Geospatial Metadata

Revising the Approach

Brian Hebert
ScribeKey, LLC
www.scribekey.com

Outline

- Geospatial metadata isn't easy.
- Imagine using RDB tables and feature classes to manage metadata.
- Why isn't metadata managed in RDB tables?
- XML Use in IT
- CSDGM Design Problems
- CSDGM Communication Problems
- Politics, Religion, and Change
- Possible Solution

This is an Opinion Paper Based On:

- Project experience related to generating 1000s of FGDC CSDGM metadata documents, and data dictionaries for large volume commercial data providers.
- Analysis and assessments of 1000's of FGDC CSDGM XML documents.
- Requires industrial strength approach.
- Based on a mix of Data Profiling, Data Warehousing, and Library Science.
- In context of ESRI ArcGIS technology

Part 1: Geospatial Metadata Isn't Easy

- Authoring and using FGDC CSDGM and ISO XML metadata files isn't easy.
- One reason is that GIS practitioners, the authors and consumers of metadata, are inherently more familiar with RDB tables than they are with XML technology.
- Changes to ESRI ArcCatalog metadata tools in ArcGIS 10 caused some confusion.
- In addition to FGDC CSDGM and ISO there is now an ESRI XML format.
- The migration from FGDC to ISO is very problematic and complicated.

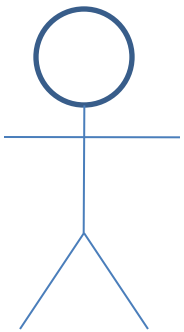
Quote from FGDC

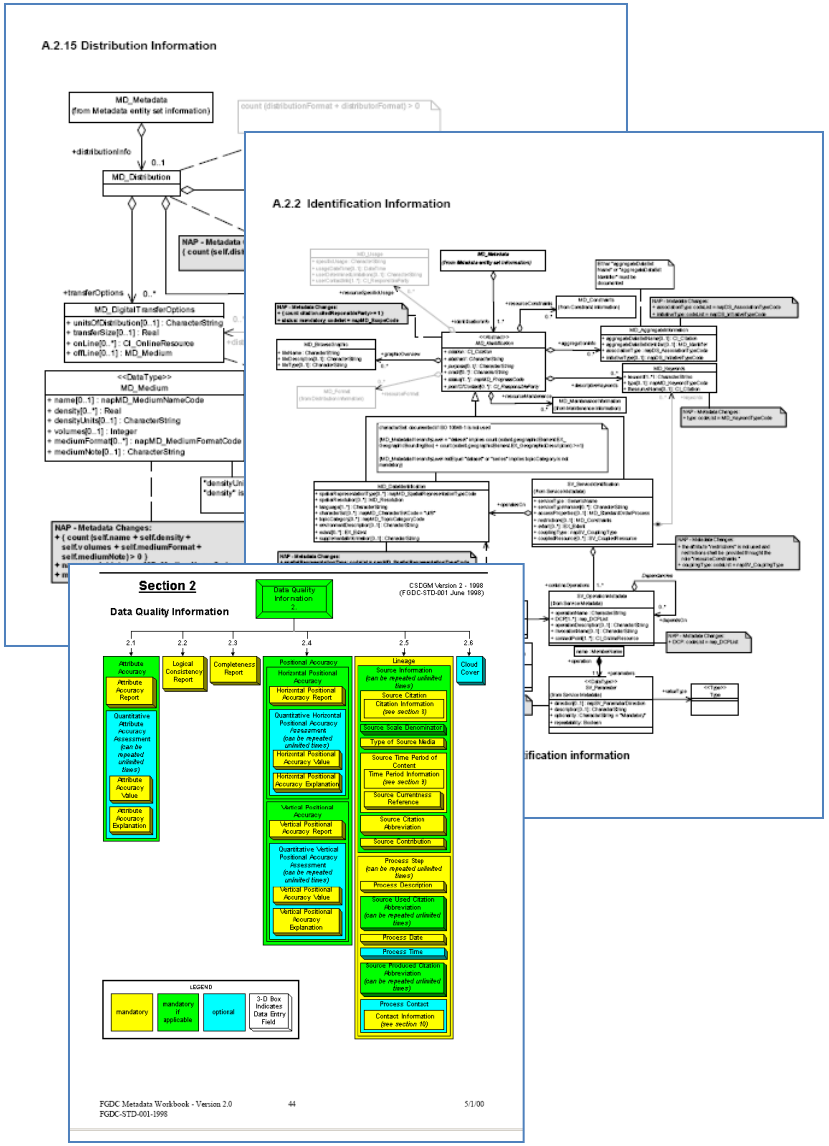
The current array of geospatial metadata standards and variations of standards has left the community somewhat **bewildered** as to which geospatial metadata standard/variant they should be utilizing. At this time the FGDC recommends that organizations currently using the CSDGM metadata standard remain to do so *unless there is some compelling reason to change standards*.

(http://www.fgdc.gov/metadata/documents/preparing-for-international-metadata-guidance.pdf)

Understanding the Standards

- Complicated for end users, what's core?
- UML, XSD, Grammar Production Rules
- Language and artifacts familiar to professional data modelers, academicians, but not end users

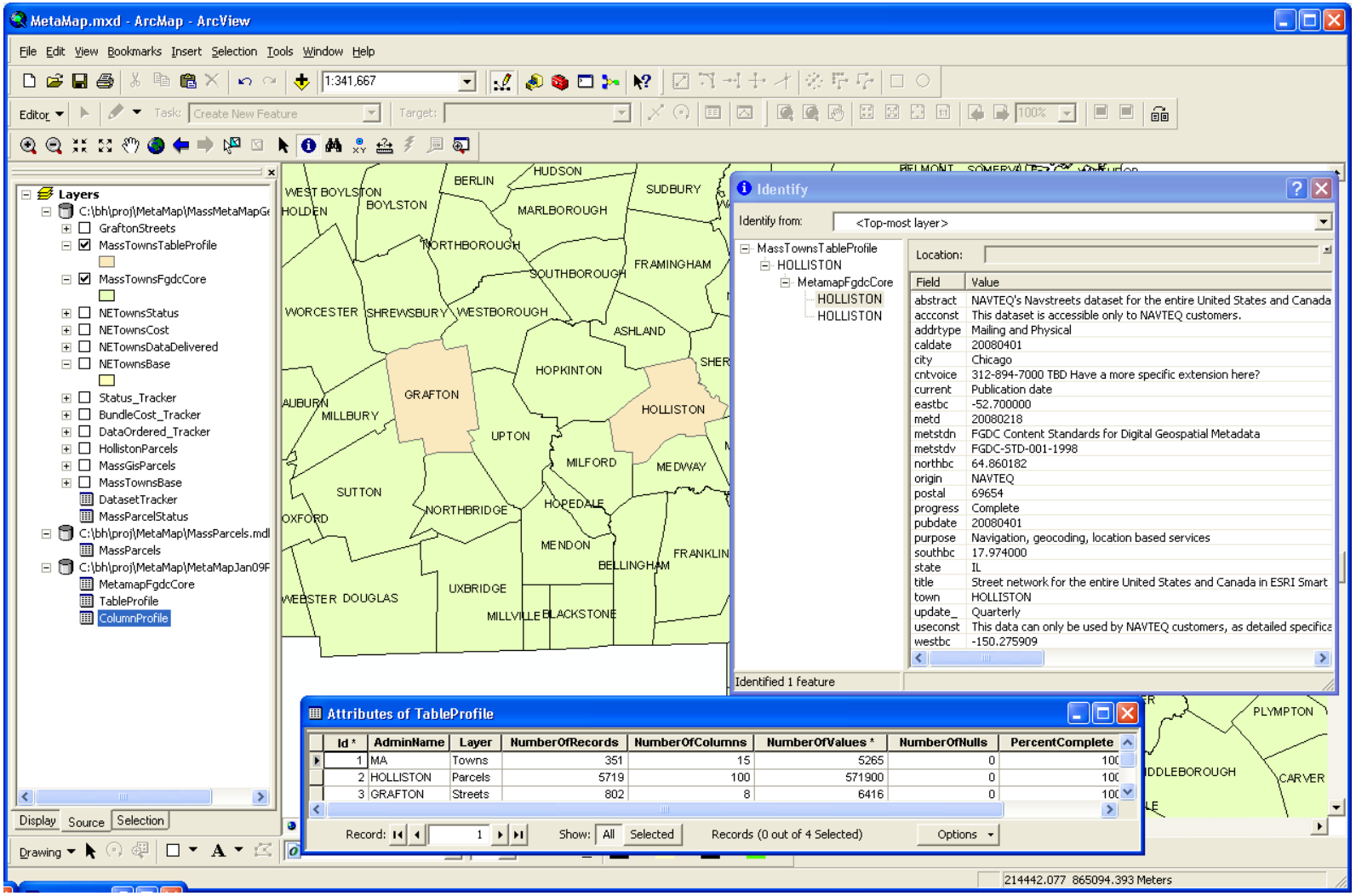




Part 2: Imagine

- Let's put the challenges associated with geospatial metadata, as it's done today, aside …
- **Imagine if RDB tables and feature classes were used to manage metadata, the same way GIS data is typically managed.**
- Consider a single metadata table holding FGDC CSDGM core elements for all the feature classes and tables in a datasets.
- You could add the table to an ArcMap .mxd and have access to list of dataset metadata records very easily.

Metadata Tables with GIS Data

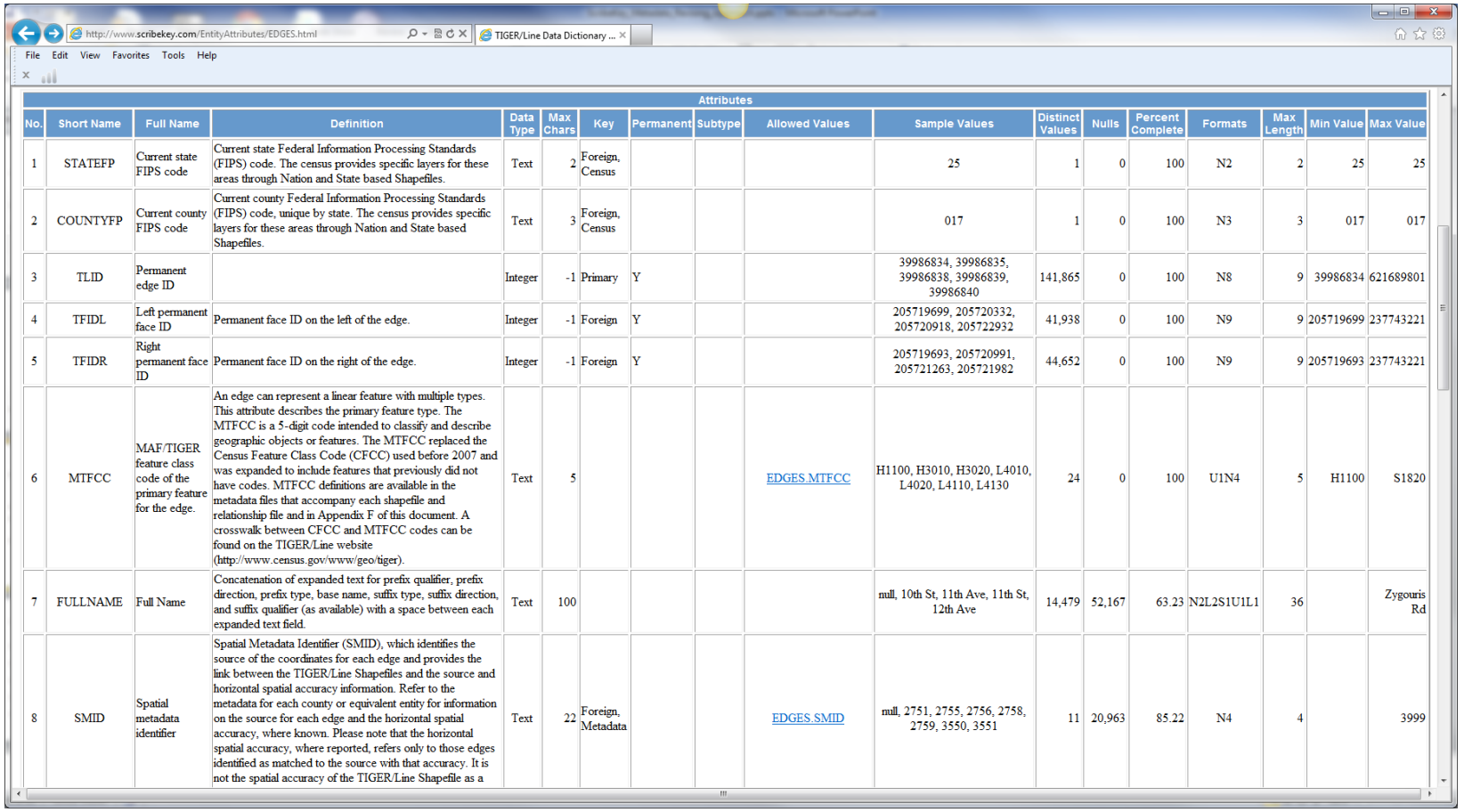


FGDC CSDGM Core metadata stored in an RDB table, integrated with the geospatial data it is describing, and accessible from ArcMap, or other applications, for review, query, sorting, filtering, etc.

Attribute and Domain Metadata

- Consider capturing metadata elements describing attributes and domains in RDB tables.
- Can also add them to ArcMap .mxd
- Can do a relate between FGDC CSDGM core and attributes
- Can do a relate between attributes and domains

Attribute Metadata Table in HTML



http://www.scribekey.com/EntityAttributes/EDGES.html — TIGER/Line Data Dictionary ...

