College of Information Studies

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

Web Characterization
Web Design

Week 3

LBSC 690

Information Technology

Why is there a Web?

- Affordable storage
  - 300,000 words/$ in 1995
- Adequate backbone capacity
  - 25,000 simultaneous transfers in 1995
- Adequate “last mile” bandwidth
  - 1 second/screen in 1995
- Display capability
  - 10% of US population in 1995
- Effective search capabilities
  - Lycos and Yahoo were started in 1995

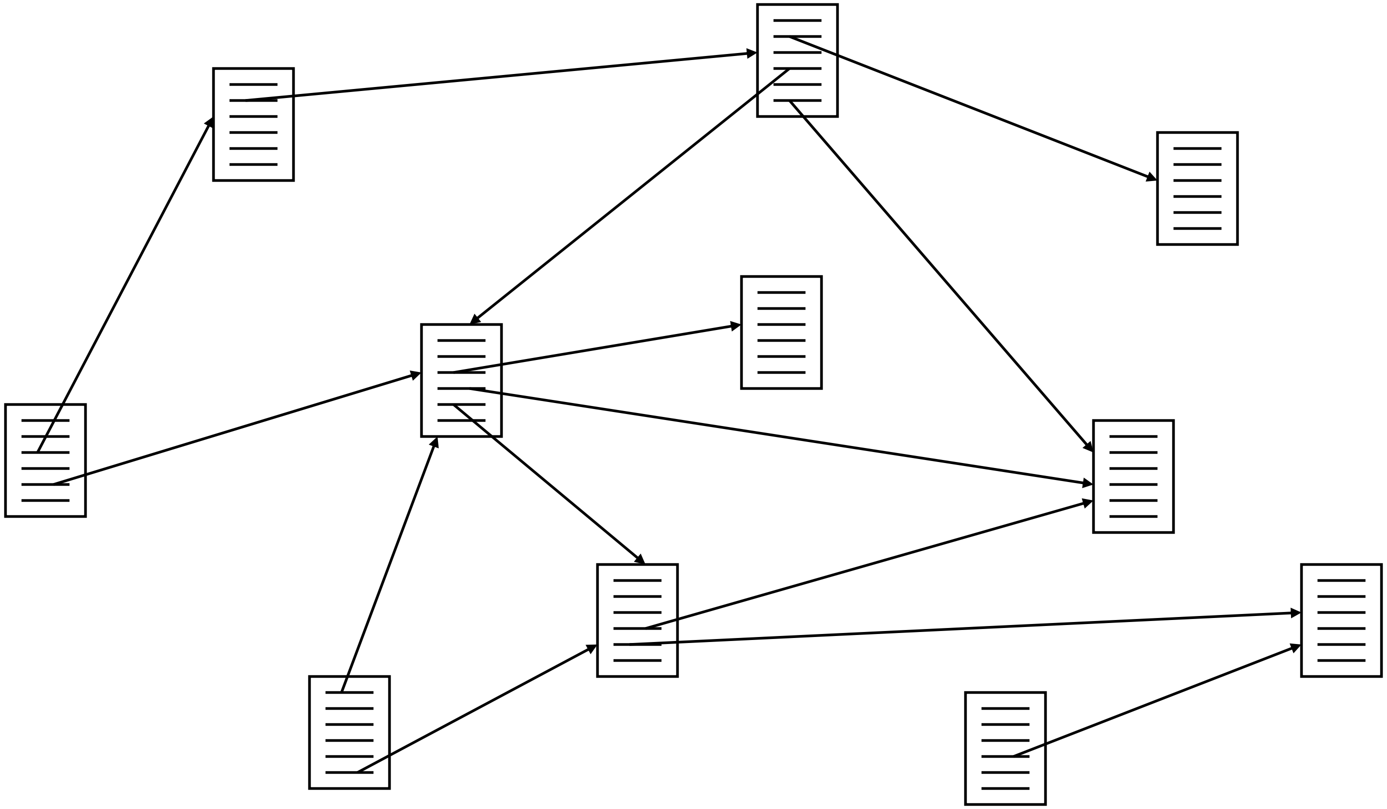
What is the Web?

