Whither Digital Libraries?

The case of a “billion-dollar” business

Yi-Tzuu Chien
School of Library and Information Science
University of Tsukuba
ytchien@slis.tsukuba.ac.jp

October 31, 2002

Outline

- Need for a business model
- Vision of digital libraries: then and now
- Making e-contents accessible, useful and profitable: Reversing the steps of research-to-applications paradigm
- An example in digital government: Turning government into a business partner and research investor
- Connections to the Knowledge Society

Digital Library

(Circa 1994)

Vision – then and now

- ***A digital network of knowledge systems*** - connecting computing, information, and people resources
- ***A set of enabling technologies*** - for creating, distributing, and using knowledge in human-centered multimedia, multi-modal environments
- ***New information services*** - in networked education, commerce, health care, transportation, government, and others, beyond those provided by traditional libraries and information sources
- ***Ubiquitous, public, and personal*** **–** open 24 hours and is accessible where the network is

DL Roadblocks

- How much information? Production outpaces consumption
- Lack of a business model and incentives for making public e-contents accessible
- Research focuses on technological innovation, not on user needs
- Commercial success in non-public domains (music, games, etc.) overshadows real DL applications in public sector
- Slow government actions in last decade, but the landscape is changing.

Information Glut

World production of data: 1999 estimates

- Magnetic 1,693,000 terabytes
  - PC disk drives, departmental servers, camcorder tape, enterprise servers
- Film 427,000
  - Photograph, X-rays cinema
- Paper 240
  - Office documents, newspapers, periodicals, books
- Optical 80
  - Music CDs, DVDs, Data CDs

Grand Total ~ 2,120,000 terabytes

Source: Lyman and Varian, UC Berkeley

Information Consumption

Total time American households spend reading, watching TV or listening to music:

1992: 3,324 hours

2000: 3,380 hours

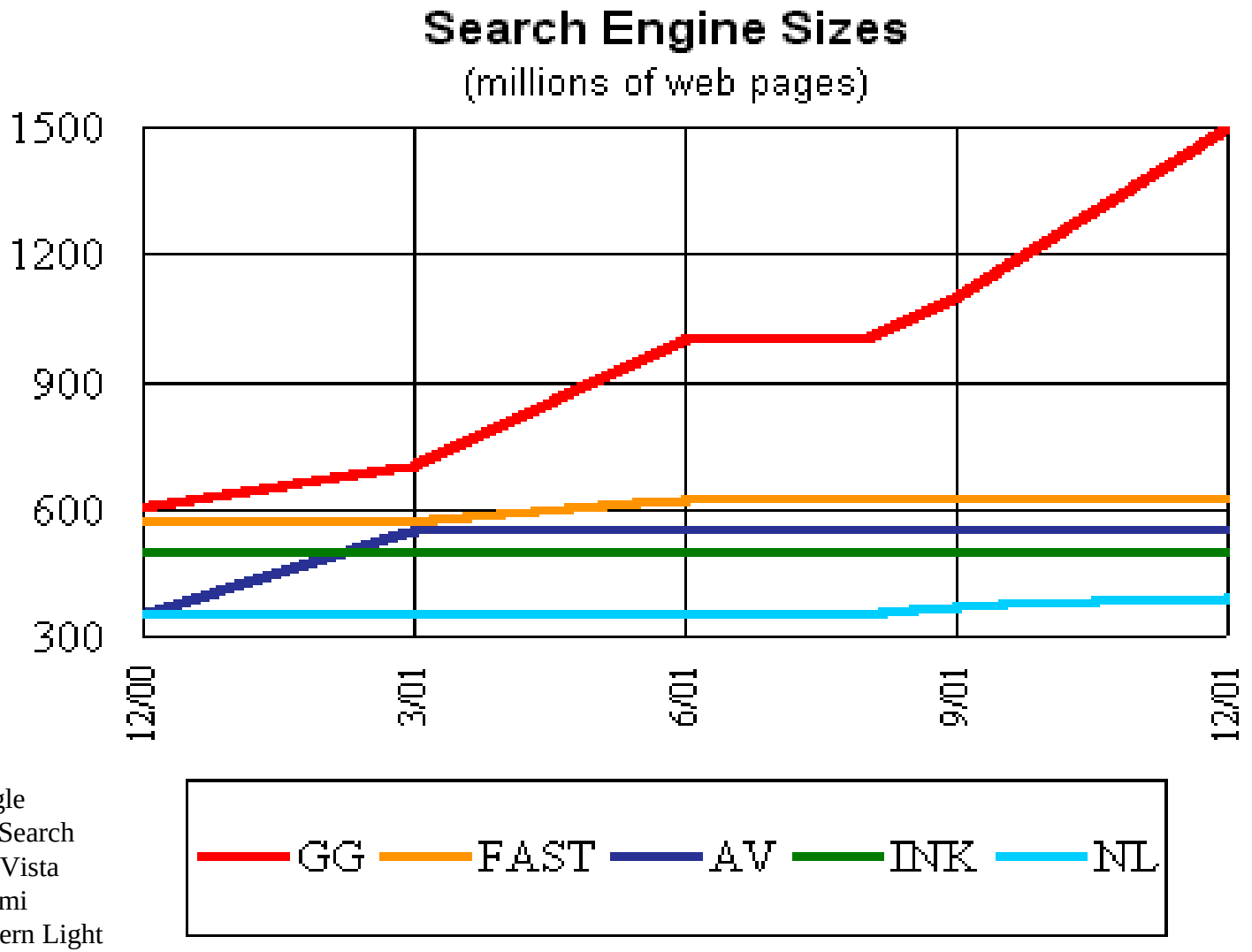
Bits consumed: 3,344,783 megabytes

or ~ 3 Terabytes

(Bits created: ~2,120,000 Terabytes)