File Edit View Favorites Tools Help

Attributes

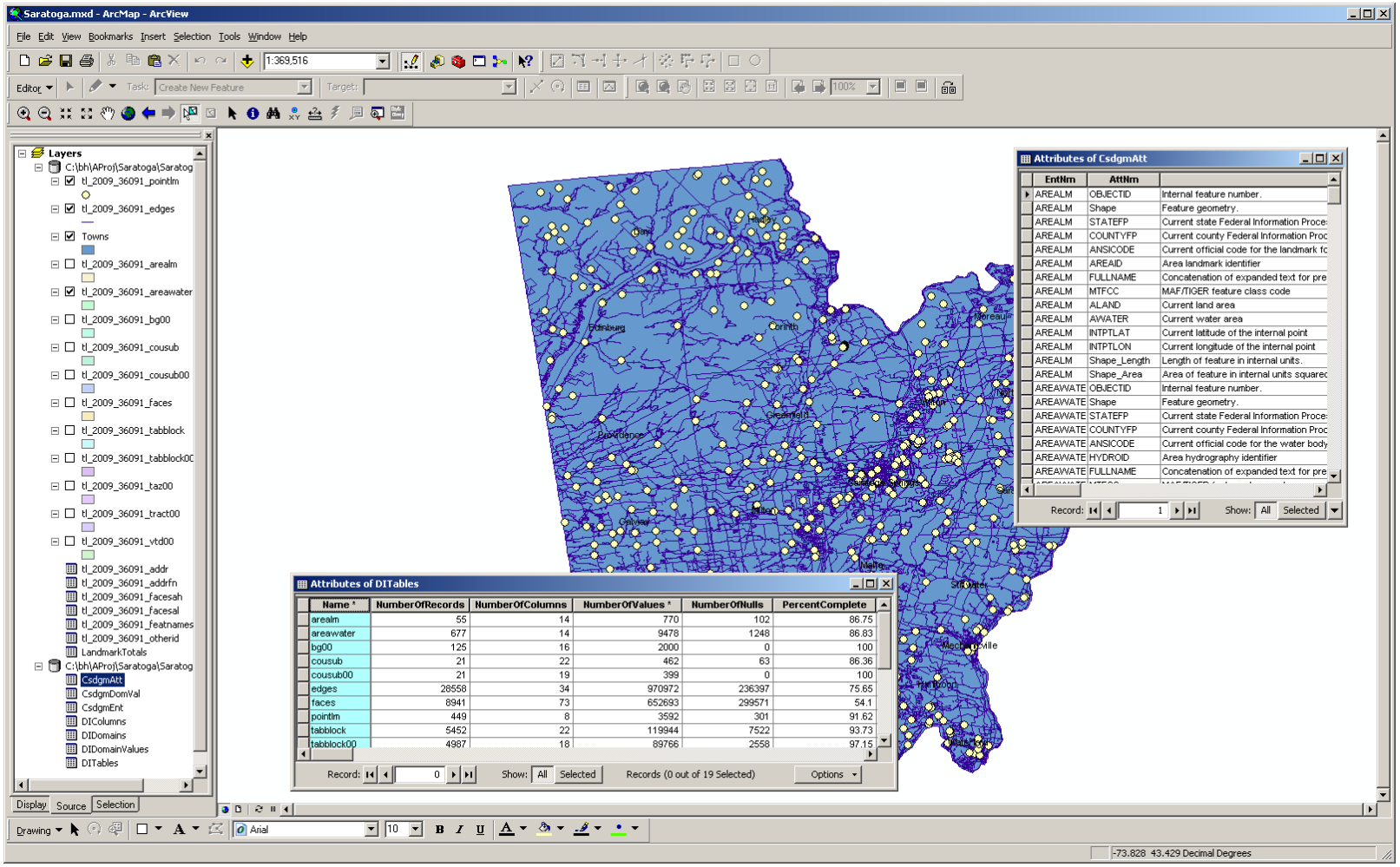
No.	Short Name	Full Name	Definition	Data Type	Max Chars	Key	Permanent	Subtype	Allowed Values	Sample Values	Distinct Values	Nulls	Percent Complete	Formats	Max Length	Min Value	Max Value
1	STATEFP	Current state FIPS code	Current state Federal Information Processing Standards (FIPS) code. The census provides specific layers for these areas through Nation and State based Shapefiles.	Text	2	Foreign, Census				25	1	0	100	N2	2	25	25
2	COUNTYFP	Current county FIPS code	Current county Federal Information Processing Standards (FIPS) code, unique by state. The census provides specific layers for these areas through Nation and State based Shapefiles.	Text	3	Foreign, Census				017	1	0	100	N3	3	017	017
3	TLID	Permanent edge ID		Integer	-1	Primary	Y			39986834, 39986835, 39986838, 39986839, 39986840	141,865	0	100	N8	9	39986834	621689801
4	TFIDL	Left permanent face ID	Permanent face ID on the left of the edge.	Integer	-1	Foreign	Y			205719699, 205720332, 205720918, 205722932	41,938	0	100	N9	9	205719699	237743221
5	TFIDR	Right permanent face ID	Permanent face ID on the right of the edge.	Integer	-1	Foreign	Y			205719693, 205720991, 205721263, 205721982	44,652	0	100	N9	9	205719693	237743221
6	MTFCC	MAF/TIGER feature class code of the primary feature for the edge.	An edge can represent a linear feature with multiple types. This attribute describes the primary feature type. The MTFCC is a 5-digit code intended to classify and describe geographic objects or features. The MTFCC replaced the Census Feature Class Code (CFCC) used before 2007 and was expanded to include features that previously did not have codes. MTFCC definitions are available in the metadata files that accompany each shapefile and relationship file and in Appendix F of this document. A crosswalk between CFCC and MTFCC codes can be found on the TIGER/Line website (http://www.census.gov/www/geo/tiger).	Text	5				EDGES.MTFCC	H1100, H3010, H3020, L4010, L4020, L4110, L4130	24	0	100	U1N4	5	H1100	S1820
7	FULLNAME	Full Name	Concatenation of expanded text for prefix qualifier, prefix direction, prefix type, base name, suffix type, suffix direction, and suffix qualifier (as available) with a space between each expanded text field.	Text	100					null, 10th St, 11th Ave, 11th St, 12th Ave	14,479	52,167	63.23	N2L2S1U1L1	36		Zygouris Rd
8	SMID	Spatial metadata identifier	Spatial Metadata Identifier (SMID), which identifies the source of the coordinates for each edge and provides the link between the TIGER/Line Shapefiles and the source and horizontal spatial accuracy information. Refer to the metadata for each county or equivalent entity for information on the source for each edge and the horizontal spatial accuracy, where known. Please note that the horizontal spatial accuracy, where reported, refers only to those edges identified as matched to the source with that accuracy. It is not the spatial accuracy of the TIGER/Line Shapefile as a	Text	22	Foreign, Metadata			EDGES.SMID	null, 2751, 2755, 2756, 2758, 2759, 3550, 3551	11	20,963	85.22	N4	4		3999

Metadata tables describing attributes and domains can easily be rendered as HTML for light weight and centralized data description

Metalayers

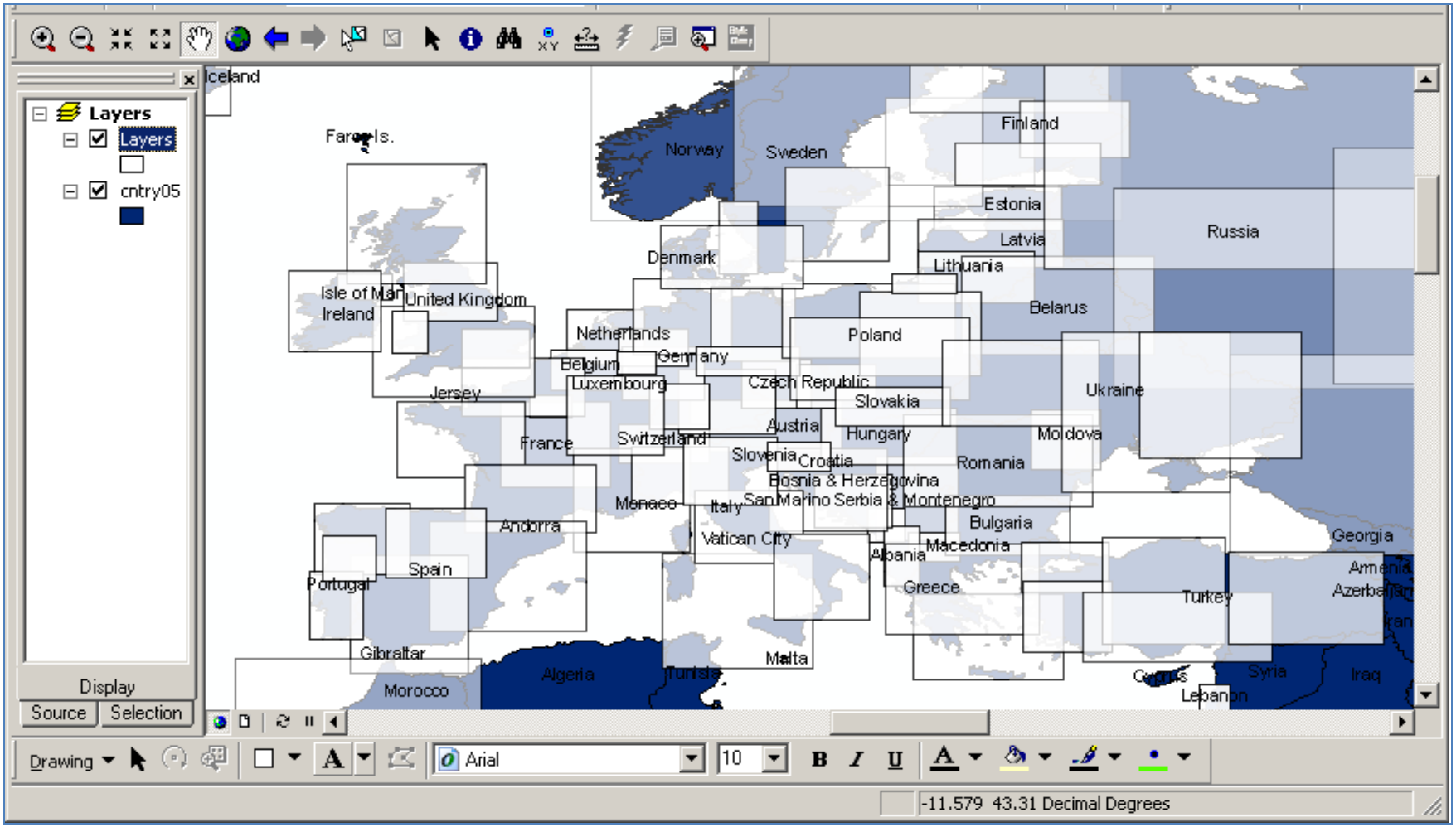
- Create bounding box polygons representing the geographic spatial extents of a set of feature classes.
- In ArcMap, link the bounding box polygons with RDB metadata tables.
- Metadata and the data being described are fully integrated, using the same physical format.
- Can use meta-features and tables in exactly the same way other GIS data is used.
- Intuitively familiar to both end users and application developers.
- No synchronization problems.

Metalayers as GIS Data



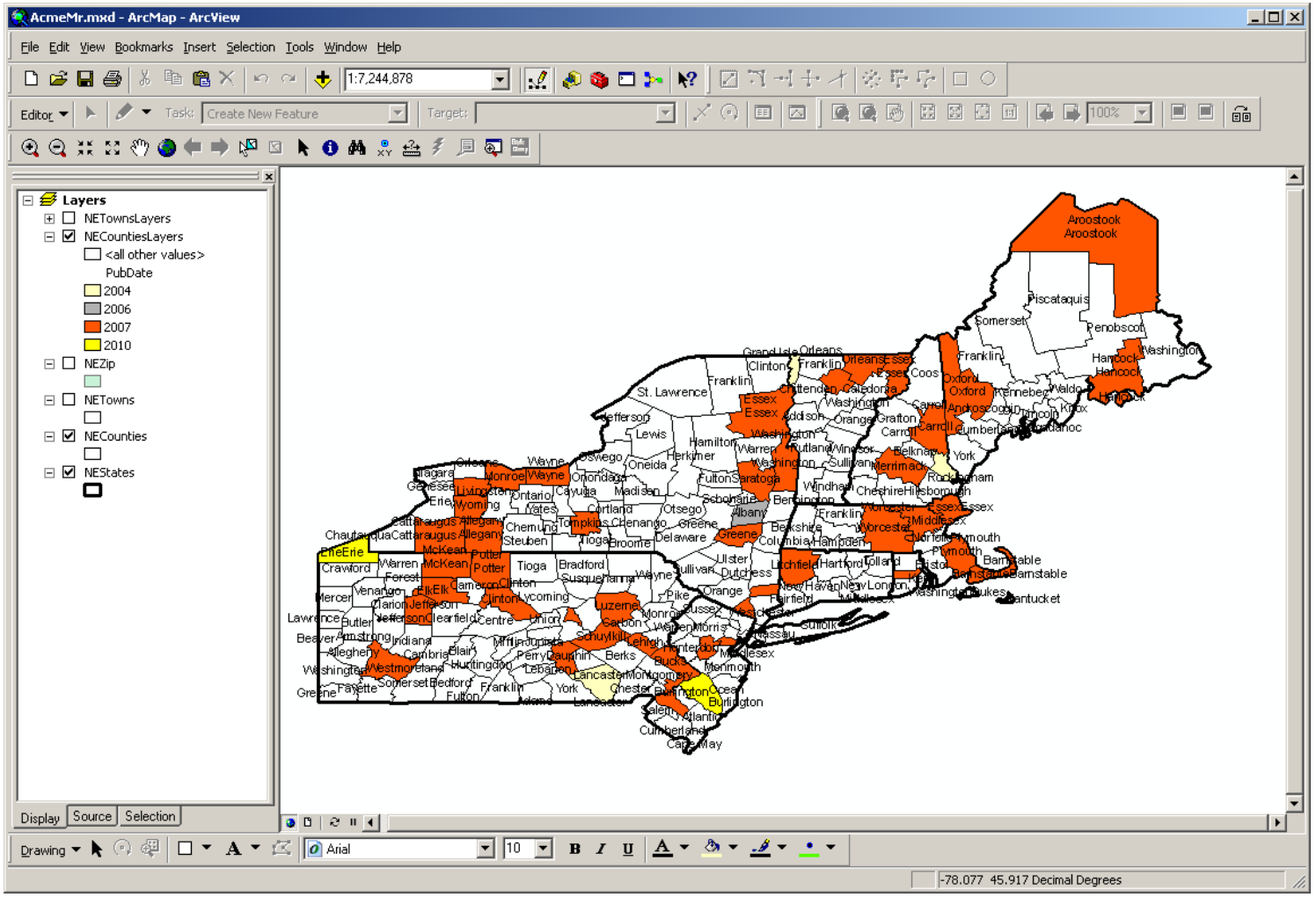
Metadata as feature classes and RDB tables, can be used in the same way as the data being described, through classification, symbolization, filtering, etc.