- Protocols
  - HTTP, HTML, or URL?
- Perspective
  - Content or behavior?
- Content
  - Static, dynamic or streaming?
- Access
  - Public, protected, or internal?

Some Perspectives

- Web “sites”
  - In 2002, OCLC counted any server at port 80
  - Total was 3 million, an undercount
    - Misses many servers at other ports
    - Some servers host unrelated content (e.g., TerpConnect)
    - Some content requires specialized servers (e.g., rtsp)
- Web “pages”
  - In 2012, Google counted any URL it has seen
  - Total was 30 trillion, an overcount
    - Includes dead links, spam, …
- Web “use”
  - Google users pose 3 billion queries a day

Crawling the Web



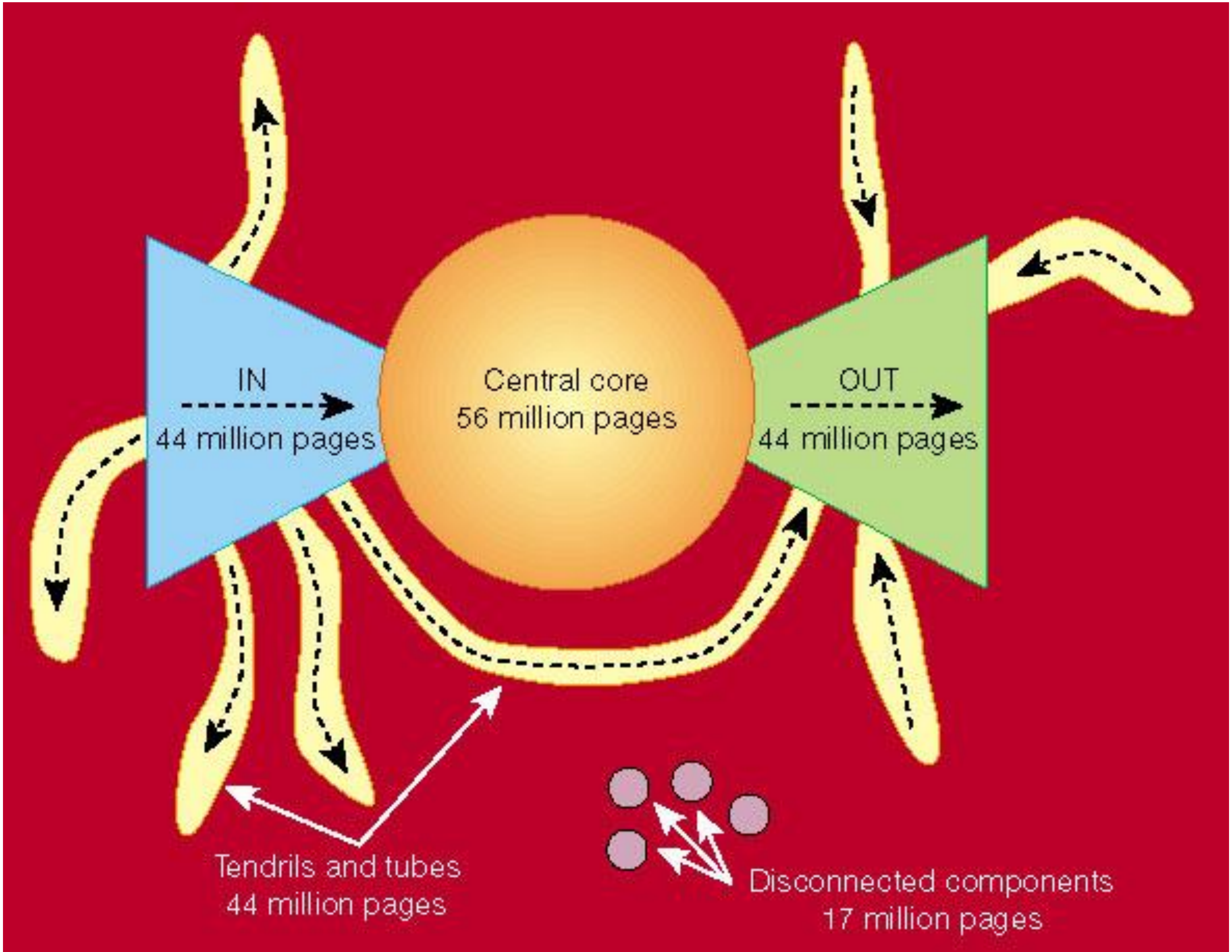
Robots Exclusion Protocol

- 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

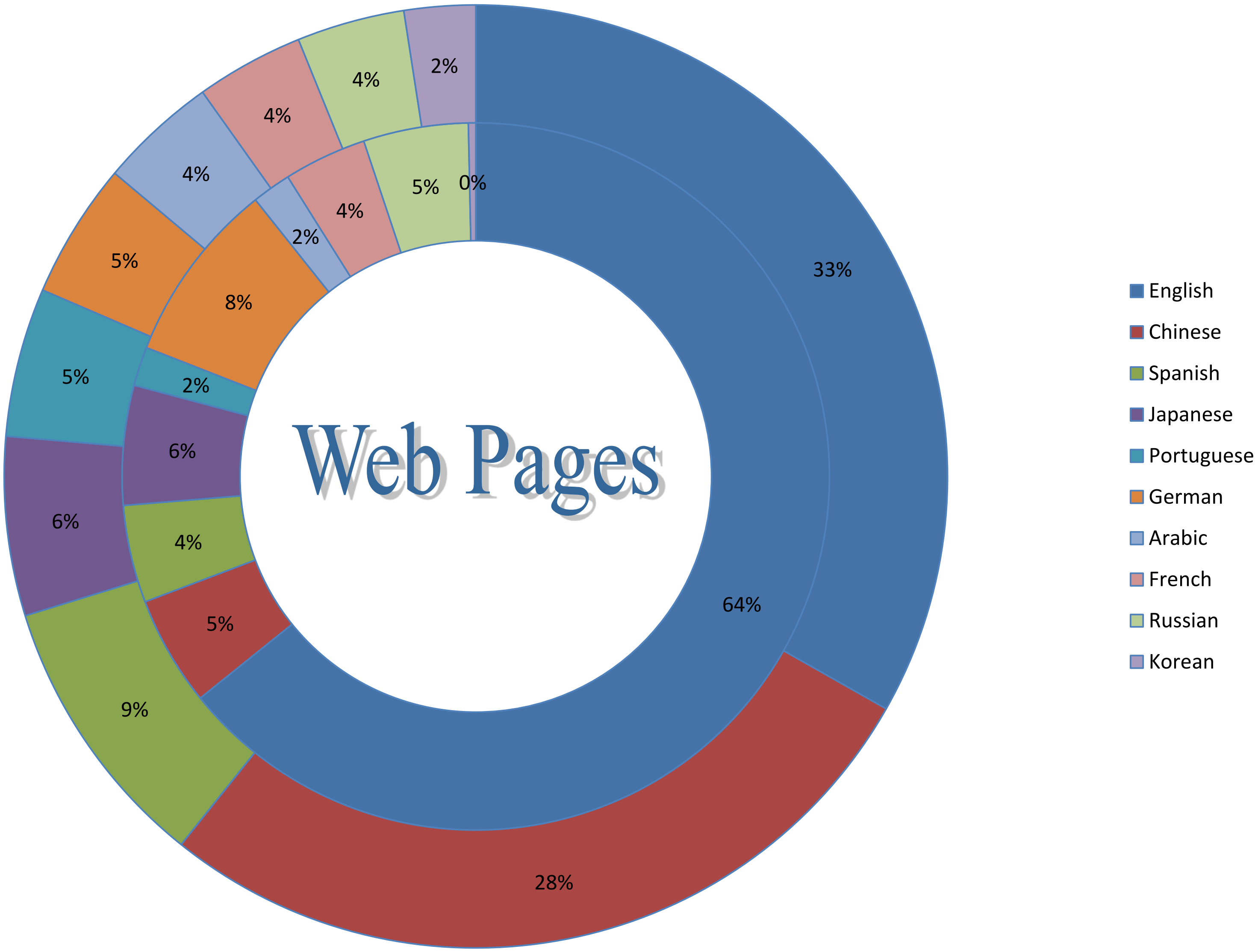
Duplicate Detection

- Structural
  - Identical directory structure (e.g., mirrors, aliases)
- Syntactic
  - Identical bytes
  - Identical markup (HTML, XML, …)
- Semantic
  - Identical content
  - Similar content (e.g., with a different banner ad)
  - Related content (e.g., translated)

