

# **Creating a Culture of Data Integration and Interoperability: Librarians Collaborate on a Geoinformatics Course**

**April 2008**

*Michael Fosmire and Chris Miller*

*Purdue University Libraries*

[fosmire@purdue.edu](mailto:fosmire@purdue.edu) [ccmiller@purdue.edu](mailto:ccmiller@purdue.edu)

# Motivation

- Why teach a course?
  - E-science requires data-savvy researchers, but instruction to create those researchers is not well developed
  - “advanced computing is no longer restricted to a few research groups in a few fields... but pervades scientific and engineering research.”

–Atkins, *NSF Blue Ribbon Report* (2003)
  - Cyberinfrastructure will “both ***demand and support*** a new level of technical competence in the science and engineering workforce...”

# Motivation

- Why Earth Science/GIS?
  - Data heavy research areas
  - GIS seen as a highly valued skill
  - Several distributed data repositories in Earth and Atmospheric sciences
  - Developing interest groups in Geoinformatics in American Geophysical Union, Geological Society of America
  - EAS Department very supportive that they understand need, not sure how to teach

# So, what is geoinformatics?

- Loosely, the use of information and communication technologies in the support of a discipline can be called X-informatics
- Major ambiguity within disciplines is:
  - Geospatial-informatics or
  - Geologic-informatics...although much overlap between the two

# Course Details

- EAS 591G: Geoinformatics
  - 3 credits
  - 2 hours lecture/week
  - 3 hour lab/week
  - Lectures provide background for informatics concepts, which are explored in lab sessions
  - Initial offering Spring 2008: 12 enrollees
  - Mainly graduate students, a few advanced undergraduates
  - Pre-assessment queried skills of enrolled students

# Semester Project

- Students asked to integrate concepts they learned into a semester research project (often research they were already doing)
- Project must include
  - Data student has acquired locally
  - Data gathered from external sources
  - Analysis/Integration of above data
  - Visualization that adds value to data
  - Submission of data to a portal

# Motivating Theme...

- In order to provide context for learning concepts in Geoinformatics, we developed a scenario
  - There has been a benzene spill on campus. You need to figure out where it occurred so you can keep it from making Purdue uninhabitable.
    - Of course it takes six weeks to solve the problem, as we add layers of concepts and skills that are necessary for a resolution

# In order to solve the spill problem...

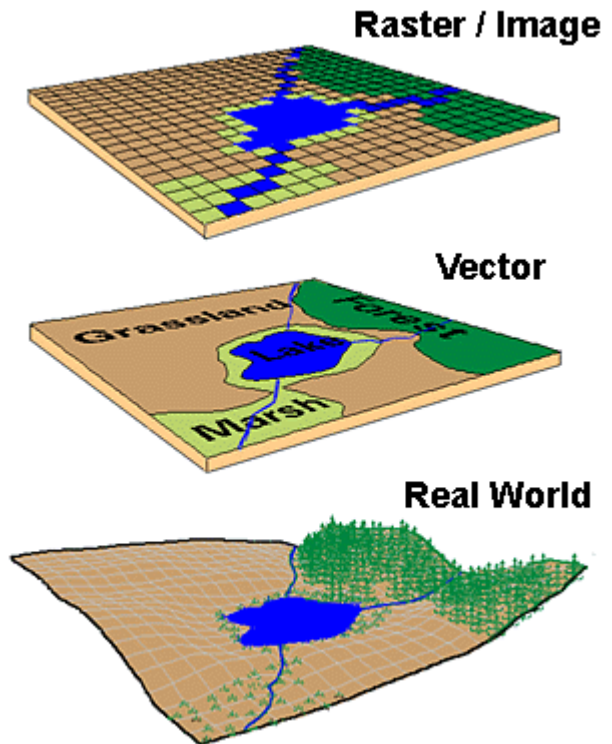
- Find well data to determine groundwater flow directions (creating own databases)
- Gather external GIS information (topography, Purdue and neighborhood shape files)
- Use ArcGIS tools to interpolate data as well as run groundwater flow simulations
- Use GPS units to collect 'spill data' in situ
- Use data management techniques (regression, statistics) to fit data to a theoretical curve and determine locus of spill
- Add layers to determine physically where spill occurred
  - Mixture of science, GIS, and data management to solve a 'real' problem

