An Enterprise GIS Warehouse Architecture

By David Lanter, Ph.D., CDM Senior GIS Specialist

Introduction

Knowledge about an enterprise's GIS data and applications assets and how to update products dependent on them is easily lost and forgotten. This paper presents an architecture for a dynamic data warehouse that learns about its contents and how to update products derived from them.

Notwithstanding the human resource and intellectual assets nurtured by urban and regional departments, the bulk of their physical GIS assets are migrating to relational database management systems (RDBMS). This paper presents a flexible and simple dynamic metadata design that supports enterprise GIS data warehouses for sharing and maintaining these assets within RDBMS. This design is a cumulative product, gleaned from a decade of experience developing dynamic metadata enabled geographic information production systems at the NSF's NCGIA, Geographic Designs, Microsoft, and Rand McNally.

Most recently, these ideas have been reengineered and applied to the development of data collection protocols for CDM's urban and regional water and sewer service clients. These include specifications for televised inspection, physical survey, dye testing, hydraulic modeling, and mapping data for integration within enterprise GIS data warehouses.

A Federated solution

The evolution of GIS is resulting in a new class of heterogeneous GIS consisting of new and old systems. Some use these systems in a stand-alone environment, others in tandem. Individually, they have been rather efficient in their respective applications, data models, and programming environments. They have been supported with different computer hardware, system software, and professional personnel, often operating in their own distinct stand-alone environments. As a practical matter, a conceptual architecture for a GIS enterprise must provide a federated solution for sharing heterogeneous GIS datasets and data processing methods unified to meet organizational sharing and reuse needs.

Federated Departments

Federated departments maintain their own data, data editing and development tools, and decision support systems. They share their data and tools with other departments via a common data warehouse and tool library. Members of the federation adopt standards to help make sharing a reality. The enterprise GIS provides benchmark measures on usage, value, currency, accuracy and ongoing need.

Coordinating Committee

A coordinating committee maintains data sharing standards, coordinates expenditures, and guides the federation by agreeing on priorities based on a simple model.

In this model, a department having a GIS related need searches the warehouse for existing data or applications programs. Failure to find what they need demonstrates a need for a data update, new kind of data or programs. Once funding is approved, the department develops and adds the update, new data, or programs to warehouse in compliance with the federation's standards.

Coordinating Committee Advantages

Next time around, the committee's support for funding will flow easier to departments doing good jobs filling the warehouse with effective reusable data and programs. In the case of departments that did not do a good job sharing their new assets well, funding will be more difficult to justify and come by.

In summary, an effective enterprise architecture promotes cooperation by allowing department autonomy, provides carrots instead of sticks to motivate effective participation, automates performance benchmarks, and supports an incremental development path.

Dynamic METADATA Architecture

The remainder of this paper presents a high level conceptual architecture for sharing heterogeneous GIS datasets and data processing methods unified to meet organizational sharing and reuse needs. As will be illustrated, two basic attributes are the glue that make the architecture work: Entity_ID is a unique persistent identifier for features classified into subthemes. Dataset_ID if a unique persistent identifier for datasets, linking datasets to sources they are based on and products they participate in, and update propagation.

The architecture supports:

- Cataloging and Browsing
- Updating

Cataloging

In earlier work, the implementation of dynamic metadata within a data catalog was illustrated that enabled users to browse, find, and reuse unknown environmental datasets (Lanter 1999; Michener, W.K., Lanter, D.P., and Houhoulis, P.F. 1997; Lanter and Essinger, 1995). In GIS data warehouses supporting dynamic information product production, it is instrumental for data to be cataloged and stored with identifications concerning their theme, source, and other products supported.

Theme Cataloging

Within the data catalog, data are organized thematically. The theme identifies which department and general geographic thematic data class the dataset belong to. Examples include: "Water Department", "Streets Department", or "Parks Department." Each theme is subdivided and associated with a set of subthemes (or general feature classes) that are developed and maintained by the department. For example, the Water department might maintain subthemes for rivers, water treatment plants, and possibly sewer pipes.

