Ensuring Data Quality using Topology and Attribute Validation in the Geodatabase

By Heather I. Stanton[1], Stephanie O'Meara[1], James R. Chappell[1],
Anne R. Poole[2], Gregory Mack[3], and Georgia Hybels[4]

[1]Colorado State University/National Park Service Cooperator
1201 Oak Ridge Drive, Suite 200
Fort Collins, CO 80525
Telephone: (970) 267-2159
e-mail: Heather_Stanton@partner.nps.gov,
Stephanie_O'Meara@partner.nps.gov, Jim_Chappell@partner.nps.gov

[2]United States Forest Service – Chippewa National Forest
e-mail: apoole@fs.fed.us

[3]National Park Service – Pacific West Region, Seattle Office
e-mail: Greg_Mack@nps.gov

[4]University of Denver/National Park Service Cooperator
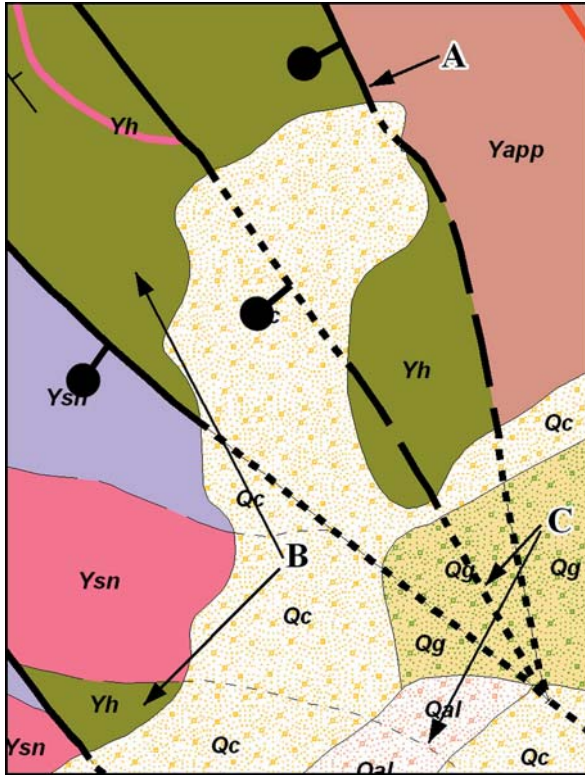e-mail: Georgia_Hybels@partner.nps.gov

INTRODUCTION

Data on paper geologic maps have certain, common-sense rules that govern how the data appear, such as spatial coincidence between certain features, and defined and limited attribution, as depicted in the map's legend (Figure 1). On a geologic map, faults are sometimes coincident with geologic contacts, and geologic contacts are always coincident with the boundaries of geologic units. A geologic unit is accompanied by a unit symbol appearing on the map, and units sharing the same symbol can also be assumed to share the same unit name, major lithology, and age, among other attributes. Features such as geologic contacts and faults have limited and defined lists of positional accuracy (for instance; known, approximate, or concealed).

Although these are common-sense rules, they are nonetheless important to retain and follow, particularly when translating the data on the paper map to a digital format. For instance, a small deviation from exact spatial coincidence between a fault and a geologic contact could have meaning – did the author intend to represent a fault that was 10 meters away from the geologic contact, or should one infer that they are coincident because at the printed map scale they appear to be coincident?

The National Park Service's (NPS) previous data model stored geologic GIS data in ESRI coverage and shapefile formats (O'Meara, et al, 2005a; O'Meara, et al, this volume). Spatial coincidence and attribute validity were ensured through the use of coverage topology (the manner in which geographic data are spatially interrelated), tables of appropriate values (domains) for certain attributes, and data capture methodology (including the use of Arc Macro Language (AML) programs designed to find and/or fix problems). However, over the lifecycle of creating digital geologic data through digitizing, editing, and quality checking (QC), it was difficult to maintain attribute validity and spatial coincidence where it was appropriate. It was not uncommon to find errors in coincidence and attribution after completion of a digital geologic map.