# What Did We Hope to Accomplish?

- Course Goals
  - Data gathering
  - Data analysis
  - Data management
  - Data visualization
  - Data curation and preservation

# Data Gathering

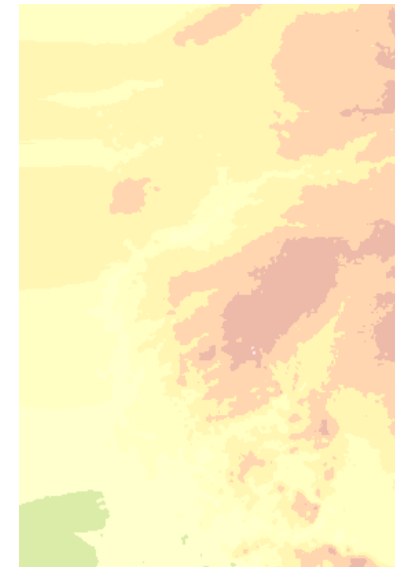
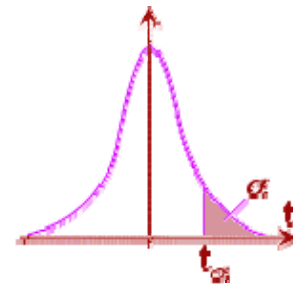
- First, how to recognize data
  - Relational databases
  - Geospatial data:
    - Vector vs. Raster
- Then, how to find it
  - Data portals
    - GIS Atlas for Indiana
    - USGS
    - EPA, etc.
- How to import it
  - Downloading files
  - Using Web Services



-[www.innovativegis.com](http://www.innovativegis.com)