Source: Lyman and Varian, UC Berkeley

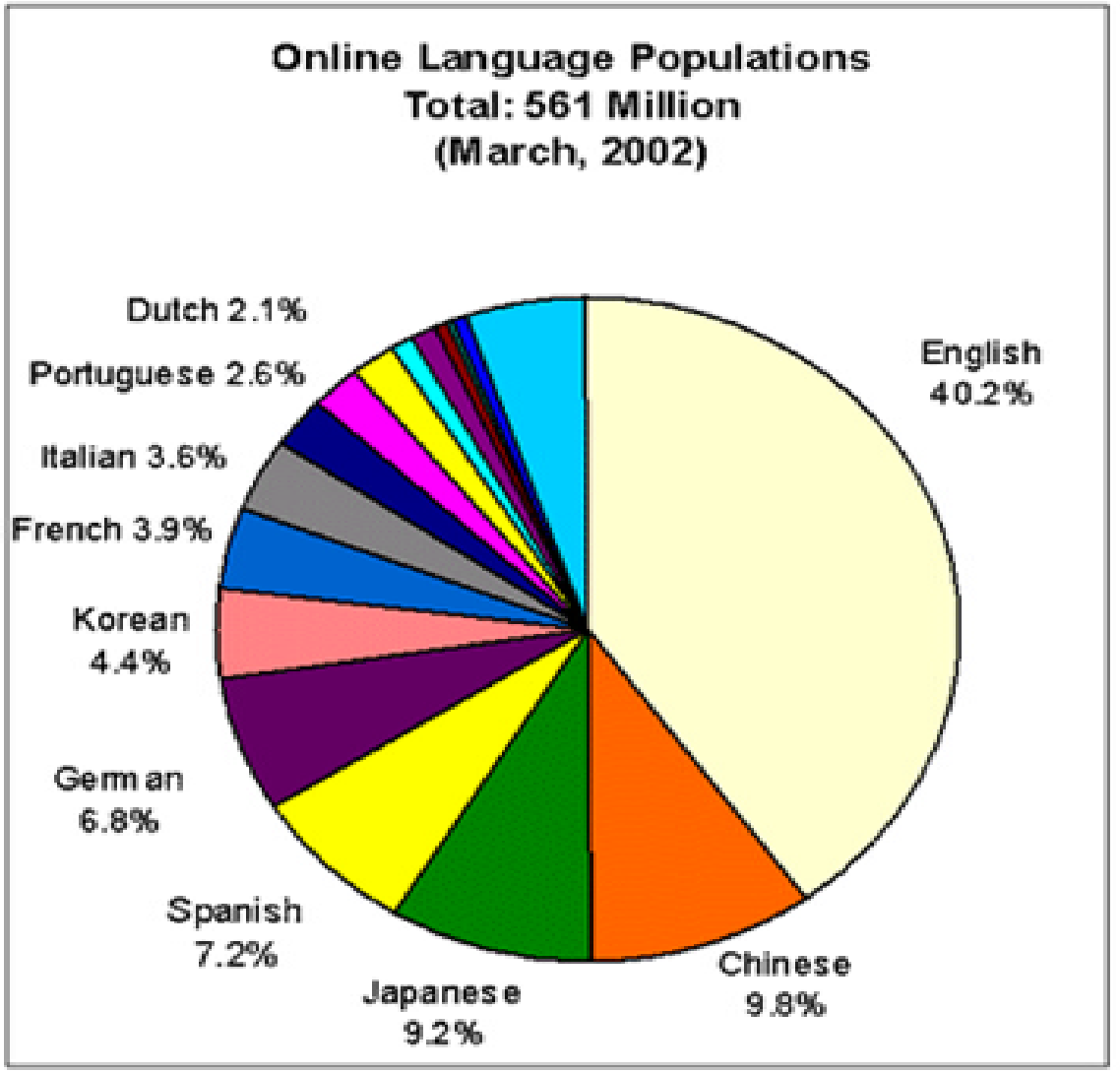
Search Information on the Internet



GG: Goggle
FAST: Fast Search
AV: Alta Vista
INK: Inktomi
NL: Northern Light

Source: Global Reach

Sharing Information on the Internet



Source: Global Reach

Where is the e-Content Business?

INFORMATION TECHNOLOGY PRODUCING INDUSTRIES

Hardware Industries

Computers and equipment
Wholesale trade of computers and equipment
Retail trade of computers and equipment
Calculating and office machines
Magnetic and optical recording media
Electron tubes
Printed circuit boards
Semiconductors
Passive electronic components
Industrial instruments for measurement
Instruments for measuring electricity
Laboratory analytical instruments

Software/Services Industries

Computer programming services
Prepackaged software
Wholesale trade of software
Retail trade of software
Computer-integrated system design
Computer processing, data preparation
Information retrieval services
Computer services management
Computer rental and leasing
Computer maintenance and repair
Computer related services, nec

Communications Equipment Industries

Household audio and video equipment
Telephone and telegraph equipment
Radio and TV communications equipment

Communications Services Industries

Telephone and telegraph communications
Cable and other TV services

** Although Radio and TV broadcasting industries were included as IT-producing industries in prior Digital Economy publications, they are not included in this report because they are now considered "content" providers, not IT infrastructure producing sectors.*

