Distributed and Federated Search of Georeferenced Cultural Information

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley

Overview

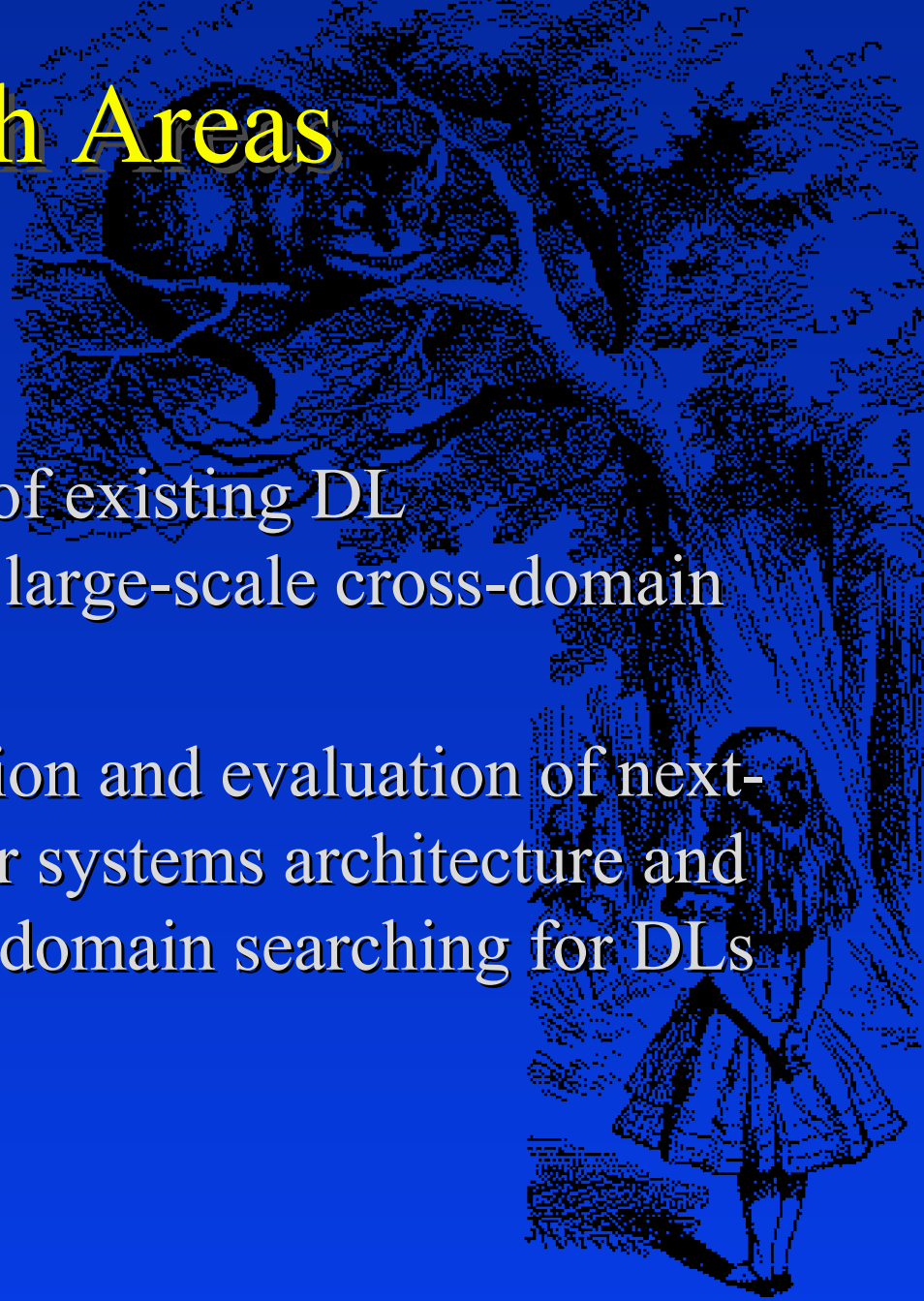
- The Cheshire approach to building distributed services
- GEO searching and georeferenced information in our collections
- NSDI Clearinghouse and Services Architecture
- Future Plans

Context

- NSF/JISC International Digital Library Grant
  - **Cross-Domain Resource Discovery: Integrated Discovery and Use of Textual, Numeric and Spatial Data**
- UC Berkeley working with the University of Liverpool/Manchester Computing with participation from
  - DeMontfort University (MASTER)
  - Art and Humanities Data Service (http://ahds.ac.uk/)
    - Especially the History Data Service (HDS) at the University of Essex
  - Consortium of University Research Libraries (CURL)
  - UC Berkeley Library (and California Digital Library)
    - Making of America II
    - Online Archive of California
  - British Natural History Museum, London
  - NESSTAR → FASTER
  - UKOLN – Resource Discovery Network
  - British Library (ZETOC at Manchester Computing)

Research Areas

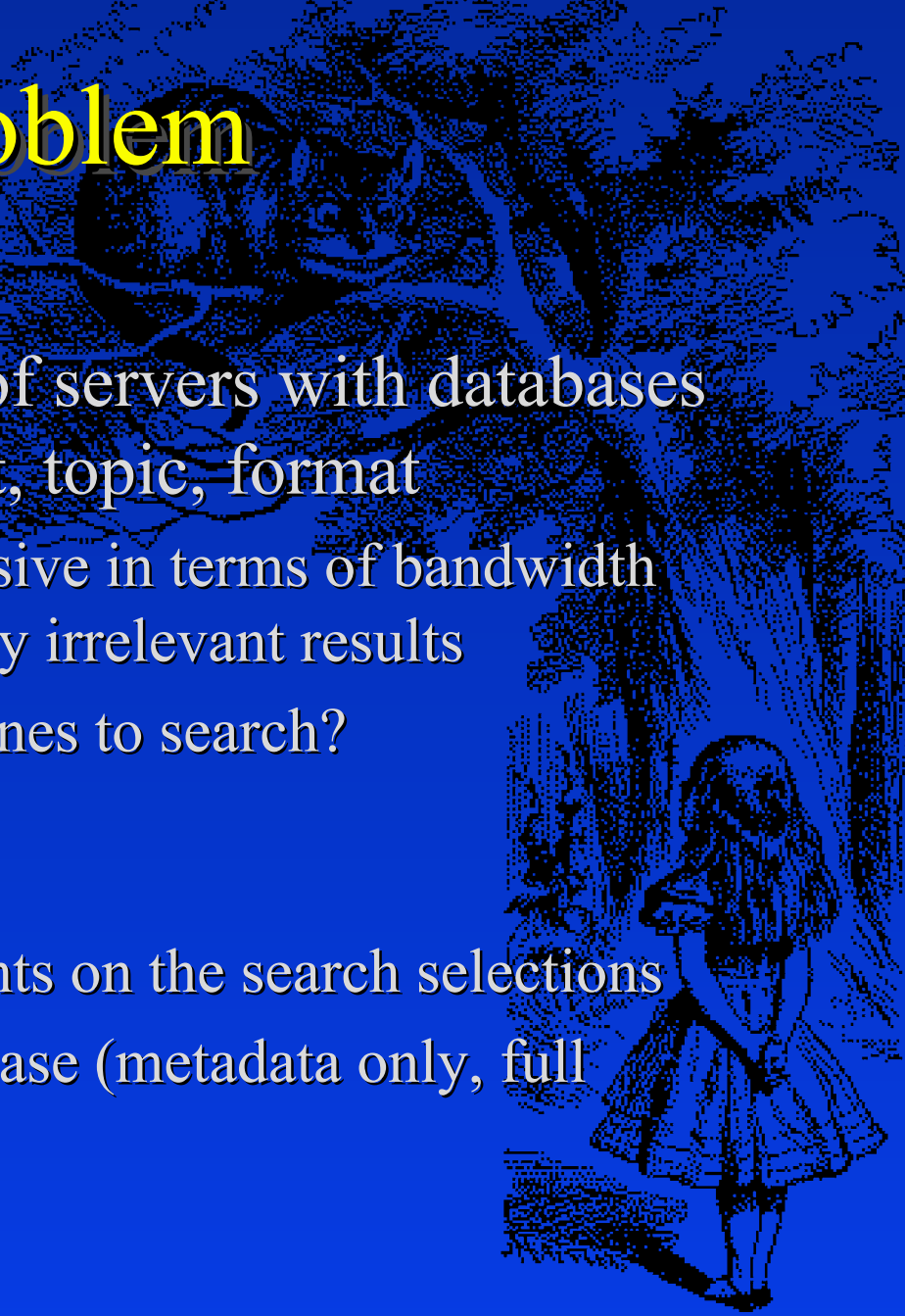
- Goals are
  - Practical application of existing DL technologies to some large-scale cross-domain collections
  - Theoretical examination and evaluation of next-generation designs for systems architecture and and distributed cross-domain searching for DLs



Approach

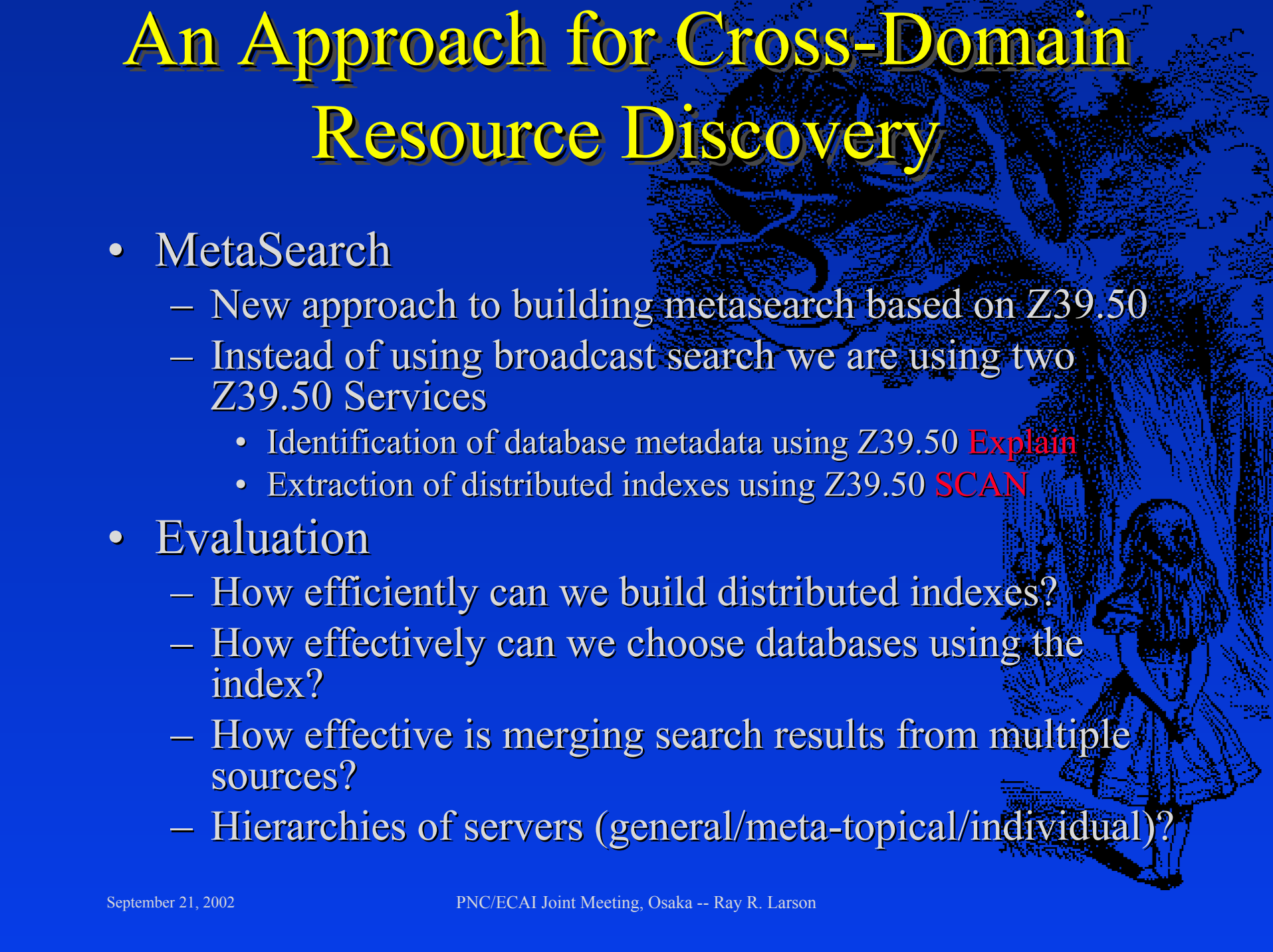
- For the first goal, we are implementing a distributed search system based on international standards (Z39.50 and SGML/XML) using the Cheshire II information retrieval system
- Databases included in original proposal:
  - HE Archives hub
  - Arts and Humanities Data Service (AHDS)
  - MASTER
  - CURL (Consortium of University Research Libraries)
  - Online Archive of California (OAC)
  - Making of America II (MOA2) (Basis of the METS Standard)