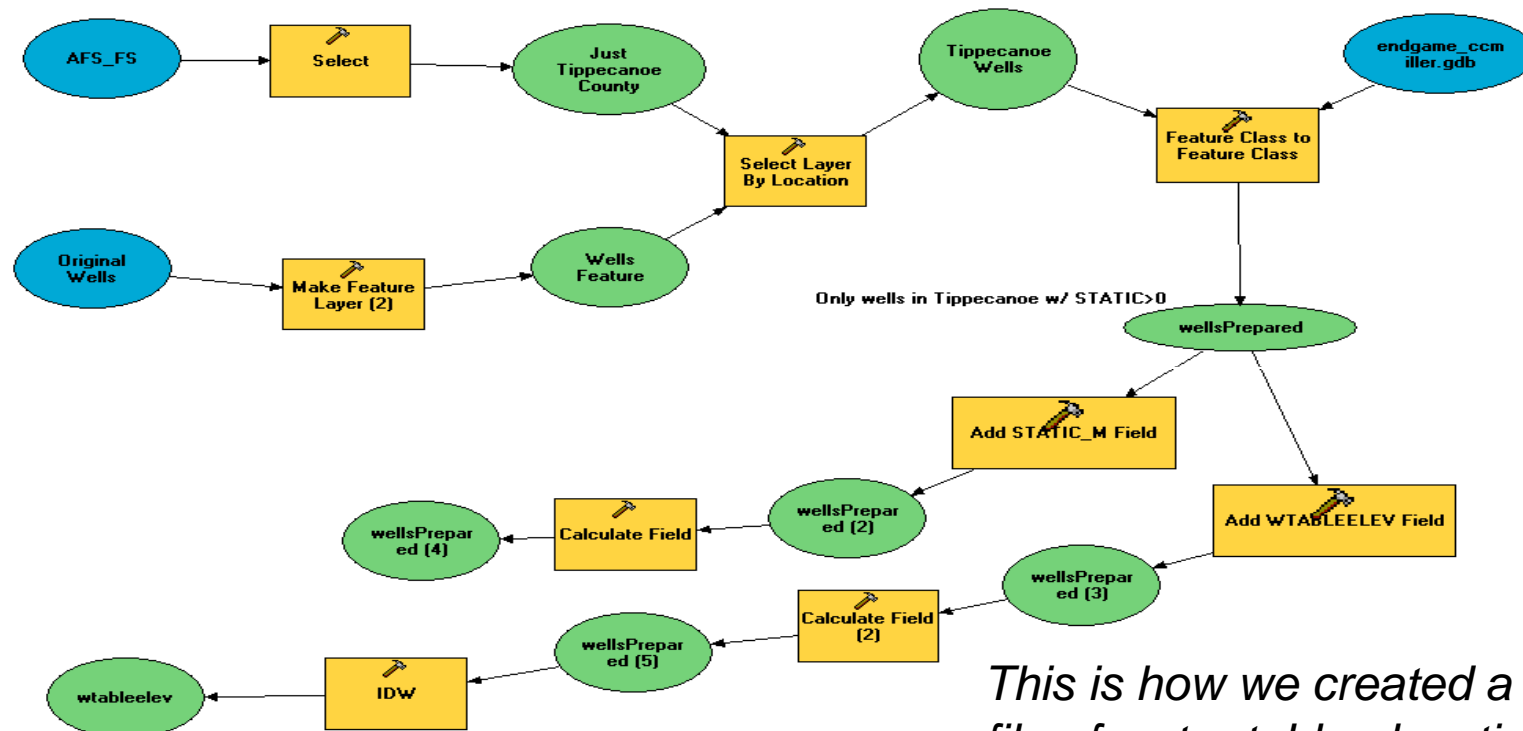
# Data Analysis

- Very basic statistics introduction
  - Means, variances, summary statistics
  - Tests of significance
  - Regressions
- Tools within ArcMap
  - Interpolation (depth to water →)
  - Simulations
  - Creating buffer areas
  - Measuring
  - Workflow tools...