New methods of ensuring the validity of attributes and spatial coincidence, where appropriate, are now available with ESRI's latest software, ArcGIS, and its new format for storing geographic and tabular data, the geodatabase. Geodatabase topology can mimic the rules available with coverages and has additional rules that were previously unavailable, including those that relate data between different geographic layers, stored as feature classes within the geodatabase. Feature classes can also be subdivided using subtypes, or breaks in the feature class based on integer values stored in a field in its associated attribute table. These subtypes can be used to enforce different rules for attribution or topology for different parts of the feature class. In addition, attribution can be controlled by linking domains of acceptable values to selected fields in the feature class and by associating

Figure 1. Excerpt of Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003) illustrating common sense geologic map rules: A) Faults and Geologic Contacts are sometimes coincident; Geologic Contacts are always coincident with Geologic Unit boundaries. B) Geologic Units with the same unit symbol share defined ages, lithologies, etc. C) Faults and Geologic Contacts have limited and defined values for positional accuracy.
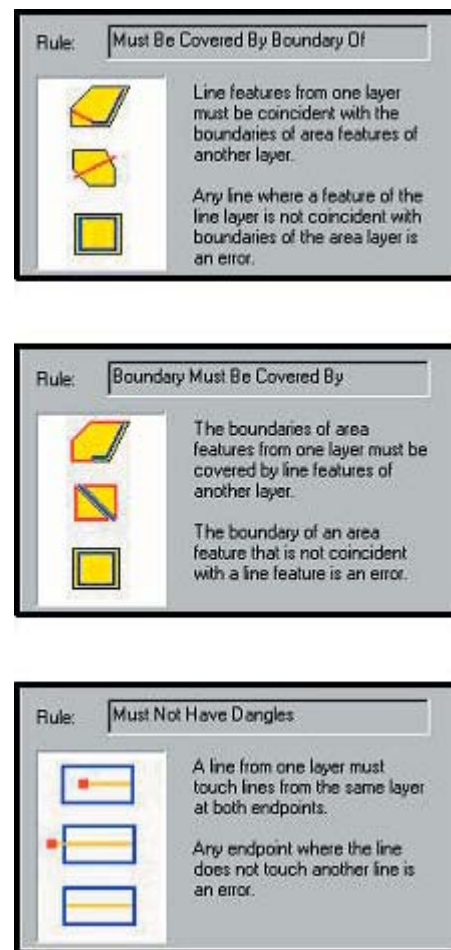
related tables of ancillary information using a key field through relationship classes, thereby reducing duplication of information throughout the geodatabase.

TOPOLOGICAL RULES IN THE GEODATABASE

To demonstrate the usefulness of ArcGIS geodatabase topological rules, we organize the rules into three classes: rules that mimic coverage topology, intra-feature class rules (within a feature class), and inter-feature class rules (between feature classes). All feature classes participating in the topology must be stored in a feature dataset, guaranteeing that they share the same spatial domain. Within the feature dataset, a topology is created where rules governing the interaction within and between feature classes are stored. Rules can then be validated in ArcCatalog or ArcMap. If any feature or features violate the topological rules, an error will be created. Errors can then be corrected using topological editing tools in an ArcMap edit session.

Examples of rules that mimic coverage topology include "Must Be Covered By Boundary Of", "Boundary Must Be Covered By" and "Must Not Have Dangles" (Figure 2). In coverages, polygonal features are stored together with their bounding arcs (lines), and the topology is inherent in the dataset. The attributes of the polygon features are stored in a .pat file (polygon attribute table) and the attributes of the associated arcs are stored in an .aat file (arc attribute table). An arc in a polygon coverage that ends without touching another arc (a dangle) is an error because its associated polygon(s) are not completely bounded by lines.

In the geodatabase, attributed line boundaries of polygons must be stored in a feature class separate from the polygon feature class, requiring that topological rules be created to maintain coincidence between the polygon features and their boundary lines. The "Must Be Covered By Boundary Of" rule ensures that all lines in a line feature class coincide with the boundaries of associated polygons,
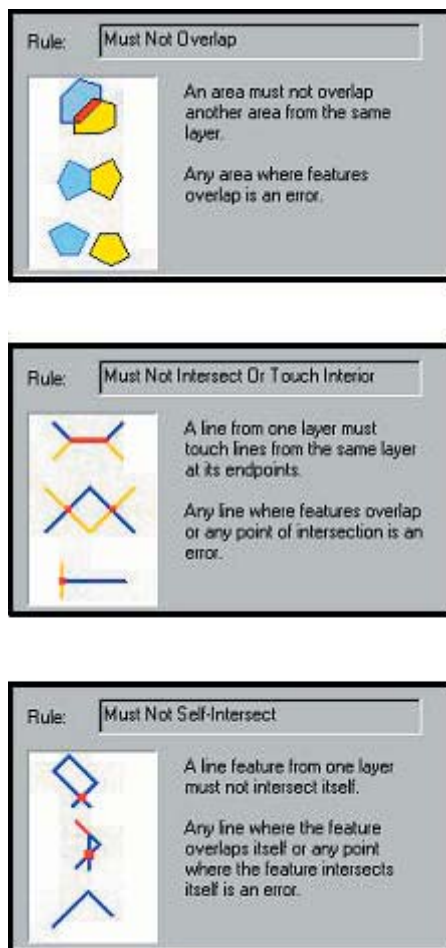