The Problem



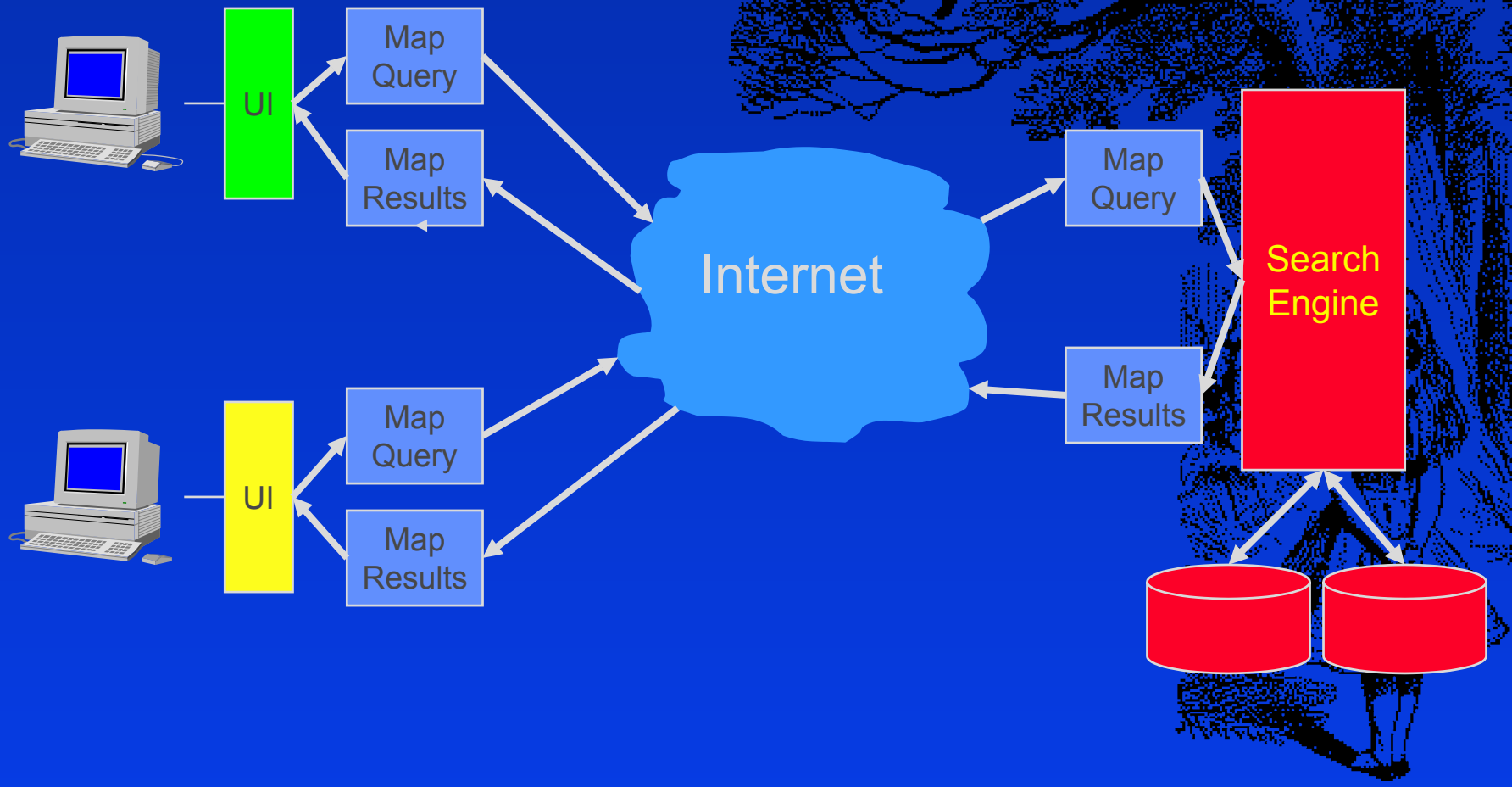
- Hundreds or Thousands of servers with databases ranging widely in content, topic, format
  - Broadcast search is expensive in terms of bandwidth and in processing too many irrelevant results
  - How to select the "best" ones to search?
    - What to search first
    - Which to search next
  - Topical /domain constraints on the search selections
  - Variable contents of database (metadata only, full text…)

An Approach for Cross-Domain Resource Discovery



- MetaSearch
  - New approach to building metasearch based on Z39.50
  - Instead of using broadcast search we are using two Z39.50 Services
    - Identification of database metadata using Z39.50 Explain
    - Extraction of distributed indexes using Z39.50 SCAN
- Evaluation
  - How efficiently can we build distributed indexes?
  - How effectively can we choose databases using the index?
  - How effective is merging search results from multiple sources?
  - Hierarchies of servers (general/meta-topical/individual)?

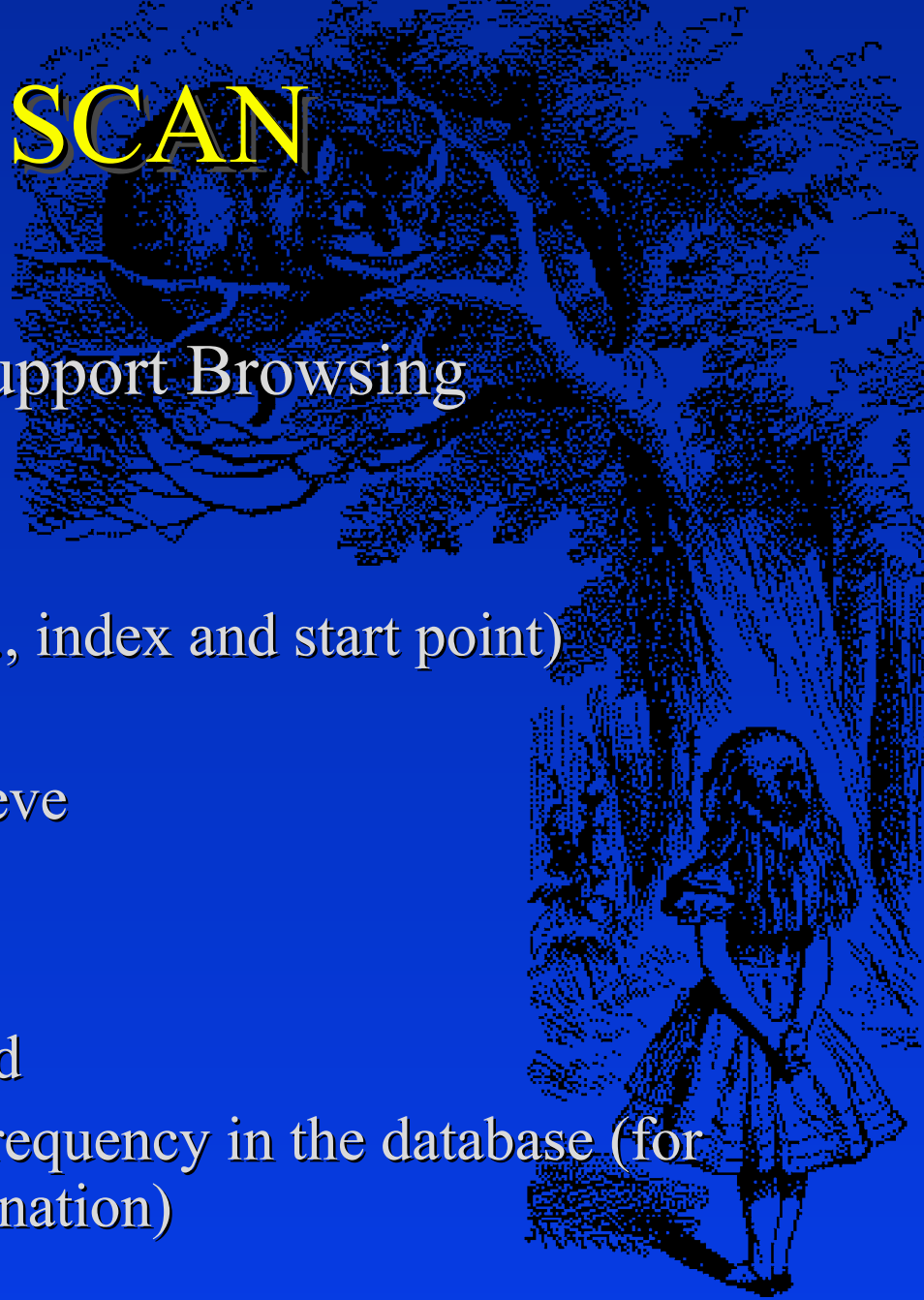
Z39.50 Overview



Z39.50 Explain

- Explain supports searches for
  - Server-Level metadata
    - Server Name
    - IP Addresses
    - Ports
  - Database-Level metadata
    - Database name
    - Search attributes (indexes and combinations)
  - Support metadata (record syntaxes, etc)

Z39.50 SCAN



- Originally intended to support Browsing
- Query for
  - Database
  - Attributes plus Term (i.e., index and start point)
  - Step Size
  - Number of terms to retrieve
  - Position in Response set
- Results
  - Number of terms returned
  - List of Terms and their frequency in the database (for the given attribute combination)

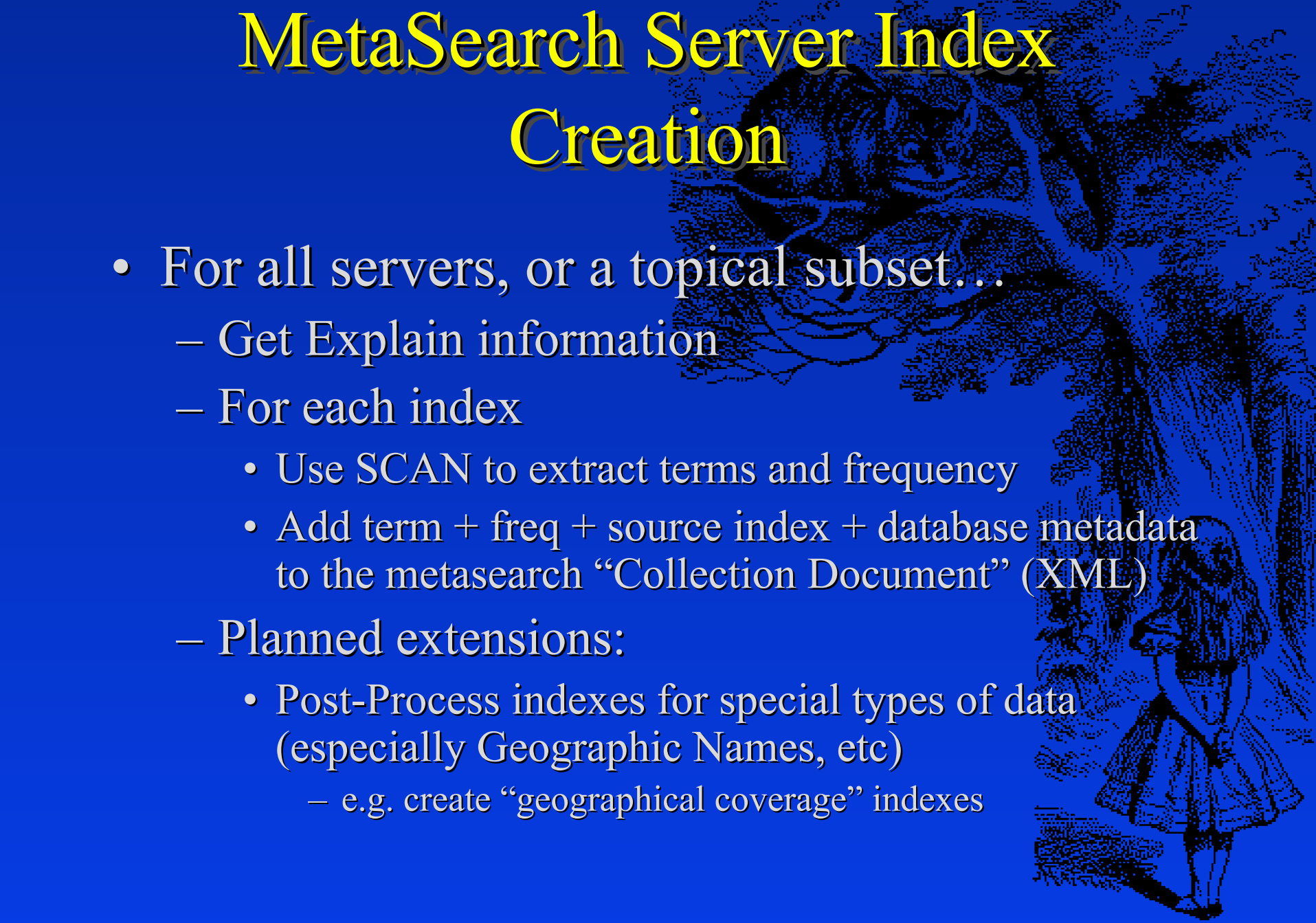
Z39.50 SCAN Results

Syntax: zscan indexname1 term stepsize number_of_terms pref_pos

```
% zscan title cat 1 20 1
{SCAN {Status 0}
{Terms 20}
{StepSize 1}
{Position 1}}
{cat 27}
{cat-fight 1}
{catalan 19}
{catalogu 37}
{catalonia 8}
{catalyt 2}
{catania 1}
{cataract 1}
{catch 173}
{catch-all 3}
{catch-up 2} ...
```