# Workflow Management

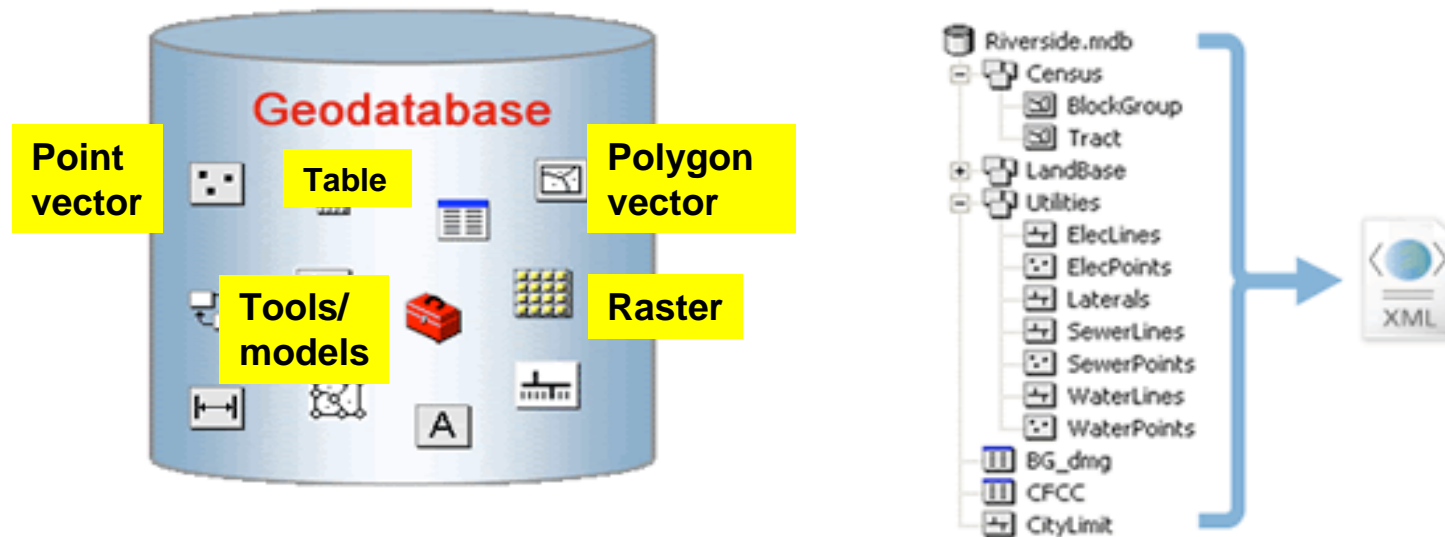
- *Model Builder in ArcGIS and Kepler*



*This is how we created a raster file of water table elevation as precursor of simulation of contaminant flow*