Meta-Layer Dataset Outlines using Boxes



Geospatial data provider coverage of Europe using bounding boxes for meta-layer dataset outlines.

Metalayers: Using Metadata as GIS Data



Using metadata the same way we use other GIS data allows wide variety of map presentations, reports, etc. to summarize and highlight datasets by metadata values.

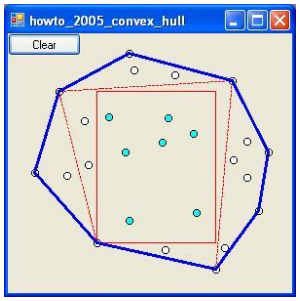
Metalayer Geometry Creation and Management

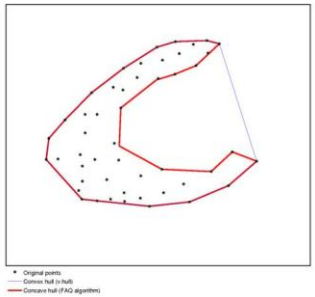
1

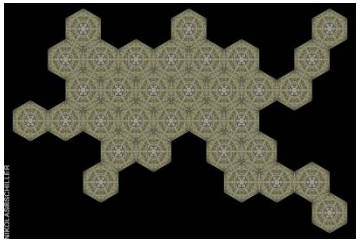
Lon/Lat Bounding Boxes

```
Spatial_Domain:
  Bounding_Coordinates:
    West_Bounding_Coordinate: -167.946360
    East_Bounding_Coordinate: 179.001991
    North_Bounding_Coordinate: 71.298141
    South_Bounding_Coordinate: 17.678360
```

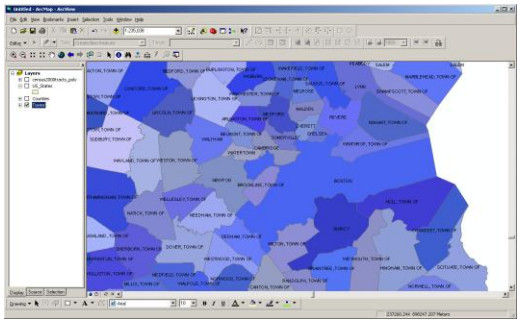
2

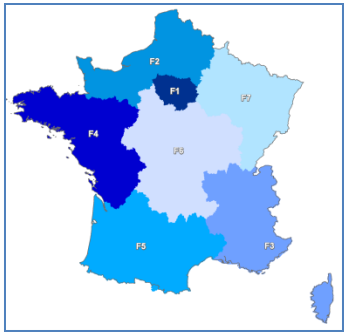






3





Three basic approaches to generating layer coverage polygons with increasing level-of-effort as 1) bounding boxes 2) convex/concave hulls, tessellations and 3) existing administrative or other polygons. Choice based on presentation and data management requirements.

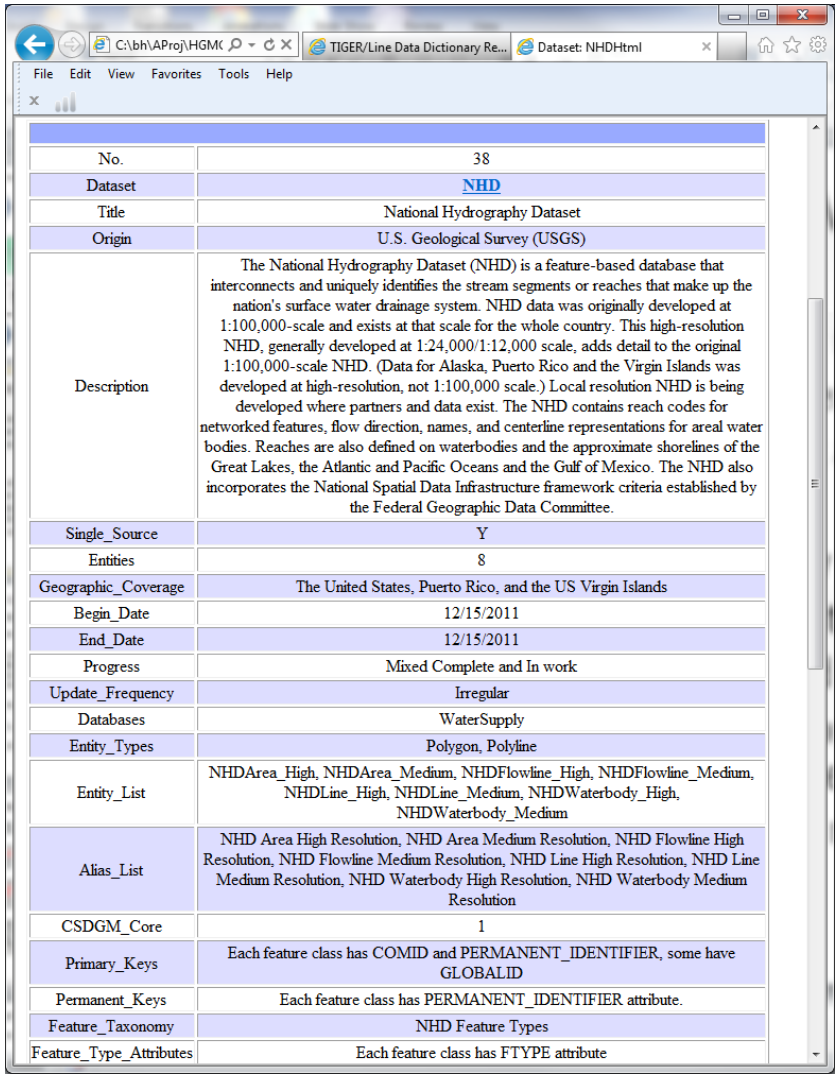
Expanded Metadata

- The basic unit of geospatial metadata today is a single XML document describing a feature class.
- The use of RDB tables for managing geospatial metadata allows for the extension of metadata in several hierarchical directions:
  - The Dataset
  - Feature Level Metadata
  - Aggregation

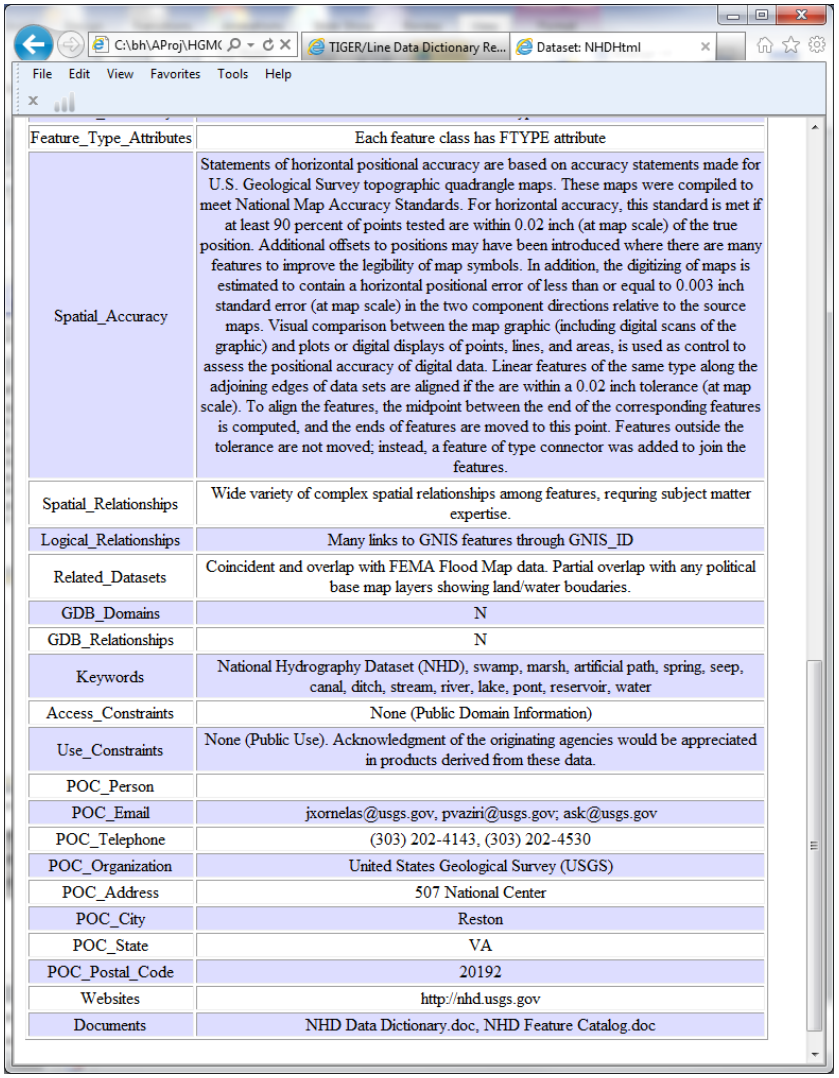
Expanded Metadata: The Dataset

- In many cases, the FGDC CSDGM core elements for the entire set of feature classes and tables comprising a dataset are the same.
- Metadata naturally falls into hierarchies including datasets, entities, attributes, domains, etc.
- In a multi-dataset data store, dataset metadata is very valuable for a quick view of data contents.
- The results of attempting this for the ISO standard, through the ‘series’ has resulted in some very confusing artifacts.

Dataset Metadata: National Hydrography Dataset



No.	38
Dataset	NHD
Title	National Hydrography Dataset
Origin	U.S. Geological Survey (USGS)
Description	The National Hydrography Dataset (NHD) is a feature-based database that interconnects and uniquely identifies the stream segments or reaches that make up the nation's surface water drainage system. NHD data was originally developed at 1:100,000-scale and exists at that scale for the whole country. This high-resolution NHD, generally developed at 1:24,000/1:12,000 scale, adds detail to the original 1:100,000-scale NHD. (Data for Alaska, Puerto Rico and the Virgin Islands was developed at high-resolution, not 1:100,000 scale.) Local resolution NHD is being developed where partners and data exist. The NHD contains reach codes for networked features, flow direction, names, and centerline representations for areal water bodies. Reaches are also defined on waterbodies and the approximate shorelines of the Great Lakes, the Atlantic and Pacific Oceans and the Gulf of Mexico. The NHD also incorporates the National Spatial Data Infrastructure framework criteria established by the Federal Geographic Data Committee.
Single_Source	Y
Entities	8
Geographic_Coverage	The United States, Puerto Rico, and the US Virgin Islands
Begin_Date	12/15/2011
End_Date	12/15/2011
Progress	Mixed Complete and In work
Update_Frequency	Irregular
Databases	WaterSupply
Entity_Types	Polygon, Polyline
Entity_List	NHDArea_High, NHDArea_Medium, NHDFlowline_High, NHDFlowline_Medium, NHDLine_High, NHDLine_Medium, NHDWaterbody_High, NHDWaterbody_Medium
Alias_List	NHD Area High Resolution, NHD Area Medium Resolution, NHD Flowline High Resolution, NHD Flowline Medium Resolution, NHD Line High Resolution, NHD Line Medium Resolution, NHD Waterbody High Resolution, NHD Waterbody Medium Resolution
CSDGM_Core	1
Primary_Keys	Each feature class has COMID and PERMANENT_IDENTIFIER, some have GLOBALID
Permanent_Keys	Each feature class has PERMANENT_IDENTIFIER attribute.
Feature_Taxonomy	NHD Feature Types
Feature_Type_Attributes	Each feature class has FTYPE attribute

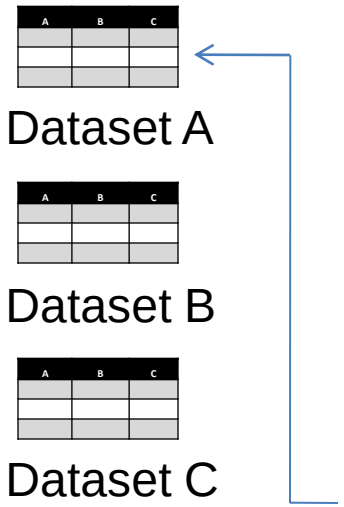


Feature_Type_Attributes	Each feature class has FTYPE attribute
Spatial_Accuracy	Statements of horizontal positional accuracy are based on accuracy statements made for U.S. Geological Survey topographic quadrangle maps. These maps were compiled to meet National Map Accuracy Standards. For horizontal accuracy, this standard is met if at least 90 percent of points tested are within 0.02 inch (at map scale) of the true position. Additional offsets to positions may have been introduced where there are many features to improve the legibility of map symbols. In addition, the digitizing of maps is estimated to contain a horizontal positional error of less than or equal to 0.003 inch standard error (at map scale) in the two component directions relative to the source maps. Visual comparison between the map graphic (including digital scans of the graphic) and plots or digital displays of points, lines, and areas, is used as control to assess the positional accuracy of digital data. Linear features of the same type along the adjoining edges of data sets are aligned if the are within a 0.02 inch tolerance (at map scale). To align the features, the midpoint between the end of the corresponding features is computed, and the ends of features are moved to this point. Features outside the tolerance are not moved; instead, a feature of type connector was added to join the features.
Spatial_Relationships	Wide variety of complex spatial relationships among features, requring subject matter expertise.
Logical_Relationships	Many links to GNIS features through GNIS_ID
Related_Datasets	Coincident and overlap with FEMA Flood Map data. Partial overlap with any political base map layers showing land/water boudaries.
GDB_Domains	N
GDB_Relationships	N
Keywords	National Hydrography Dataset (NHD), swamp, marsh, artificial path, spring, seep, canal, ditch, stream, river, lake, pont, reservoir, water
Access_Constraints	None (Public Domain Information)
Use_Constraints	None (Public Use). Acknowledgment of the originating agencies would be appreciated in products derived from these data.
POC_Person	
POC_Email	jxornelas@usgs.gov, pvaziri@usgs.gov; ask@usgs.gov
POC_Telephone	(303) 202-4143, (303) 202-4530
POC_Organization	United States Geological Survey (USGS)
POC_Address	507 National Center
POC_City	Reston
POC_State	VA
POC_Postal_Code	20192
Websites	http://nhd.usgs.gov
Documents	NHD Data Dictionary.doc, NHD Feature Catalog.doc

Single metadata record for a Dataset: a set of related feature classes and tables

Multi-Source Data Layers

- Some data layers are the result of a merge between multiple input layers from different datasets.
- To track this, each record needs to carry a link back to the original dataset.
- Necessary if swap-out updates are part of workflow.



