Data Access and Data Warehousing

Michael F. Goodchild
University of California
Santa Barbara
Outline

- Geolibraries
- Contrasting world views
- Object-oriented design
- Sharing information
The digital library

- The digital catalog
  - author, title, subject
  - ca 1985
- $10^6$ books in a major research library
- $10^6$ text characters per book
- $10^{12}$ total storage requirement
  - ca 1992
  - preWeb
The digital map library

- $10^6$ maps and images
- $10^8$ bytes per map or image
- $10^{14}$ bytes total
- Geographic location as the primary key
  - the geolibrary
- A physical geolibrary is infeasible
  - one-dimensional, discrete author/title/subject
  - multidimensional, continuous location
ADL Search Results

The query that produced these results can be found at the bottom of this page.

1. DRG 033117g6, Digital Raster Graphic of BLACK STAR CANYON, CA.
   [HIGHLIGHT IN MAP] [COMPLETE DESCRIPTION] [BROWSE GRAPHIC] [ACCESS/DOWNLOAD]

2. DRG 033117g7, Digital Raster Graphic of ORANGE, CA.
   [HIGHLIGHT IN MAP] [COMPLETE DESCRIPTION] [BROWSE GRAPHIC] [ACCESS/DOWNLOAD]

3. DRG 033117g8, Digital Raster Graphic of ANAHEIM, CA.
   [HIGHLIGHT IN MAP] [COMPLETE DESCRIPTION] [BROWSE GRAPHIC] [ACCESS/DOWNLOAD]
The Alexandria Digital Library

- $10^7$ items, $10^{13}$ bytes
- Three ways of specifying location
  - pointing to a map
  - latitude and longitude
  - gazetteer
- The library paradigm
  - putting data in the user’s hands
  - a gatekeeper
- Beyond maps and images
  - information with a geographic footprint
Instances of geolibraries

- www.alexandria.ucsb.edu
- National Geospatial Data Clearinghouse
  - www.fgdc.gov
- www.geographynetwork.com
NRC report

- "Distributed Geolibraries: Spatial Information Resources", 1999

www.nap.edu
“Imagine, for example, a young child going to a Digital Earth exhibit at a local museum. After donning a head-mounted display, she sees Earth as it appears from space. Using a data glove, she zooms in, using higher and higher levels of resolution, to see continents, then regions, countries, cities, and finally individual houses, trees, and other natural and man-made objects. Having found an area of the planet she is interested in exploring, she takes the equivalent of a ‘magic carpet ride’ through a 3-D visualization of the terrain.”
Perspectives on Digital Earth

- High-end visualization
  - an immersive environment
  - specialized hardware
  - massive bandwidth requirements

- Spin, zoom, pan
  - "fly-by" technology

- 4 orders of magnitude zoom
  - 10km to 1m
Does DE scale?

- 500,000,000 sq km
  - 5 million at 10km resolution
  - 500,000,000,000,000 at 1m resolution
Transmitting Digital Earth

- 1m resolution at T1 (order 10 megabits/sec)
  - 69.4 working years
- 1m resolution at 56k
  - done in 12,400 years
- The Internet-killer
What resolution do we really need?

- Whole Earth at 10km
  - California at 1km
  - Santa Barbara County at 100m

- L/S (extent divided by resolution)
  - order $10^3$ or $10^4$
  - ratio for computer screen
  - ratio for human retina
The Internet can support DE

- 1 refresh per second, 1 megapixel images (L/S=10³)
  - T1 rates without compression
  - 10+ refreshes per second with compression
  - sufficient for zoom, pan, flyby
Research challenges for geolibraries

- Defining footprints
  - fuzzy, vernacular
- Access for the child of ten
  - scale
- Search over a distributed archive
  - search engines
  - object-level metadata (OLM)
  - collection-level metadata (CLM)
CLM of the Alexandria Digital Library
Research challenges (2)

- Approaches to CLM
  - by data type
    - ortho.mit.edu
  - by area of the globe
    - SRI's Digital Earth
  - the one stop shop
    - www.fgdc.gov
  - a new generation of search engines
    - identifying footprints
Stages of problem solving

- Problem definition
- Design
- Data acquisition
- Integration
- Analysis
- Interpretation
- Presentation
Why does it take so long?