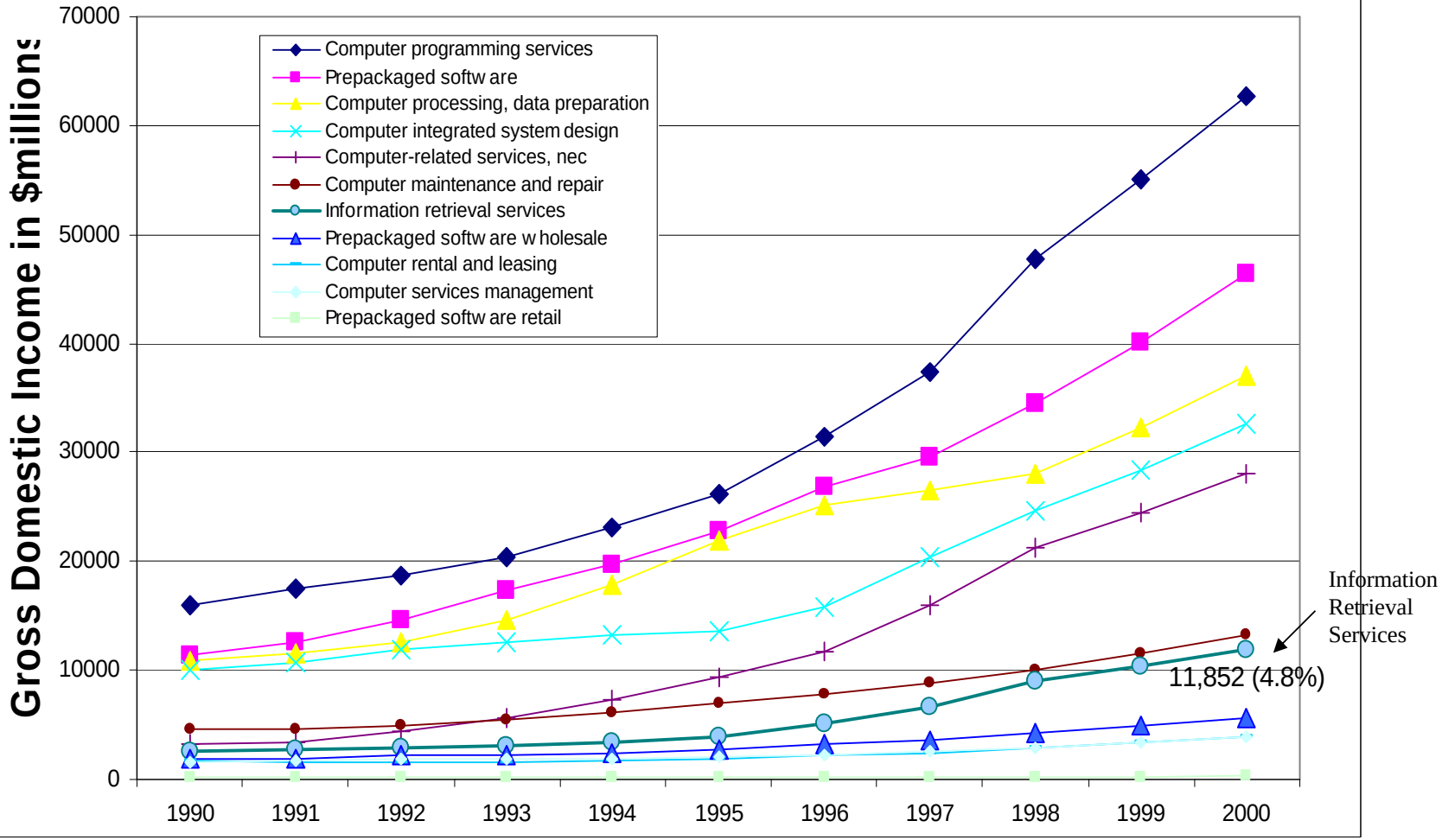
Source: U.S. Department of Commerce Report
"Digital Economy 2002"

U.S. Information Technology Producing Industries

Gross Domestic Income 2000, $Millions

Computing Hardware	251,655
Software and Services	245,656
Communications (hw&services)	299,256
Total IT-producing Industries	796,567
Total National GDI	10,003,400
IT share of economy	8.0%