Name	Value
ObjectId	88197
SOURCE ID	**134**
Name	Acme GIS
Address1	100 Elm St.
Address2	Suite 100
City	Northampton
State	MA
Zip	10160

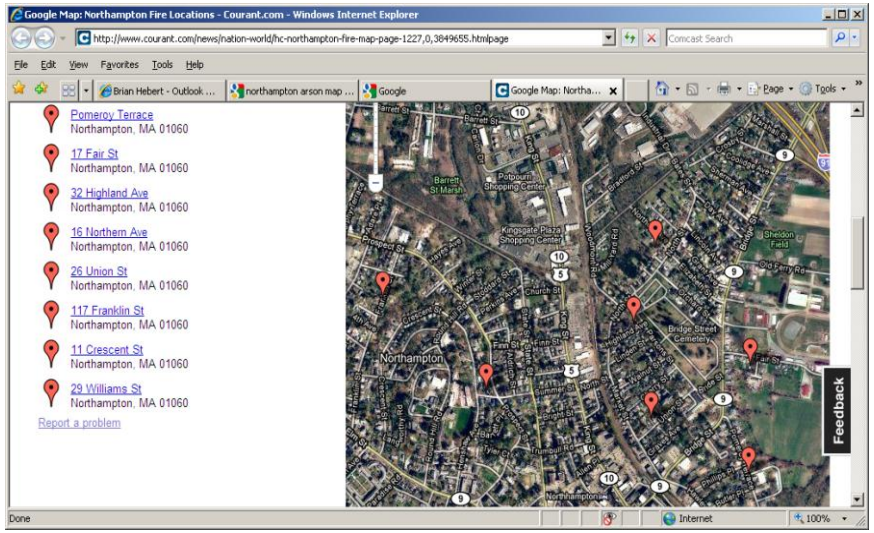
Expanded Metadata: Feature Level

- Lots of GIS data today has feature level metadata, with who, what, when, where, etc. fields found directly on each feature record.

- The use of RDB tables is useful and flexible for handling the hierarchy of metadata elements, e.g., datasets, entities, feature level metadata, using the principle of overrides.

Feature Level Metadata (cont.)

- Current geospatial metadata standards describes the group of records comprising a feature class as a single entity.
- Some end users want metadata at the individual record level.
- This would be a real challenge for FGDC CSDGM or ISO metadata, where metadata is stored in separate XML documents

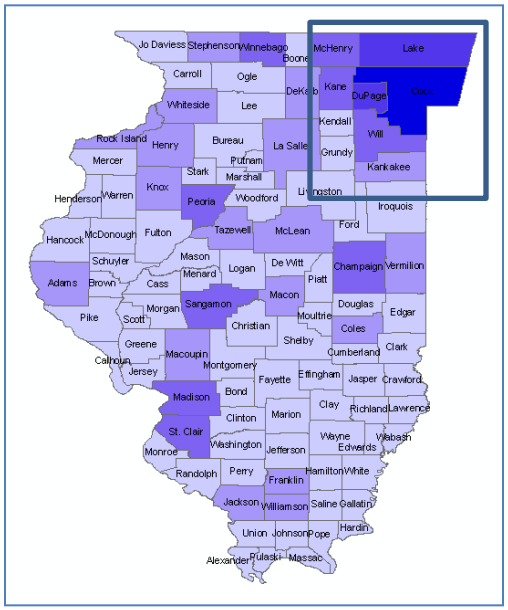


Name	Value
Contact How:	Telephone
Contact Date	11-May-10
Location Confirmed	Y
Moved Geocoded	Y
Accuracy	Building Footprint
Notification	N
Site Image	N
Status	Closed

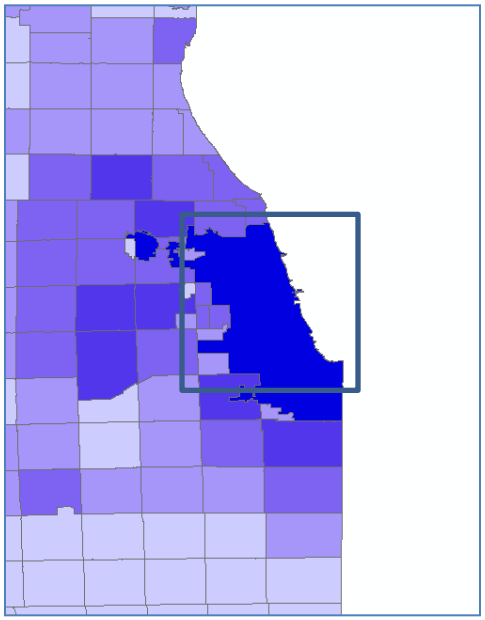
Expanded Metadata: *Georollup* Aggregation

- Metadata in RDB form also facilitates the development of OLAP for GIS
- Aggregate feature counts, lengths, and areas can be aggregated by type, data source, time frame, etc. for enhanced data warehousing based data query.
- Supports drill down and drill through.
- Based on data warehousing, business intelligence, multi-dimensional cube technology.

Metalayer Drilldown and Rollup

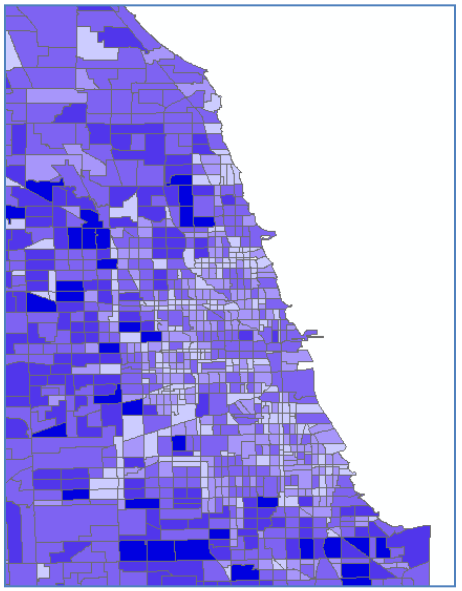


COUNTY



TOWN

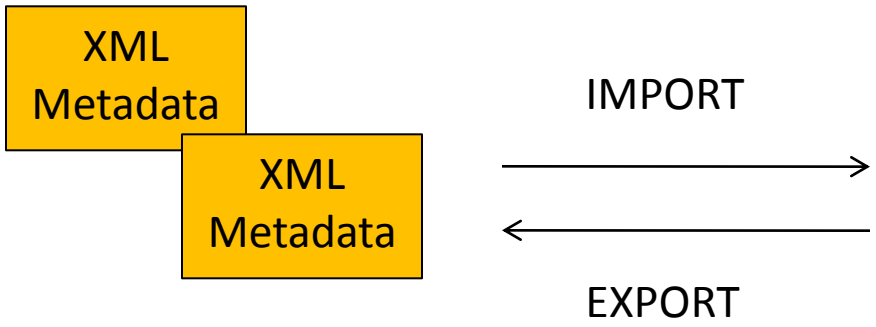
Increasingly detailed views



CENSUS TRACT

Applying Pivot Table like view and **Drilldown and Rollup** with hierarchical geography units

How to Start: FGDC/ISO XML Metadata into the RDB



When this metadata is imported into an RDB, the full flexibility of SQL is available for very flexible management and querying a large collection of metadata as a set.

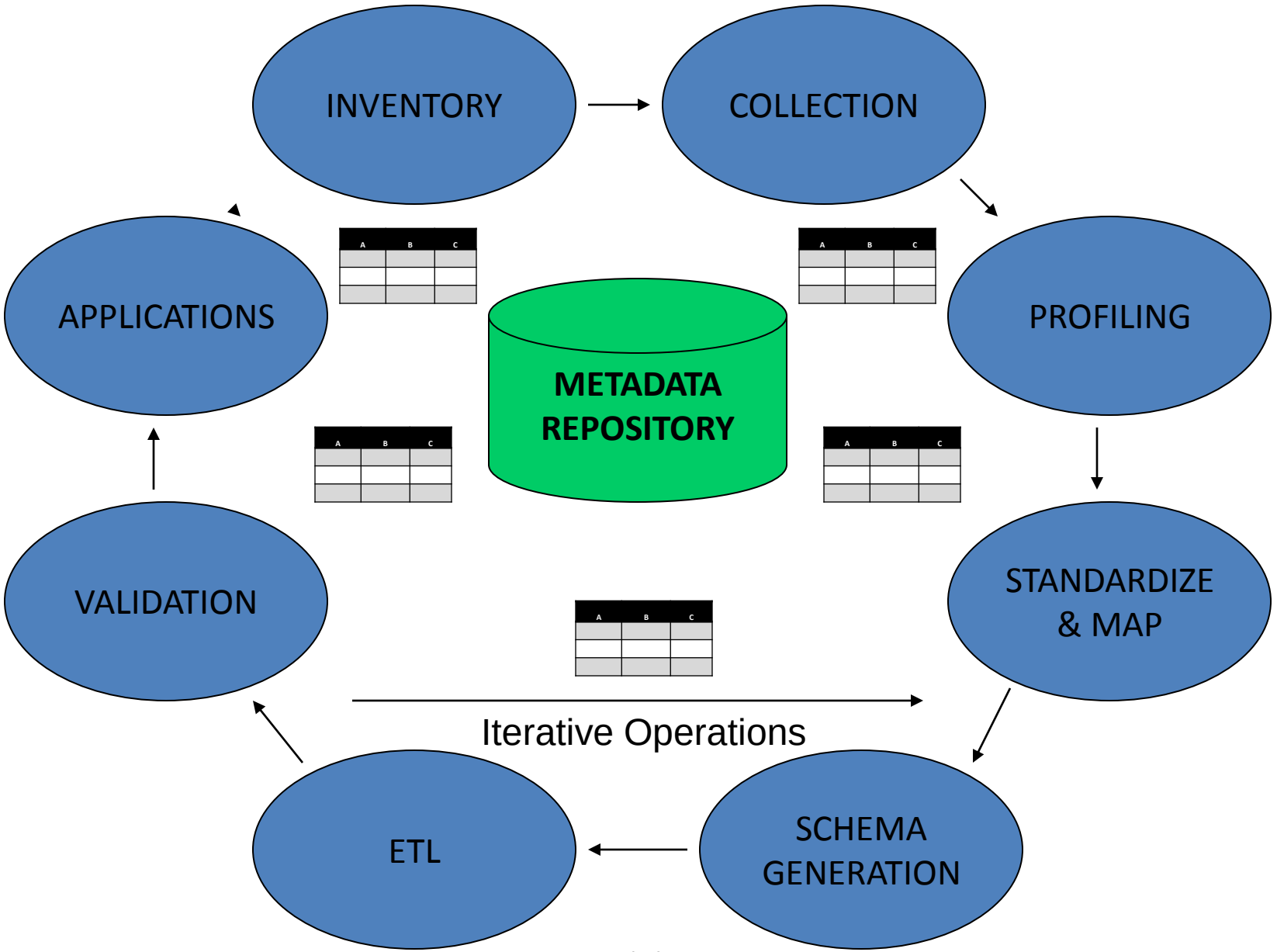
It's easy to exchange data between XML and RDB

NUM	ELEMENT
1	Originator
2	Publication_Date
3	Title
4	Abstract
5	Purpose
6	Calendar_Date
7	Currentness_Reference
8	Progress
9	Maintenance_and_Update_Frequency
10	West_Bounding_Coordinate
11	East_Bounding_Coordinate
12	North_Bounding_Coordinate
13	South_Bounding_Coordinate
14	Theme_Keyword_Thesaurus
15	Theme_Keyword
16	Access_Constraints
17	Metadata_Date
18	Contact_Person
19	Address_Type
20	Address
21	City
22	State_or_Province
23	Postal_Code
24	Contact_Voice_Telephone
25	Metadata_Standard_Name
26	Metadata_Standard_Version

Benefits to Application Developers

- The current set of NSDI (National Spatial Data Infrastructure) server nodes and the applications providing access to metadata are based on sets of FGDC CSDGM documents , as database.
- The code written to use this data is quite different from what we think of as GIS data application code.
- If metadata were stored in the same physical format as the data it was describing, the same code could be used to write applications for accessing and viewing it.

The RDB Supports a Wide Variety of Data Description and Integration Tasks



Consequences of XML to RDB Mapping

- Losses
  - Arbitrarily nested elements
  - Variable length elements
  - Can’t look at in a browser
- Gains
  - Familiarity
  - Ease of authoring and access
  - Integrated data and metadata
  - Reuse of presentation and application development technology
  - Read only models can relax normalization

Part 3: Why Don't We Use Feature Classes and RDB Tables for Geospatial Metadata?