# Data Management

- Creating geodatabases

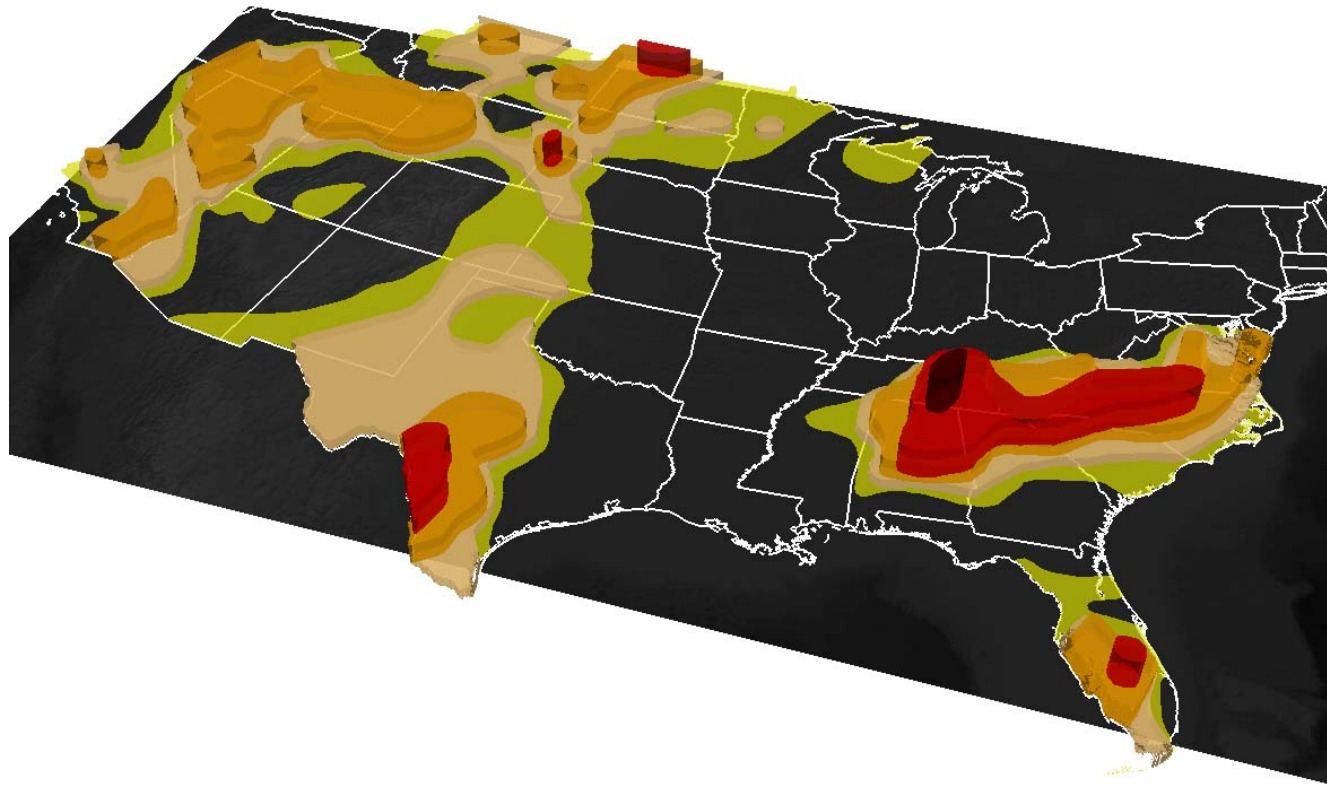


- Ability to 'wrap' up data and export via XML
- A way to share data and processes

# Data Visualization

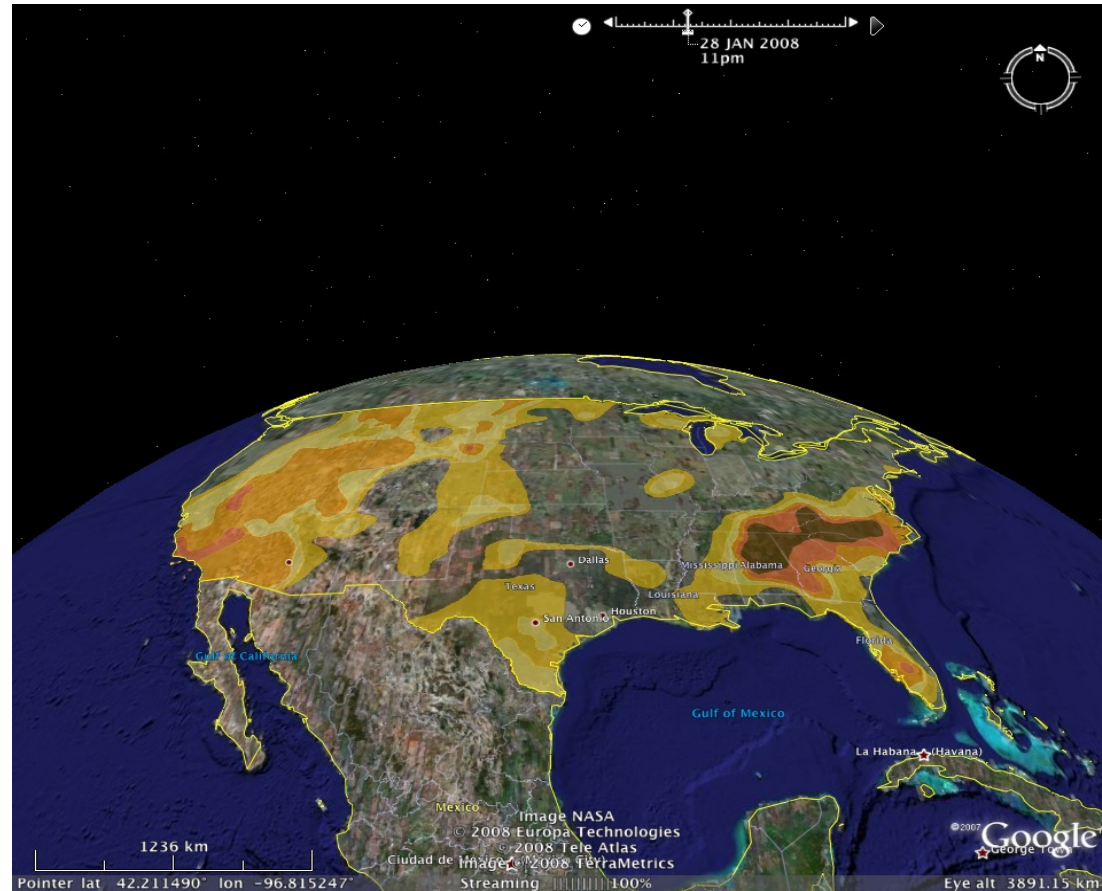
- Introduction to purpose of visualization
  - Not just about flashy animations, but rather adding value to the data—insights into meaning
- Practice with different kinds of visualization techniques:
  - ArcScene
  - Google Earth
  - NASA WorldWind