Figure 2. Screenshots from ArcGIS topology properties, illustrating the "Must Be Covered By Boundary Of", "Boundary Must Be Covered By" and "Must Not Have Dangles" rules. These are examples of topological rules that mimic the topology inherent in coverages.

and the "Boundary Must Be Covered By" rule ensures that all polygons have lines that coincide with their boundaries. The "Must Not Have Dangles" rule ensures that lines in the line feature class touch at least one other line at their endpoints, so that their associated polygons are completely bounded by lines. Following these rules is especially important when data are to be exported to coverages.
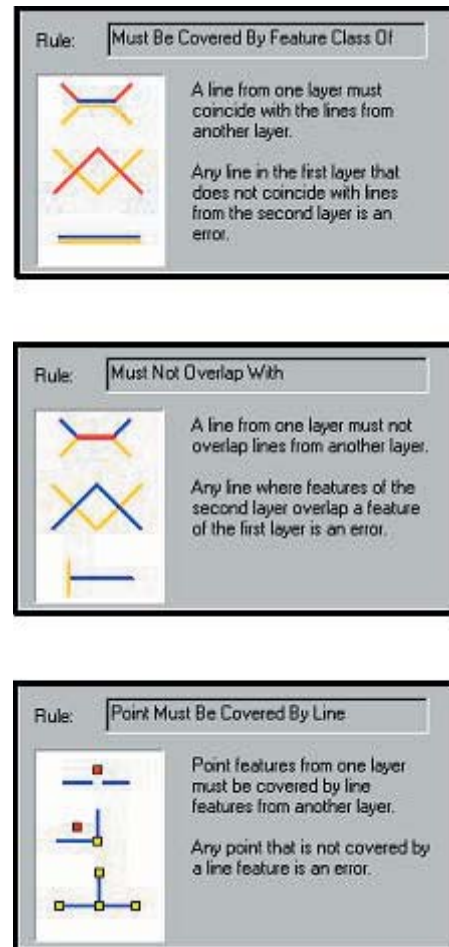
Intra-feature class rules include "Must Not Overlap", "Must Not Have Gaps", "Must Not Intersect or Touch Interior" and "Must Not Self-Intersect" (Figure 3). In order to maintain a complete mosaic of polygons in a polygon feature class where polygons do not overlap each other or have gaps between them, the rules "Must Not Overlap" and "Must Not Have Gaps" are used. "Must Not Intersect Or Touch Interior" ensures that lines do not intersect or touch other lines in the same feature class without the line being broken at the point of intersection, whereas "Must Not Self-Intersect" ensures that a line feature does

not intersect itself. These rules can also be used to ensure that, when exported to coverages, polygons do not violate polygon coverage topology.

Inter-feature class rules demonstrate the real power of geodatabase topology, allowing feature classes within a feature dataset (or geographic grouping of feature classes) to maintain defined spatial relationships, such as coincidence (Figure 4). "Must Be Covered By Feature Class Of" is a rule that maintains coincidence between two feature classes, whereas "Must Not Overlap With" ensures that features from two different feature classes that should not be coincident are not. "Point Must Be Covered By Line" is used to ensure that point features from one feature class are coincident with lines in another feature class. While coincidence can be checked using other methods when data is stored in coverages and shapefiles, it cannot be enforced. The automated methods of topological validation and the tools provided by ArcMap to edit topological

Figure 3. Screenshots from ArcGIS topology properties, illustrating the "Must Not Overlap", "Must Not Intersect Or Touch Interior", and "Must Not Self-Intersect" rules. These are examples of intra-feature class topological rules, applicable to features within a single feature class.

Figure 4. Screenshots from ArcGIS topology properties, illustrating the "Must Be Covered By Feature Class Of", "Must Not Overlap With" and "Point Must Be Covered By Line" rules. These are examples of inter-feature class topological rules, making it possible to ensure spatial relationships between feature classes.

errors ensure that errors can be found and fixed without requiring user-created scripts and tools.

TOPOLOGICAL RULES IN THE NPS GEOLOGY-GIS GEODATABASE DATA MODEL

In the NPS Geology-GIS Geodatabase Data Model (O'Meara, et al, 2005b, O'Meara et al, this volume) topological rules are used to enforce the common-sense spatial coincidence rules that apply to data on geologic maps, as well as to enforce attribution in certain feature classes. A summary of the rules present in the data model can be found in Figure 5. Rules can be grouped based on:

1. whether a line feature class is associated with a polygon feature class (geologic contacts and geologic units; rules C, D, and H),
2. line feature classes overlap with other line feature classes (geologic contacts and faults; rules A and B),
3. line feature classes are spatially unrelated to other line and polygon feature classes (folds; rule G),
4. polygon feature classes are allowed to have gaps (surficial units, or recent deposits covering bedrock geologic units; rule G), or

5. not allowed to have gaps (bedrock geologic units, rule F).