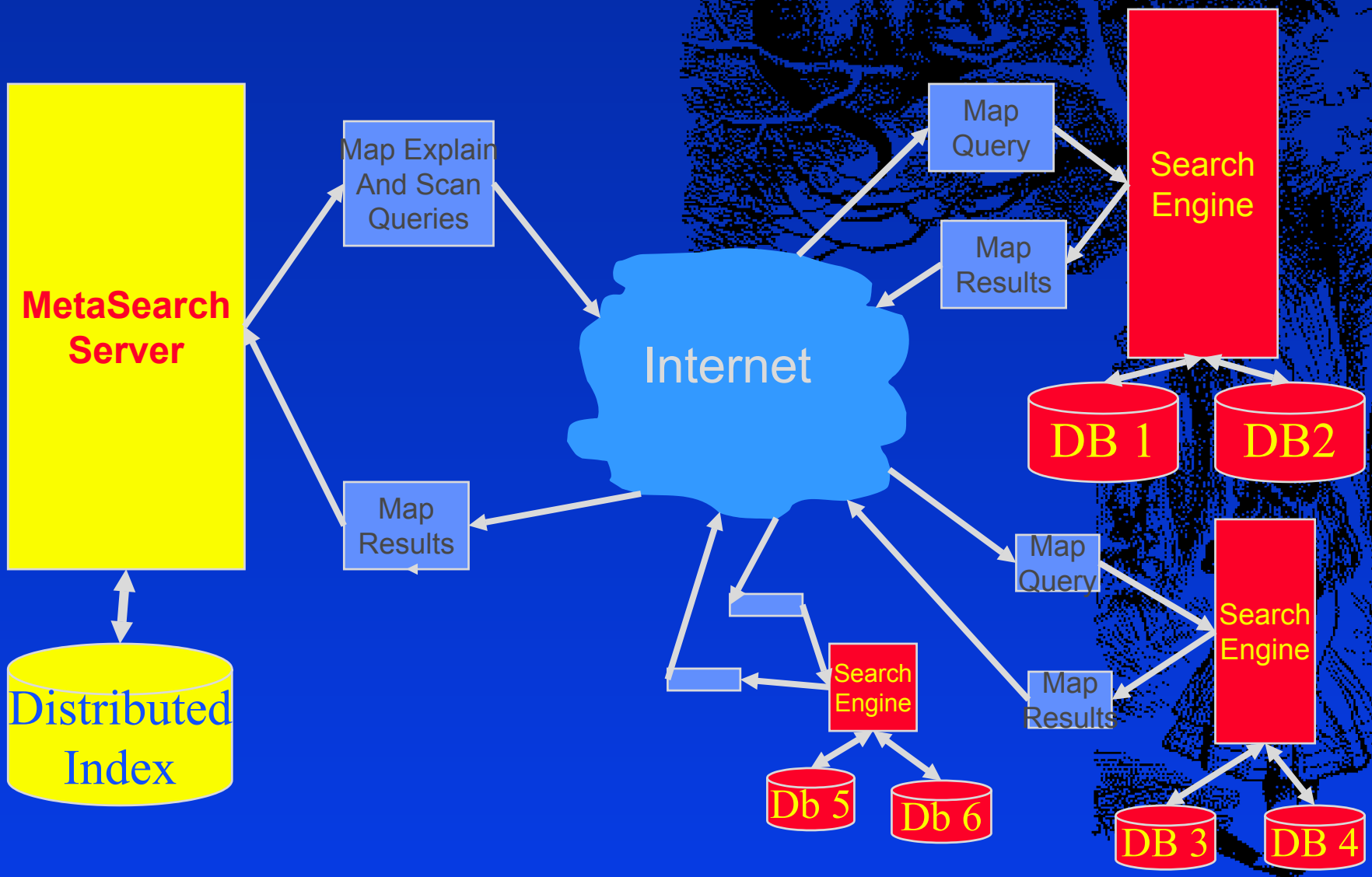
```
zscan topic cat 1 20 1
{SCAN {Status 0}
{Terms 20}
{StepSize 1}
{Position 1}}
{cat 706}
{cat-and-mouse 19}
{cat-burglar 1}
{cat-carrying 1}
{cat-egory 1}
{cat-fight 1}
{cat-gut 1}
{cat-litter 1}
{cat-lovers 2}
{cat-pee 1}
{cat-run 1}
{cat-scanners 1} ...
```

MetaSearch Server Index Creation

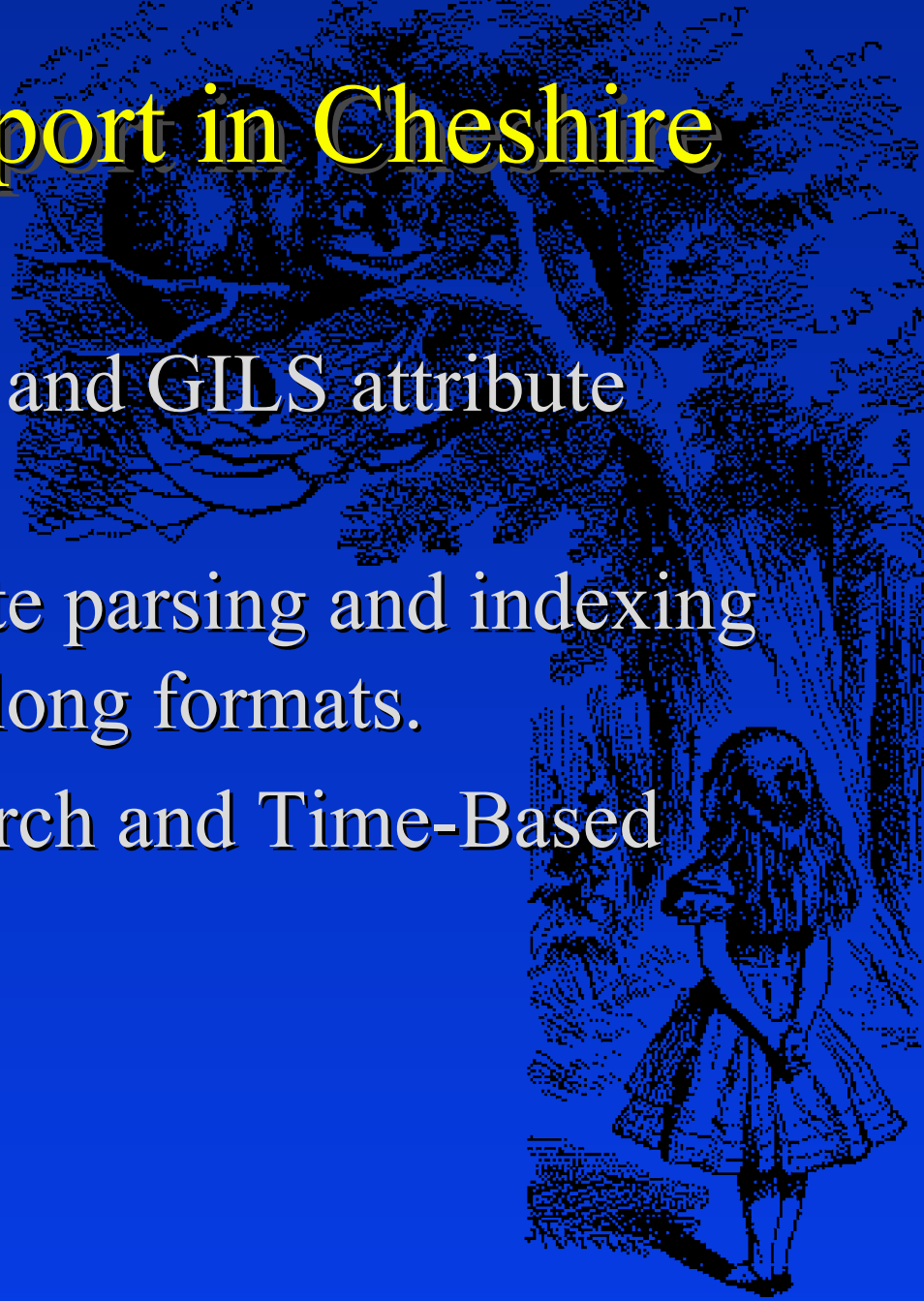


- For all servers, or a topical subset…
  - Get Explain information
  - For each index
    - Use SCAN to extract terms and frequency
    - Add term + freq + source index + database metadata to the metasearch "Collection Document" (XML)
  - Planned extensions:
    - Post-Process indexes for special types of data (especially Geographic Names, etc)
      - e.g. create "geographical coverage" indexes

MetaSearch Approach

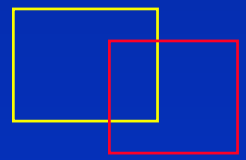


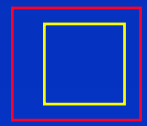
Geographic Support in Cheshire

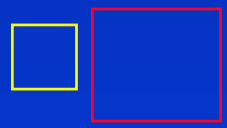


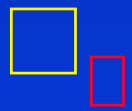
- Support for the GEO and GILS attribute sets
- Support for coordinate parsing and indexing from a variety of lat-long formats.
- Support for GEO search and Time-Based operations