# ArcScene



Allows extrusions (above), animations...production quality visualizations

# Google Earth



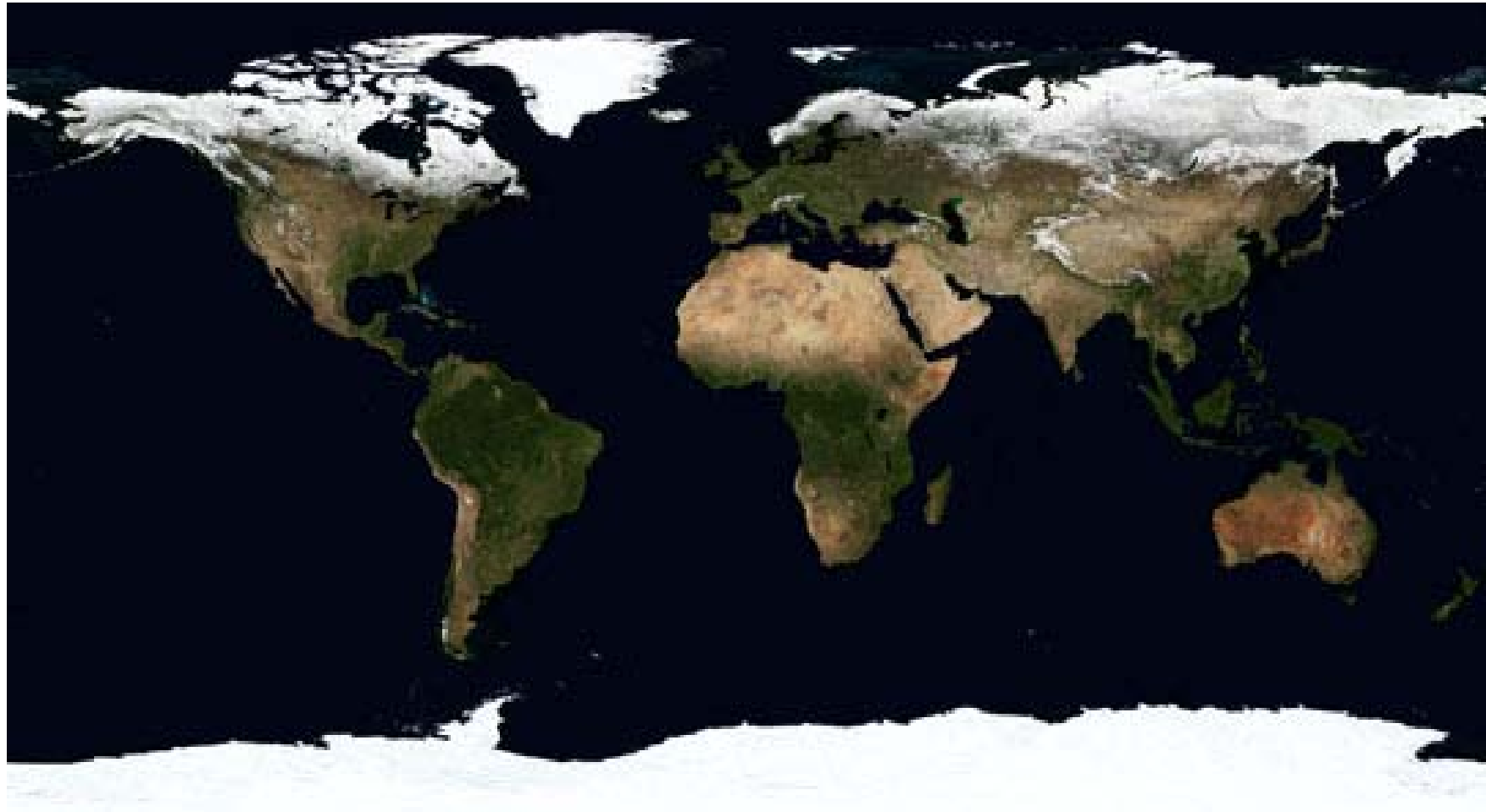
-Adding KML scripts to add your layers to Google Earth infrastructure  
(and from there mashup with other applications)