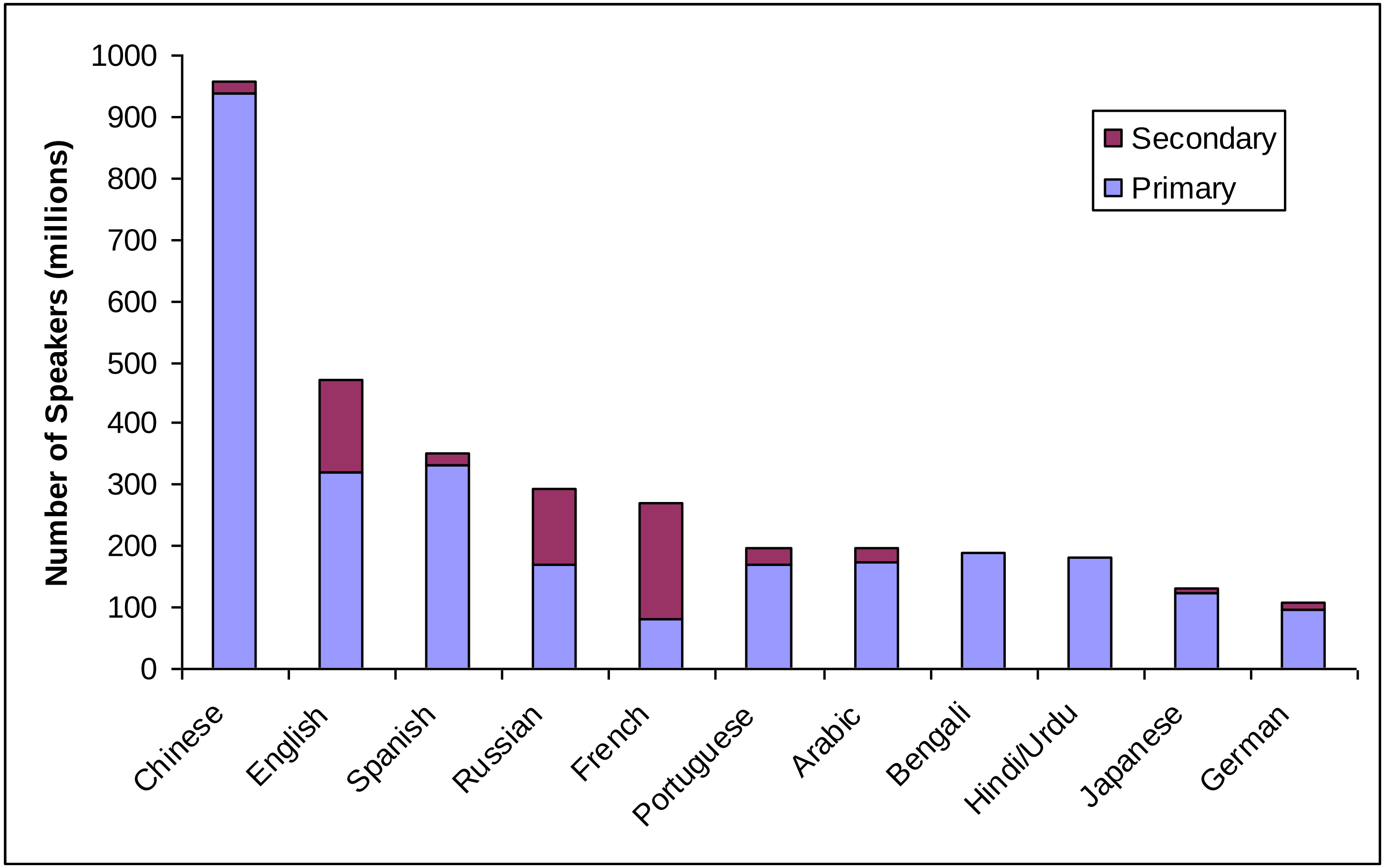
Hands on: The Internet Archive

- 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



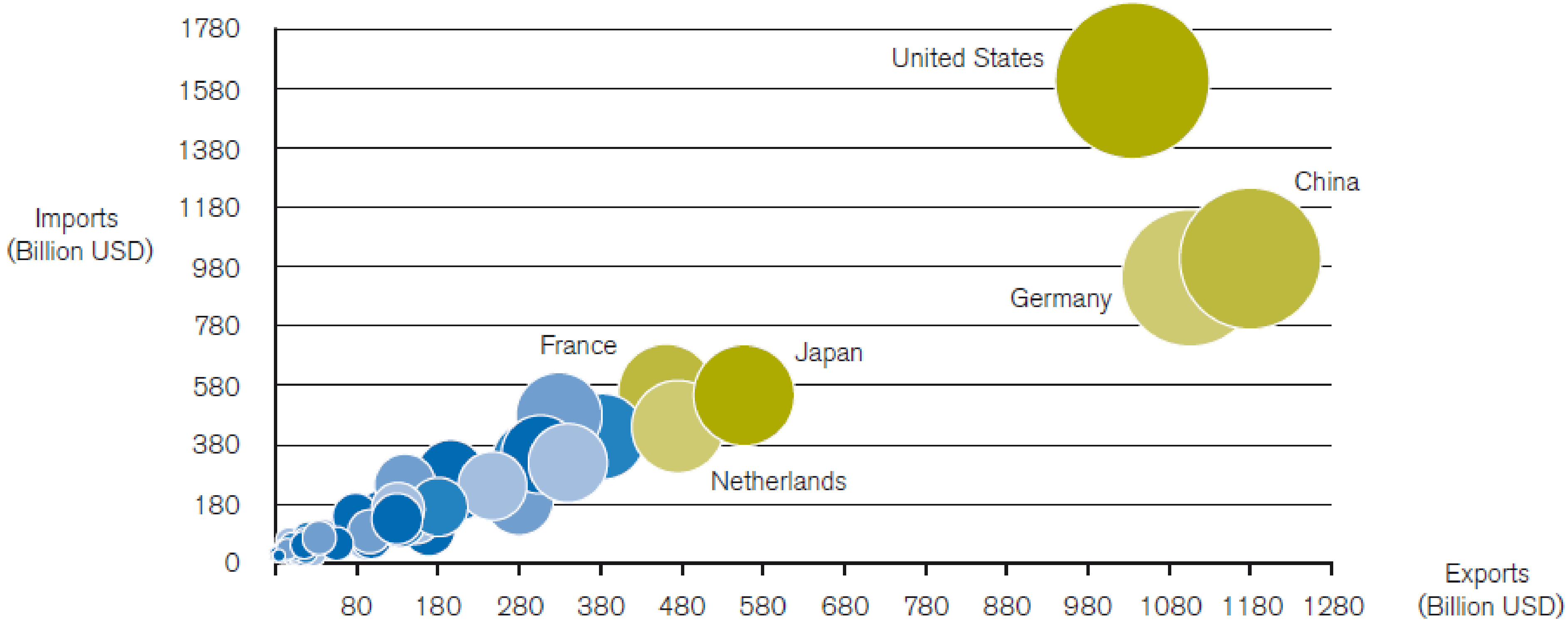
