College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Acquisition

Week 2

LBSC 671

Creating Information Infrastructures

Muddiest Points

- My “5 levels” of metadata
- Moore’s Law
  - And other technical stuff
- Life cycle models

Five Levels of Metadata

- Framework
  - Functional Requirements for Bibliographic Records (FRBR)
- Schema
  - Dublin Core
- Vocabulary
  - Library of Congress Subject Headings (LCSH)
- Representation
  - Resource Description Framework (RDF)
- Serialization
  - RDF in eXtensible Markup Language (RDF/XML)

Adapted from Dante Alighieri, *Comedia* (c. 1321)

Some Metadata Process Examples

Created By

		Human	Machine
Used By	Human	**Indexing**	**Machine-assisted indexing**
	Machine	**HTML metadata field**	**Search engine**

Two Basic Technologies

- Print
  - Physicality closely couples collection and access
  - Cost structure shapes production and use
  - Management of scarcity
- Digital
  - Collection and access are more easily separated
  - Cost structure shapes production and use
  - Management of abundance

Tonight

- Accessioning, appraisal and deaccessioning in archives
- Selection, acquisition and weeding in libraries
- Crawling by Web search engines

Selection and Acquisition Criteria

- LAC [Libraries and Archives Canada] will develop:
  - a comprehensive collection of published Canadiana that documents the published heritage of Canada and materials published elsewhere of interest to Canada, and that supports the creation of a comprehensive national bibliography to make that heritage known and accessible,
  - records holdings sufficient to document the functions and activities of the Government of Canada, and
  - a representative collection of records of heritage value that document the historical development and diversity of Canadian society.

LAC Digital Collection Development Policy, 2006

Some Types of “Archives”

- Government
  - Legal, cultural
- Institutional
  - Liability, institutional memory
- Manuscript repositories
  - Research, preservation

Some Sources for Collections

- Institutional components
  - Transferred from records management
- Donors
  - Typically deed of gift specifies terms
- Purchase

National Archives Records Schedules

Schedule 1. Civilian Personnel Records
Schedule 2. Payrolling and Pay Administration Records
Schedule 3. Procurement, Supply, and Grant Records
Schedule 4. Property Disposal Records
Schedule 5. Budget Preparation, Presentation, and Apportionment Records
Schedule 6. Accountable Officers' Accounts Records
Schedule 7. Expenditure Accounting Records
Schedule 8. Stores, Plant, and Cost Accounting Records
Schedule 9. Travel and Transportation Records
Schedule 10. Motor Vehicle Maintenance and Operations Records
Schedule 11. Space and Maintenance Records
Schedule 12. Communications Records
Schedule 13. Printing, Binding, Duplication, and Distribution Records
Schedule 14. Information Services Records
Schedule 15. Housing Records
Schedule 16. Administrative Management Records
Schedule 17. Cartographic, Aerial Photographic, Architectural, and Engineering Records
Schedule 18. Security and Protective Services Records
Schedule 20. Electronic Records
Schedule 21. Audiovisual Records
Schedule 23. Records Common to Most Offices Within Agencies
Schedule 24. Information Technology Operations and Management Records
Schedule 25. Ethics Program Records
Schedule 26. Temporary Commissions, Boards, Councils and Committees
Schedule 27. Records of the Chief Information Officer

Collection Development Policies

- **Mission**
  - – Intended (“statement of purpose”): 92%
  - – Emergent (“strengths of holdings”): 53%
- **Scope**
  - – Subject: 84%
  - – Geographic: 84%
  - – Time frame: 57%
- **Anticipated use**
  - – Users: 59%
  - – Activities: 53%

Cynthia Sauer, Ding the Best We Can, (2001)

Basis for Exceptions

- Donor relationship: 70%
- Implicit broadening of scope
  - Risk of destruction:` 49%
  - Exceptional opportunity: 30%
- Prestige
  - Publicity value: 15%
  - Attract future resources: 12%
  - Institutional competition: 6%

Cynthia Sauer, Ding the Best We Can, (2001)

Evolutionary Policy

- Envision
  - Available materials, future use, existing alternatives
- React
  - Establish decision basis for individual cases
- Evolve
  - Changing mission, resources, opportunities, pressures

Codify

- Decide which parts to put in writing (and why!)

Why Codify?

- Develop shared vision with stakeholders
  - Keep resources in line with requirements
  - Minimize unintended policy drift
- Facilitate appropriate donations
  - Solicit in-scope donations
  - Communicate limitations to donors
- Facilitate referrals
- Foster continuity in the decision process

Appraisal

- Value
  - Evidential
  - Informational
- Costs
  - Storage, arrangement, description, preservation, …
- Stakeholder interests
  - Primary: Institutional needs
  - Primary: Accountability
  - Secondary: Other future record users

Deaccessioning

- Space limits
- Policy changes
- Technology changes

Tonight

- Accessioning, appraisal and deaccessioning in archives

➢ Selection, acquisition and weeding in libraries

- Crawling by Web search engines

A Collection Development Policy

Customer use is the most powerful influence on the Library’s collection. …The other driving force is the Library’s strategic plan.