This approach allows for the application of a set of topological rules to a feature class based on spatial/geometric relationships to other feature classes, rather than being specific to a single feature class. For instance, fold axes, structural contours, and glacial feature lines (such as moraine crests) have the same topological rules applied to them because they are spatially unrelated to other geologic features on the map; while surficial units, deformation areas, and dike swarm areas can have the same rules applied to them because they all have line feature classes that are coincident with their boundaries and they are allowed to have gaps between their individual polygons. Point feature classes, such as attitude measurements or symbology, have topological rules associated with them only if they are associated with line feature classes such as faults or folds.

Topological rules in the NPS Geology-GIS Geodatabase Data Model are often used in association with subtypes. Subtypes subdivide elements in a feature class into groups so that different rules for attribution or topology can be applied to those groups. For instance, faults in the data model are subdivided into either a "Fault" subtype or a "Fault/Contact" subtype. The "Fault" subtype refers to lines in the faults feature class that are not coincident

	Geologic Contacts	Faults	Surficial Contacts	Linear Dikes	Folds	Geologic Units	Surficial Units
Geologic Contacts	H	1-A-1; 0-B-0	-	-	-	C	-
Faults	1-A-1	G	-	-	-	-	-
Surficial Contacts	0-B, 2-A-0, 3-A-1	0-B, 1-A-0	H	-	-	-	C
Linear Dikes	0-B, 2-A-0, 3-A-1	0-B, 1-A-0	-	G	-	-	-
Folds	-	-	-	-	G	-	-
Geologic Units	D	-	-	-	-	F	-
Surficial Units	-	-	D	-	-	-	F

Rules between feature classes that are specific to the geodatabase

Rules that apply only within a single feature class; mimic coverage rules

Rules between feature classes that mimic coverage rules

A — Must Be Covered By Feature Class Of

B — Must Not Overlap With

C — Must Be Covered By Boundary Of

D — Boundary Must Be Covered By

E — General Polygon Rules: Must Not Overlap

F — General Polygon Rules 2: Must Not Overlap, Must Not Have Gaps

G — General Line Rules: Must Not Intersect Or Touch Interior, Must Be Single Part, Must Not Self-Intersect

H — General Line Rules 2: General Line Rules Plus Must Not Have Dangles

Line Subtypes:
Faults - 0: Fault, 1: Fault/Contact
Geologic Contacts - 0: Contact, 1: Contact/Fault
Others - 0: Self, 1: Self and Fault, 2: Self and Geologic Contact, 3: Self and Fault and Geologic Contact

Figure 5. Topological rules in the NPS Geology-GIS Geodatabase Data Model are represented in the table with the source or origin feature classes in rows and the destination feature classes in columns (a similar figure is present in O'Meara, et al, this volume). For instance, the Faults feature class (source) is related to the Geologic Contacts feature class (destination) using the rule 1-A-1, where the first "1" is the Fault/Contact subtype (see Line Subtypes below table) of the Faults, the "A" represents the topological rule (see list on right), and the second "1" is the Contact/Fault subtype of the Geologic Contacts. In other words, the Fault/Contact subtype "Must Be Covered By the Feature Class Of" or must coincide with the Contact/Fault subtype. If there is no number to represent a subtype for the source and/or destination feature class, the rule applies to the entire source and/or destination feature class regardless of subtype. Note that one or more rules may apply to a given source and destination feature class.

with geologic contacts, whereas the "Fault/Contact" sub-type refers to the lines that are coincident with geologic contacts. The geologic contacts feature class is similarly subdivided into "Contact" and "Contact/Fault" subtypes.