GEO Search Operators



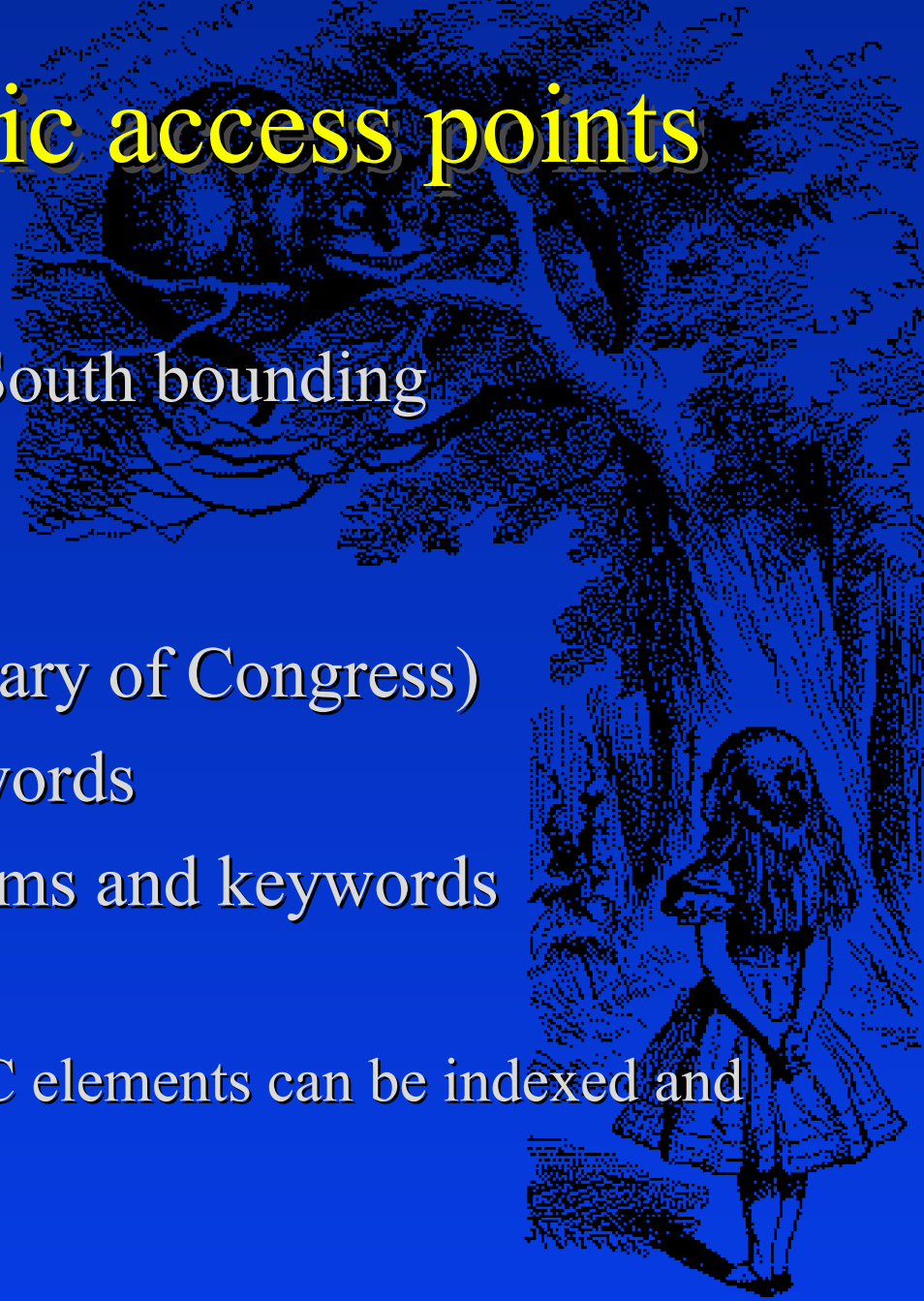






- Overlaps
- Fully Enclosed within
- Encloses
- Outside of
- Near
- In addition, the conventional operators
  - != = <= < > >= can be used with coordinate indexes.

Other Geographic access points

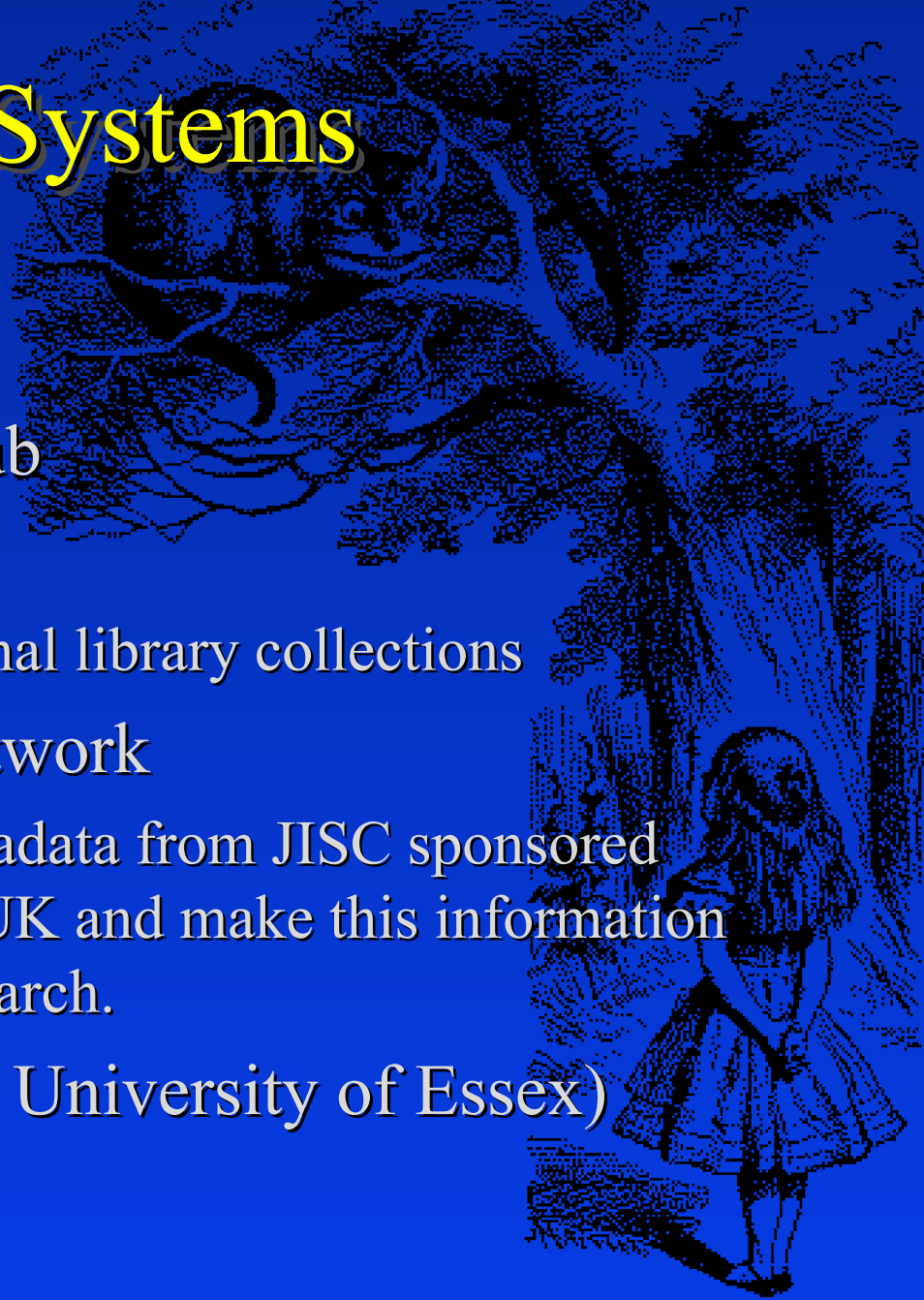


- East, West, North, and South bounding coordinates
- Geographic names
- Geographic Codes (Library of Congress)
- Theme and Theme keywords
- Placename thesaurus terms and keywords
- Etc.
  - Virtually all of the FGDC elements can be indexed and made access points

GEO Time Operators

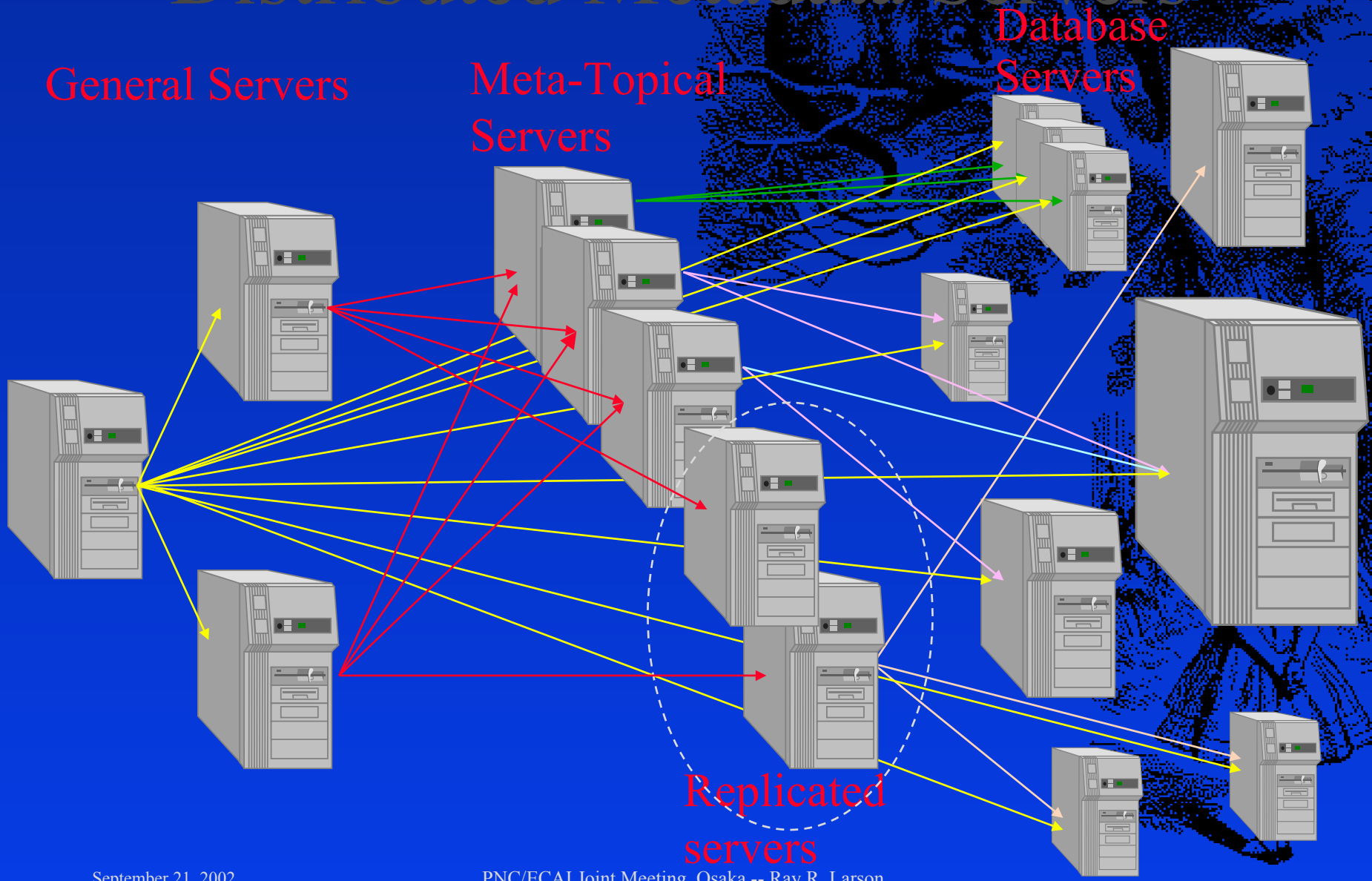
- Before
- Before or During
- During (within or overlapping time range)
- During or After
- After

Current Systems



- Archives Hub
- Distributed Archives Hub
- WARM
  - Liverpool area and national library collections
- Resource Discovery Network
  - Uses OAI to harvest metadata from JISC sponsored services throughout the UK and make this information available via Cheshire search.
- History Data Service (at University of Essex)