5/5/2008

16



# NASA WorldWind– Blue Marble



5/5/2008

--<http://earthobservatory.nasa.gov/Newsroom/BlueMarble/>

18

# Data Curation

- Ontology and Thesaurus Construction
- Reading and Creating metadata
- Deposition of data files from semester project on GEON portal to share with class

# Data Curation Overview

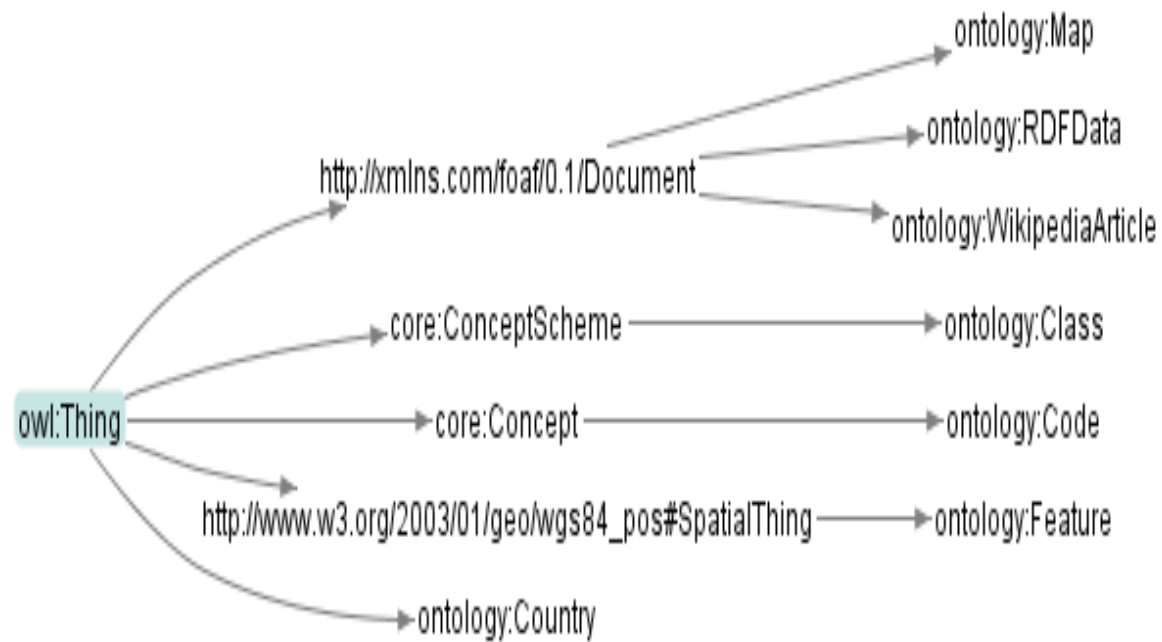
**Curation** is the activity of **managing and promoting** the use of data, starting from the point of creation, to ensure its fitness for contemporary purposes and availability for discovery and re-use.

**Archiving** is a curation activity which ensures that data is properly selected and stored, can be **easily accessed** and that its logical and physical integrity is maintained over time.

**Preservation** is an archiving activity in which specific items of data are **maintained over time** so that they can still be accessed and understood through succession and obsolescence of technologies.

Lord, P. Macdonald, Lyon & Giaretta (2004) "From data deluge to data curation." Proceedings of the UK e-Science All Hands Meeting 2004, 31st August - 3rd September, Nottingham UK.