Subtheme development

Each subtheme consists of a collection of persistently identified features. Examples include Delaware River, JFK Boulevard, and Manhole 2857. In addition to its association with its subtheme (and transitively with its theme), each feature is identified with a persistent unique identifier (for example "Entity-ID"). GIS data entities are associated with their geospatial data representations (i.e. geometry and attributes. They may also be associated with media (e.g. pictures, movies, or sound files) and any dynamic editorial event information (e.g. "Televised Inspection planned for 11/23"). When appropriate addresses are included. Hospitals, precinct houses, commercial facilities, governmental offices or landmarks are all likely to have addresses. For named street features (e.g. JFK Boulevard), the geometry components corresponding to the street subnetworks (adjacent connected set of street segments) are uniquely identified and persisted, as are the individual street segments that comprise these subnetworks.

Product Cataloging

Within the data catalog, geographic information products are identified, as are users of the products. The latter is important to support identifying those potentially affected by data updates. Product catalog details will not be discussed in detail in this paper. Products are associated with the informational subthemes, features, data types, and datasets (discussed later in the paper) that comprise them.

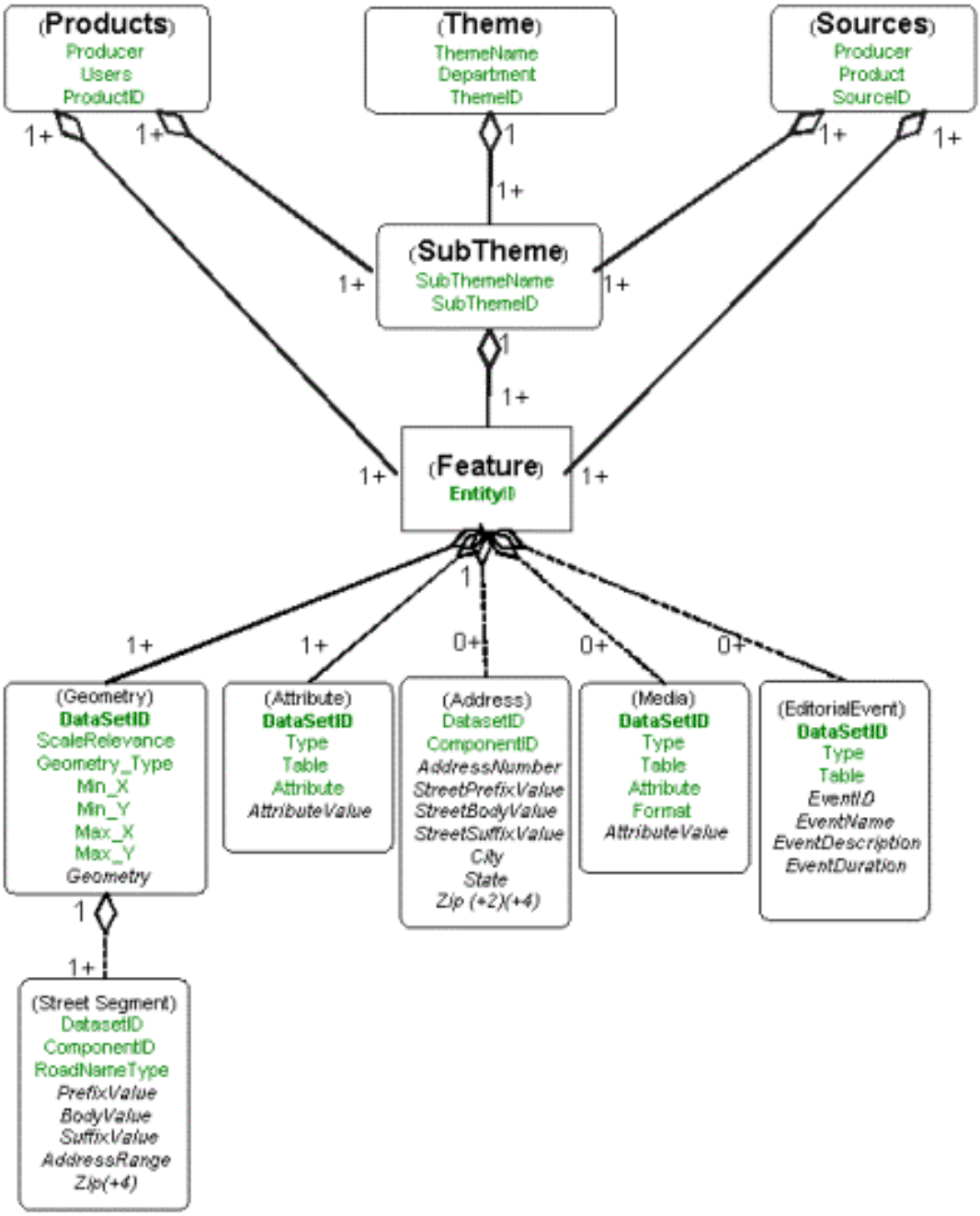
Source Cataloging

Within the data catalog, data sources are identified as well as the producing agencies are described. Specific details of the source catalog are not discussed in this paper. What we are concerned with here is the relationship the source has to the data themes, features and ultimately to their transformation and storage within the warehouse.

A conceptual object model of the organization of themes and database tables that will ultimate store them is presented in Figure 1 below:

FIGURE 1

DATA CATALOGED BY THEME, PRODUCTS, AND SOURCES

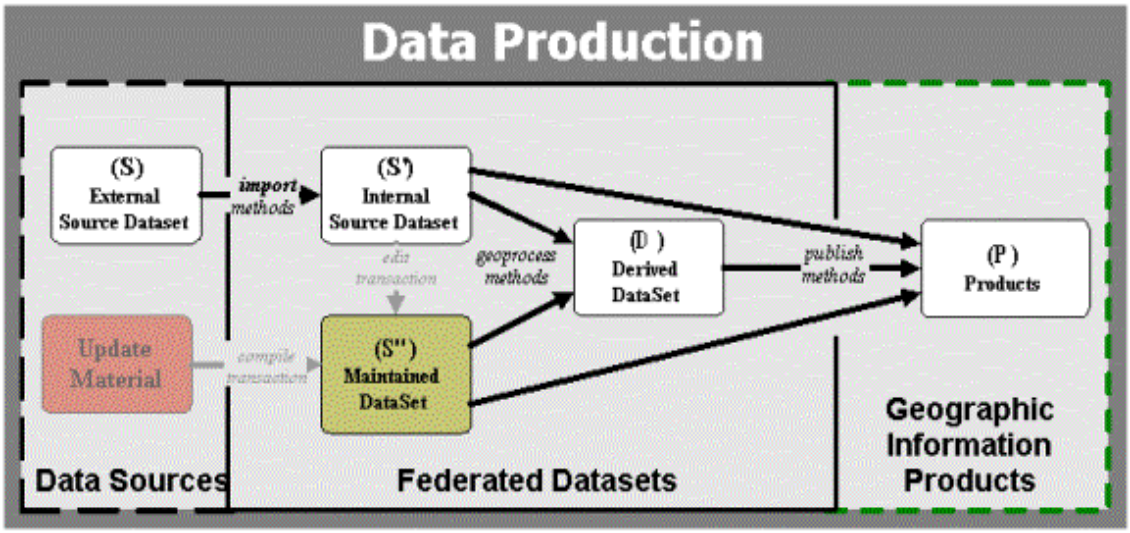


Updating

Much has been written about dynamic data lineage metadata, representation and inference (Veregin, H. and Lanter, D.P. 1995; Lanter, D.P. and R. Essinger 1995; Lanter, D.P. 1994a, b, c; 1993a, b, 1992a, b, c; Lanter, D.P. and Veregin, H. 1992; Essinger, R. and Lanter, D.P. 1992; Lanter 1991a, b; 1990). Figure 2 (below) illustrates the data flows, transformational import, geoprocessing and publishing methods, and input/output dependencies existing among various data sources, maintained data, derived data and dependent data:

FIGURE 2

DATA LINEAGE FLOWS

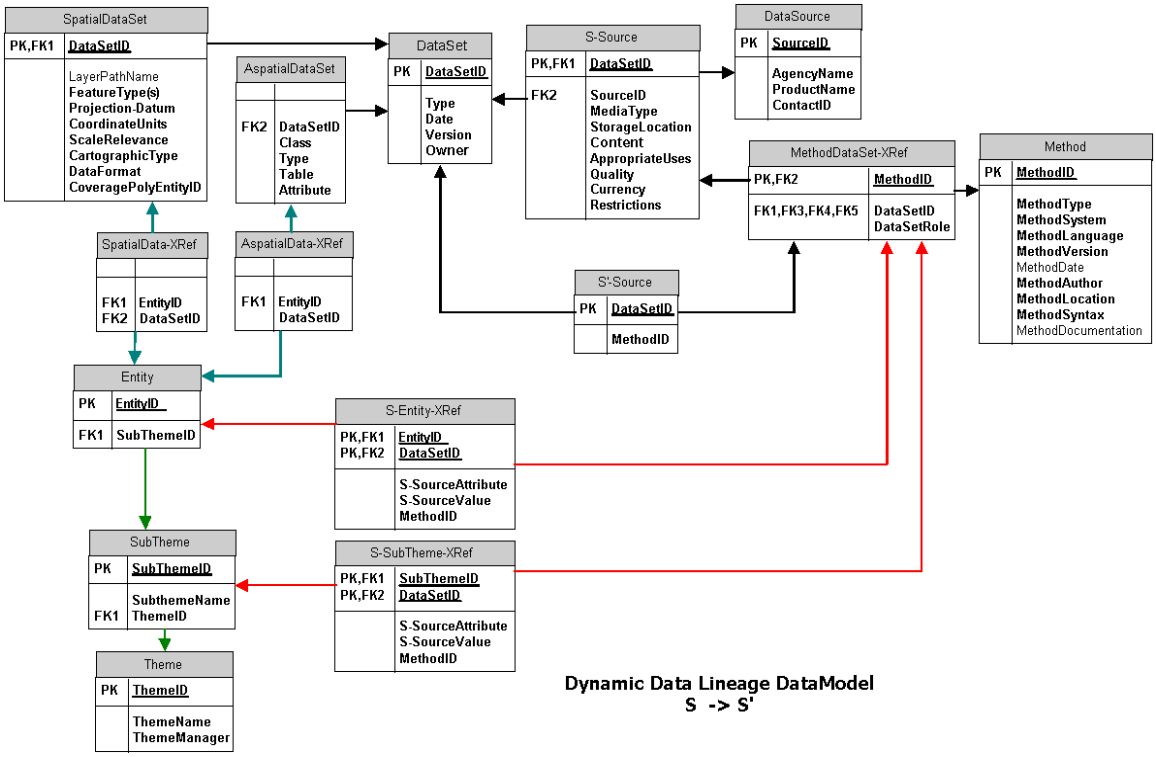


Data entered into the warehouse are accompanied with thematic, source, and product metadata in the data catalog. Lineage metadata is assembled for detailing the import methods applied to the data sources, geoprocessing methods for deriving new data, and publishing methods for extracting data and using with geographic information products. This lineage metadata is setup along with thematic crosswalks between source subtheme and entity identifications and their corresponding references in the enterprises catalog. This metadata is dynamic or active. It enables import, derivation, and publishing methods to be rerun on updated source materials to propagate these updates to affected internal data sources, derived datasets, and ultimately to products. This dynamic metadata processing supports plugging 'n playing new data source updates and propagating ("rippling") updates throughout the warehouse.

Figure 3 illustrates how metadata can be assembled to support source thematic data cataloging, source to warehouse subtheme and entity identification translation crosswalks, and imports of updates to the warehouse to update dependent data:

FIGURE 3

SOURCE DATA CATALOG AND CROSSWALK

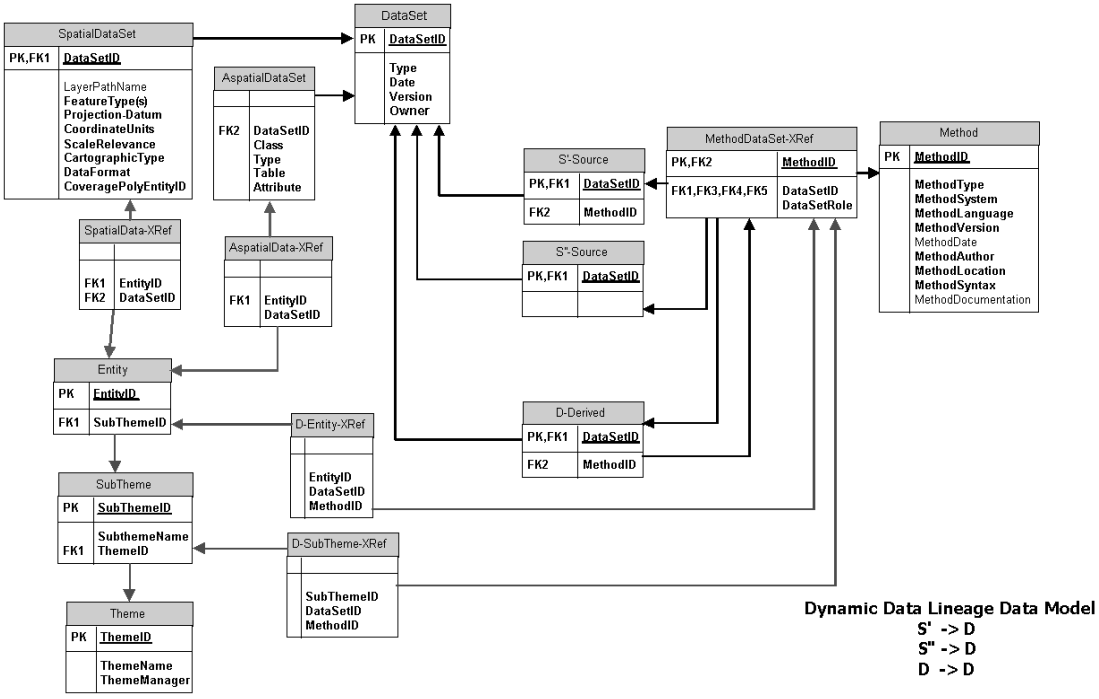


Each department in an enterprise has needs for its own specific kinds of geographic data. As a result, they focus on collecting and developing data as well as applications ("methods") to process these data and meet their needs. These departments often depend on data from other departments to support their own work. Data developed within and among different departments, however, may be heterogeneous with respect to data format (e.g. coverage, shapefile, geodatabase, GRID, TIN, mid-mif, etc.). The methods to process these data are also often correspondingly heterogeneous with respect to software language, dialect, and computational environment. (e.g. AML, Avenue, AXL, C Shell, MapBASIC, VBA, VB, COM, etc.).

Figure 4 illustrates how metadata can be assembled to track and reapply these applications to support interoperability of these data and the transformational methods they flow among. This logical model supports source update propagation to dependent derived datasets. It also, supports the development of heterogeneous applications that span across multiple data formats and applications environments.

