Geospatial Preservation: State of the Landscape

A Quick Overview

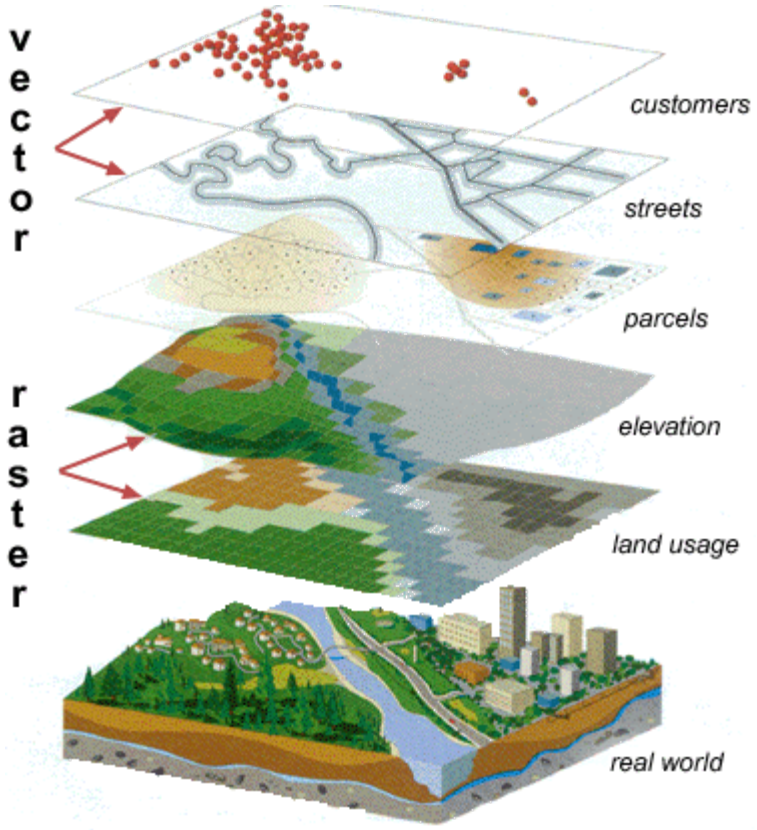
Steve Morris
NCSU Libraries

SAA 2011
August 26, 2011

Brief Overview of the Problem

Variety of Geospatial Data Types

- Variety of producers
  - Government (federal/state/local)
  - Commercial
  - Academic
  - Not-for-profit
- Variety of products
  - Datasets
  - Data projects
  - Data packages
  - Data representations (e.g. cartographic)



Some Domain-Specific Challenges

- Complex and proprietary data formats
  - No widely-supported *open* format for vector data
- Increasing use of spatial databases for data management
  - Complex: whole is greater than the sum of the parts
- Temporal versions of content
  - How often to capture?
- “Data” versus “data representations”
  - What is the record?
- More reliance on web services- or API-based access
  - Data and documents becoming more ephemeral