Most Widely-Spoken Languages



Source: Ethnologue (SIL), 1999

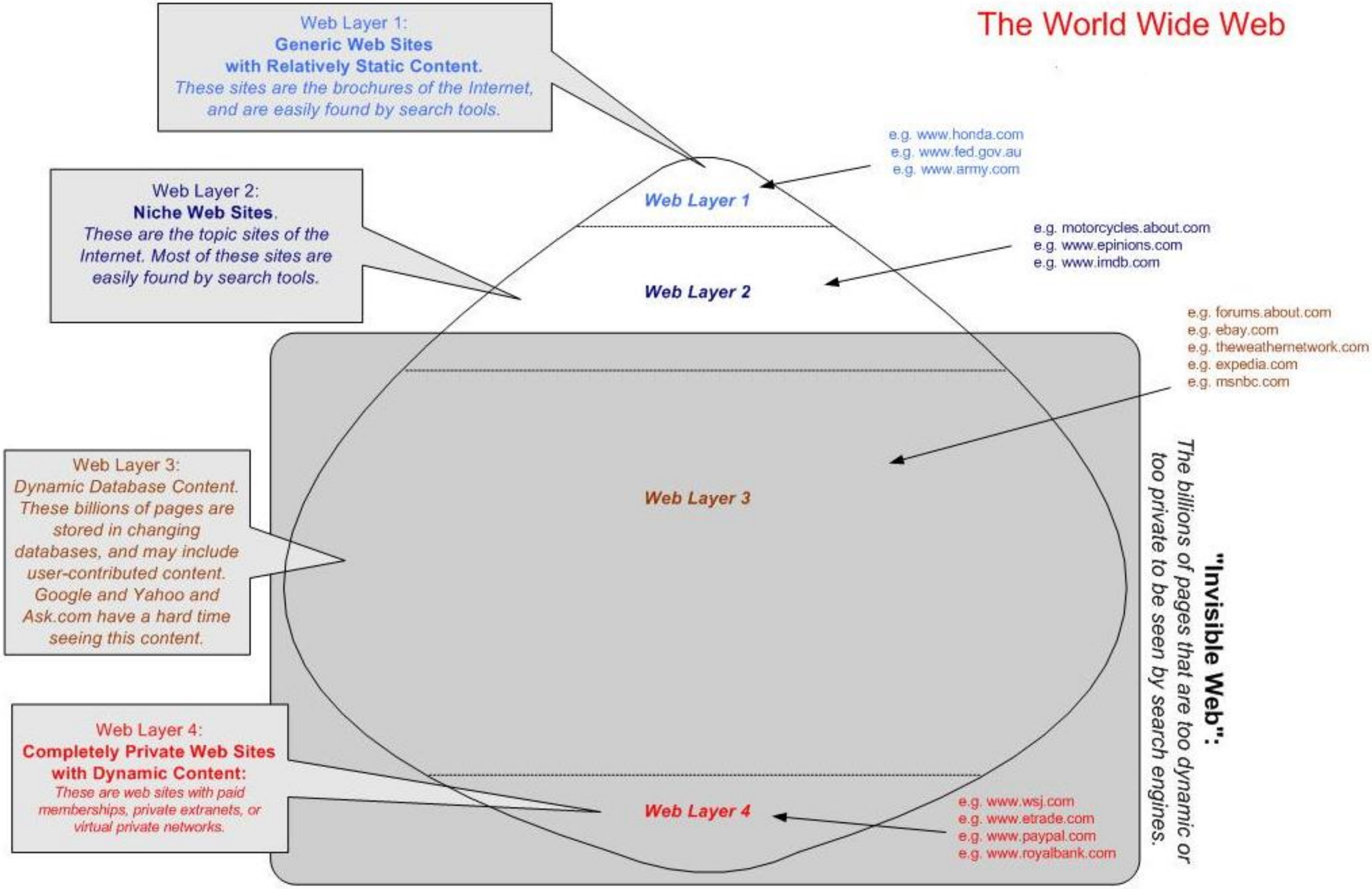
Global Trade

Leading economies of merchandise trade, 2009



Source: World Trade Organization 2010 Annual Report

The “Deep Web”



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.