Distributed Metadata Servers

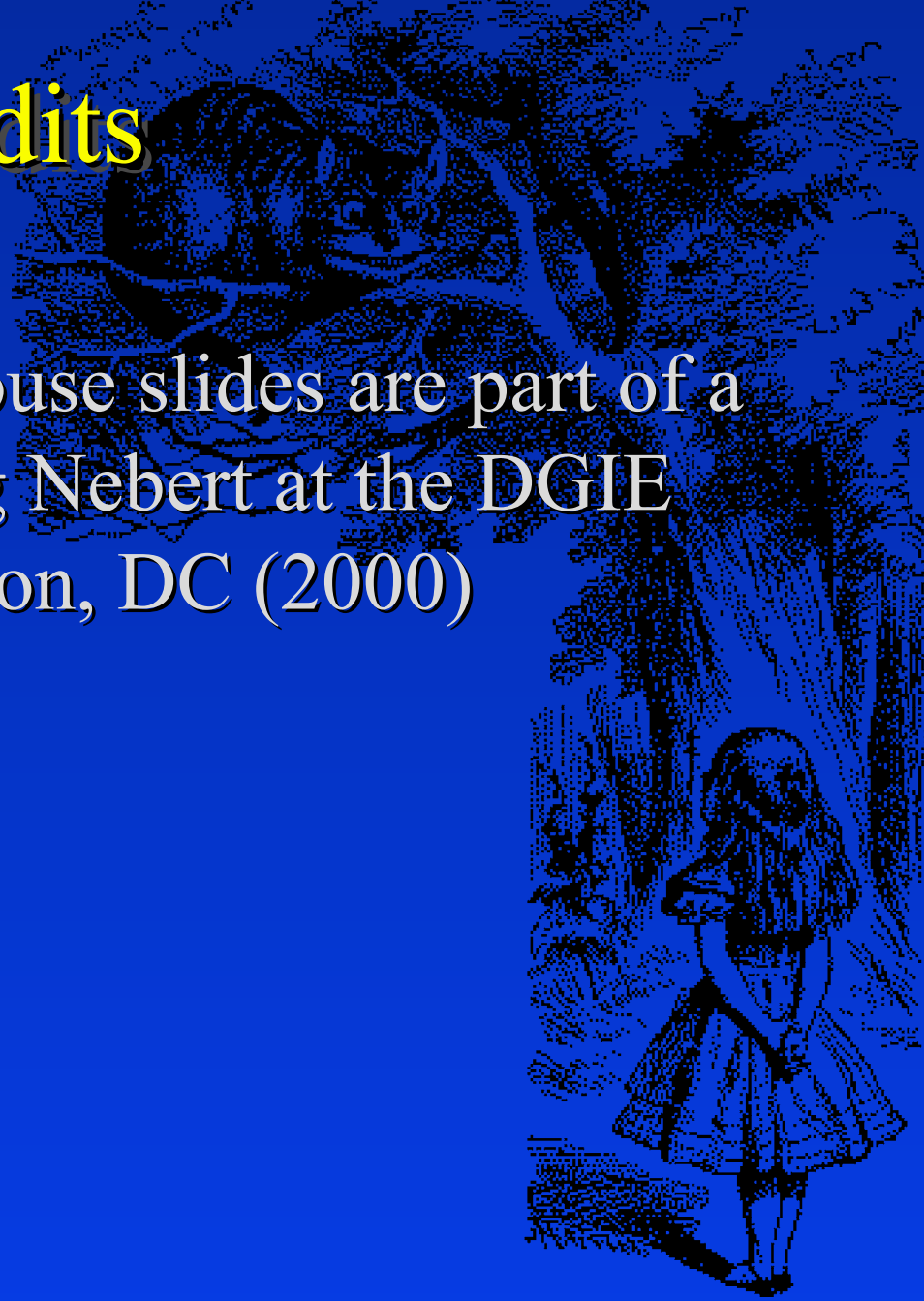


NSDI

- Brief overview of the NSDI

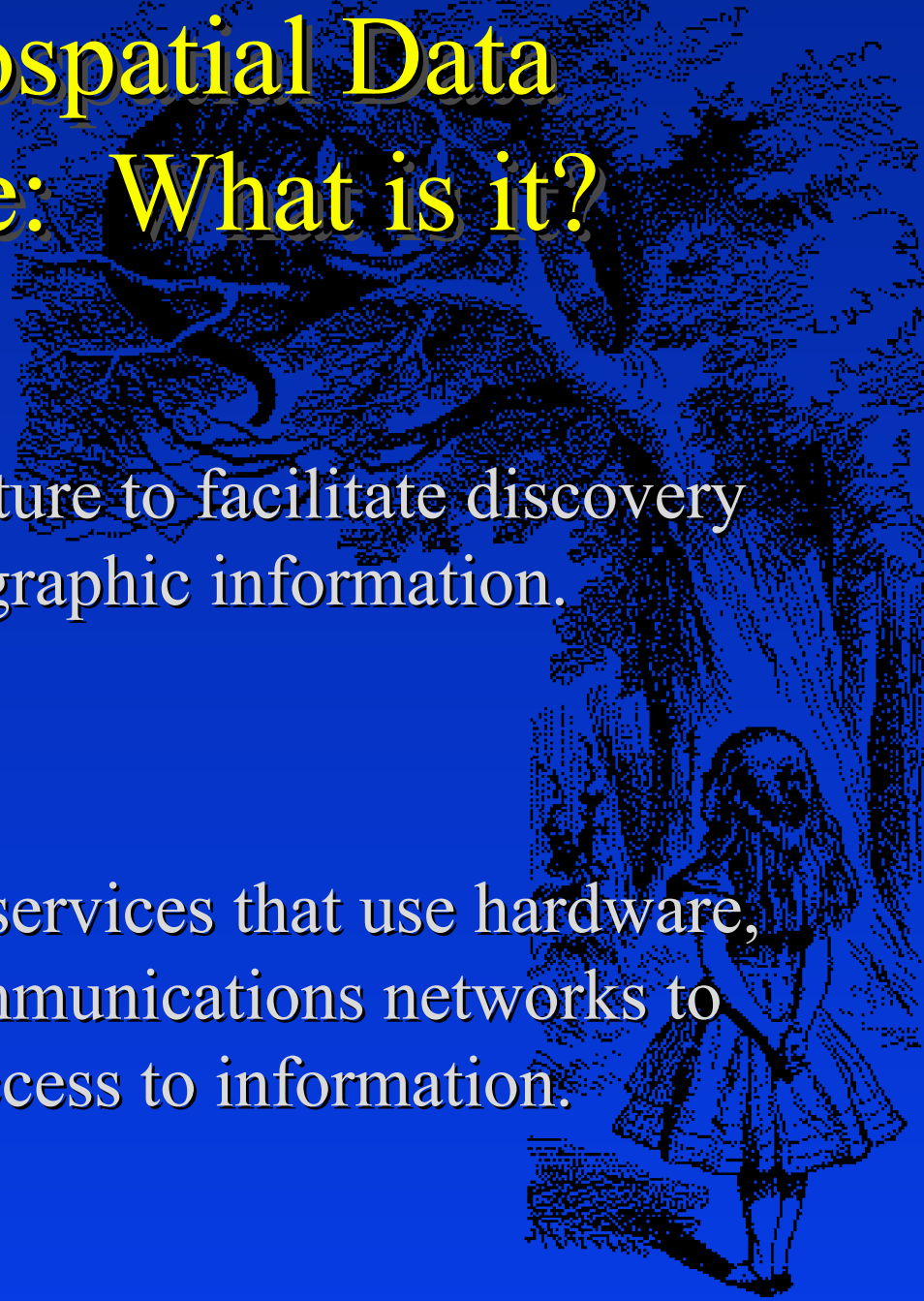
Credits

- The NSDI clearinghouse slides are part of a presentation by Doug Nebert at the DGIE Meeting in Washington, DC (2000)



National Geospatial Data Clearinghouse: What is it?

- Institutional view
  - People and infrastructure to facilitate discovery of who has what geographic information.
- Technical view
  - A set of information services that use hardware, software, and telecommunications networks to provide searchable access to information.



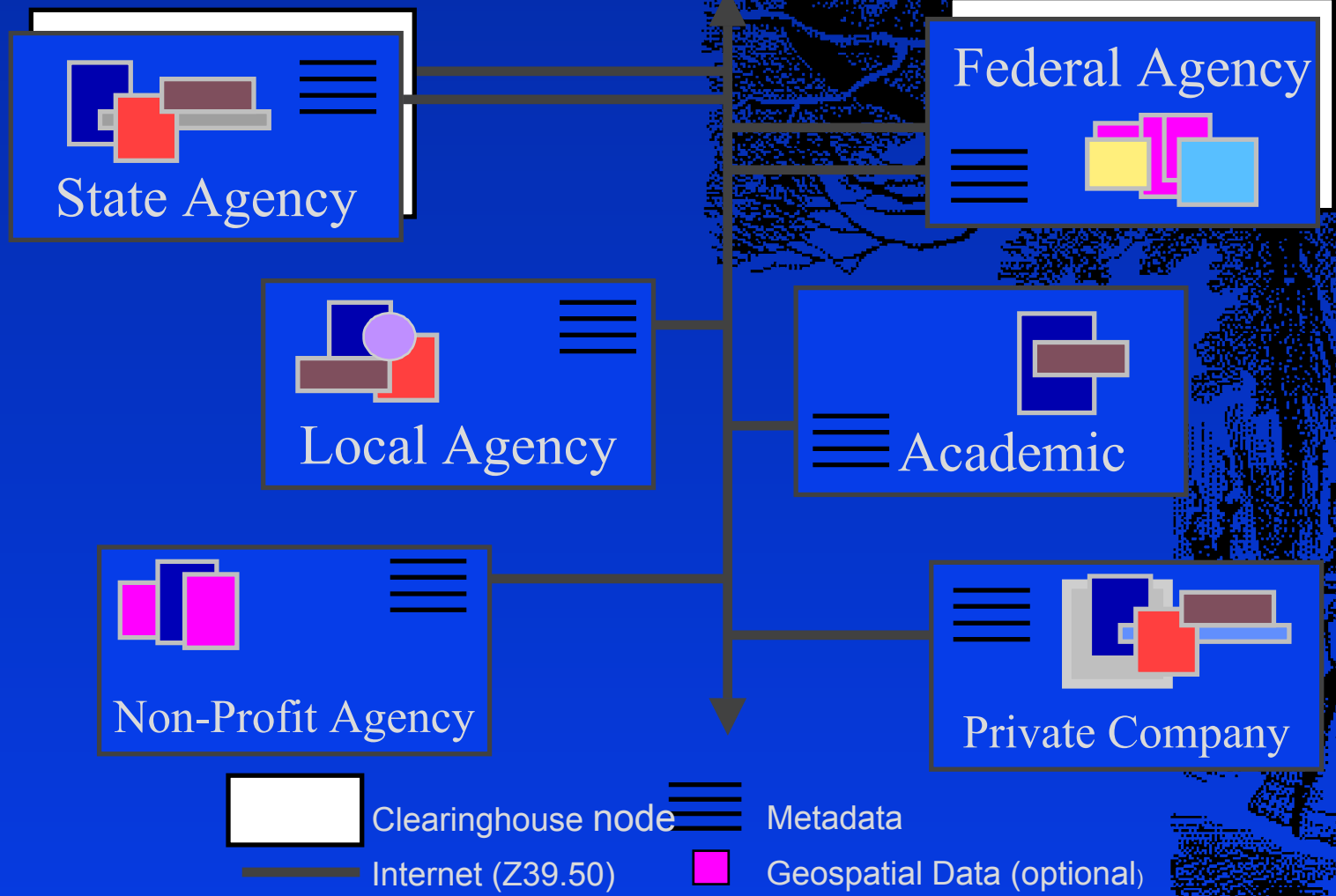
Credit: Doug Nebert

National Geospatial Data Clearinghouse

- Distributed data producers and users.
- Key components:
  - Data documentation (metadata)
  - Networking (Internet)
  - Serving, searching, and accessing software
    - Z39.50 Search and Retrieve Protocol
    - WWW - World Wide Web