- XML became the de facto standard for GIS metadata implementation
- FGDC did not explicitly state that XML would be used for implementing the standard.
- FGDC CSDGM did explicitly state that the standard provides the content, not the implementation or encoding.
- XML was a fad at the time, particularly in a web based context.
- XML was a better HTML, and great for configuration files.
- ISO does explicitly use XML for *encoding* of standard.

FGDC CSDGM Physical Implementation Guidelines

- The FGDC/CSDGM standard clearly states that the standard describes content, and not physical implementation. From the CSDGM Workbook:

*The standard specifies information content, but **not how to organize this information in a computer system** or in a data transfer, or how to transmit, communicate, or present the information to a user. There are several reasons for this approach:*

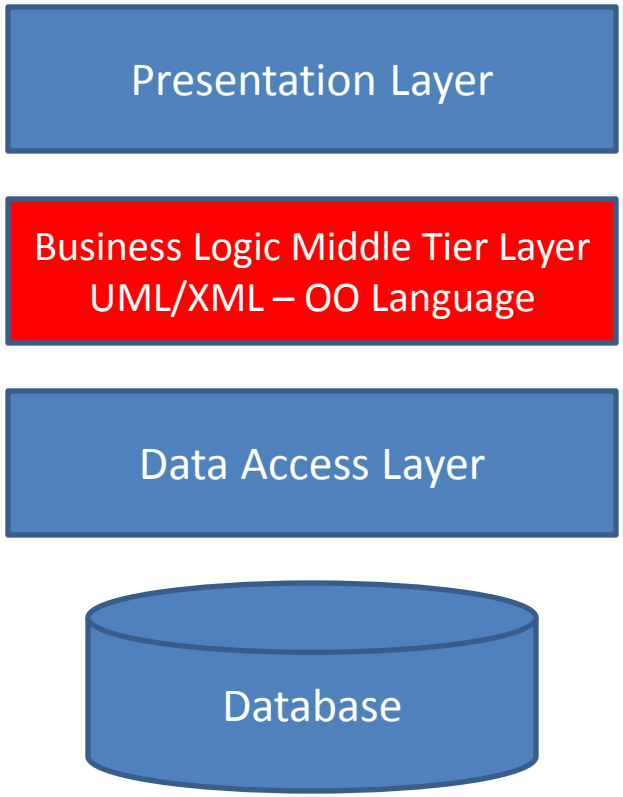
*There are many means by which metadata could be organized in a computer. **These include incorporating data as part of a geographic information system, in a separate data base**, and as a text file. Organizations can choose the approach which suits their data management strategy, budget, and other institutional and technical factors.*

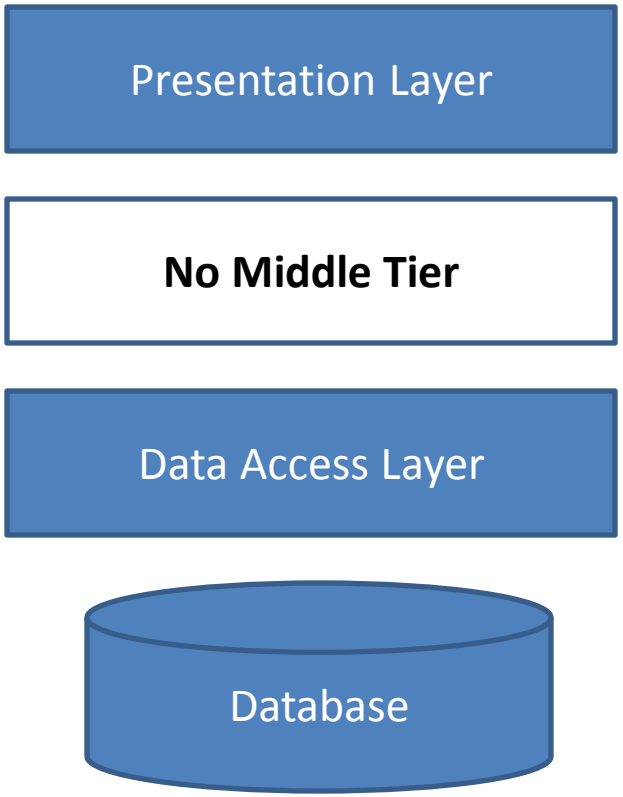
In spite of these statements, geospatial metadata implementation has not been approached using industrial strength RDBMS data access technology, but rather relies on sets of separate XML files, using an entirely different data access and management paradigm than that used by the data it is describing.

Database Models in IT

- Broadly speaking, there are 2 basic kinds of databases and related applications:
  - OLTP: On-Line-Transaction-Processing
  - OLAP: On-Line-Analytical-Processing
- GIS is closest to OLTP.
- (OO databases never made it.)
- Best practice design of OLTP systems involves 3 tiers:
  - Back end database
  - Middle tier business logic, OO language
  - Front end presentation tier, web pages or thick desktop windows
- OLAP typically has a 2 tier system where the query language, e.g., MSQL, is used as an interface between the 2.

OLTP/OLAP Design Differences and the Middle Tier

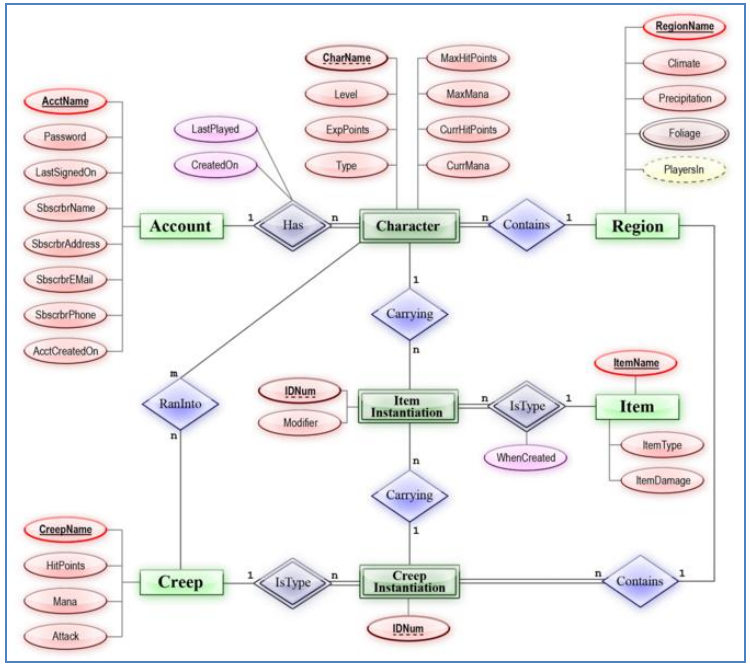


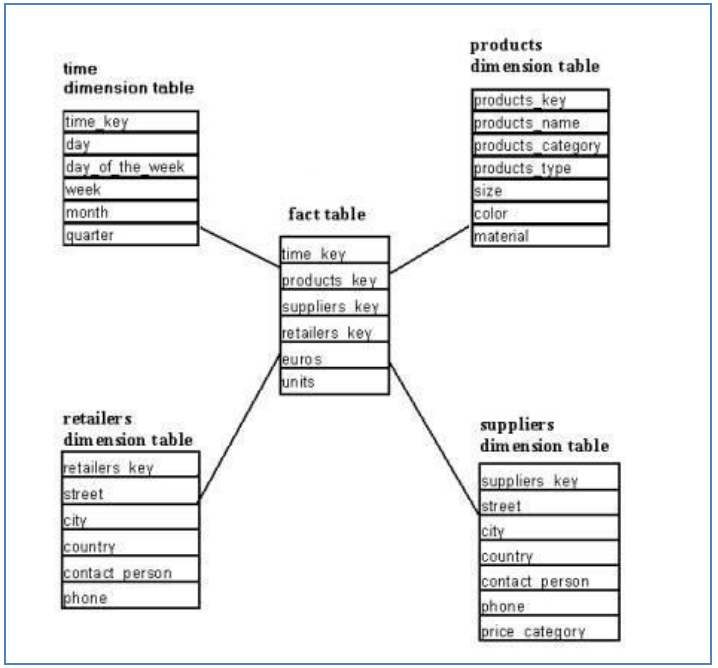


Production **OLTP** database solutions typically use a middle tier for representing higher level business objects and rules. This middle tier is often designed using UML and implemented with an Object Oriented programming language.

Decision Support **OLAP** database solutions typically have no Middle tier. They present and access data directly through query language behind pivot tables and report generators.

Data Model Differences: Production vs. Decision Support





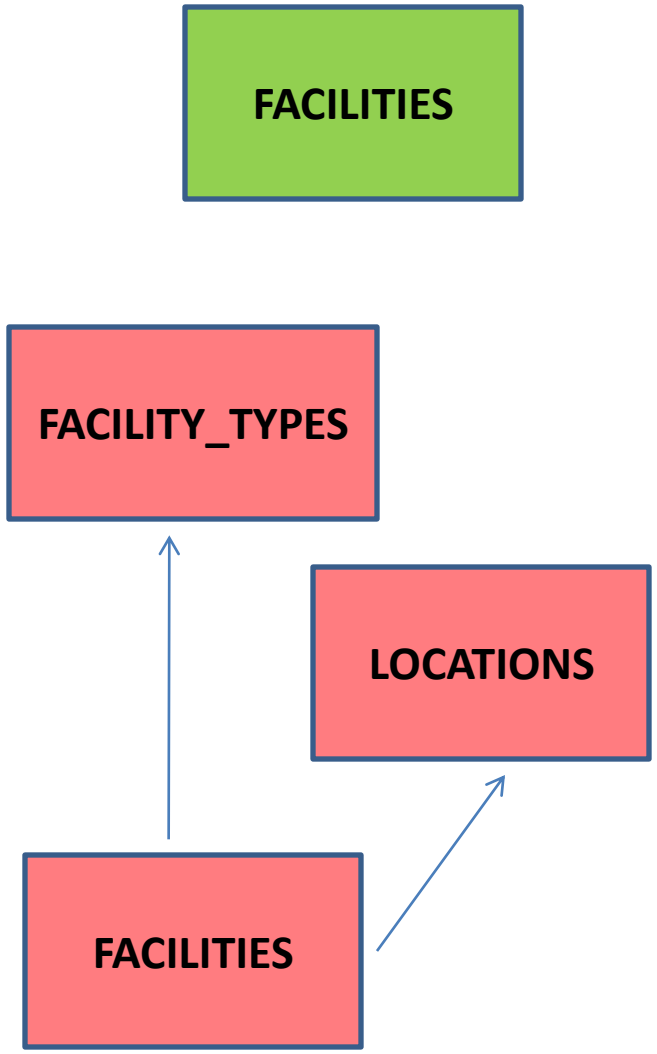
Normalized for referential integrity, **complex and slower performing queries**, data is edited

De-normalized for **easily formed and faster performing queries**, data is read-only

The data models and supporting tools used in data warehousing are significantly different from those found across the geospatial community. **Geospatial data modelers tend to incorrectly use production models for decision support databases.**

De-Normalization Makes Queries Easier

- 1 De-Normalized Table: SELECT TYPE, LOCATION FROM FACILITIES
- 3 Normalized Tables: SELECT FACILITY_TYPES.TYPE, LOCATIONS.LOCATION FROM (FACILITIES INNER JOIN FACILITY_TYPES ON FACILITIES.TYPE = FACILITY_TYPES.ID) INNER JOIN LOCATIONS ON FACILITIES.LOCATIONID = LOCATIONS.ID;
- NAVTEQ SDC data is a good example. De-normalized, e.g., County Name and FIPS, highly indexed, very fast and easy to use



XML is Great as an OO Design Tool

- XML is very useful, when used with UML, etc. for generating Object Oriented, in memory, code-based middle tier models.
- Can easily handle complex variable length, nested data constructs.
- However this is only a single tier, a part, of the entire solution space.
- There is a great deal of technology now used, e.g., Hibernate, to handle the mapping between middle tier, in-memory, object oriented data stores, and back end databases.
- The FGDC CSDGM would be typical of a design for an in-memory, object oriented middle tier supporting an OLTP system.

Backend Data Tier Example: The Shapefile

- ESRI Shapefile is an open source standard
- Many different applications can read and write Shapefile data.
- There in-memory models are vastly different
- But the key to flexible exchange is the common, open source, persistent format.
- The notion that exchanging data packaged in a middle tier OO model vs. using backend RDB table based data storage is questionable.
- The meta-model of XML is considerably more complex than the simple table, row, column meta-model used in RDB tables.

Intentions of XML Use

- XML was not meant to be a replacement for RDB
- It's a better HTML.
- One primary use was to exchange system neutral data over the internet.
- Great for settings files, as mini-databases.
- Many characteristics of an OLTP or OLAP DBs can not easily be implemented with XML technology:
  - Multi-user access
  - Links between data entities
  - Indexing
  - Record locking
  - Rollbacks
  - Aggregation

Common Coordinate Storage in XML

- Coordinate geometry values stored in lots of geospatial XML do not really use XML data element storage.