Subtypes, in association with topological rules, can help to identify errors in attribution in the feature classes to which they are applied. For instance, a line in the faults feature class may be incorrectly coded as a "Fault/Contact" subtype because there is no associated coincident geologic contact (Figure 6). When the data model topological rules are validated, this line will violate the "Must Be Covered By Feature Class Of" rule because it does not coincide with a "Contact/Fault" line in the geologic contacts. The resulting topological error would alert the user that a change in attribution (coding the line as a "Fault") is needed.
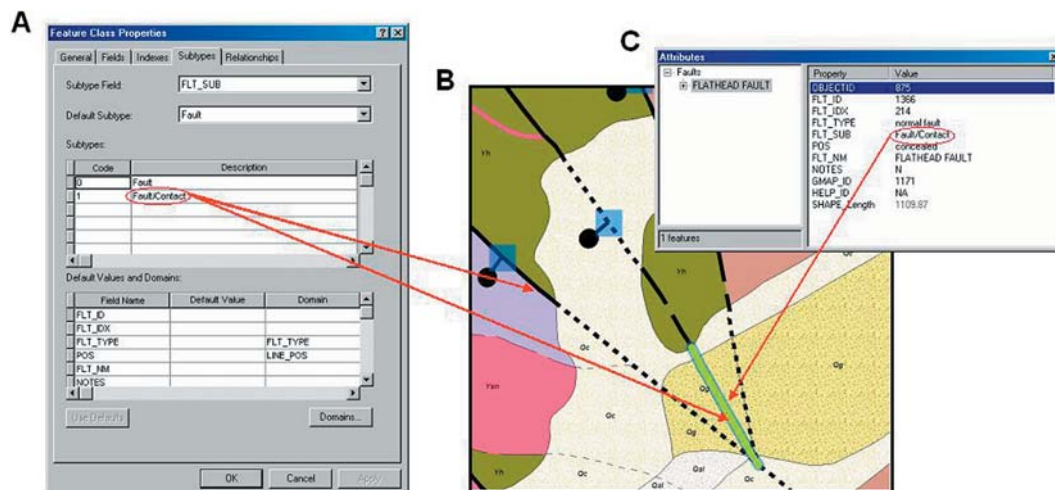
Topological rules can also be used with subtypes to identify features within one feature class that should be coincident with features in another feature class. For instance, at a glance, a fault and a geologic contact may appear to be coincident at map scale, but at larger scales it may become apparent that they are not (Figure 7). In the realm of geologic GIS data, these discrepancies are crucial to identify, and the errors generated by the validation of topological rules help to locate the discrepancies and fix them. In the NPS Geology-GIS Geodatabase Data Model, lines coded as "Contact/Fault" in the geologic contacts feature class and lines coded as "Fault/Contact" in the faults feature class that are not coincident with each other will be flagged as errors because they violate the "Must Be Covered By Feature Class Of" rule.

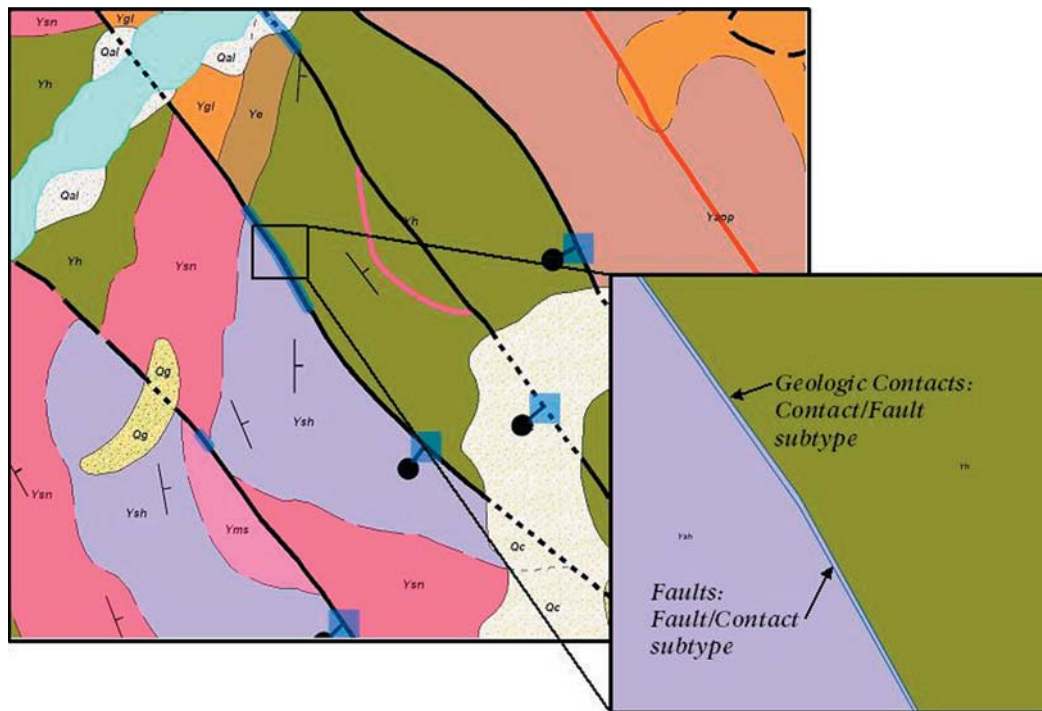Another advantage of setting up a topology in the

geodatabase is that coincident features from different feature classes can be moved as one entity using the Topology Edit tool in ArcMap (Figure 8). An example of this is editing the location of a fault and its associated point symbology that is also coincident with a geologic unit boundary and a geologic contact. With coverages and shapefiles, or feature classes not participating in a topology, each of the coincident features would have to be moved separately. Even with the snapping function set, there is a risk of losing coincidence because each individual vertex in a line has to be moved separately, and it is difficult to identify areas that are no longer coincident since often the lack of coincidence is not visible without examining lines in great detail.