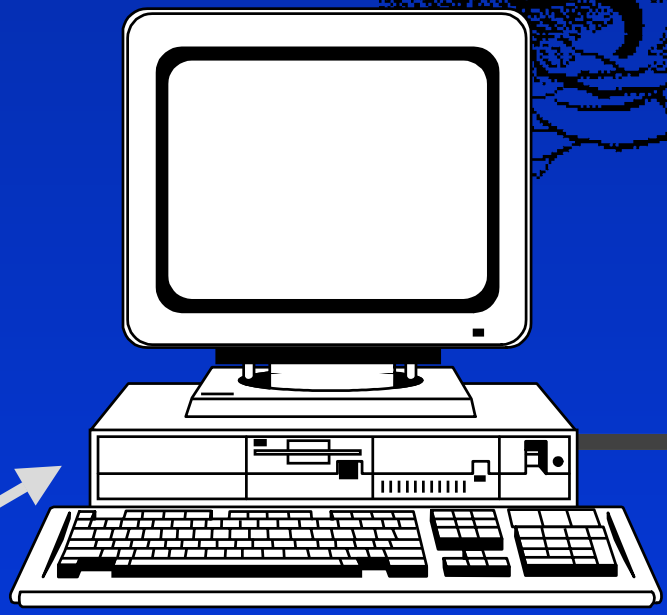
Credit: Doug Nebert

Clearinghouse postal metaphor



Credit: Doug Nebertc

National Geospatial Data Clearinghouse - Client Side

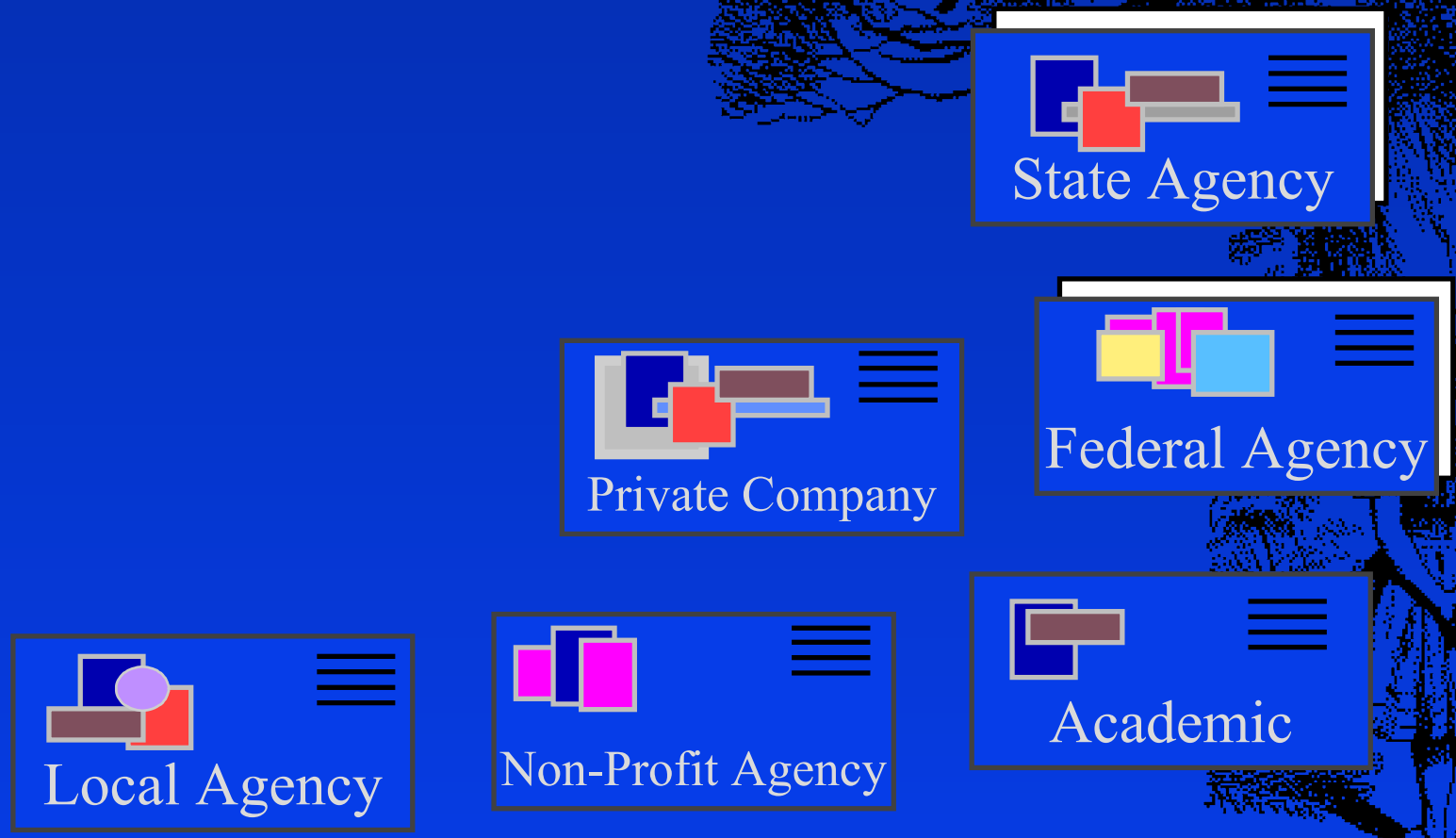


Hardware and Software
(commercially available
Web browser software on
PC or UNIX computer)

Access Method
(http / Z39.50)

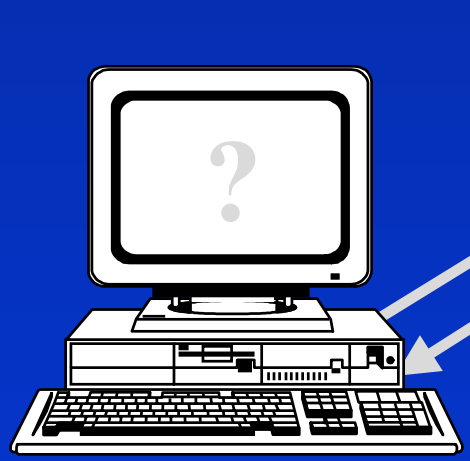
Credit: Doug Nebert

Conceptual design



Credit: Doug Nebert

Client consults server registry



?

Clearinghouse
Server Registry
HTML/Java

list of links

State Agency

Federal Agency

Private Company

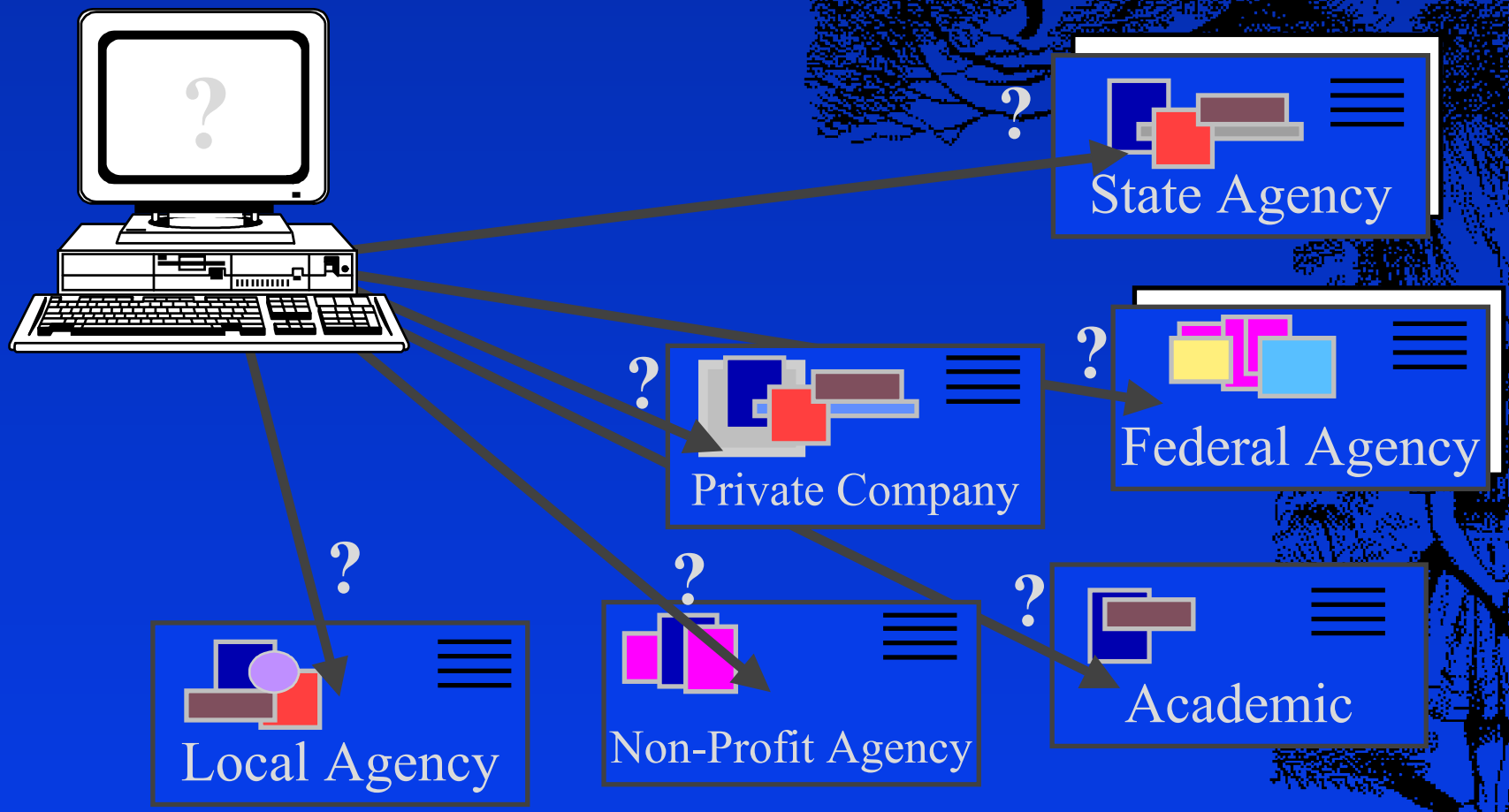
Academic

Non-Profit Agency

Local Agency

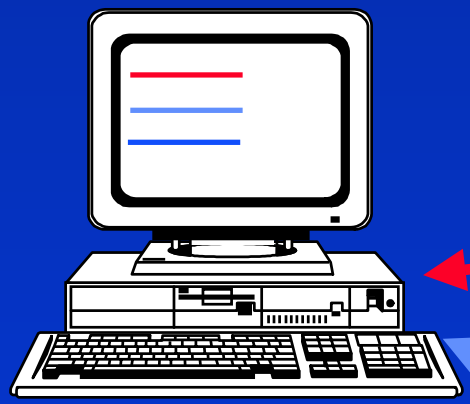
Credit: Doug Nebert

Distributed query is passed to servers



Credit: Doug Nebert

Results returned as "headlines"