- GML (XML) Coordinates are typically stored as delimited text strings like this:

  <gml:LineString>**45.67 88.56 55.56 89.44**</gml:LineString >

- Otherwise they would look like this as individual elements:

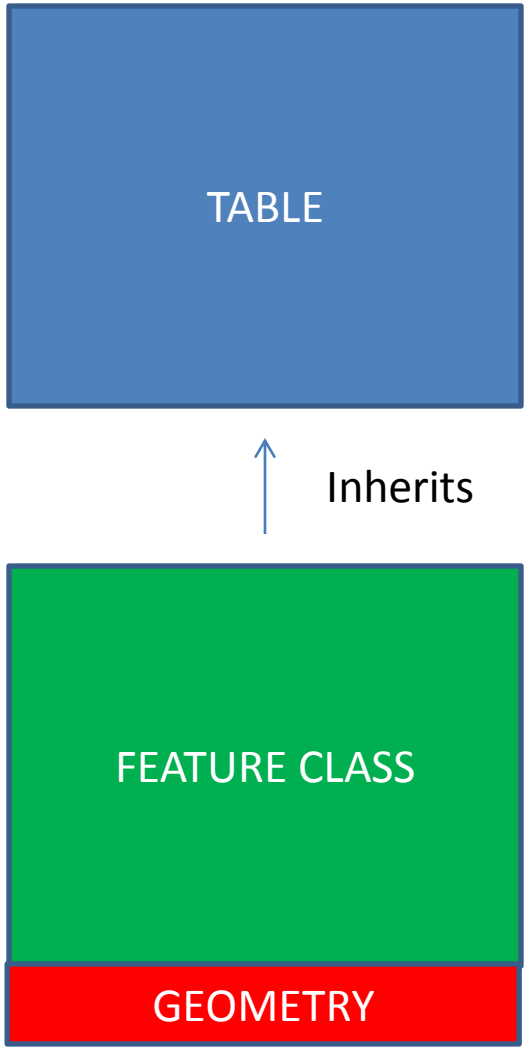
  <x>**45.67**</x><y>**88.56**</y><x>**55.56**</x><y>**89.44**</y>

Facilitating Data Exchange and Description

- Table to table data exchange is, and has been for a long time, the primary method for moving data from one database to another.
- FME, ArcGIS, MS Access, SQL Server, Oracle
- Use of RDB for data exchange and collection: CDC Cancer Records, FBI Crime Records, etc.
- GIS Example: NAVTEQ uses a number of table based metadata stores.

Part 4: Design Problems with CSDGM

- These observations are independent of physical implementation through XML.
- A basic OO design consideration, that GIS feature classes are RDB tables with binary blobs for holding geometry, was not used in the design of FGDC CSDGM.
- A feature class is a special type of, and extends the definition, of a table
- It was used in ArcGIS/ArcObjects
- BUT, there is no explicit support for tables in the FGDC CSDGM standard.
- **HOW DID THEY MISS THIS?**
- In SDTS based entity type domain, there is no such thing as a table.



Design Problems with CSDGM (cont.)

- The FGDC CSDGM element for holding a record count is hard coded into a geometry only construct.
- So there is no place to put the number of records for a table.
- Domains don’t have names and can’t be shared, they are part of an attribute
- This is also not the case with ESRI Geodatabases.
- Domains can’t hold more than 2 values.

Design Problems with CSDGM (cont.)

- There are no relationships.
- There are no full names.
- Column data type, length are optional
- There are no data domain patterns, e.g., regular expressions
- Horizontal position is optional and not standardized, in spite of widespread geocoding results classification and American Map Accuracy Standards
- Highly nested, mandatory, optional, etc. elements are very confusing.

Part 5: Communication Problems with FGDC CSDGM

- When first trying to learn about FGDC CSDGM through the Workbook, users are required to learn about very abstract concepts focused on how compilers are written, using what are called *production rules*.
- Equivalent to asking SQL developers to master Cartesian algebra.
- Efforts to deal with this, e.g., BLM color coded nested boxes have gone a long way to help, but it's still very complicated.

Production Rules from the FGDC CSDGM Standard

A production rule specifies the relationship between a compound element, and data elements and other (lower-level) compound elements. Each production rule has a left side (identifier) and a right side (expression) connected by the symbol "=", meaning that the term on the left side is replaced by or produces the term on the right side. Terms on the right side are either other compound elements or individual data elements. By making substitutions using matching terms in the production rules, one can explain higher-level concepts using data elements. The symbols used in the production rules have the following meaning:

Symbol Meaning

=	is replaced by, produces, consists of
+	and
[]	selection - select one term from the list of enclosed terms (exclusive or). Terms are separated by "|"
m{}n	iteration - the term(s) enclosed is(are) repeated from "m" to "n" times
()	optional - the term(s) enclosed is(are) optional

Examples:

a = b + c	"a consists of b and c"
a = [b | c	"a consists of one of b or c"
a = 4{b}6	"a consists of four to six occurrences of b"
a = b + (c)	"a consists of b and optionally c"

Interpreting the production rules:

The terms bounded by parentheses, "(" and ")", are optional and are provided at the discretion of the data producer. If a producer chooses to provide information enclosed by parentheses, the producer shall follow the production rules for the enclosed information. For example, if the producer decides to provide the optional information described in the term:

(a + b + c)

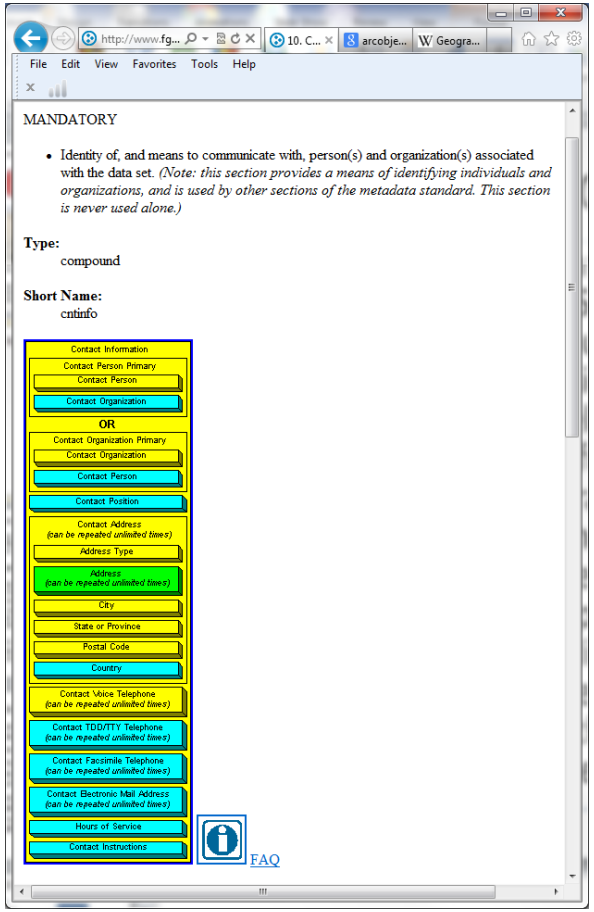
the producer shall provide a and b and c.

Only for terms bounded by parentheses does the producer have the discretion of deciding

Should it be necessary for a geospatial metadata author to start by needing to learn the basics of compiler writing?

Color Coded Nested Diagrams

- Susan Stitt of the USGS Biological Resources Division developed a very helpful diagraming method for presenting very complex, conditional, nested, variable length, data model.



http://www.fgdc.gov/csdgmgraphical/index.html

A Communication Gap

Layers
Attributes
Symbols
…

?

Production
Rules, UML,
XSD
GML, ISO

GIS Users

The Tower of Babel

Data Modelers
Standards Bodies

GIS end users think of data and data models in terms of layers, tables, attributes. Geospatial standards developers think of data and models in terms of Object Oriented UML, XSD, XML, etc.

CSDGM Design/Communication Conclusions

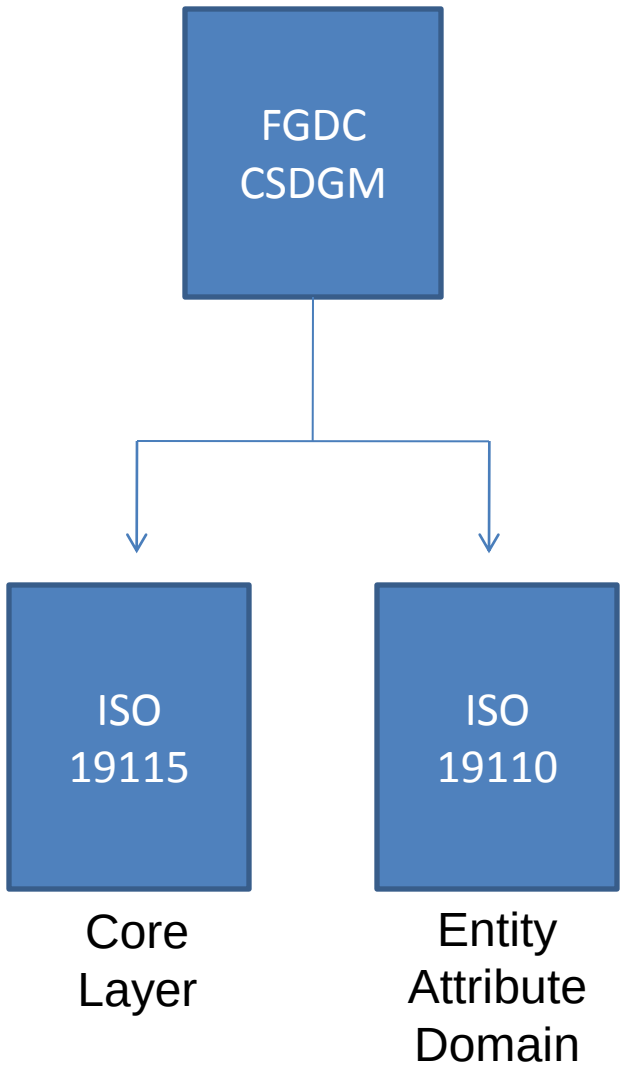
- The communities designing these standards do not contain the full assortment of solution roles required to build comprehensive information management systems, e.g., OLTP, OLAP.
- **Critical IT Problem: No separation between design and implementation.**
- Standards are designed that are difficult if not impossible to implement using OLTP or OLAP technology.
- (The architects have built the building.)
- If a top-notch IT solutions team were asked to develop a system for managing large volumes of any kind of data, would they use something like FGDC CSDGM or ISO for the design of backend data store or the middle tier?
- Result: GIS metadata is not easy
- Lots of really good metadata won't pass the USGM Metadata Parser (MP) test, the Keyword Thesaurus element, etc.

The ISO Standard and Migration

- The ISO standards are incredibly complex, and basically inaccessible to GIS practitioners.
- There was a schema split as well.
- We see similar complexity explosions in other areas, e.g., compare the first version of earlier and later versions of GML.
- The FGDC CSDGM to ISO migration basically involves mapping one object oriented database model to another, with ample use of nested, variable length constructs, using different names.
- There are countless efforts to address the complexity of this transition, including cross references, training, applications, etc.
- Is this really necessary?

ISO 19115/19110 Split

- FGDC CSDGM included both core metadata about layer and entity, attribute, domain info
- ***ISO 19115 Geographic Information – Metadata*** doesn't include this basic database-centric metadata
- ***19110 Feature Catalog*** does contain entity, attribute, domain info, but mixed in with a great deal of other material
- Presents a significant challenge to migration, for tool providers and users alike



ESRI ArcGIS Metadata Technology

- ESRI's dominance in the GIS market, along with the XML based approach used by standards bodies, had framed the way we think about metadata.
- ESRI took the FGDC CSDGM standard very literally in designing their tool set.
- This is an XML as database approach.
- A different approach, in which an RDB were used, with XML available as a format for reports was not used.
- This RDB approach was used by some other vendors, e.g., Intergraph.
- The NSDI nodes, the geospatial portal toolkit, still end up using databases to store metadata.

GIS Technology in Broader IT Context

- GIS technology developed in a somewhat isolated manner
- Binary blobs for coordinate values invited the development of custom, proprietary systems, e.g., Intergraph IGDS, AutoCAD, Microstation, pre-ArcGIS ArcView, Smallworld, etc.
- The merger with RDB technology had enormous consequences, ArcGIS, Oracle Spatial, PostGIS, etc. but there are still many differences, particularly with metadata management. RDBs have system tables containing metadata describing tables, columns, etc.
- Result: GIS practitioners are typically unfamiliar with a vast array of mainstream IT data management and application development paradigms.

Part 6: Why do we create metadata?

Rudyard Kipling Poem from The Elephant's Child

I KEEP six honest serving-men - (They taught me all I knew);
Their names are **What** and **Why** and **When** - And **How** and **Where** and **Who**.
I send them over land and sea, - I send them east and west;
But after they have worked for me, - I give them all a rest.

I let them rest from nine till five, - For I am busy then,
As well as breakfast, lunch, and tea, - For they are hungry men.
But different folk have different views; - I know a person small—
She keeps ten million serving-men, - Who get no rest at all!

She sends'em abroad on her own affairs, - From the second she opens her eyes—
One million Hows, two million Wheres, - And seven million Whys!