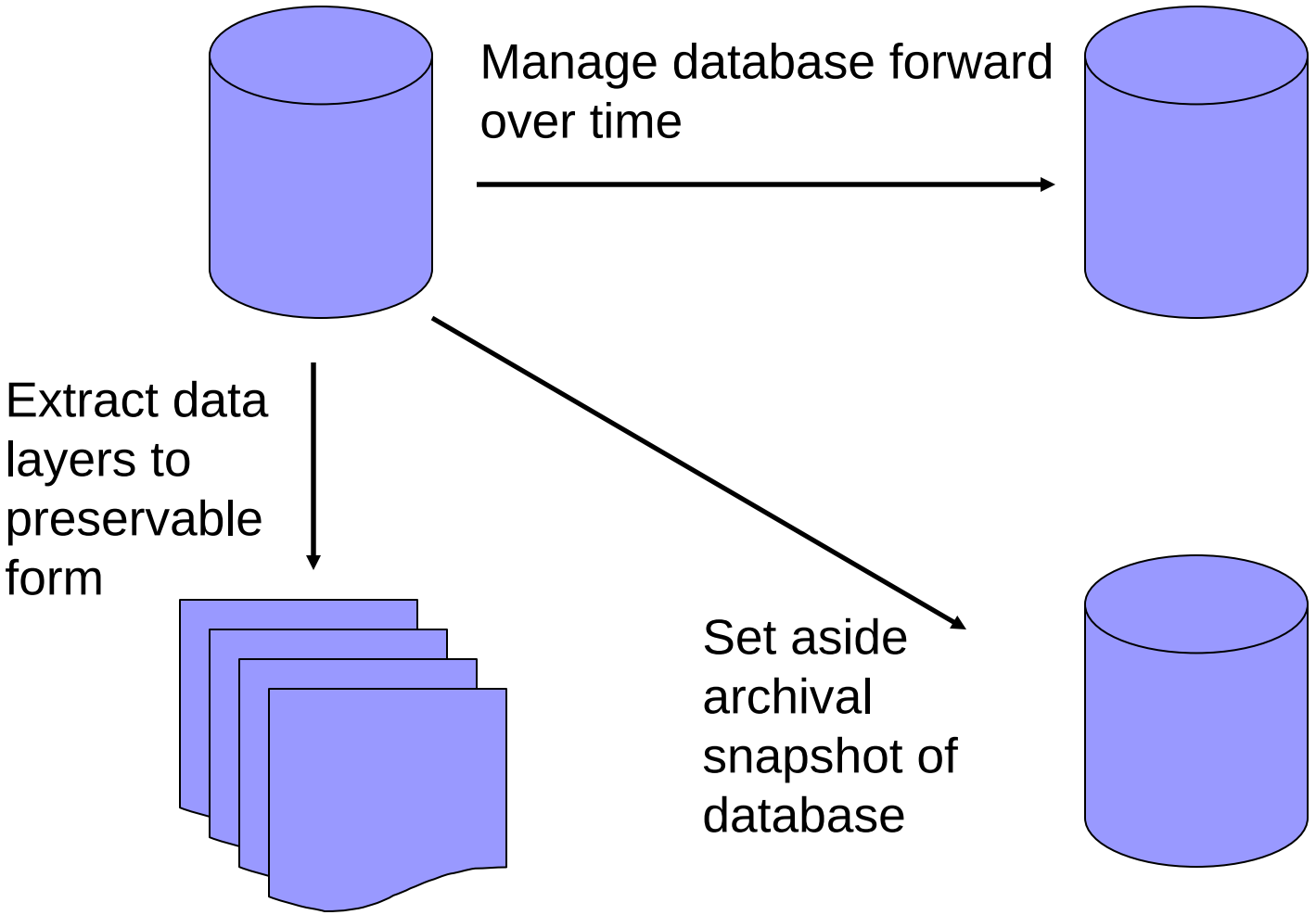
Format Strategy: One Size Does Not Fit All

- What format to archive?
  - Shapefile?
  - Geodatabase?
  - GeoPDF?

GeoMAPP State Approaches to Archival Transfer

- **Kentucky**: Transferring Geodatabases
- **North Carolina**: Transferring Shapefiles
- **Utah**: Transferring Shapefiles and Geodatabases *and* creating geospatial PDF documents