FIGURE 4

DERIVED DATA CATALOG AND CROSSWALK

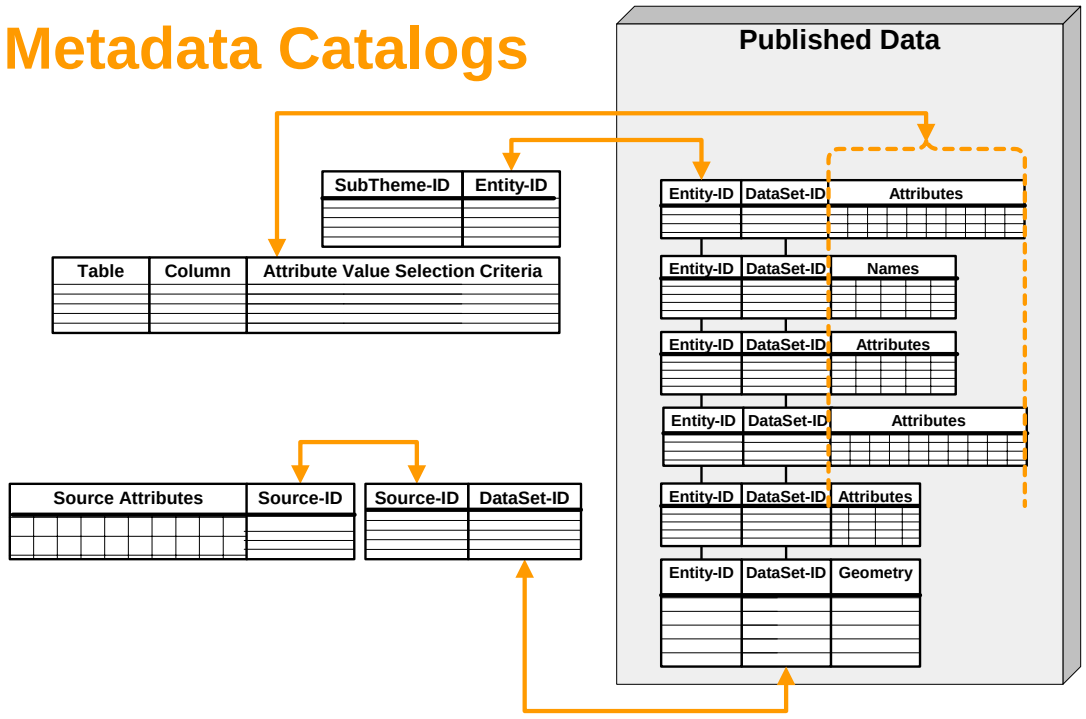


Data Selection

During geographic information product (map and report) production, thematic data is identified and selected from the catalog by subtheme and/or feature attributes. The specific representation (i.e. version) of the data is selected using source metadata (e.g. scale, date, quality, use restrictions, etc.). This is illustrated in Figure 5

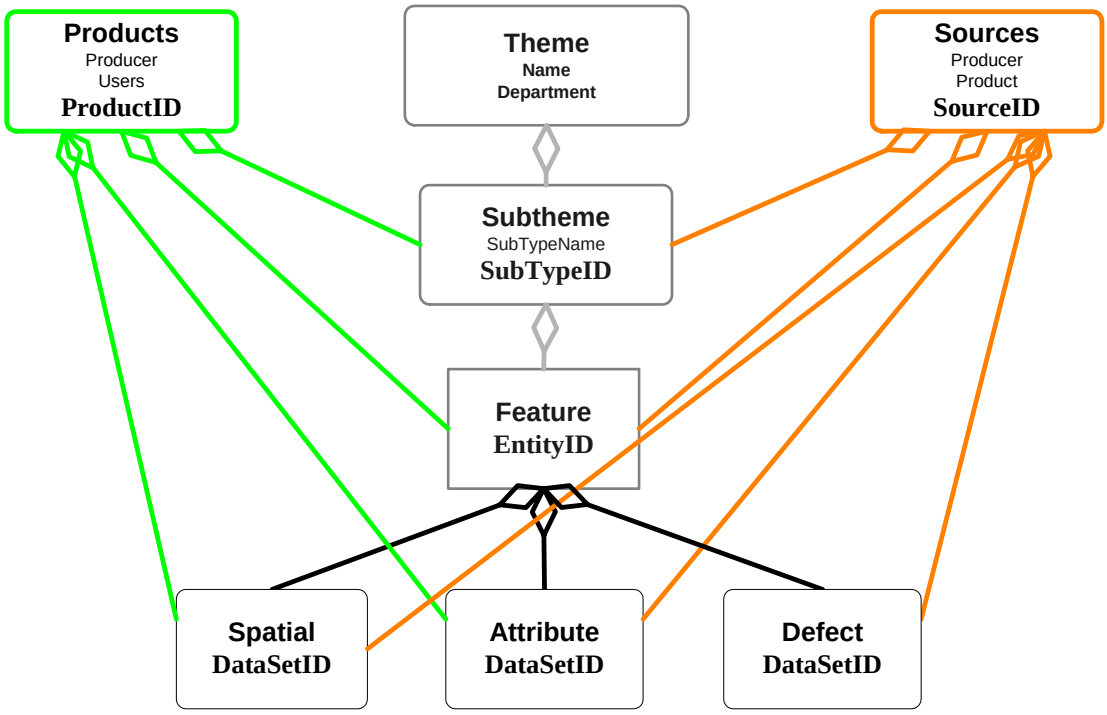
FIGURE 5

DATA SELECTION FROM A GIS DATA WAREHOUSE



Data selection queries are formulated by combining subtheme and/or entity, with data source specifications. These identify and bring data from the warehouse in for product (i.e. map or report) development. Such queries are stored as the link between warehouse and product - and serve to close the gap in resolving product update needs and processes.

