Serow

Gazetteers as Components of Digital Library Services

PNC Conference

Osaka 2002

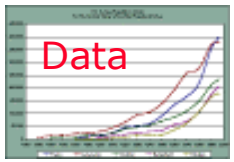
Linda L. Hill

University of California, Santa Barbara

Outline

- Basics of digital gazetteers
- Alexandria Digital Library gazetteer development
- Worldwide gazetteer development community
- Closing discussion points

Place-based information challenge



Papers

Maps

Books

Harvested Webpages

Georeferencing by placename and by spatial footprint

Cataloging – Metadata Creation

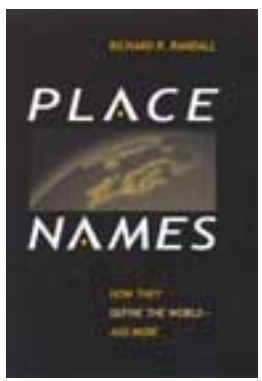
Metadata

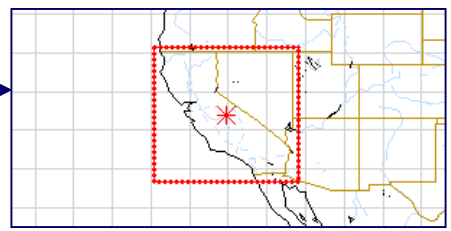
```
<!ENTITY % geographic-coordinate "(#PCDATA)">
<!-- a geographic latitude in degrees north of the equator or
        geographic longitude in degrees east of the Greenwich
        meridian, e.g., "-121.025" -->
<!ELEMENT west_bounding_coor %geographic-coordinate;>
<!ELEMENT east_bounding_coor %geographic-coordinate;>
<!ELEMENT south_bounding_coor %geographic-coordinate;>
<!ELEMENT north_bounding_coor %geographic-coordinate;>
<!ELEMENT measurement_begin_date %calendar-date;>
```

Translation needed between placenames - locations

GIS datasets

Oral histories





Gazetteers

Search Engines

Where is …?
What's there?
What happened there?

Digital Library Model

DATA STORE OF OBJECTS

CATALOG OF METADATA

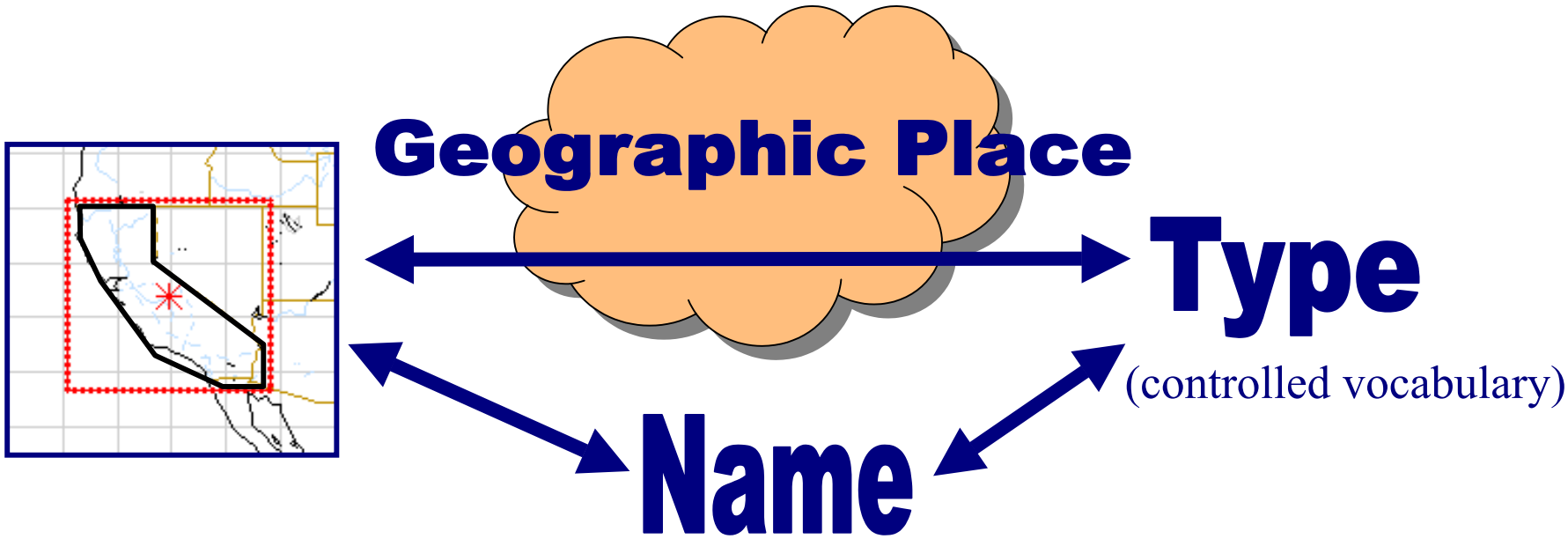
SERVICES

- CATALOGING
- DIGITIZING
- ACCESSING
- SEARCHING
- RETRIEVING
- VISUALIZING
- USING
- ARCHIVING

KNOWLEDGE ORGANIZATION SYSTEMS

- THESAURI
- AUTHORITY FILES
- CONCEPT SPACES
- GAZETTEERS
- DICTIONARIES
- GLOSSARIES
- SUBJECT HEADING SETS
- CLASSIFICATION SYSTEMS
- ONTOLOGIES

Digital gazetteer essentials



- None of these elements are unique identifiers of a particular place

What's a gazetteer?