… selections are made to provide depth and diversity of viewpoints to the existing collection and to build the world-class Western History/Genealogy and African American Research Library collections. …

… The Library provides materials to support each individual’s journey, and does not place a value on one customer’s needs or preferences over another’s. …

Materials for children and teenagers are intended to broaden their vision, support recreational reading …

Denver Public Library, 2012

Why Libraries Collect

- Access
  - Current users
  - Future users
  - Social responsibility

- Prestige

Selection

- Scope
  - Demographics, research focus, …
- Quality metrics
  - Publisher, author, impact factor, …
- Practical factors
  - Cost, language, availability elsewhere, …
- Use
  - Circulation, inter-library loan, requests, …

Publishing Infrastructure

- Publishers
  - Intermediation on behalf authors
- Vendors
  - Intermediation on behalf of libraries
  - Value added services
    - Electronic Data Interchange (EDI)
    - Stock profiles (on approval)
    - Shelf-ready books

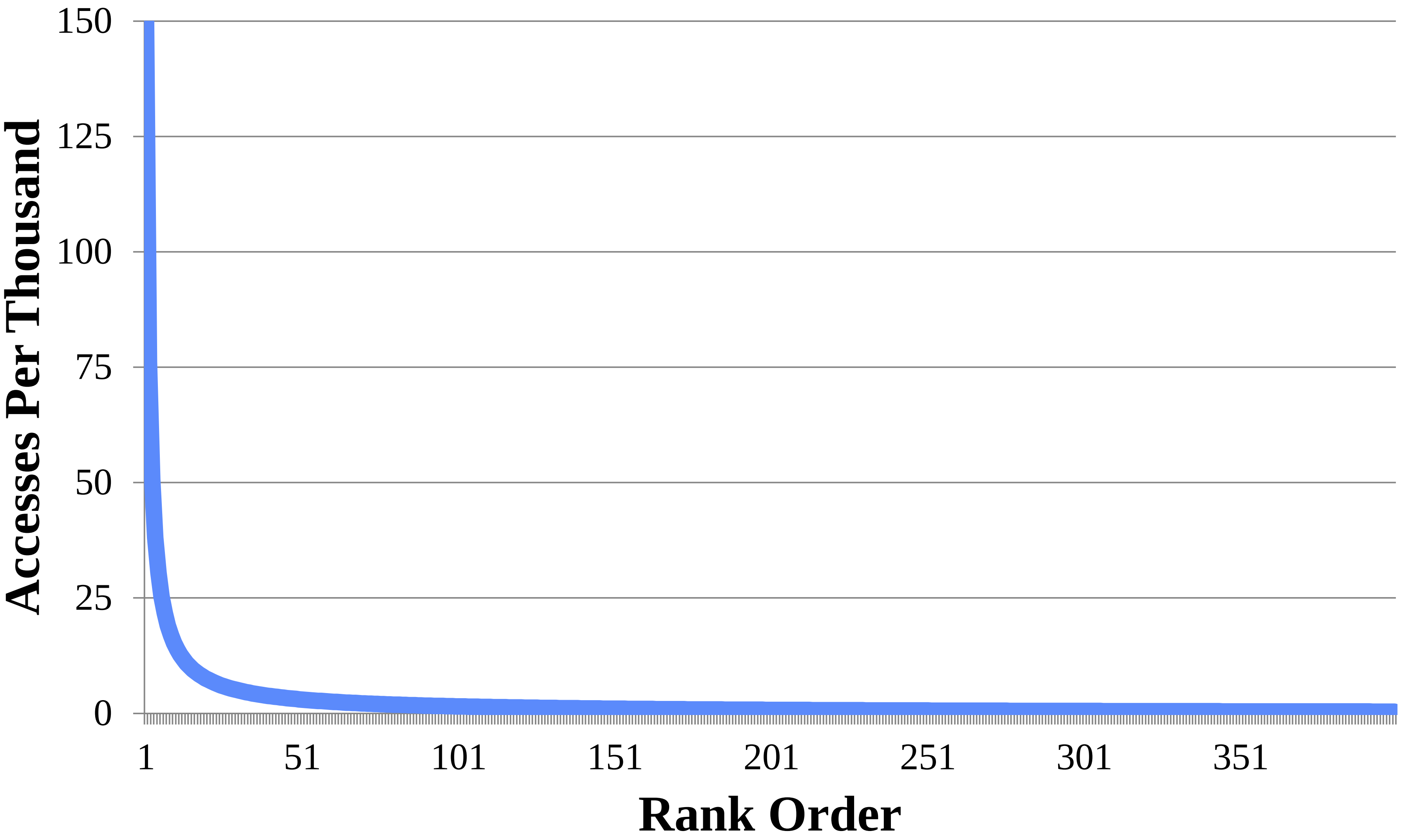
Access models

- Ownership (“just in case”)
  - Unlimited use for an unlimited period
  - Right of first sale vs. license restrictions
- Subscription
  - Unlimited (or limited) use for a defined period
  - Single vs. multiple users
- Pay-per-view (“just in time”)

Use-Driven Acquisition

- Online catalog includes unpurchased items
- First few access requests cause rental each time
- Next request results in unlimited-use subscription (or ownership)
- Transfers some risk to vendor
  - Lowers cost of low-use items
  - Somewhat raise cost of high use items

Zipf's Law


150
125
100
75
50
25
0
Accesses Per Thousand
1
51
101
151
201
251
301
351
Rank Order

The “Big Deal”

- Bundled access (usually to serials)
  - Vendor goal: cross-sell lower-demand items
  - Incentive: Access to much more content
    - Sometimes with some delay (e.g., 1 year)
- Risks:
  - Future access to subscription content
  - Future price increases

Open Access

- Self-archiving
  - Personal Web sites
  - Institutional repositories

- Publishing
  - Author pays
  - Volunteer labor

Weeding (“Library Hygiene”)