Spatial Databases Archiving Approaches



Versioned Data: How Often to Capture

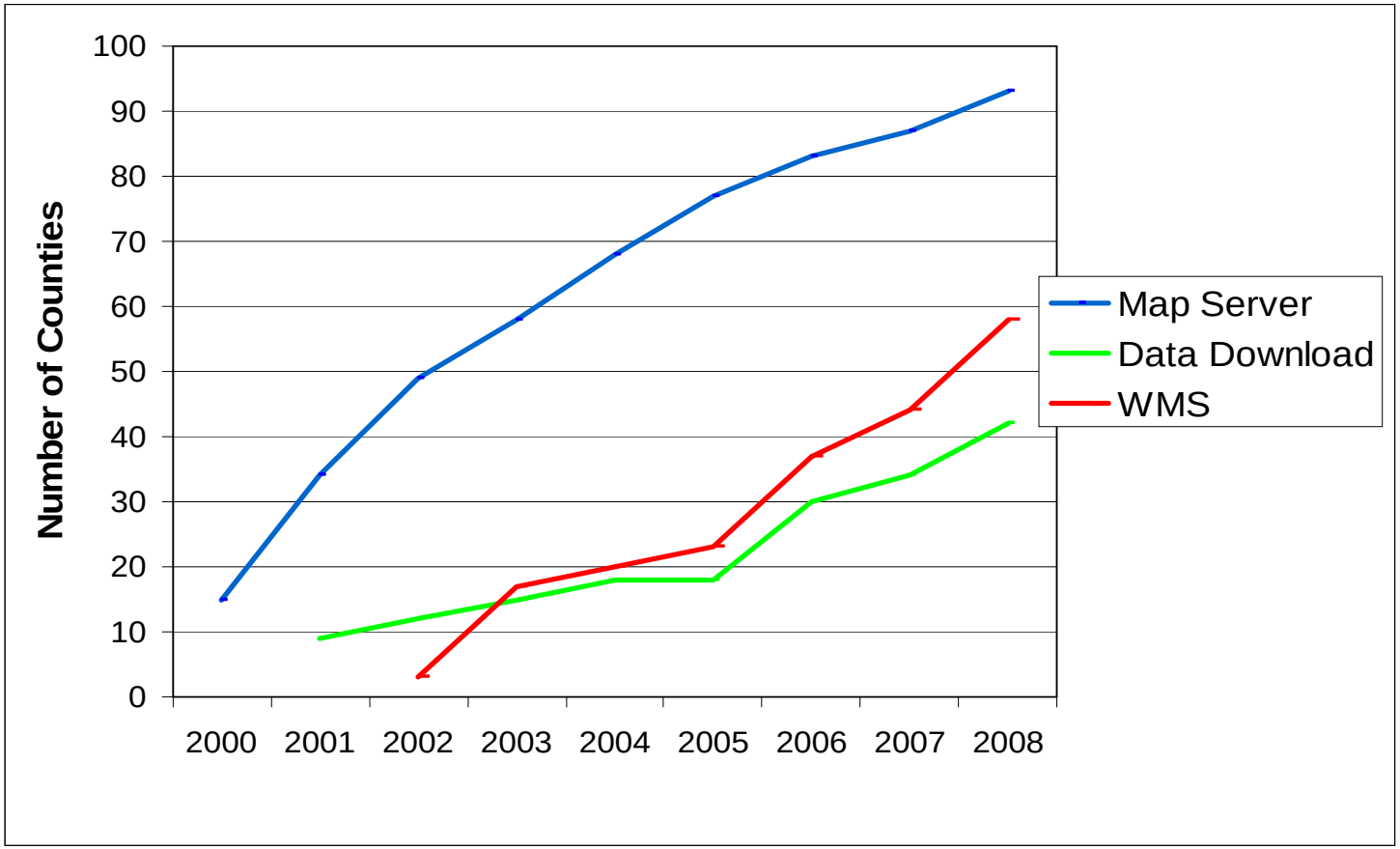
- Not all datasets will be captured with same rate
  - Rapid data change: land parcels, streets
  - Moderate data change: zoning, jurisdictional boundaries
- Consider frequency that the data changes
  - Consider regularly scheduled captures for frequently changing data (e.g. once/quarter)
  - For “stable” datasets capture once/year (or what makes sense)