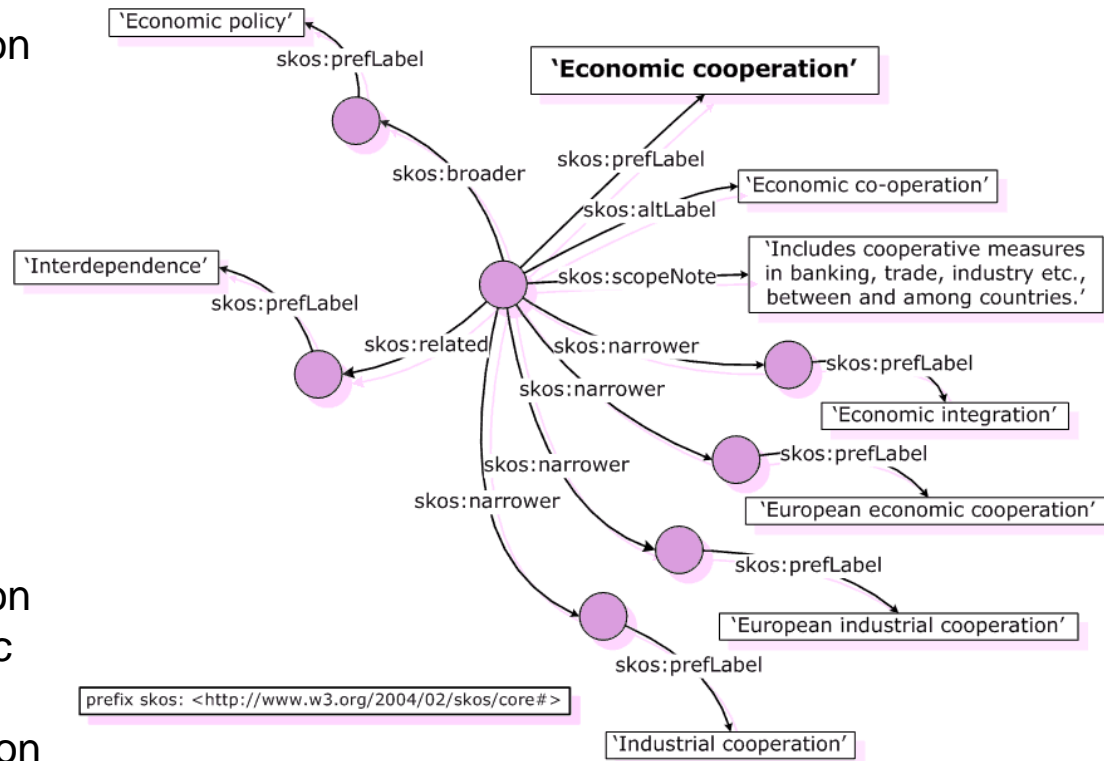
# Ontologies



--GeoNames.org ontology visualized at knoodl.com

# Thesauri using SKOS

- Economic cooperation
- Scope Note:  
Includes cooperative measures...
- UF: economic co-operation
- BT:  
Economic Policy
- NT:  
Economic integration  
European economic integration  
Industrial cooperation
- RT:  
Interdependence



# Conclusions

- **Success?**
  - Several student semester projects were part of their thesis/dissertation projects
  - Students demonstrated ability to manipulate GIS applications and integrate into their research
  - Diverse disciplines of student body... not so useful to treat as geologic-informatics vs. geospatial informatics
  - In process of gathering feedback on concepts/activities that were most useful and important
  - Need to start from scratch with data and informatics skills—can't assume much background (even with databases and programming languages)
- **Acknowledgements:** *The authors thank CSIG 2007 (Cyberinfrastructure Summer Institute for Geoscientists) for ideas and exposure to cyberinfrastructure in the Earth Sciences and the GEON portal.*