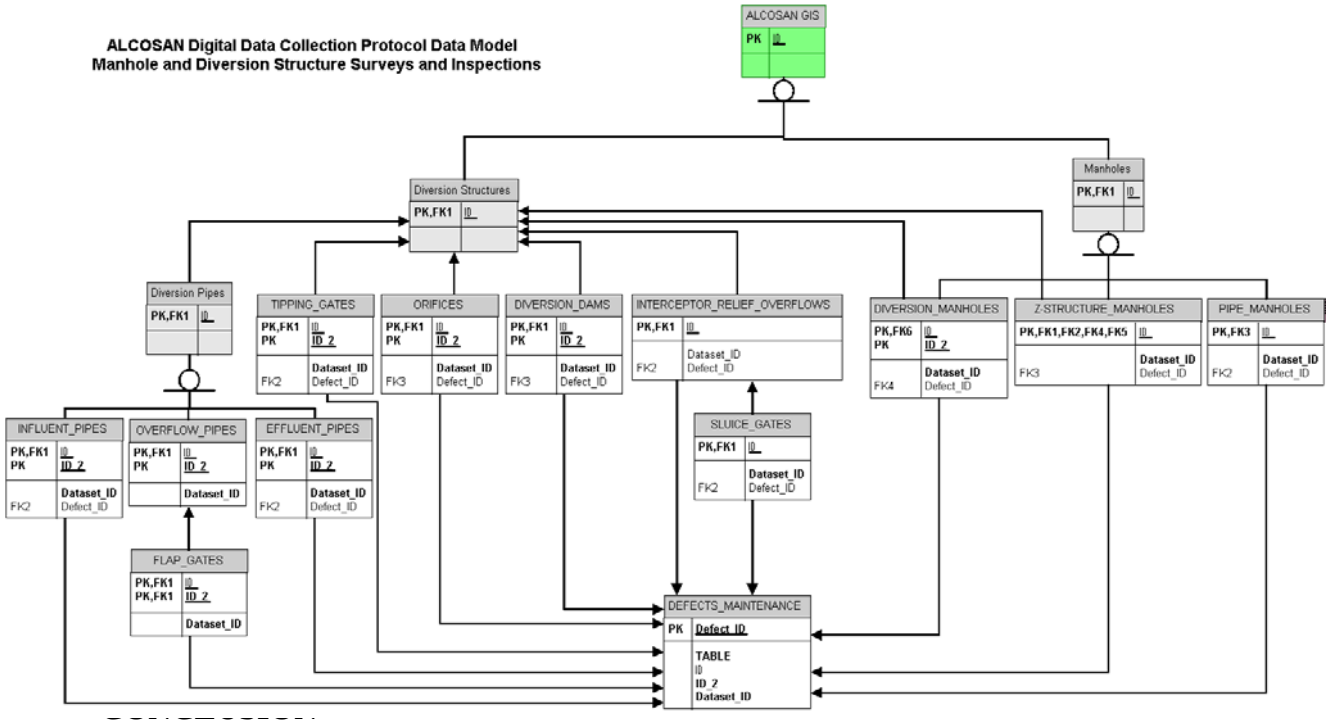
FIGURE 6
DATASETS LINKED TO THEME, PRODUCTS, and sources,
DEFECTS LINKED TO FEATURES AND SOURCES
AND TRANSITIVELY TO PRODUCTS



Data Selection By Defect

Figure 7 illustrates the use of ID (Entity_ID) as a way to associate a cartographic feature representation of a sewer interceptor system with attributes describing its constituent manhole, and diversion structure facility components recorded during physical inspections. These in turn, are linked and associate with information concerning found defects.