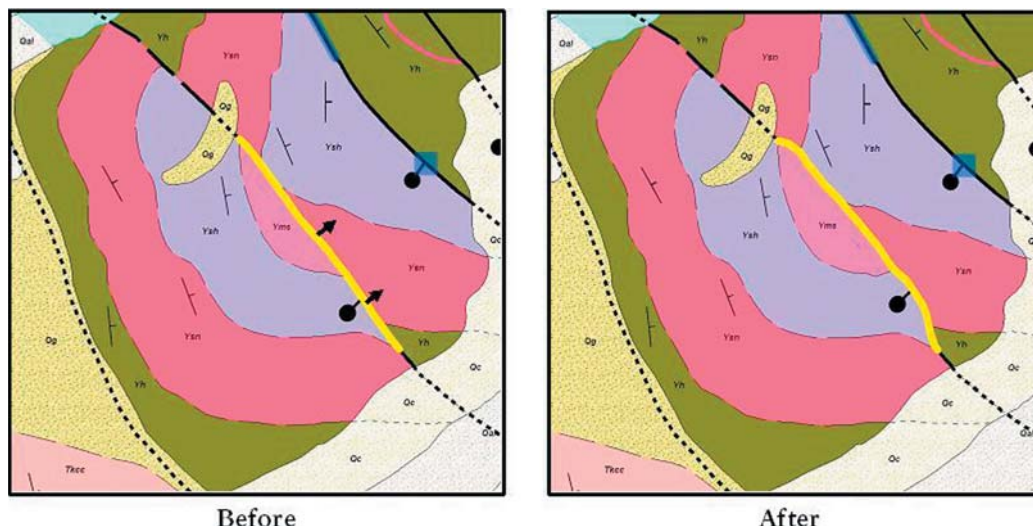## ATTRIBUTE VALIDATION IN THE NPS GEOLOGY-GIS GEODATABASE DATA MODEL

In the NPS Geology-GIS Geodatabase Data Model, the consistency and quality of attribution are not just controlled by topological rules, but also in two other ways. Domains (lists of acceptable values) and default values are used along with subtypes to define attribution for various fields (including positional accuracy, fault types, and attitude measurement types) in the feature classes in the data model. Also, relationship classes between feature classes and tables are used to avoid duplication of data and to limit attribution for certain fields (for instance, a geologic unit information table related to the geologic units feature class).

Figure 6. Excerpt of the Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003). A) Feature class properties for the faults feature class, illustrating that "Fault" and "Fault/Contact" subtypes are stored in the FLT_SUB field. B) The fault highlighted on this map has violated the "Must Be Covered By Feature Class Of" rule because it is coded as a "Fault/Contact" and is not coincident with a geologic contact. C) The line should be coded as a "Fault" in the FLT_SUB field.

Figure 7. Excerpt of the Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003). Both the fault and geologic contact present in this example have violated the "Must Be Covered By Feature Class Of" rule because, although they are coded correctly in their subtype fields, they are not spatially coincident. The problem can be fixed by editing the fault's vertices, snapping them to the geologic contact.

Figure 8. Excerpt of the Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003). Without topology, in order to move the highlighted shared edge, each feature class (faults, geologic unit boundaries, geologic contacts and fault symbols) would have to be moved separately, risking lost spatial coincidence. Using the Topology Edit Tool, coincident features can be moved together.