Web Layer 1

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

Web Layer 2

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

Web Layer 3

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

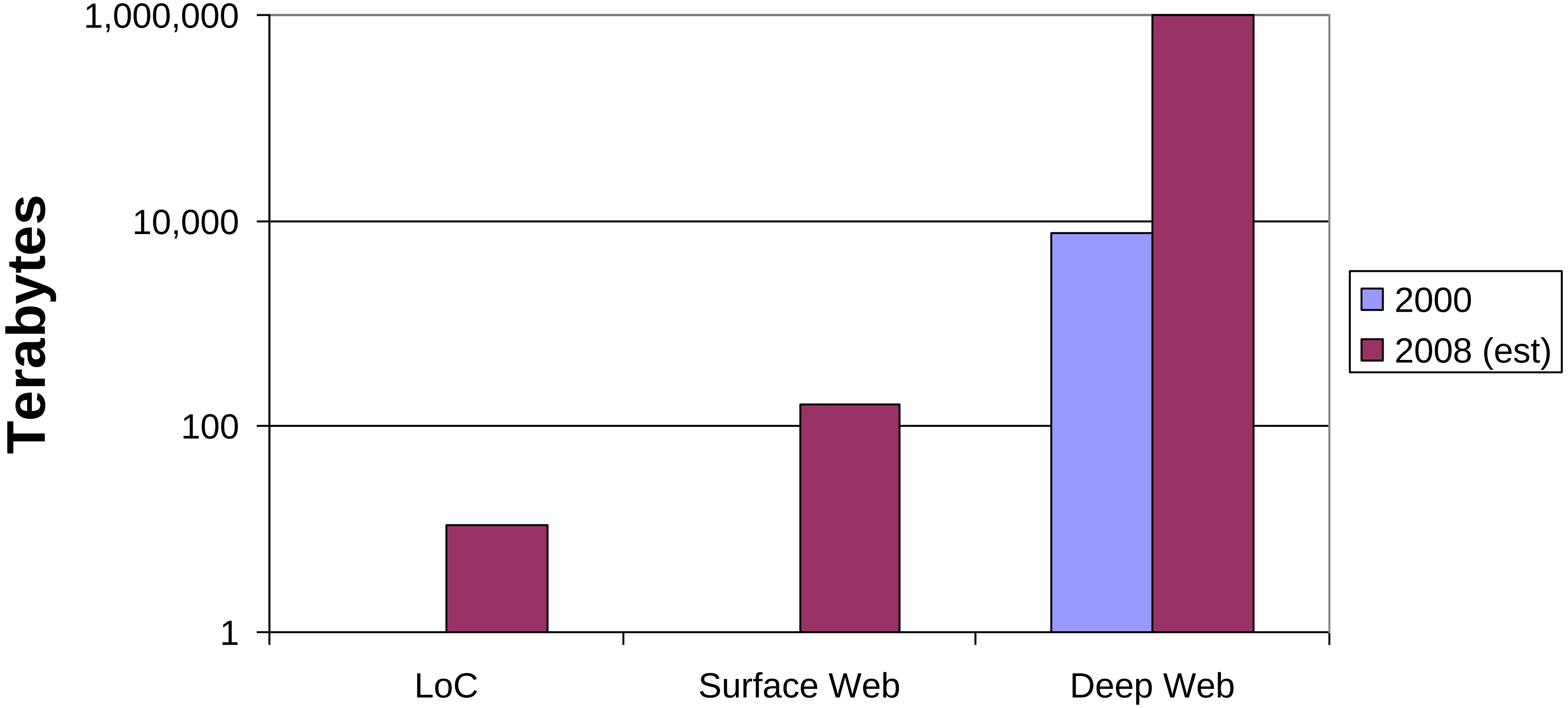
Web Layer 4

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”

- Dynamic pages, generated from databases
- Much larger than surface Web
- Not easily discovered using crawling


Terabytes
LoC
Surface Web
Deep Web
2008 (est)