FIGURE 7
SEWER INTERCEPTOR SYSTEM COMPONENT ATTRIBUTE TABLES LINKED BY ID (ENTITY_ID) TO DEFINE FACILITY STRUCTURE, TO CARTOGRAPHIC FEATURE FOR MAPPING, AND TO OBSERVED DEFECTS



This paper has provided an overview of how dynamic metadata can be setup and used within an enterprise warehouse to catalog, develop, update, and select datasets and dependent products. Illustrations of this architecture have been provided to demonstrate how the constituent metadata concerning sources of data, warehouse datasets, processing methods (import, geoprocessing, and publishing), and products can be assembled in a RDBMS. This is the basis for data federation both within and across large GIS using organizations.

REFERENCES

Lanter, D. 1999, "Environmental Data Explorer - An Intelligent Interface for Exploring Unfamiliar Environmental Data Sets", Papers and Proceedings of the Applied

Geography Conferences, Volume 22, October 1999, Charlotte, University of North Carolina at Charlotte, North Carolina.

Michener, W.K., Lanter, D.P., and Houhoulis, P.F. 1997. "Geographic Information Systems for Sustainable Development: A Review of Applications and Research Needs", Sustainable Development in the South Eastern Coastal Zone, Editors: F. J. Vernberg et al. University of South Carolina Press. pp. 89-110.

Veregin, H. and Lanter, D.P. 1995. "Data Quality Enhancement Techniques in Layer-Based Geographic Information Systems", Computers, Environment and Urban Systems, Vol. 19, No. 1.

Lanter, D.P. and R. Essinger 1995, "Object-Oriented Exploration of Environmental Data Sets", GIS/LIS '95, Nashville, TN.

Lanter, D.P. 1994a. "A Lineage Metadata Approach to Removing Redundancy and Propagating Updates in a GIS Database", Cartography and Geographic Information Systems, Vol. 21, No. 2, pp. 91-98.

Lanter, D.P. 1994b. "Comparison of Spatial Analytic Applications of GIS", in Environmental Information Management and Analysis: Ecosystem to Global Scales, Editors: Michener, W.K. et al., London: Taylor & Francis, pp. 413-425.

Lanter, D.P. 1994c. "The Contribution of ARC/INFO's Log File to Metadata Analysis of GIS Data Processing", Proceedings of the Fourteenth Annual ESRI User Conference, Palm Springs, California.

Lanter, D.P. 1993a. "A Lineage Meta-Database Approach Towards Spatial Analytic Database Optimization", Cartography and Geographic Information Systems, Vol. 20, No. 2, pp.112-121.

Lanter, D. 1993b. "Scale Independent Analysis of Spatial Analytic Applications of GIS", Environmental Information Management and Analysis: Ecosystem to Global Scales, May 20-22. Albuquerque, New Mexico.

Lanter, D. 1992a. "Intelligent Assistants for Filling Critical Gaps in GIS", Technical Publication 92-4, National Center for Geographic Information and Analysis, Santa Barbara, CA.

Lanter, D. 1992b. "GEOLINEUS: Data Management and Flowcharting for ARC/INFO", Technical Software Series S-92-2, National Center for Geographic Information and Analysis, Santa Barbara, CA.

Lanter, D.P. 1992c. "Propagating Updates by Identifying Data Dependencies in Spatial Analytic Applications", Proceedings of the Twelfth Annual ESRI User Conference, Palm Springs, California.

Lanter, D.P. and Veregin, H. 1992. "A Research Paradigm for Error Propagation in Layer-Based GIS", Photogrammetric Engineering and Remote Sensing, Vol. 58, No.6. pp. 825-833.

Essinger, R. and Lanter, D.P. 1992. "User-Centered Software Design in GIS: Designing An Icon-Based Flowchart That Reveals The Structure of ARC/INFO Data Graphically", Proceedings of the Twelfth Annual ESRI User Conference, Palm Springs, California.

Lanter, D.P. 1991a. "Design of a Lineage-Based Meta-Database for GIS", Cartography and Geographic Information Systems, Vol. 18 No. 4. pp. 255-261.

Lanter, D.P. 1991b. "GEOLINEUS: A Graphical User Interface for GIS", Proceedings of the Eleventh Annual ESRI User Conference, Palm Springs, California.

Lanter, D.P. 1990. "Lineage in GIS: The Problem and a Solution", Technical publication 90-6. National Center for Geographic Information and Analysis, Santa Barbara, CA.