Domains and default values are useful in a number of ways. In the NPS Geology-GIS Geodatabase Model, coded value domains contain an alias for each domain member, for instance, a positional accuracy of 1 is aliased to "known." A change in subtype in a given feature class changes the domains and default values that are accessed. For instance, when the subtype is changed during attribution to "Planar Measurements – Vertical" in the attitude measurements feature class, a domain restricted to vertical attitude types is accessed, the dip field is assigned a default value of 90 and a domain that has 90 (Vertical) as its only member is accessed. The default value provides automated attribution, and if a value other than 90 is later entered into the dip field, it will be flagged as an error during attribute validation because it is outside the domain for measurements with vertical dip.
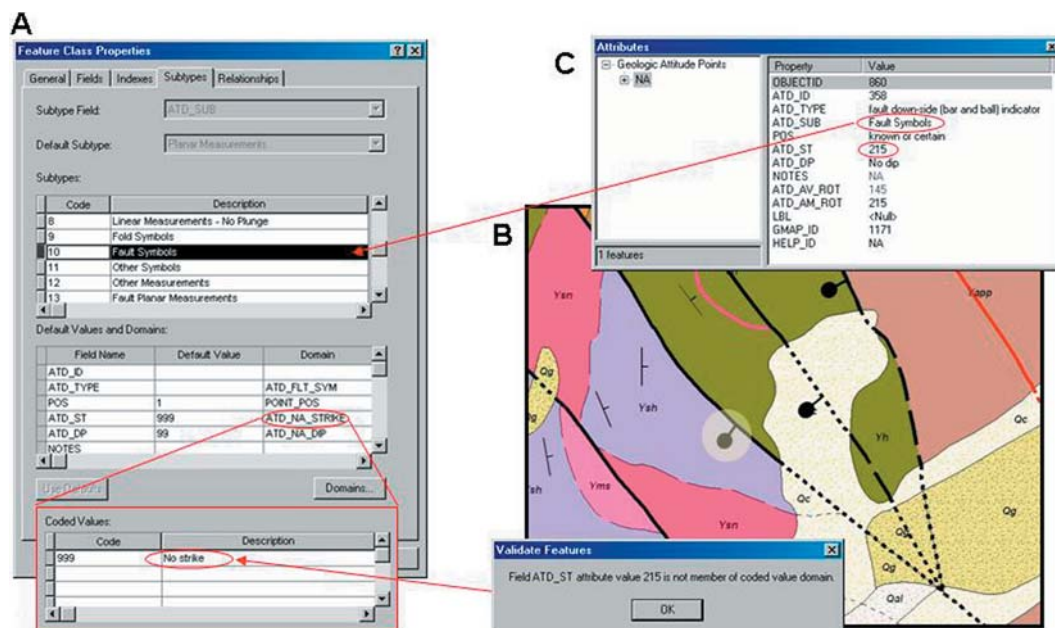
For third-party users of the data, domains provide a built-in data dictionary in the geodatabase that is easily understood. Domains also provide ease of attribution by allowing the user to pick a description (strike and dip of beds) rather than a number referenced to an outside list of acceptable values. Default values increase this ease by automatically placing a commonly-used value in a field of the user's choice when a feature is created. During QC, domains are useful in yet another way. Using the built-in functionality of attribute validation in ArcGIS, values stored in a feature class's fields are compared to the domains assigned to those fields. If a value lies outside a domain because it was somehow misattributed, an error is generated and the feature(s) in error are identified (Figure 9) and can then be corrected.
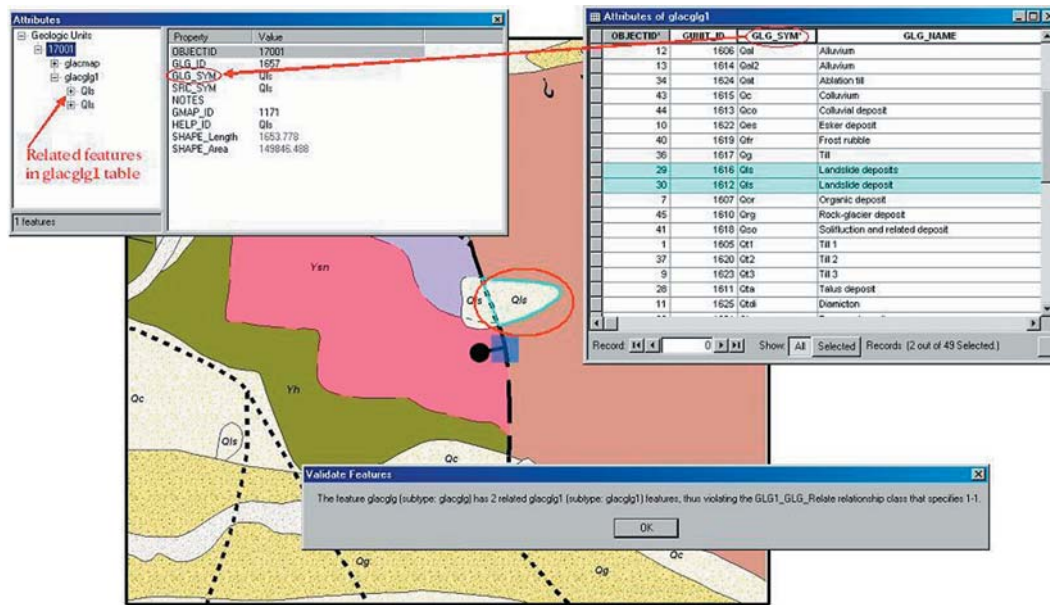
In the NPS Geology-GIS Geodatabase Data Model, a geologic units table is used to store the name, age, description, and other information related to a given geologic unit. The unit information table is related to the geologic units feature class based on the geologic unit symbol using a relationship class. This avoids repeating multiple fields with the same information for every unit with the same map symbol in the feature class. During the initial attribution of geologic unit polygons, the table can be used to select which unit to apply to a series of polygons if desired. During attribute validation, the polygons in the geologic units feature class can be checked against the table for errors. If the geologic unit does not exist in the table, or if there are multiple entries in the table for a single unit on the map, an error will be generated, alerting the user to change attribution in the geologic units feature class or to edit the geologic units table (Figure 10).