- Presumes some limited asset
  - e.g., shelf space, browsing time, …
- Anticipated future use
  - Reshelving and circulation statistics
  - Historical value
  - Sufficiency of single copies
  - Last copy doctrine
- Condition
  - Preservation costs

Tonight

- Accessioning, appraisal and deaccessioning in archives

- Selection, acquisition and weeding in libraries

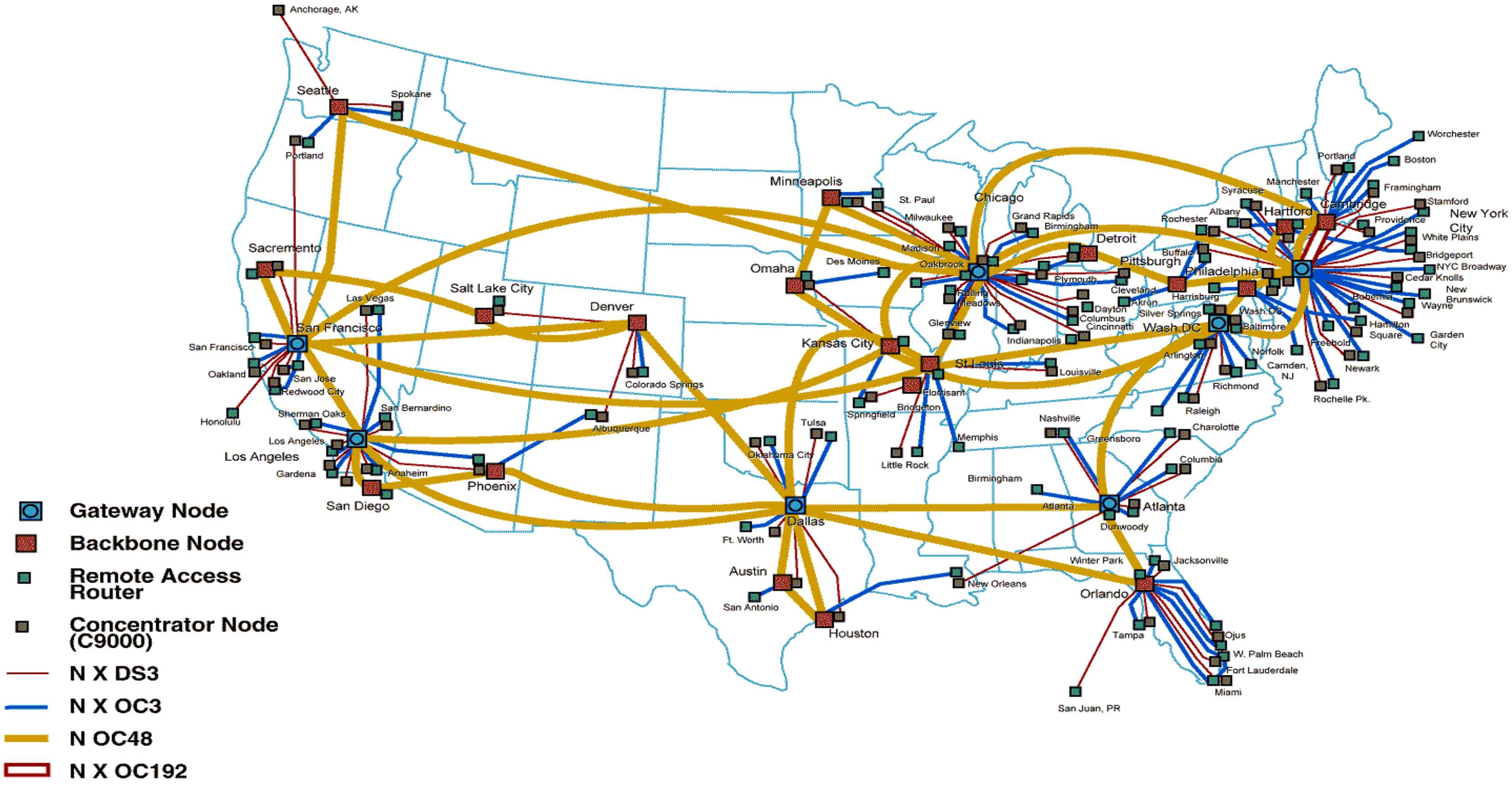
➢ Crawling by Web search engines

The Internet

AT&T

AT&T IP BACKBONE NETWORK

2Q2000

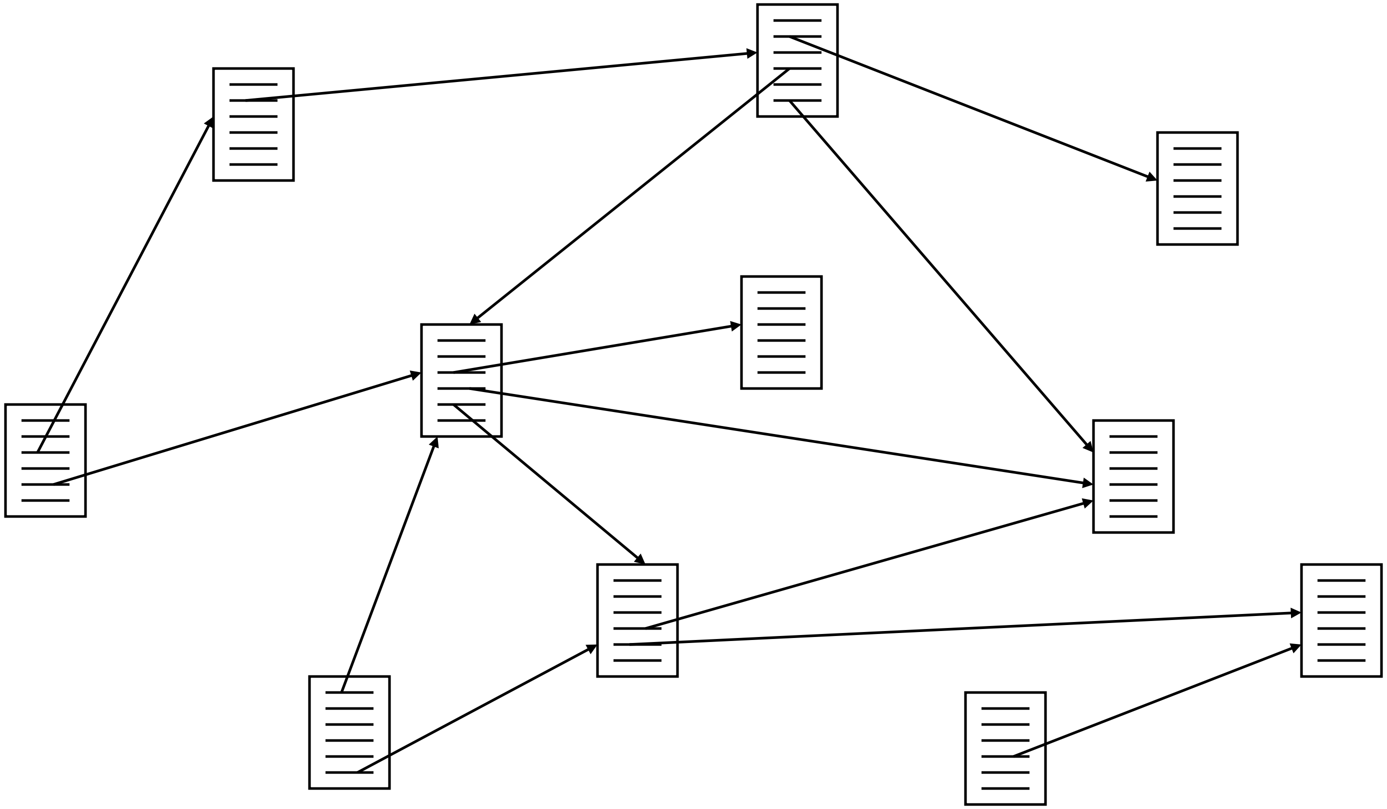


Note: map is not to scale.

The Web

- The Protocols
  - – Uniform Resource Locator (URL)
  - – Hypertext Markup Language (HTML)
  - – Hypertext Transport Protocol (HTTP)
- Content types
  - – Static, dynamic, streaming, transactional
- Access
  - – Public, protected, or intranet?

