U.K., pp. 91–102.

Yeide, N.H., K. Akinsha, and A.L. Walsh. 2001. *The AAM Guide to Provenance Research*. Washington, DC: American Association of Museums.

Yue, P., J. Gong, and L. Di. 2010a. Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Computers & Geosciences* 36: 270–281.

Yue, P., J. Gong, L. Di, L. He, and Y. Wei. 2010b. Semantic provenance registration and discovery using geospatial catalogue service. *Proceedings of the Second International Workshop on the role of Semantic Web in Provenance Management (SWPM 2010), Shanghai, China*.

Yue, P., Y. Wei, L. Di, L. He, J. Gong, and L. Zhang. 2011. Sharing geospatial provenance in a service-oriented environment. *Computers, Environment and Urban Systems* 35(4): 333–343.

Yue, P. and L. He. 2009. Geospatial data provenance in cyberinfrastructure. In *Proceedings of the 17th International Conference on Geoinformatics*, Fairfax, VA.