GeoMAPP State Approaches

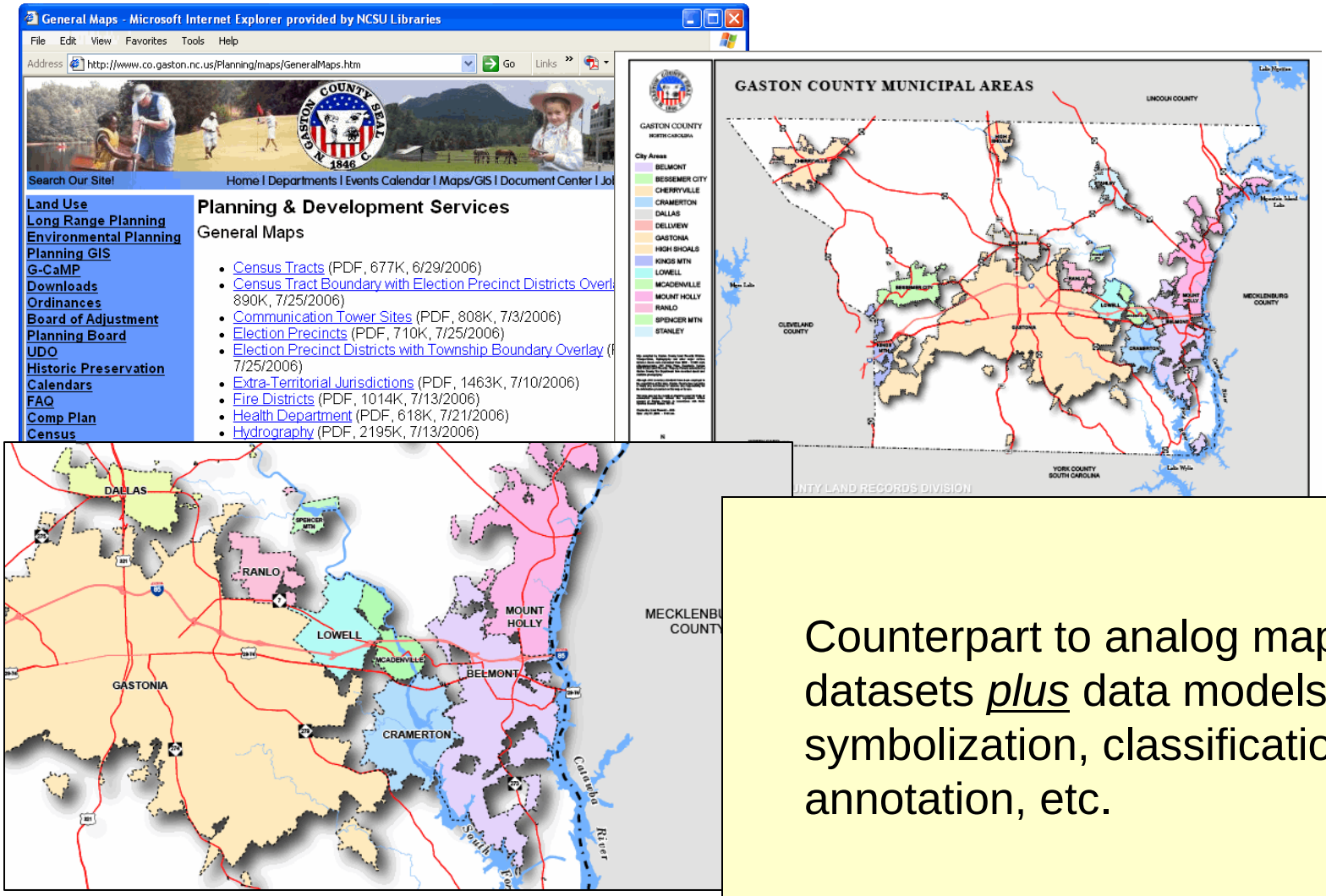
- Kentucky**:** Captures all KYGEONET datasets quarterly in geodatabase snapshot
- North Carolina**:** Captures older datasets as superseded
- Utah**:** Frequency depends on record series. Most annual snapshot, except parcels quarterly

Increasing Reliance on Geospatial Web Services

GIS Data Services Among the 100 Counties in North Carolina



What is the Record? Data vs. Representation



Counterpart to analog map = datasets *plus* data models, symbolization, classification, annotation, etc.

Original Data vs. Desiccated Data

Complex data representations can be made more preservable (and less useful) through simplification

Some Organizational and Cultural Challenges

- Industry focus on “latest and greatest” data
- “Kill and fill” as a common approach to data management (past versions of vector data lost)

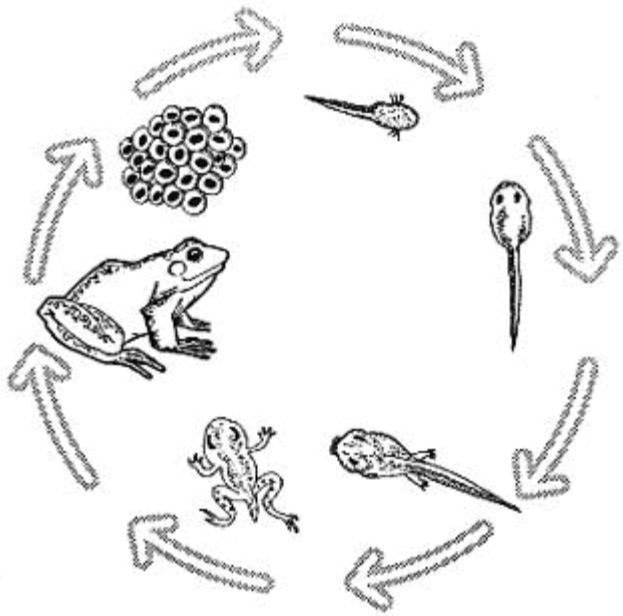
Also loss of *memory* about the data

- Older data not made accessible
- Older data not reported in inventories
- Older data not available through web services
- Data inventories not saved

What to Do?