- Analysis at the speed of light
- Why can't we solve problems in real time?
- How can we make it faster?
Dear Waldo,

View of the Outer Banks of North Carolina from Apollo 9

This photograph was taken on March 12, 1969 at 4:10:00 a.m. EST, from an altitude of about 620 miles.

Posted at the old sea-faring village of Hatteras, I know this card, with its complete and accurate address will get to you.

A pinhole shows you where we are.

Yours Geographically,

[Signature]

Elizabeth City News Co., Elizabeth City, North Carolina

25 USA

POST CARD

Professor Waldo Tabler
34° 26' 41" N
119° 48' 26" W
The Geography Network is a global community of data providers who are committed to making geographic content available. This content is published from many sites around the world, providing you immediate access to the latest maps, data, and related services. This portal to the Geography Network enables you to discover this content and share your own.
Objectives of interoperability

- Using technology to overcome differences
  - rather than imposing uniformity
  - enabling rather than intrusive
  - specifications not standards
- Bridging information communities
- Speeding and easing access to data
Major forces in spatial data interoperability

- National Spatial Data Infrastructure
  - Federal Geographic Data Committee
- Open GIS Consortium
  - industry, government, academic
- National, regional, and international standards organizations
GIS in the field

- Mobile, traveling with the user
- Ubiquitous, operating anywhere
- Augmenting the senses with information from digital representations
  - of the past
  - of what is beyond the senses
  - of the future
CharmIT™ Developer's Kit

- CharmIT™ is built on the PC/104 specification, which has been an industry standard for embedded computing for nearly ten years.
- Hundreds of companies manufacture a wide variety of PC/104 hardware.
- Majority of components are low power and ruggedized.
- CharmIT™ Developer's Kit is lower cost (approximately $2000), low power (approximately 7 watts with Jumptec 266) and offers enough computing power for most everyday wearable tasks.
Head-mounted displays

ClipOn Display ($2500)
- evaluation kit comes with a belt-worn, VGA interface
  box connected to the display by a 4’ cable
Display format: 640x480, 24-Bit color, 60 Hz refresh rate
Field of View: Approximately 16 degrees horizontal

Integrated Eyeglass displays ($5000)
The Twiddler2 chorded keyboard is designed for one-handed input with an array of 12 finger keys and six thumb keys.

Frequent users can enter text at close to two-hand touch-typing speeds.
User interface for augmented vision

Augmented: see-through map plus locator

Viewed reality
View options
Field-work applications

- Finding oneself in the field
  - recovering past sample locations
- Accessing previously collected data
  - the previous census
- Analyzing data continuously
  - progressive formation of geographic knowledge
Location as attribute

- The data table
  - Census summary table
  - County-level health statistics
- What value is location as an explanatory variable?
- Linking the table to a boundary file
  - Enabling maps of summary data
<table>
<thead>
<tr>
<th>Tract</th>
<th>Pop</th>
<th>Location</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3786</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2966</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5001</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4983</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4130</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3229</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4086</td>
<td>x,y</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3979</td>
<td>x,y</td>
<td></td>
</tr>
</tbody>
</table>
Abstraction of geographic space

- Cartograms
- Invariance under rotation, displacement
- Reconstruction from a distance matrix
- Reconstruction from ranked distances
  - ordered metric data (Coombs)
Space as a matrix

$W$ where $w_{ij}$ is some measure of interaction
- adjacency
- decreasing function of distance
- invariant under rotation, displacement
- readily obtained from a GIS
Applications of the $W$ matrix

- Spatial regression
  - add spatially lagged terms weighted by $W$
  - Anselin’s SPACESTAT
- Moran and Geary indices of spatial dependence

\[
C = \frac{(n - 1)\sum \sum w_{ij} (x_i - x_j)^2}{2 \sum \sum w_{ij} \sum (x_i - a)^2}
\]
The location-as-attribute world view

- **Objective**: scientific explanation, understanding of social processes
  - is location an explanatory factor?