Crawling the Web



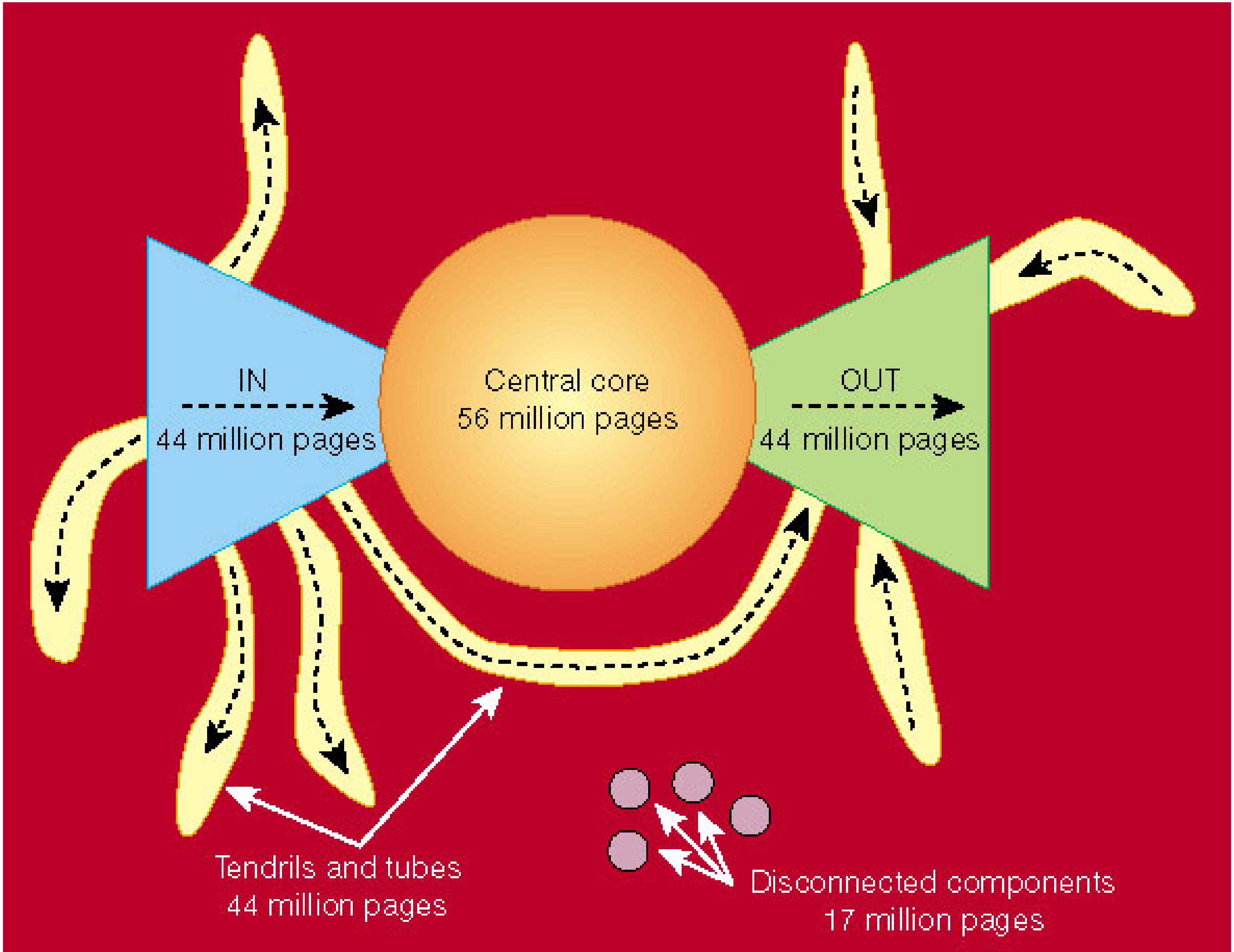
Robots Exclusion Protocol

- Requires voluntary compliance by crawlers
- Exclusion by site
  - Create a robots.txt file at the server's top level
  - Indicate which directories not to crawl
- Exclusion by document (in HTML head)
  - Not implemented by all crawlers

    <meta name="robots“ content="noindex,nofollow">

Link Structure of the Web

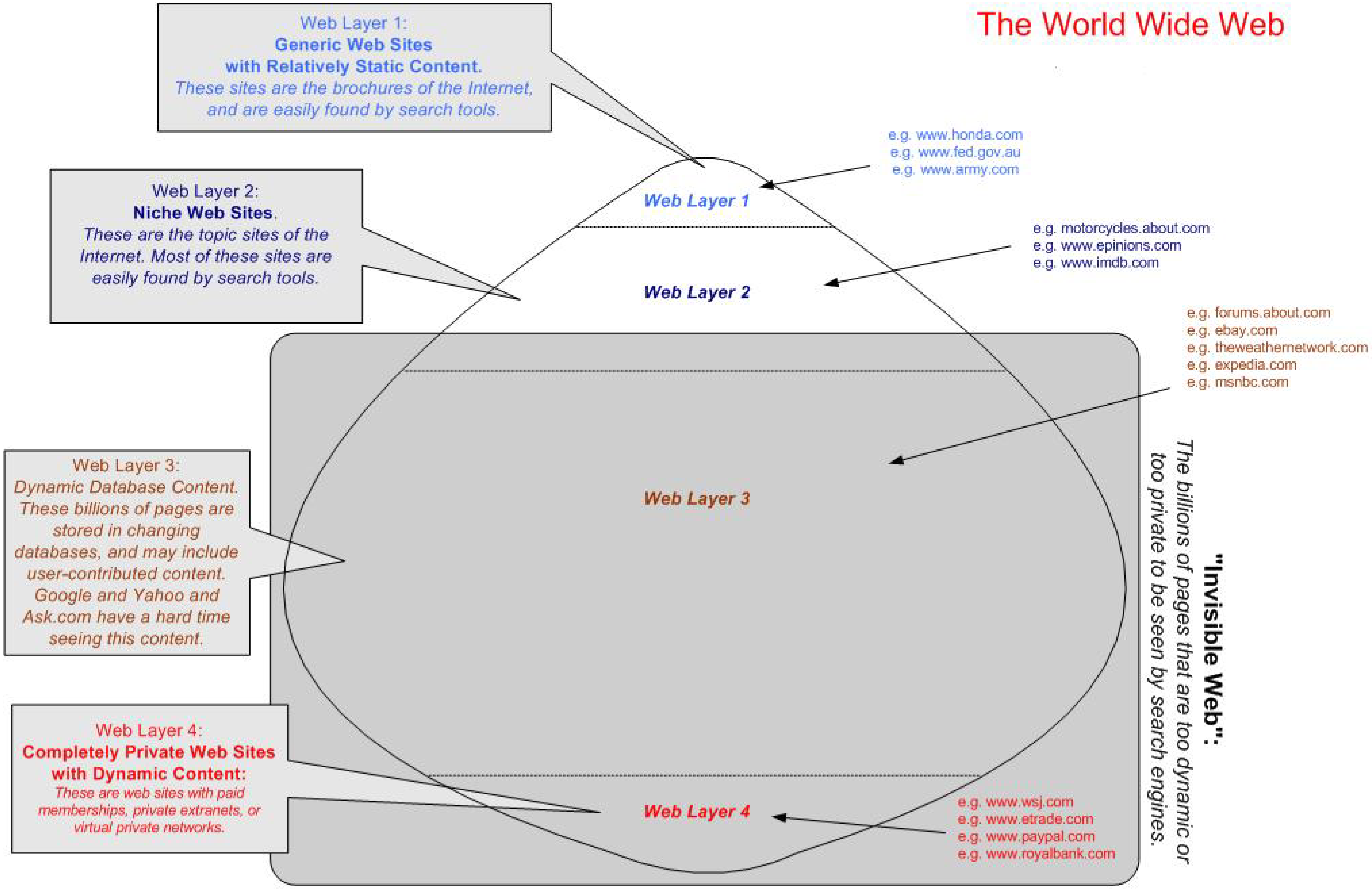
Nature **405**, 113 (11 May 2000) | **doi:10.1038/35012155**



Web Crawl Challenges

- Discovering “islands” and “peninsulas”
- Duplicate and near-duplicate content
  - 30-40% of total content
- Link rot
  - Changes at ~1% per week
- Network instability
  - Temporary server interruptions
  - Server and network loads
- Dynamic content generation

The World Wide Web



Web Layer 1:
Generic Web Sites
with Relatively Static Content.
These sites are the brochures of the Internet, and are easily found by search tools.