Kipling Metadata

No.	Metadata Element	Metadata Value
1	What	EDGES - The All Lines shapefile contains visible linear features such as roads, railroads, and hydrography, as well as non-feature edges, non-visible Current boundaries, or superseded Census 2000 boundaries.
2	Why	In order for others to use the information in the Census MAF/TIGER database in a geographic information system (GIS) or for other geographic applications, the Census Bureau releases to the public extracts of the database in the form of TIGER/Line Shapefiles.
3	When	2009
4	How	TIGER/Line Shapefiles are extracted from the Census MAF/TIGER database by nation, state, county, and entity. Census MAF/TIGER data for the nation, state, county, and entity are then distributed among 58 shapefiles each containing attributes for line, polygon, or landmark geographic data.
5	Where	United States, U.S., County or Equivalent Entity, St. Louis, 29510
6	Who	U.S. Department of Commerce, U.S. Census Bureau, Geography Division

- GIS metadata doesn't have to be that difficult. Imagine creating a simple table with these 6 basic elements.
  - Extend these to match the Dublin Core
- Map between FGDC CSDGM core elements and Dublin Core

Goal: Maximize Understanding of Data

FGDC Metadata

Data Profiles

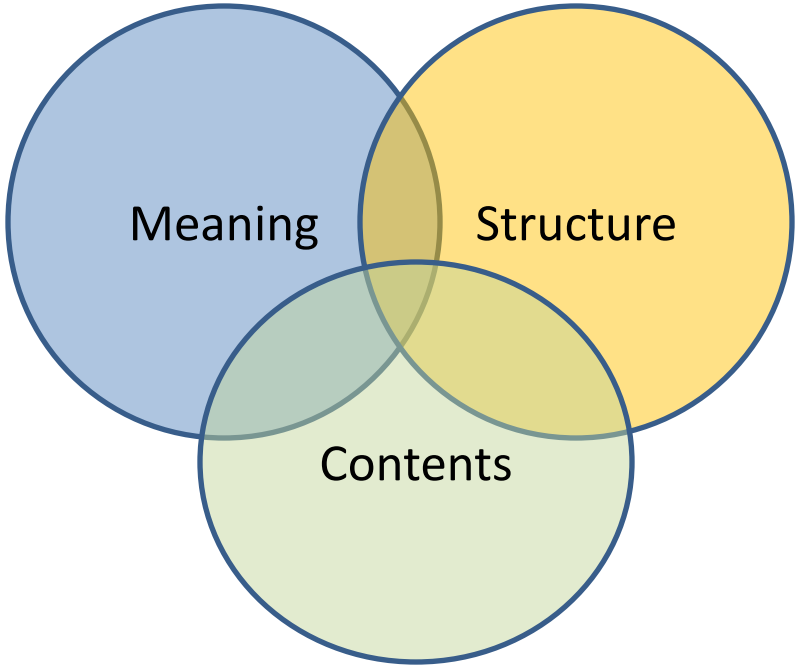
Data Quality Assessments

Cross Referenced Terms

Keywords, Aliases, Indexes

Table of Contents

Glossary



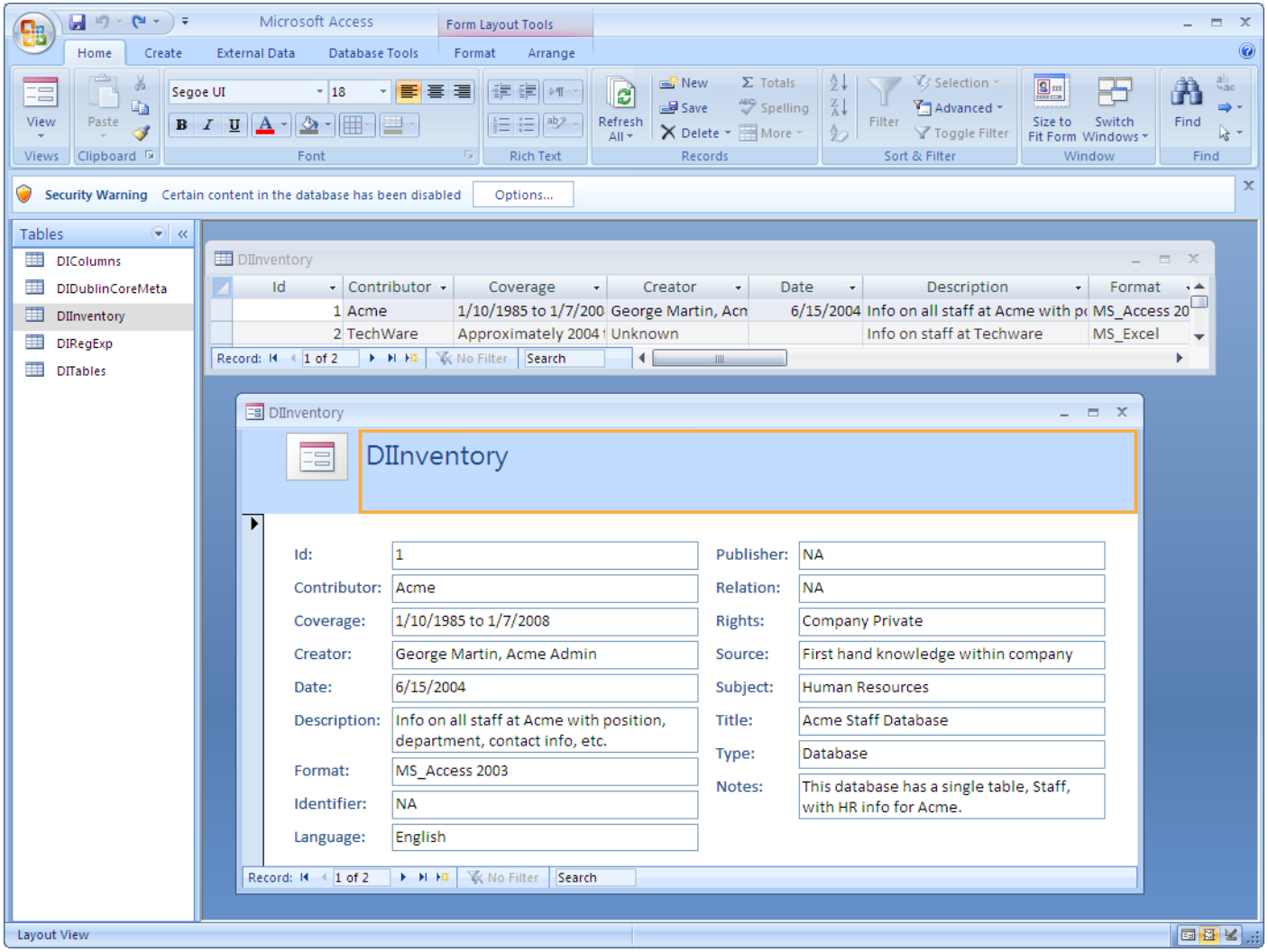
Complete metadata describes Meaning, Structure, and Contents.
Maximize understanding by end user and help write applications.
Help with variety of data description and integration tasks.

The Dublin Core

NUM	ELEMENT	DEFINITION
1	Contributor	An entity responsible for making contributions to the resource.
2	Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
3	Creator	An entity primarily responsible for making the resource.
4	Date	A point or period of time associated with an event in the lifecycle of the resource.
5	Description	An account of the resource.
6	Format	The file format, physical medium, or dimensions of the resource.
7	Identifier	An unambiguous reference to the resource within a given context.
8	Language	A language of the resource.
9	Publisher	An entity responsible for making the resource available.
10	Relation	A related resource.
11	Rights	Information about rights held in and over the resource.
12	Source	The resource from which the described resource is derived.
13	Subject	The topic of the resource.
14	Title	A name given to the resource.
15	Type	The nature or genre of the resource.

http://dublincore.org/documents/dces

Dublin Core Example: Similar to CSDGM Core



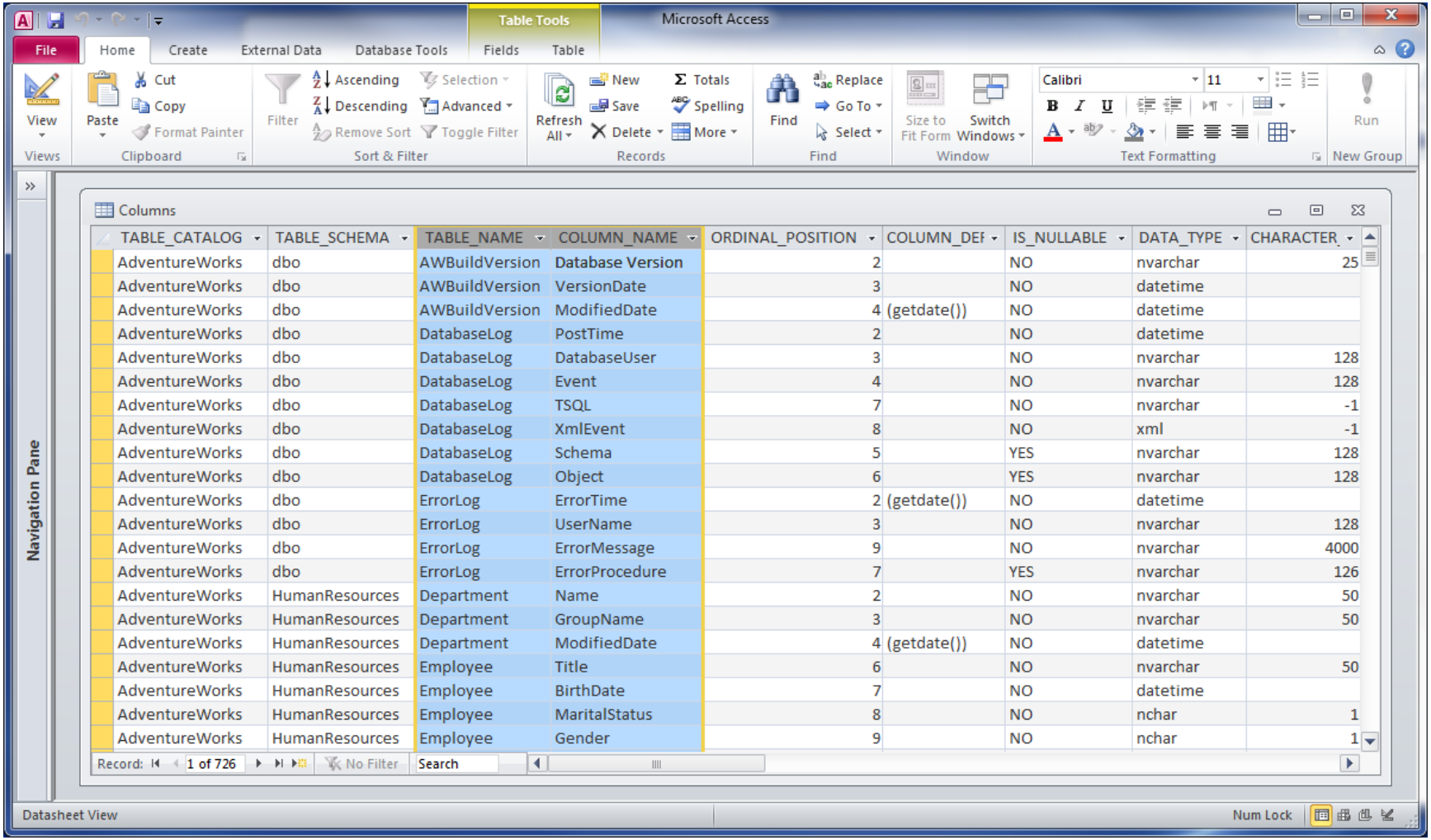
Metadata authoring does not have to be difficult.

Part 7: RDB Metadata as Tables

- RDB systems provide metadata, as tables, through standardized data access API's, e.g., ODBC, JDBC, OLE.DB, etc.
- RDB systems also typically have system tables listing tables, columns, constraints, domains, etc.

Database	Metadata Column
sqlserver	TABLE_CATALOG
sqlserver	TABLE_SCHEMA
sqlserver	TABLE_NAME
sqlserver	TABLE_TYPE
oracle	OWNER
oracle	TABLE_NAME
oracle	TYPE
mysql	TABLE_CATALOG
mysql	TABLE_SCHEMA
mysql	TABLE_NAME
mysql	TABLE_TYPE
dbf	TABLE_CATALOG
dbf	TABLE_SCHEMA
dbf	TABLE_NAME
dbf	TABLE_TYPE
access	TABLE_CATALOG
access	TABLE_SCHEMA
access	TABLE_NAME
access	TABLE_TYPE

Metadata Table Example: SQL Server



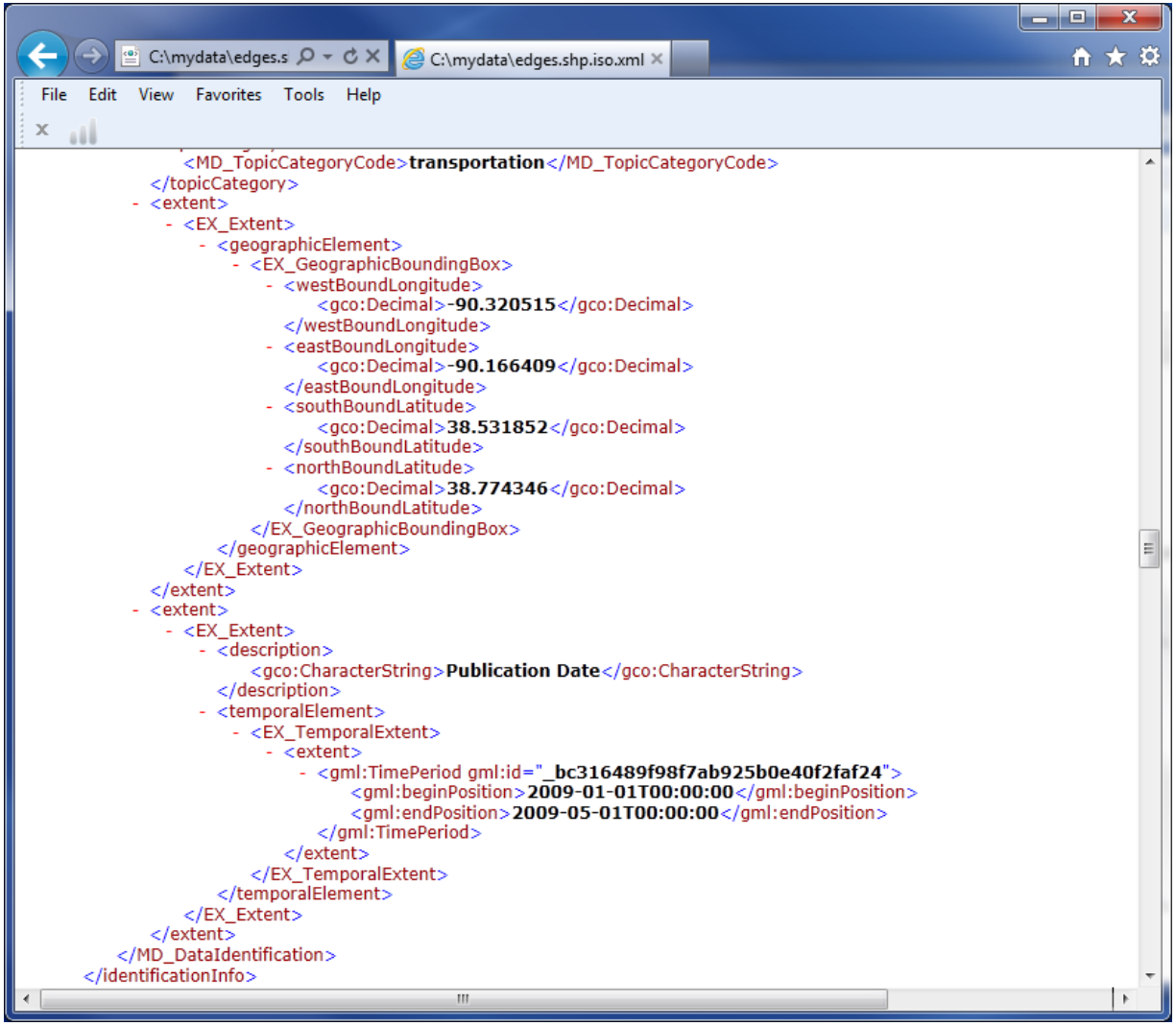
TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	COLUMN_NAME	ORDINAL_POSITION	COLUMN_DEF	IS_NULLABLE	DATA_TYPE	CHARACTER
AdventureWorks	dbo	AWBuildVersion	Database Version	2		NO	nvarchar	25
AdventureWorks	dbo	AWBuildVersion	VersionDate	3		NO	datetime	
AdventureWorks	dbo	AWBuildVersion	ModifiedDate	4	(getdate())	NO	datetime	
AdventureWorks	dbo	DatabaseLog	PostTime	2		NO	datetime	
AdventureWorks	dbo	DatabaseLog	DatabaseUser	3		NO	nvarchar	128
AdventureWorks	dbo	DatabaseLog	Event	4		NO	nvarchar	128
AdventureWorks	dbo	DatabaseLog	TSQL	7		NO	nvarchar	-1
AdventureWorks	dbo	DatabaseLog	XmlEvent	8		NO	xml	-1
AdventureWorks	dbo	DatabaseLog	Schema	5		YES	nvarchar	128
AdventureWorks	dbo	DatabaseLog	Object	6		YES	nvarchar	128
AdventureWorks	dbo	ErrorLog	ErrorTime	2	(getdate())	NO	datetime	
AdventureWorks	dbo	ErrorLog	UserName	3		NO	nvarchar	128
AdventureWorks	dbo	ErrorLog	ErrorMessage	9		NO	nvarchar	4000
AdventureWorks	dbo	ErrorLog	ErrorProcedure	7		YES	nvarchar	126
AdventureWorks	HumanResources	Department	Name	2		NO	nvarchar	50
AdventureWorks	HumanResources	Department	GroupName	3		NO	nvarchar	50
AdventureWorks	HumanResources	Department	ModifiedDate	4	(getdate())	NO	datetime	
AdventureWorks	HumanResources	Employee	Title	6		NO	nvarchar	50
AdventureWorks	HumanResources	Employee	BirthDate	7		NO	datetime	
AdventureWorks	HumanResources	Employee	MaritalStatus	8		NO	nchar	1
AdventureWorks	HumanResources	Employee	Gender	9		NO	nchar	1