State Agency

Federal Agency

Private Company

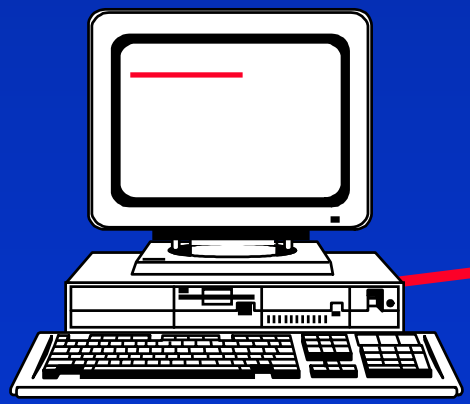
Academic

Non-Profit Agency

Local Agency

Credit: Doug Nebert

Client requests a metadata entry



State Agency

Federal Agency

Private Company

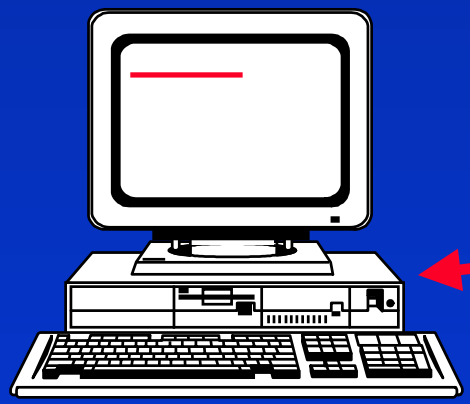
Academic

Non-Profit Agency

Local Agency

Credit: Doug Nebert

Metadata and/or data are downloaded



State Agency

Federal Agency

Private Company

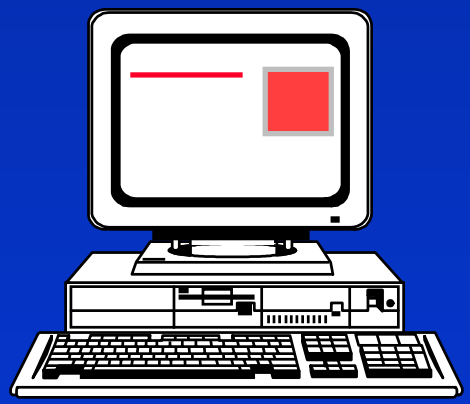
Academic

Local Agency

Non-Profit Agency

Credit: Doug Nebert

Metadata and/or data are downloaded



State Agency

Federal Agency

Private Company

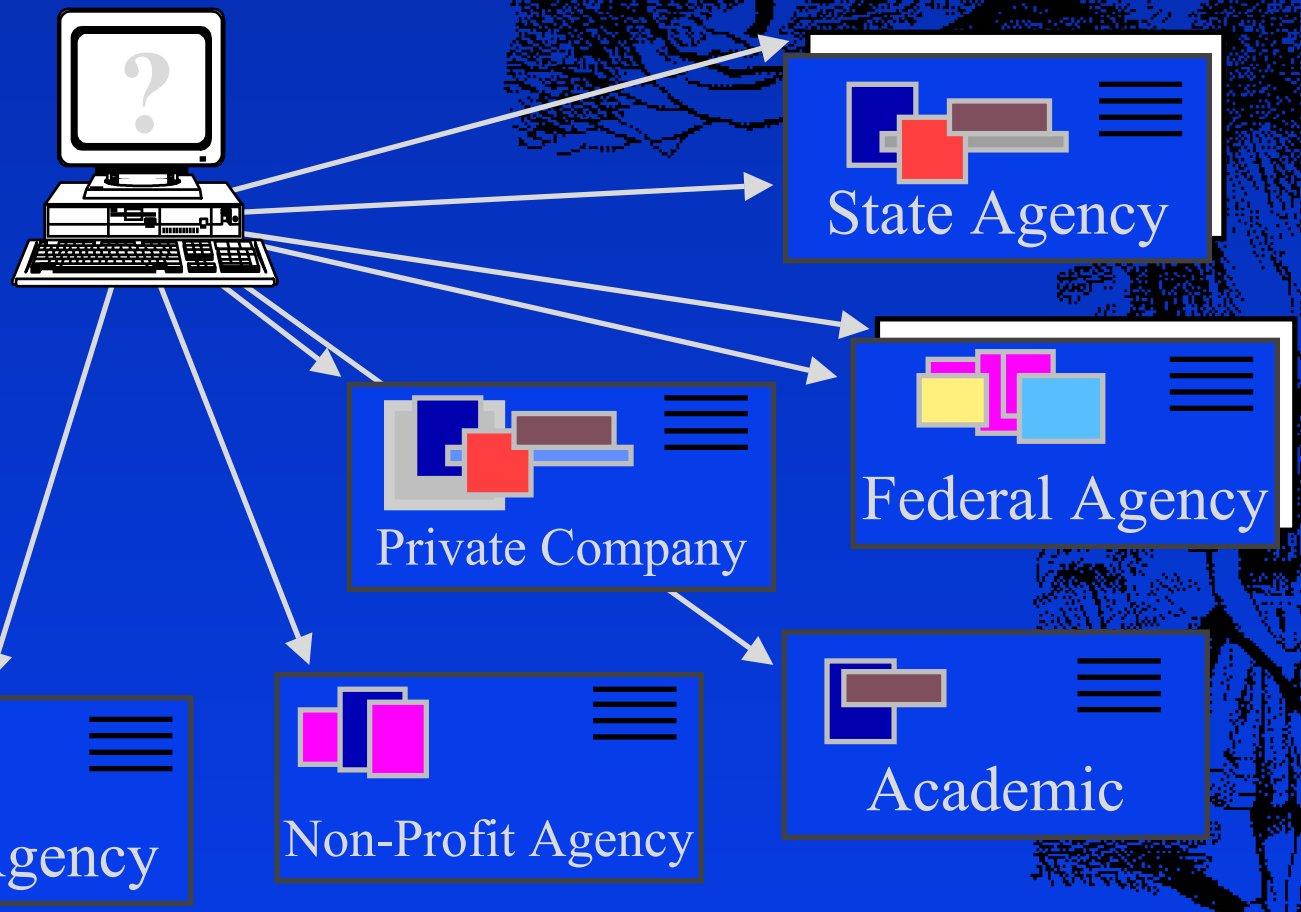
Academic

Local Agency

Non-Profit Agency

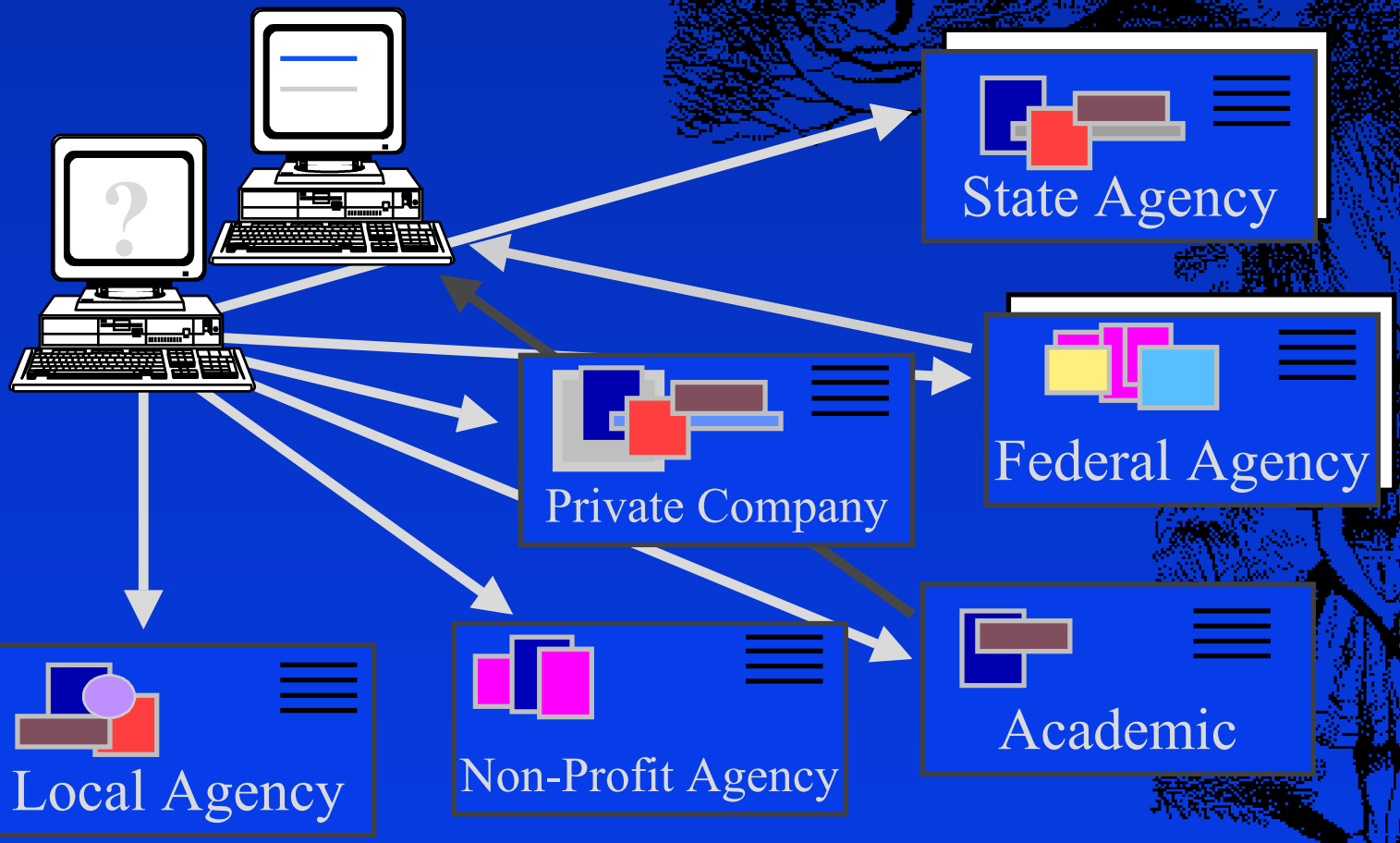
Credit: Doug Nebert

Clearinghouse assumes...



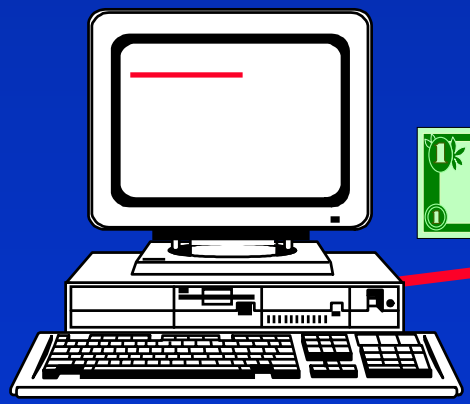
Credit: Doug Nebert

multiple concurrent accesses



Credit: Doug Nebert

Nothing precludes use of Clearinghouse...



State Agency

Federal Agency

Private Company

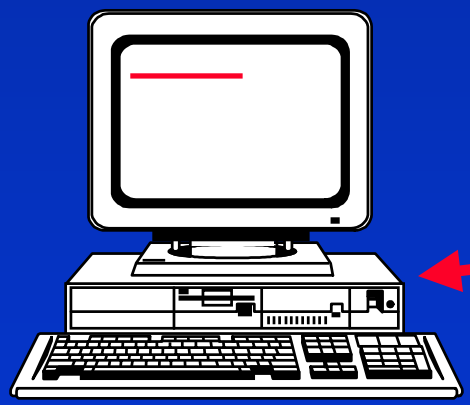
Academic

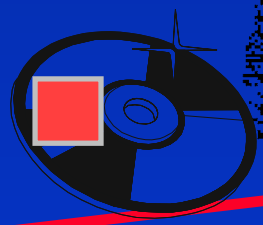
Non-Profit Agency

Local Agency

Credit: Doug Nebert

for order or sale of spatial data





State Agency

Federal Agency

Private Company

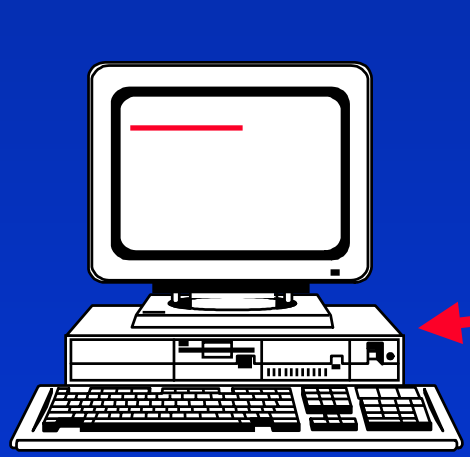
Academic

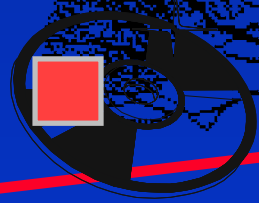
Local Agency

Non-Profit Agency

Credit: Doug Nebert

for regular or special delivery





State Agency

Federal Agency

Private Company

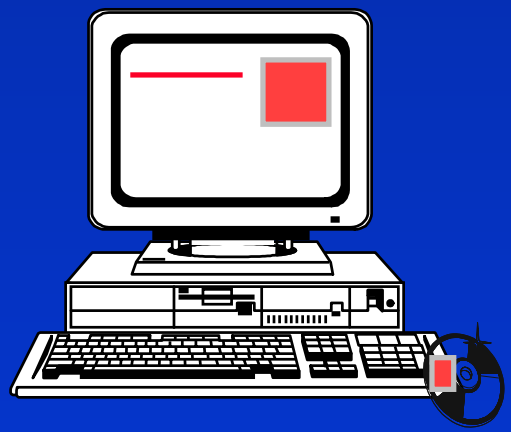
Academic

Non-Profit Agency

Local Agency

Credit: Doug Nebert

CD-ROM complements on-line access



State Agency

Federal Agency

Private Company

Academic

Local Agency

Non-Profit Agency

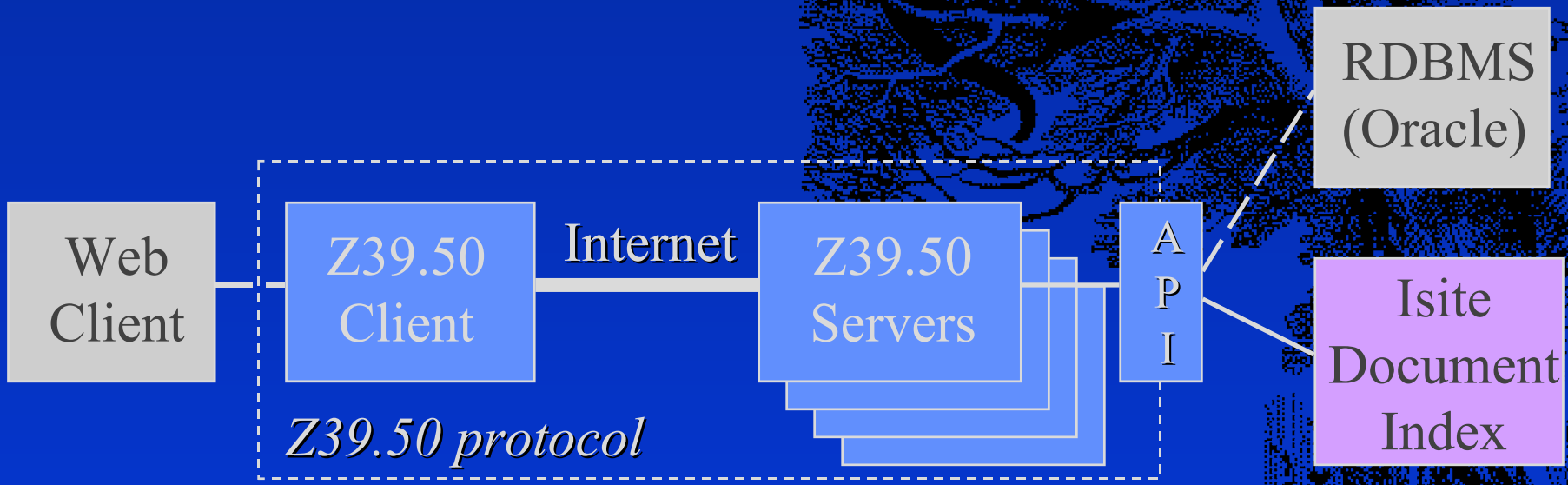
Credit: Doug Nebert

Clearinghouse approach

- Use International voluntary-consensus standards
- Develop free reference implementations and software for public and commercial re-use
- Promote a common vocabulary for geospatial data discovery on the Internet