CONCLUSION

The coverage/shapefile-based data model that was used prior to the development of the NPS Geology-GIS Geodatabase Data Model had a number of methods for ensuring attribute validity and coincidence. However, it was difficult to maintain these checks over the life cycle

Figure 9. Excerpt of the Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003). A) Feature class properties for bedding attitudes. Attribution for the "Fault Symbols" subtype is controlled by a number of domains, as well as default values. The value for ATD_ST (strike) for "Fault Symbols" should always be 999 or 'No strike'. B) When 'Validate Features' is carried out, the highlighted symbol, which is coded as a "Fault Symbols" subtype, is found to be in error because C) the ATD_ST (strike) value, 215, is not a member of the associated domain.

Figure 10. Excerpt of the Digital Geologic Map of Glacier National Park and Vicinity (O'Meara, et al, 2003). When 'Validate Features' is carried out on the circled geologic unit, an error occurs because there are two entries for the same GLG_SYM (unit symbol) in the related table. This violates the one-to-many simple relationship class set up between the geologic units and geologic unit information (glacglg1) table using GLG_SYM as the key field.

of a digital geologic map project. Geodatabases can not only reproduce coverage topology and programmatic methods for ensuring attribute validity, but they also have features that provide further assurances of spatial and attribute integrity. Attribute quality is improved through the use of subtypes in combination with domains and relationship classes that access records from related tables during feature capture. Spatial coincidence can also be ensured where appropriate, not only with rules mimicking coverage topology, but with added rules that interrelate different feature classes and the ability to apply rules to parts of feature classes using subtypes. Finally, as an added benefit, ArcGIS's automation for attribution and attribute and topology validation increase the ease and speed of attribution and QC for digital geologic maps.

In the future, we will continue to refine topological and attribution rules for the NPS Geology-GIS Geodatabase Data Model. We also will be creating programmatic methods for defining, and fixing, common attribution and topological errors. Finally, we will develop methods for updating and quality-checking our legacy data when migrating to the latest data model.

REFERENCES

O'Meara, S.A., Gregson, J., Poole, A.R., Mack, G., Stanton, H.I., and Chappell, J., 2005a, National Park Service Geologic Resources Evaluation Geology-GIS Coverage/Shapefile Data Model, available at http://science.nature.nps.gov/im/inventory/geology/GeologyGIS DataModel.htm.

O'Meara, S.A., Stanton, H.I., Chappell, J., Mack, G., Poole, A.R., and Hybels, G., 2005b, National Park Service Geologic Resources Evaluation Geology-GIS Geodatabase Data Model (Draft v. 1.2), available at http://science.nature.nps.gov/im/inventory/geology/GeologyGISDataModel.htm.

O'Meara, S.A., deWolfe, V., Johnson, R., and Schaeffer, M., 2003, Digital Geologic Map of Glacier National Park and Vicinity, Montana, digitized from Whipple (1992) and Carrara (1990): National Park Service Geologic Resources Evaluation (GRE) program, unpublished, 1:100,000 scale [digitized from Whipple, J.W., 1992, Geologic Map of Glacier National Park, Montana, U.S. Geological Survey Map I-1508-F, 1:100,000 scale; and Carrara, P.E., 1990, Surficial Geologic Map of Glacier National Park, Montana, U.S. Geological Survey Map I-1508-D, 1:100,000 scale].

SOFTWARE REFERENCES

ArcGIS 8.3 and 9.0 (ArcCatalog and ArcMap) – Environmental Systems Research Institute (ESRI) Inc., 380 New York St., Redlands, CA92373, http://www.esri.com.

Geodatabase Diagrammer – Developed by Michael Zeiler, Environmental Systems Research Institute (ESRI) Inc., 380 New York St., Redlands, CA 92373, see http://arcscripts.esri.com.

Geodatabase Designer v2 – Developed by Richie Carmichael, Environmental Systems Research Institute (ESRI) Inc., 380 New York St., Redlands, CA 92373, see http://arcscripts.esri.com.

Microsoft Office Visio Professional 2003, Microsoft Corporation, http://www.microsoft.com/.