The Geoarchiving Process Lifecycle

1. Establishing key relationships
2. Inventory
3. Appraise
4. Data Preparation
5. Transfer
6. Ingest
7. Preservation
8. Access
9. Business planning for sustainability

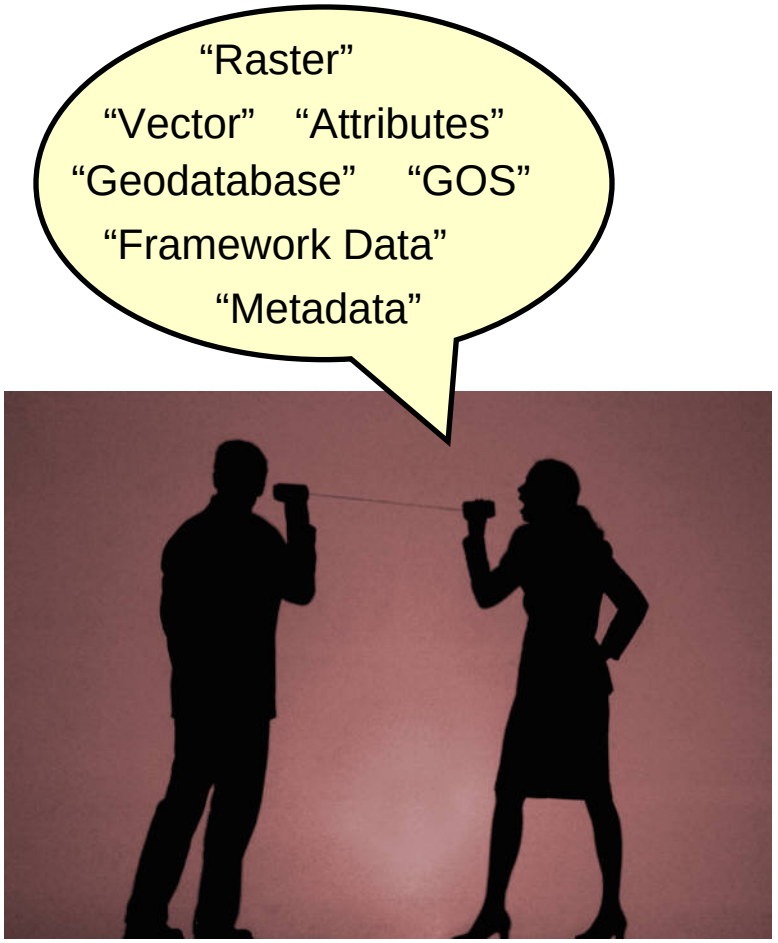


From: Geospatial Multistate Archive and Preservation Partnership (GeoMAPP)
http://www.geomapp.com

Leverage Geospatial Data Infrastructure

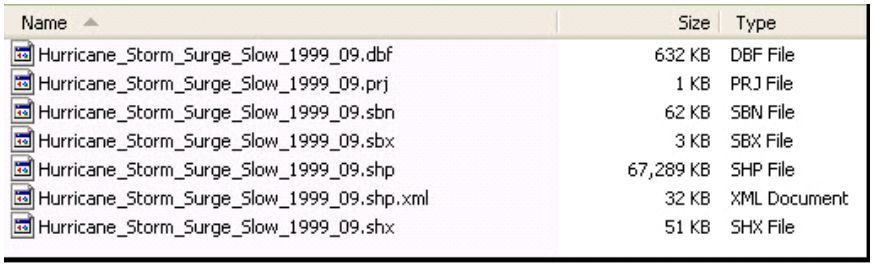
- Data inventories support content identification
- Metadata standards and best practices support discoverability and use
- Content standards support data interoperability over time and help eliminate semantic confusion
- Data exchange networks:
  - Minimize need to make contact
  - Add technical, administrative, descriptive metadata
  - Establish rights and provenance

First: Get the Conversation Started



Making Data Preservation-Ready

- File naming
  - Descriptive title
- Attributes
  - Logical name
  - Explanation in metadata record
- Metadata
  - Ideally standards compliant (FGDC CSDGM, ISO 19115/19139)
  - Important fields: Title, Abstract, Publication Date, Contact Info, Process steps, Attributes description
- Format awareness
  - Versioning and file type