- **Relative location as expressed in the $W$ matrix**
  - a surrogate for spatial interaction
  - reflecting costs of transport, probability of interaction and acquaintance, probability of migration or travel, probability of seed dispersal
The Data Documentation Initiative

- A multinational multidisciplinary effort
- A standard for description of dataset contents
- Social science focus
- Geography working group
  - joint meeting August 02
Two world views

- Location as continuum
  - the FGDC metadata standard
  - ISO 19115
  - attribute tables as part of GIS

- Location as attribute
  - the DDI standard

- Reinforced by technical GIS design
  - the hybrid model
  - attributes in an RDBMS
  - geometry in a specialized file structure
  - ARC/INFO
Object-oriented design

- Objects as instances of classes
- Classes inherit properties of more generalized classes (*inheritance*)
- Methods associated with classes (*encapsulation*)
Specialized GIS data models

- The basic elements built into the GIS
  - points, lines, areas
  - the GIS mainstream
- How these elements are specialized in application domains (vertical markets)
  - railroad track as a class of transportation link
  - transportation link as a class of line
Unified Modeling Language

- Visual representation of a data model
  - conventional symbols
  - implemented in Visio

- Creation of database layout
  - use CASE tools
  - build tables
  - populate tables with data
Helping transportation users of ArcGIS by providing a database framework that includes familiar elements:
- contains the core items
- is easy to extend and specialize
- add new attributes
- add specialized classes
ArcGIS Transportation Data Model

Feb. 21, 2001
But what about metadata?

- At the class (table) level?
- At the database level?
- Granularity
  - an unresolved issue
A comprehensive view of information

- To be sharable, information must be digital
- Many types of information
  - maps, images
  - tables
  - text
  - methods
  - simulation models
- What is the relative value per bit?
  - Windows XP >> a cloudy Ikonos image
  - academic paper >> survey data used
The infrastructure of information sharing

- Metadata standard
- Archives
- Interoperability
- To date, almost exclusively about data
  - much less about methods
Towards an infrastructure for dynamic models

- Infrastructure for sharing
  - search
  - discovery
  - evaluation of fitness for use
  - acquisition
  - execution

- Server-side or client-side execution
Falling through the cracks

- **Text-sharing infrastructure**
  - libraries, bookstores, books, journals, WWW, search engines

- **Data-sharing infrastructure**
  - metadata schema, archives, clearinghouses, data centers

- **Model-sharing infrastructure**
  - models are the highest form of sharable knowledge of the Earth system
Current status

- Some archives
  - some pre-WWW
- No standards
- No clearinghouses
- www.ncgia.ucsb.edu/~scott
Research in Metadata for Computer Models

Models available over the Web
Model research and articles
Metadata & Cataloging: Examples, Ideas & Articles

Meetings
Interviews
Readings

Reporting Model 'Fitness of Use' or 'Validation' in Metadata
Comparison chart for Model Metadata
An Easier Method for Metadata Collection
Creating a Computer Model Metadata Standard

A Special thanks to the people at The National Science Foundation for their support of the Research Education for Undergraduates program.

link to Geography 5, Fall 2000
Building a metadata standard for describing models

- A model is a transformation
  - characterized by metadata for inputs and outputs
- Write down the key elements
  - compare FGDC CSDGM
- How do humans do it?
  - we’ve been doing it for decades
- A first-draft standard
Conclusions

- Instant geospatial data now feasible
  - access to distributed archives

- CLM is a problem
  - only experts know where to look
  - the one-stop shop will not happen

- Geohealth embodies two distinct world views
  - location as continuum
  - location as attribute

- Effective sharing of data must somehow embrace both
  - perhaps through new data-modeling technologies
Conclusions (2)

- Data may not be the most valuable type of information
  - other types may justify greater investment in sharing infrastructure
  - a holistic approach to information sharing is needed in geohealth