Web Layer 2:
Niche Web Sites.
These are the topic sites of the Internet. Most of these sites are easily found by search tools.

Web Layer 3:
Dynamic Database Content. These billions of pages are stored in changing databases, and may include user-contributed content. Google and Yahoo and Ask.com have a hard time seeing this content.

Web Layer 4:
Completely Private Web Sites with Dynamic Content:
These are web sites with paid memberships, private extranets, or virtual private networks.

e.g. www.honda.com
e.g. www.fed.gov.au
e.g. www.army.com

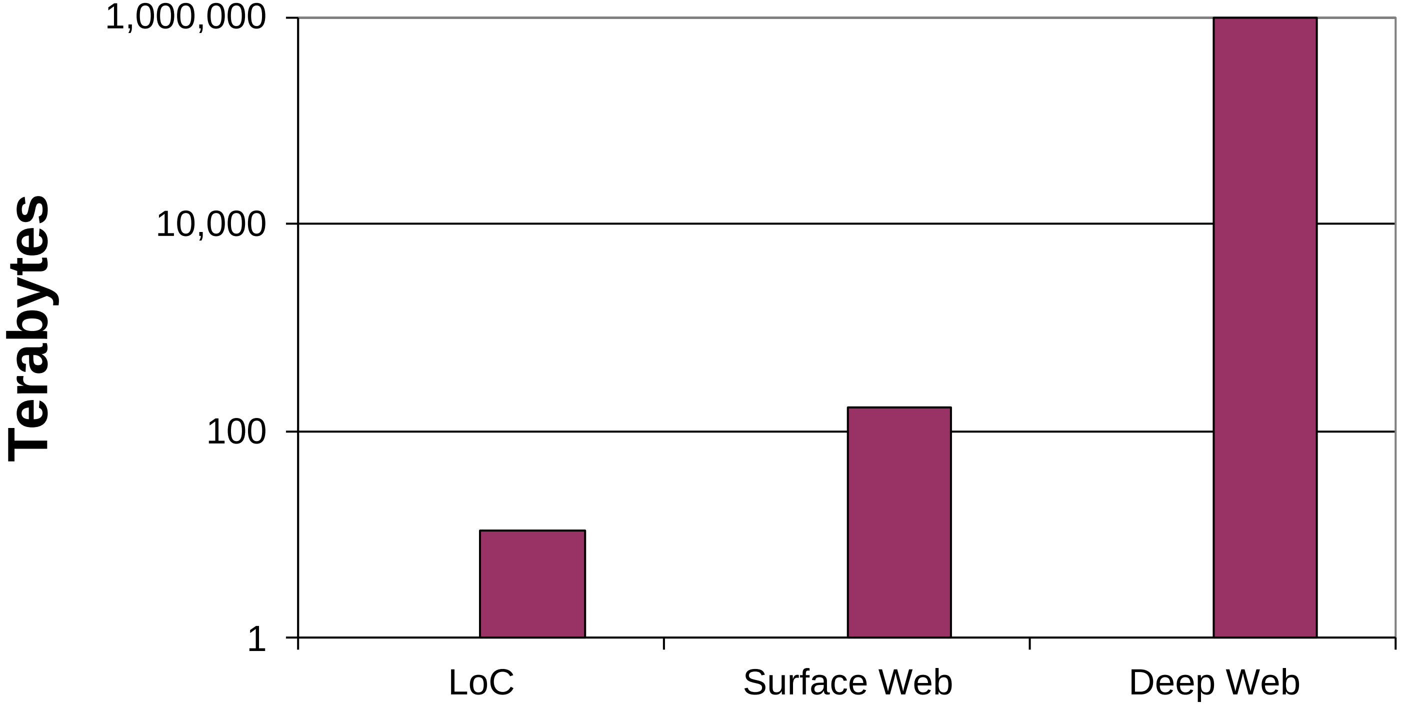
e.g. motorcycles.about.com
e.g. www.epinions.com
e.g. www.imdb.com

e.g. forums.about.com
e.g. ebay.com
e.g. theweathernetwork.com
e.g. expedia.com
e.g. msnbc.com

e.g. www.wsj.com
e.g. www.etrade.com
e.g. www.paypal.com
e.g. www.royalbank.com

"Invisible Web":
The billions of pages that are too dynamic or too private to be seen by search engines.