Credit: Doug Nebert

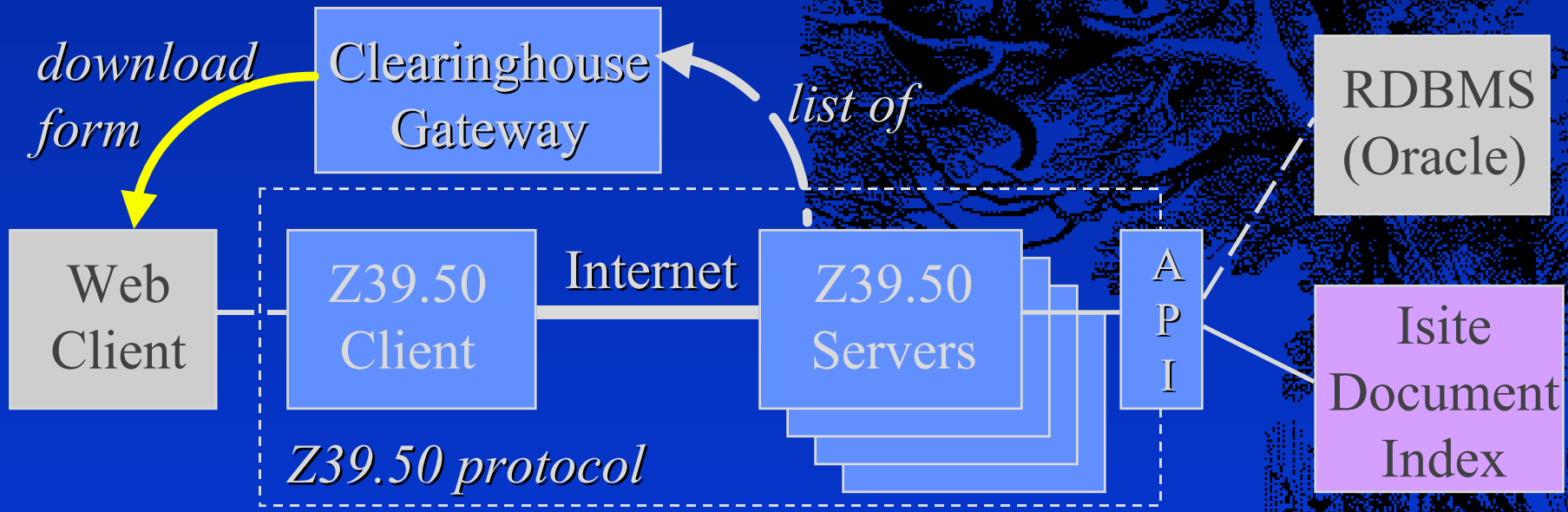
Clearinghouse Implementation



Current Clearinghouse reference implementation allows users to query multiple servers on the Internet using a Web browser. This is accomplished via a Web-HTTP "gateway" on a server on installed at the client.

Credit: Doug Nebert

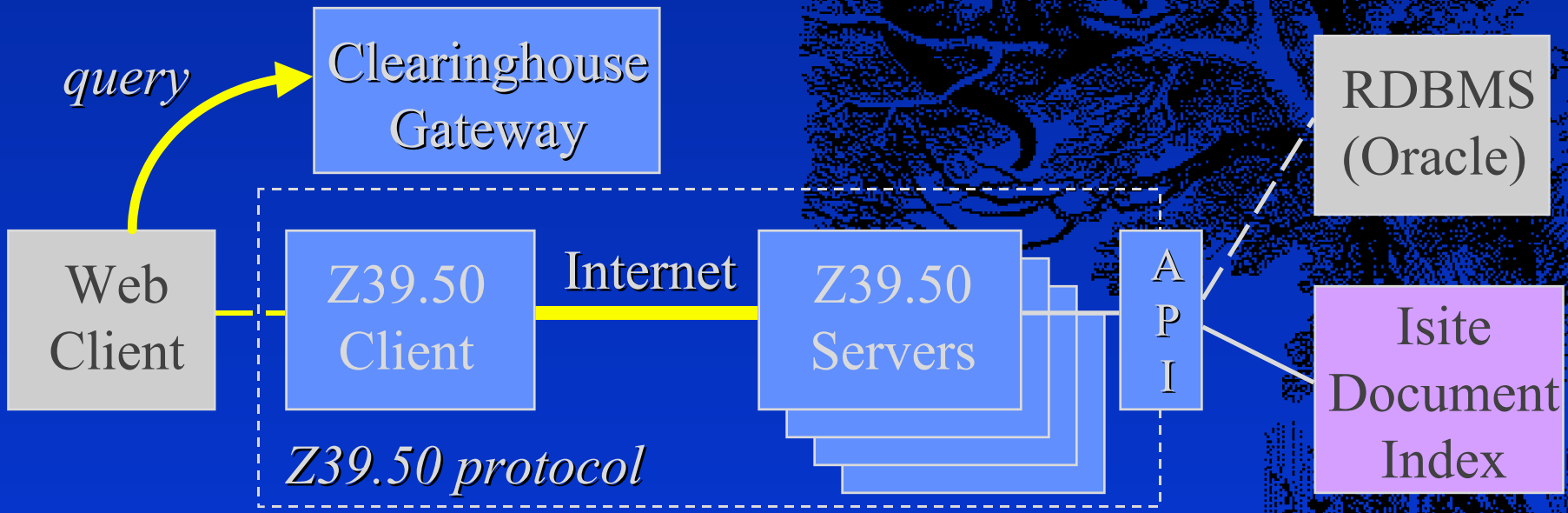
Clearinghouse Implementation



User starts consulting Clearinghouse by downloading a form from one of the gateways that includes all the query fields and a current list of servers maintained by the FGDC.

Credit: Doug Nebert

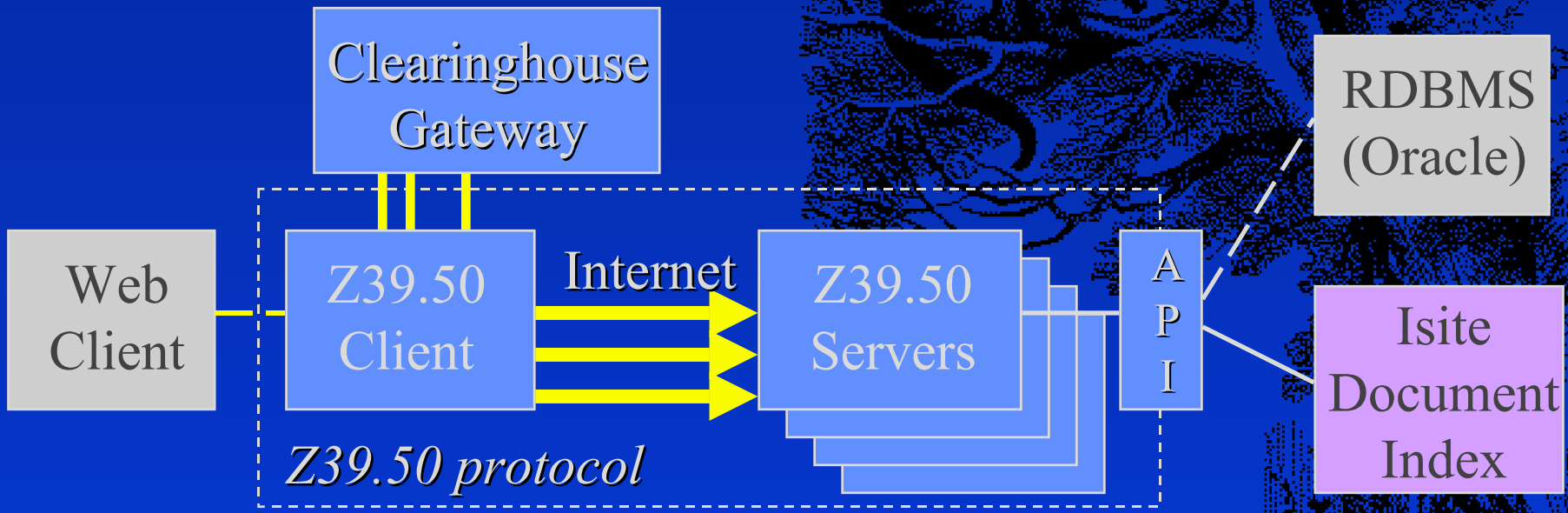
Clearinghouse Implementation



User composes a query in the HTML or Java interface and passes it back through the gateway.

Credit: Doug Nebert

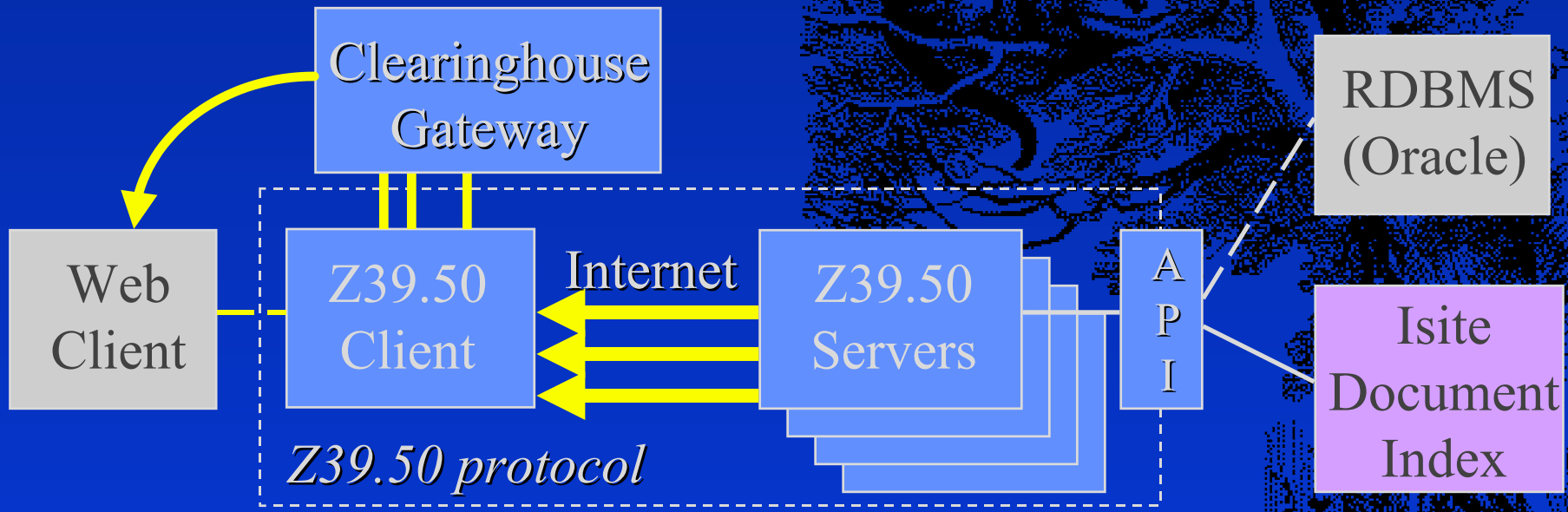
Clearinghouse Implementation



The gateway host spawns connections to the various Z39.50 servers using the client side of the protocol

Credit: Doug Nebert

Clearinghouse Implementation



The servers respond with a set of headlines and pointers to return to the client, all collated into a single set of results presented in HTML. The user selects one entry and can connect to the spatial data by the Online_Linkage element.

Credit: Doug Nebert

Future Developments

- Use of Timemap to display representations of information located in distributed databases
- Interfaces to permit Cheshire Query formulation from inside Timemap
- New Java-based version of Cheshire (Cheshire III).

Further Information

- Full Cheshire II client and server source is available ftp://cheshire.berkeley.edu/pub/cheshire/
  - Includes HTML documentation
- Project Web Site http://cheshire.berkeley.edu/