Name	Size	Type
Hurricane_Storm_Surge_Slow_1999_09.dbf	632 KB	DBF File
Hurricane_Storm_Surge_Slow_1999_09.prj	1 KB	PRJ File
Hurricane_Storm_Surge_Slow_1999_09.sbn	62 KB	SBN File
Hurricane_Storm_Surge_Slow_1999_09.sbx	3 KB	SBX File
Hurricane_Storm_Surge_Slow_1999_09.shp	67,289 KB	SHP File
Hurricane_Storm_Surge_Slow_1999_09.shp.xml	32 KB	XML Document
Hurricane_Storm_Surge_Slow_1999_09.shx	51 KB	SHX File

Creative Approaches to Record Disposition

- Archives may be challenged by:
  - Extremely large size of data collections
  - Complexity of some data and associated technology

Example Approach: North Carolina

DISPOSITION INSTRUCTION:

GIS dataset: Permanent. Create a snapshot of dataset annually.

Either:

Transfer snapshot to NCOneMap according to established procedures, complying with standards and procedures adopted by the ***North Carolina Geographic Information Coordinating Council. (See Geospatial Records, page __)***

Or,

If retained in office permanently, your agency must comply with standards (for metadata, file naming, data sharing, and long term preservation) and procedures adopted by the ***North Carolina Geographic Information Coordinating Council. (See Geospatial Records, page __)***

Making the Business Case

1993

1998

1999

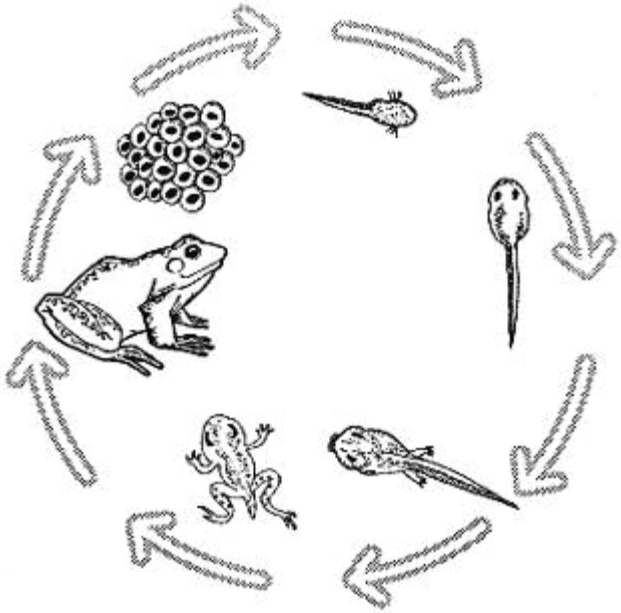
2002

2005

Use case:
Land use and impervious surface change analysis

The Geoarchiving Process Lifecycle

1. Establishing key relationships
2. Inventory
3. Appraise
4. Data Preparation
5. Transfer
6. Ingest
7. Preservation
8. Access
9. Business planning for sustainability



From: Geospatial Multistate Archive and Preservation Partnership (GeoMAPP)
http://www.geomapp.com

Engaging Industry in the Challenge

- Current data sharing needs drive infrastructure improvements that help archiving
- For vendors: Turn temporal data management and data archiving into a customer problem
- “Addressing the needs of temporal data analysis” vs. “data preservation”
- Borrow from state and local archiving innovations

Recent post to NC Local Government GIS Listserv

We are receiving request for GIS data and maps for layers (zoning, parcels, etc...) to reflect a specific instance in time.
Ex. Request for a map that shows what the zoning layer looked like January 1st 2005 vs. August 23rd of 2007.

How are you archiving past data (in-house scripts, 3rd party extension, ESRI core software)?

What's your retention schedule?

Each day we have scripts that copy the Enterprise GDB to a PGDB and maintain three years of these copies (several terabytes) so we can create maps from an instance in time.

Thanks,
xxxxxxxxxxxxxxxxxx
GIS Coordinator
City of xxxxxxxxxxxxxxxxxx

Thank You!

Contact:

Steve Morris
Head, Digital Library Initiatives
North Carolina State University Libraries
Steven_Morris @ncsu.edu

GeoMAPP
http://www.geomapp.com

NCGDAP
http://www.lib.ncsu.edu/ncgdap