The “Deep Web”



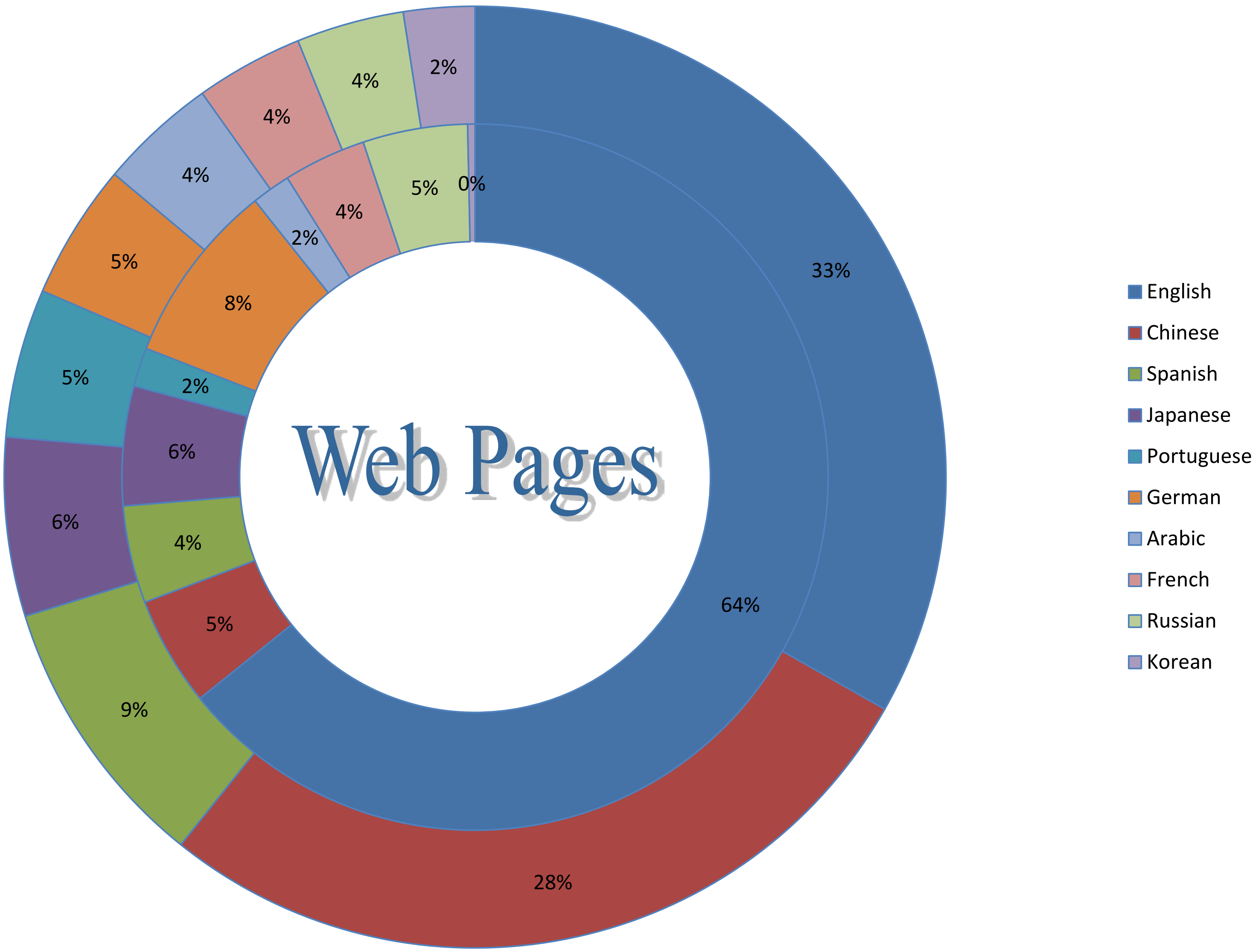
Estimates for 2008

Hands on:
The Internet Archive

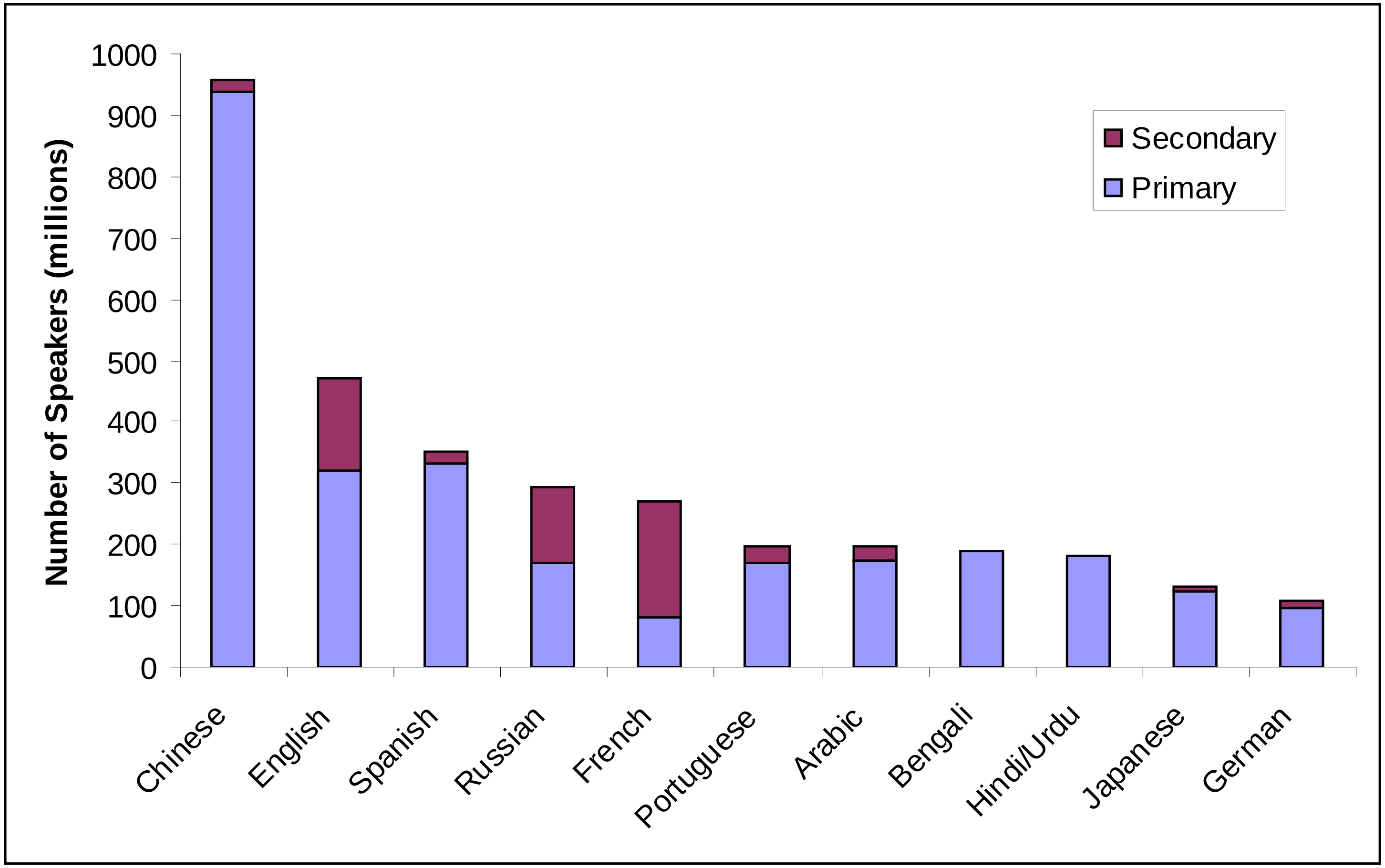
- alexa.com Web crawls since 1997
  - http://archive.org