Example of database column metadata stored in RDB table, SQL Server, AdventureWorks

Equivalent Contents

- Accepting the notion that metadata can be stored in RDB tables is dependent on a notion of Equivalent Contents
- Regardless of it's physical storage mechanism, a data element always has:
  - Name, Data Type, Meaning, Value
- Consider the 4 floating point values used to describe the bounding box of a geospatial feature class.
- Stored as XML, CSV, part of a Shapefile, RDB, HTML, etc.
- Regardless of the actual physical storage mechanism, the values mean the same thing.
- **An essential notion in successful IT systems relies on separating logical end user views from underlying physical implementation.**
- This is not the case with current geospatial metadata practices that rely on exposing relatively complicated XML schemas to authors and consumers.

Excessive Nesting in ISO Standard



What's wrong with this picture?

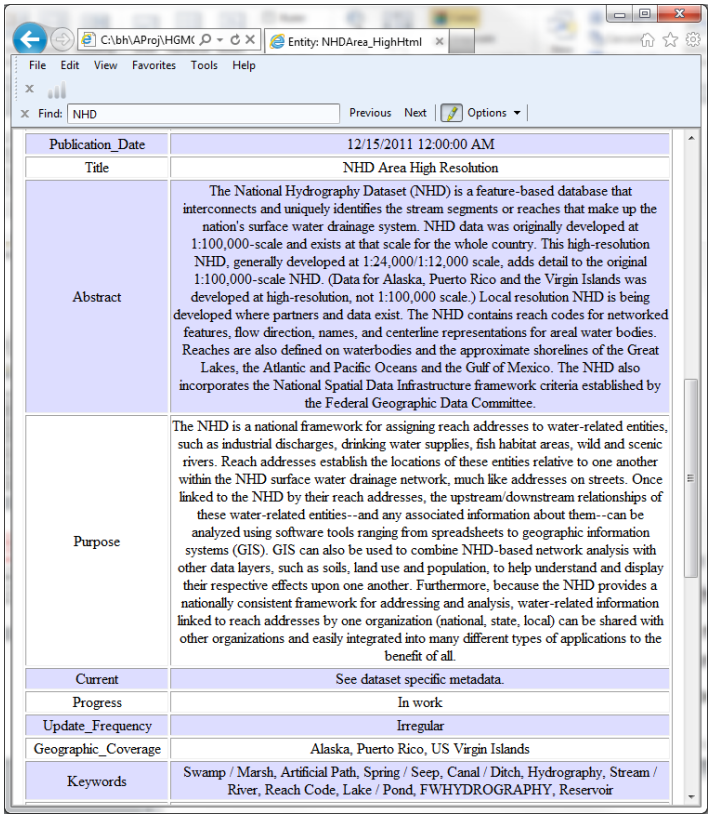
Part 8: Politics, Religion, and Change

- The idea that GIS metadata should not be managed using XML is heresy to some.
- Very strong beliefs related to the use of XML, in FGDC, ISO, OGC, etc.
- This is unlikely to change quickly in the GIS world.
- BUT, changes toward simpler more robust approaches do eventually win out, e.g., SQL vs. hierarchical or network databases, RESTFUL web services vs. bloated XML SOAP, json vs. XML, etc.
- In the meantime, can anything be done?

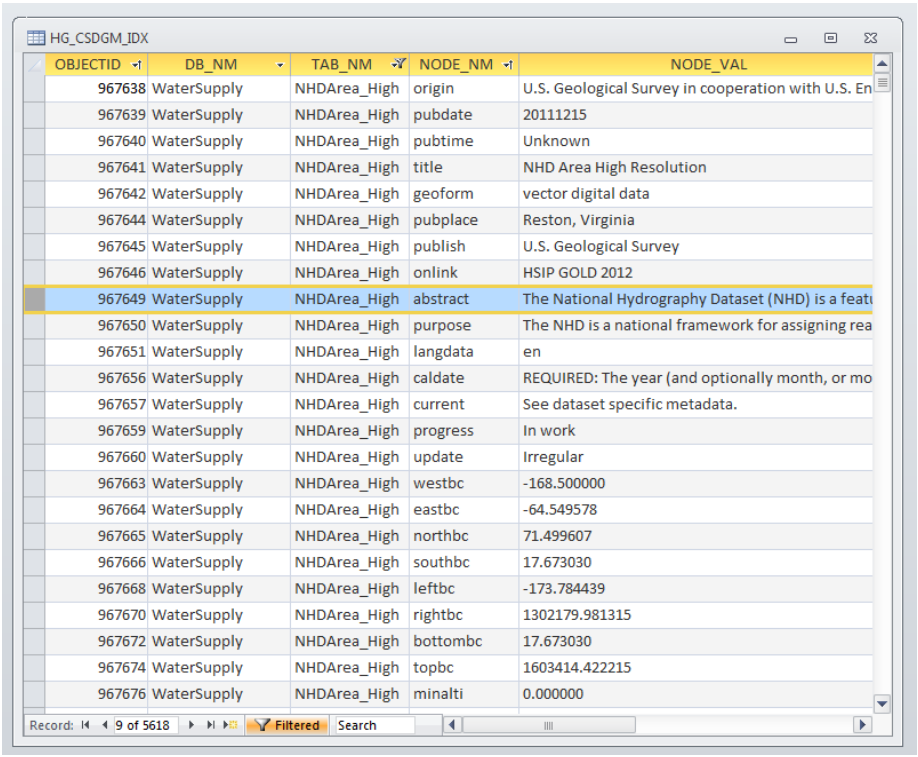
Elements of a Possible Solution

- This is not an either-or suggestion. The installed based of NSDI nodes using XML databases is very important.
- Imagine if, in addition to the XML encodings the standards bodies developed, they **also** developed RDB implementations.
- Imagine if RDB housed GIS metadata was accepted as compliant with a standard, if it was **delivered with a cross reference table indicating how XML content elements corresponded with RDB based content** elements.
- This would necessarily leave out some of the container-nesting constructs. Would this still be considered to mean the same thing?

RDB/XML Cross Reference Table or Link



Publication_Date	12/15/2011 12:00:00 AM
Title	NHD Area High Resolution
Abstract	The National Hydrography Dataset (NHD) is a feature-based database that interconnects and uniquely identifies the stream segments or reaches that make up the nation's surface water drainage system. NHD data was originally developed at 1:100,000-scale and exists at that scale for the whole country. This high-resolution NHD, generally developed at 1:24,000/1:12,000 scale, adds detail to the original 1:100,000-scale NHD. (Data for Alaska, Puerto Rico and the Virgin Islands was developed at high-resolution, not 1:100,000 scale.) Local resolution NHD is being developed where partners and data exist. The NHD contains reach codes for networked features, flow direction, names, and centerline representations for areal water bodies. Reaches are also defined on waterbodies and the approximate shorelines of the Great Lakes, the Atlantic and Pacific Oceans and the Gulf of Mexico. The NHD also incorporates the National Spatial Data Infrastructure framework criteria established by the Federal Geographic Data Committee.
Purpose	The NHD is a national framework for assigning reach addresses to water-related entities, such as industrial discharges, drinking water supplies, fish habitat areas, wild and scenic rivers. Reach addresses establish the locations of these entities relative to one another within the NHD surface water drainage network, much like addresses on streets. Once linked to the NHD by their reach addresses, the upstream/downstream relationships of these water-related entities--and any associated information about them--can be analyzed using software tools ranging from spreadsheets to geographic information systems (GIS). GIS can also be used to combine NHD-based network analysis with other data layers, such as soils, land use and population, to help understand and display their respective effects upon one another. Furthermore, because the NHD provides a nationally consistent framework for addressing and analysis, water-related information linked to reach addresses by one organization (national, state, local) can be shared with other organizations and easily integrated into many different types of applications to the benefit of all.
Current	See dataset specific metadata.
Progress	In work
Update_Frequency	Irregular
Geographic_Coverage	Alaska, Puerto Rico, US Virgin Islands
Keywords	Swamp / Marsh, Artificial Path, Spring / Seep, Canal / Ditch, Hydrography, Stream / River, Reach Code, Lake / Pond, FWHYDROGRAPHY, Reservoir



OBJECTID	DB_NM	TAB_NM	NODE_NM	NODE_VAL
967638	WaterSupply	NHDArea_High	origin	U.S. Geological Survey in cooperation with U.S. En
967639	WaterSupply	NHDArea_High	pubdate	20111215
967640	WaterSupply	NHDArea_High	pubtime	Unknown
967641	WaterSupply	NHDArea_High	title	NHD Area High Resolution
967642	WaterSupply	NHDArea_High	geoform	vector digital data
967644	WaterSupply	NHDArea_High	pubplace	Reston, Virginia
967645	WaterSupply	NHDArea_High	publish	U.S. Geological Survey
967646	WaterSupply	NHDArea_High	onlink	HSIP GOLD 2012
967649	WaterSupply	NHDArea_High	abstract	The National Hydrography Dataset (NHD) is a featu
967650	WaterSupply	NHDArea_High	purpose	The NHD is a national framework for assigning rea
967651	WaterSupply	NHDArea_High	langdata	en
967656	WaterSupply	NHDArea_High	caldate	REQUIRED: The year (and optionally month, or mo
967657	WaterSupply	NHDArea_High	current	See dataset specific metadata.
967659	WaterSupply	NHDArea_High	progress	In work
967660	WaterSupply	NHDArea_High	update	Irregular
967663	WaterSupply	NHDArea_High	westbc	-168.500000
967664	WaterSupply	NHDArea_High	eastbc	-64.549578
967665	WaterSupply	NHDArea_High	northbc	71.499607
967666	WaterSupply	NHDArea_High	southbc	17.673030
967668	WaterSupply	NHDArea_High	leftbc	-173.784439
967670	WaterSupply	NHDArea_High	rightbc	1302179.981315
967672	WaterSupply	NHDArea_High	bottombc	17.673030
967674	WaterSupply	NHDArea_High	topbc	1603414.422215
967676	WaterSupply	NHDArea_High	minalti	0.000000

- RDB Table Column to XML Element Map
- Should the full XML Path and nesting be required?
- Put another way, could RDB housed metadata be accepted as compliant equivalent contents if a valid XML document could be produced, as specified by references like this?
- **How could this proposal be made to standards bodies?**

The Use of Standards

- This is not a proposal not to use the standard(s).
- Rather, the suggestion entails using a different core data management technology, while being able to produce standards compliant output.
- There is an enormous consequence for relying on a standard for building comprehensive, industrial strength IT solutions. A great deal of time is being spent trying to use ISO standards as a basis for geospatial data infrastructures.
- Would a company base it's business model solely on the ISO 9000 series?
- Would the development of a tax accounting and payment system software be based solely on government forms?
- The standards all indicate that the goal is to facilitate information exchange, but look for concrete examples where this is the case.

Recap and Take-Aways

- **How much easier would it be if we used tables and feature classes to manage metadata, the same way we do with other GIS data?**
- **Geospatial Metadata IS Geospatial Data!**
- There would be much more and better metadata.
- Think of how easy it would be if we just had to fill out a set of simple forms, which populated underlying RDB tables.

THANK YOU

Q&A

Geospatial Metadata

Revising the Approach

Brian Hebert
ScribeKey, LLC
www.scribekey.com