- Preferred definition

  Spatial dictionary of named and typed places

- Originally (in the simplest case)
  - setof (name, location)
    - the "index" in an atlas
    - a "geographical dictionary"
    - feature labels for GIS data layers
- Redefined
  - setof (name, type, location)
- Extended
  - Time-stamped names, extents, and relationships
  - Descriptive information about names and places
  - Merging of information about a place from multiple sources

Operational ADL Gazetteer

- ADL Gazetteer
  - 4.2 million entries
  - Worldwide coverage
- Content
  - Merged US federal gazetteers plus other sets
  - Largest gazetteer with common typing scheme
- Online access
  - ADL Gazetteer Server
  - ADL-California Digital Library Interface
  - ADL Gazetteer Protocol Server

Protocol server code

- Full distribution is 358K; 290K of that is Java source code
- Requires
  - Java 1.3 or greater (it may well work with 1.2)
  - Informix JDBC driver (or another database link)
  - Apache Xerces parser version 2.0.2 or higher (could conceivably work with other versions)
  - Apache Ant build tool
- Tested with
  - Apache Tomcat 3.3 and 4.1.x
  - Informix, with the table layout, and SpatialWare and Verity Text search modules
- Configuration files
  - For example, you can insert your own XML-to-SQL translator through the properties file

Protocol functionality summary

- Get capabilities request
- Query
  - Name
  - Name and Class
  - Name and Footprint
  - Class and Footprint
  - Identifier (unique ID of gazetteer entry)
  - Relationship and Class
  - Name and Indirect Footprint
- Report
  - Standard
  - Extended

Protocol operators

- Name queries
  - Equals
  - Contains-all-words
  - Contains-any-words
  - Contains-phrase
  - Matches-pattern (e.g., * to mask characters)
- Footprint queries
  - Region
    - Polygon
    - Box
    - Identifier
  - Operators
    - Within
    - Contains
    - Overlaps

Example protocol queries

- Name: *'osaka'*
- Name and class: *'osaka' and class equals 'mountains'*
- Name and footprint: *'osaka' within '128,30 146,46'*
- Class and footprint: *'streams' within '128,30 146,46'*
- Identifier: 'adlgaz-1-77-3f'
- Relationship and class: 'streams' that are 'part of' 'adlgaz-1-77-3f
- Name and indirect footprint: 'osaka' within 'japan' (that is, 'adlgaz-1-77-3f')

Query by name

```
<gazetteer-service
   xmlns="http://www.alexandria.ucsb.edu/gazetteer"
   version="1.1">
  <query-request>
   <gazetteer-query>
     <name-query operator="contains-phrase"
       text="osaka"/>
   </gazetteer-query>
   <report-format>standard</report-format>
  </query-request>
</gazetteer-service>
```

Published Components

- Gazetteer Content Standard
  - XML schema for gazetteer entries
    - Translated into Chinese by Academia Sinica
  - XML schema for source entries
- Feature Type Thesaurus
  - 210 preferred terms
  - 946 non-preferred terms
- Gazetteer Service Protocol
  - XML-based query and response
- Placename List
  - 5.9 million placenames with ADL Gaz IDs

Current R&D

- The "duplicate" problem
  - Given that there will be *one entry* in a gazetteer for *a place*
    - Merging information from different sources is necessary
  - But, there is no single attribute of *a place* that is unique
    - Multiple placenames for the same place
    - Same placename for different places
    - Multiple coordinate representations of location
    - Variant types of categorization
  - How then do you determine that two pieces of information are about the same *place*
  - Solution: Evaluating the combination of evidence about potential duplicate entries

Other current issues and research

- Developing a reference database schema for the ADL Gazetteer Content Standard
  - For archival storage
  - For query support
- Developing a database implementation that supports frequent updating
- Developing an ingest protocol and software for the ingest process
- Textual-Geospatial Integration Project
  - NSDL project

Worldwide Gazetteer Community

- Recent Digital Gazetteer Workshop at the Joint Conference on Digital Libraries (JCDL)
  - ECAI
  - Taiwan
  - United Kingdom
  - Germany
  - Canada
  - United States
- Gazetteer discussion list
- High priorities
  - Demonstrate value of distributed gazetteers and gazetteer services
  - Demonstrate applications of digital gazetteers for geoparsing and integrate digital library services

Closing discussion points

- Gazetteers bridge the text and the geospatial domains

TEXT

- Toponyms and their histories
- Historical scholarship involving places and their roles
- Geoparsing of text to derive geospatial locations
- Fuzzy place references
- Links to textual documents about places

GEOSPATIAL

- Geospatial standards
- GIS data layers and their labels
- Administrative applications
- Precision place representations
- Links to map representations of places

Closing discussion points

- How to move forward to sharable gazetteer data and networked gazetteer services
  - Value of tapping into local and specialized knowledge of places
  - Harvesting polygonal footprints from GIS datasets to enrich gazetteer footprints
  - Shared and customized typing schemes
  - Multilingual and multi-script support
  - Demonstration projects

www.alexandria.ucsb.edu/gazetteer

It has been a pleasure to be here

Thank you