- Check out the iSchool’s Web site from 1998!
  - http://www.clis.umd.edu

Global Internet Users



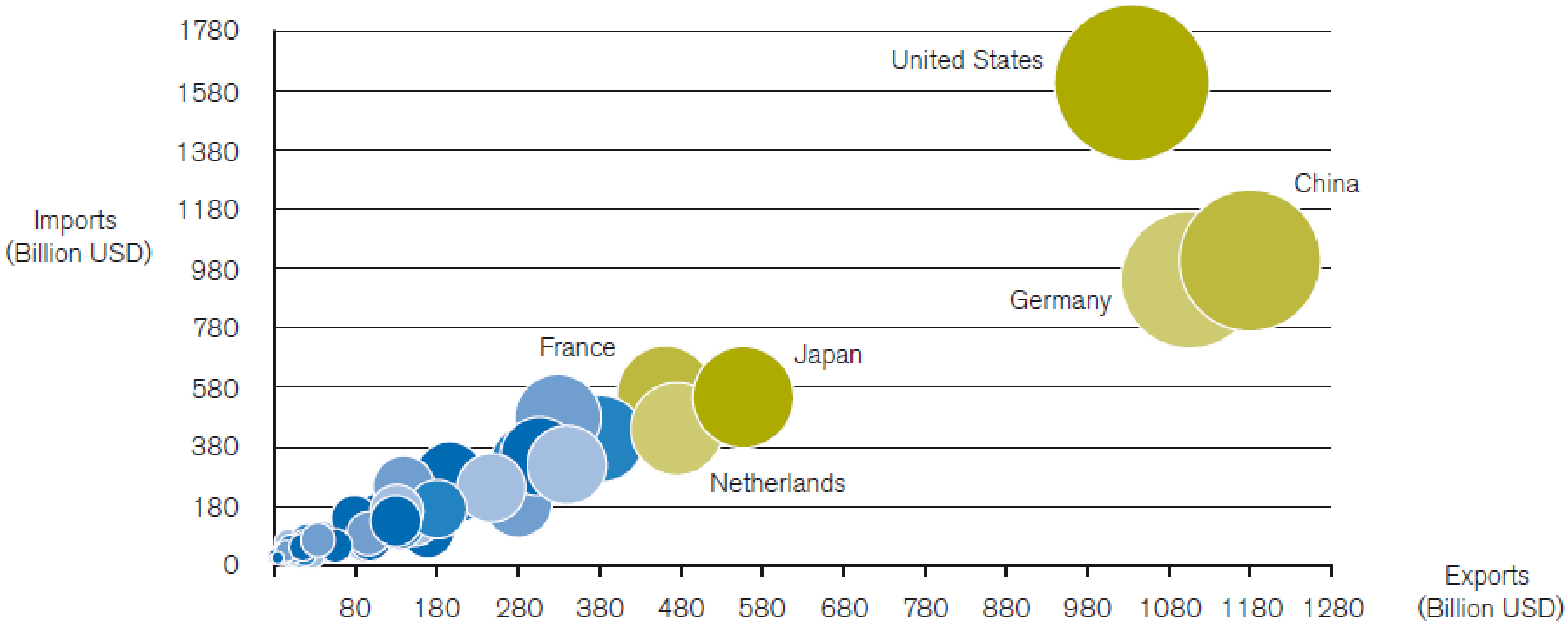
Most Widely-Spoken Languages



Source: Ethnologue (SIL), 1999

Global Trade

Leading economies of merchandise trade, 2009



Source: World Trade Organization 2010 Annual Report

Homework G3

- Life Cycle Analysis of your collection
  - Choose no more than 5 content types
- Creation
- Use
- Evolution
- Disposition

Before You Go

On a sheet of paper, answer the following (ungraded) question (no names, please):

What was the muddiest point in today’s class?