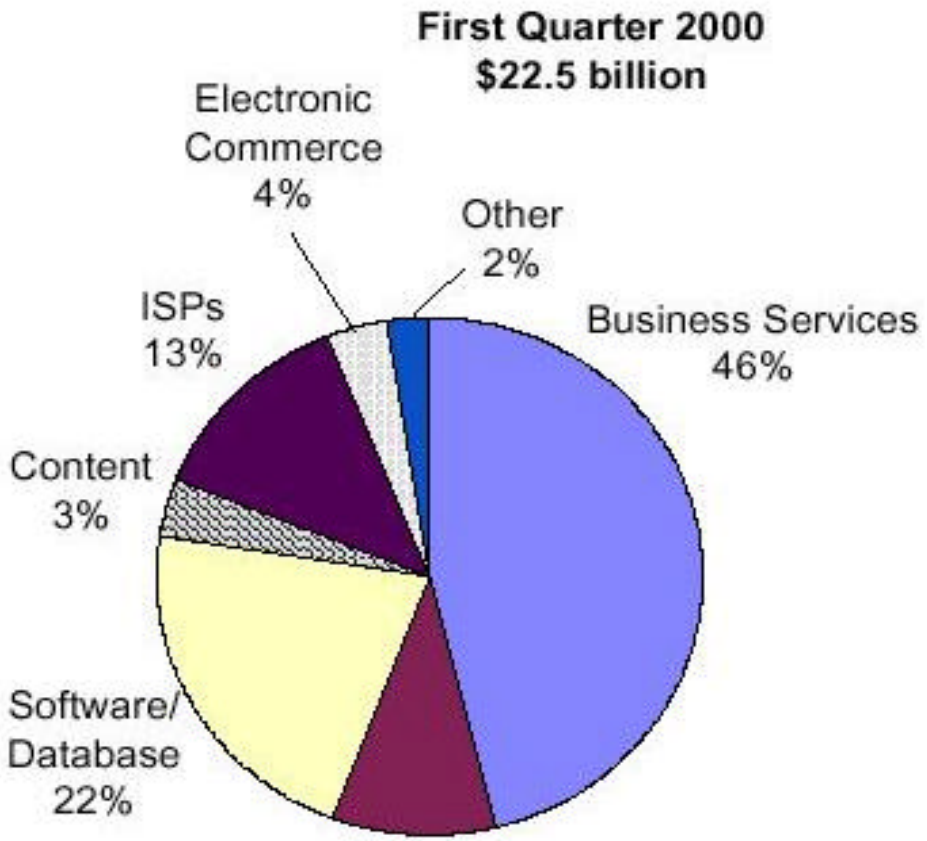
Trends in Software and Servicves

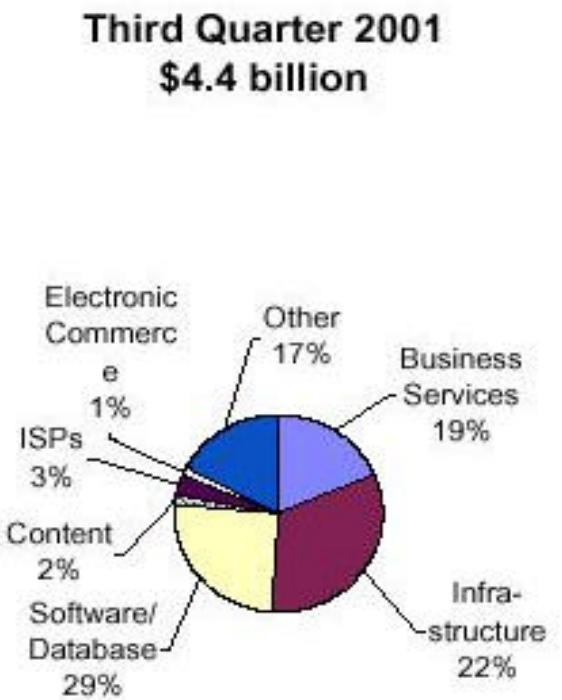


Source: U.S. Commerce Report
"Digital Economy 2002"

EQUITY FINANCING FOR VENTURE-BACKED COMPANIES BY TYPE OF INTERNET BUSINESSES

First Quarter 2000 and Third Quarter 2001





Source: U.S. Department of Commerce Report
"Digital Economy 2002"

Making e-Content a Business an European model

- Focus of Activity
  - improving access to and expending use of public sector information
  - enhancing content production in a multilingual and multicultural environment
  - increasing dynamism of the digital content market
- An ambitious, multi-year r&d program designed to take the lead in e-content business worldwide
  - research grants, demonstration projects, forging private-public partnerships, building tools and infrastructure, seeking new market spaces
- Addresses several of the DL roadblocks

Accessing Public e-Content

Beyond the walls of libraries

- Thematic areas of e-Content
  - traditional arts, cultural heritage, archives, museums, libraries
  - legal, administrative, and institutional data
  - financial, economic, and commerce data
  - entertainment, tourism, traffic/transportation information
  - geographic, agricultural, and environmental data
  - location-based services at the regional or national levels (education, health, crisis management, etc.)
  - data relating to health, safety, and consumer protection including emergency services
  - scientific and technical information (e.g., research publications, patents, data banks, standards, experimental testbeds, sharable software)
- Infrastructures for e-Content
  - Collections, platforms, networks, organizations, standards, middleware services, etc.

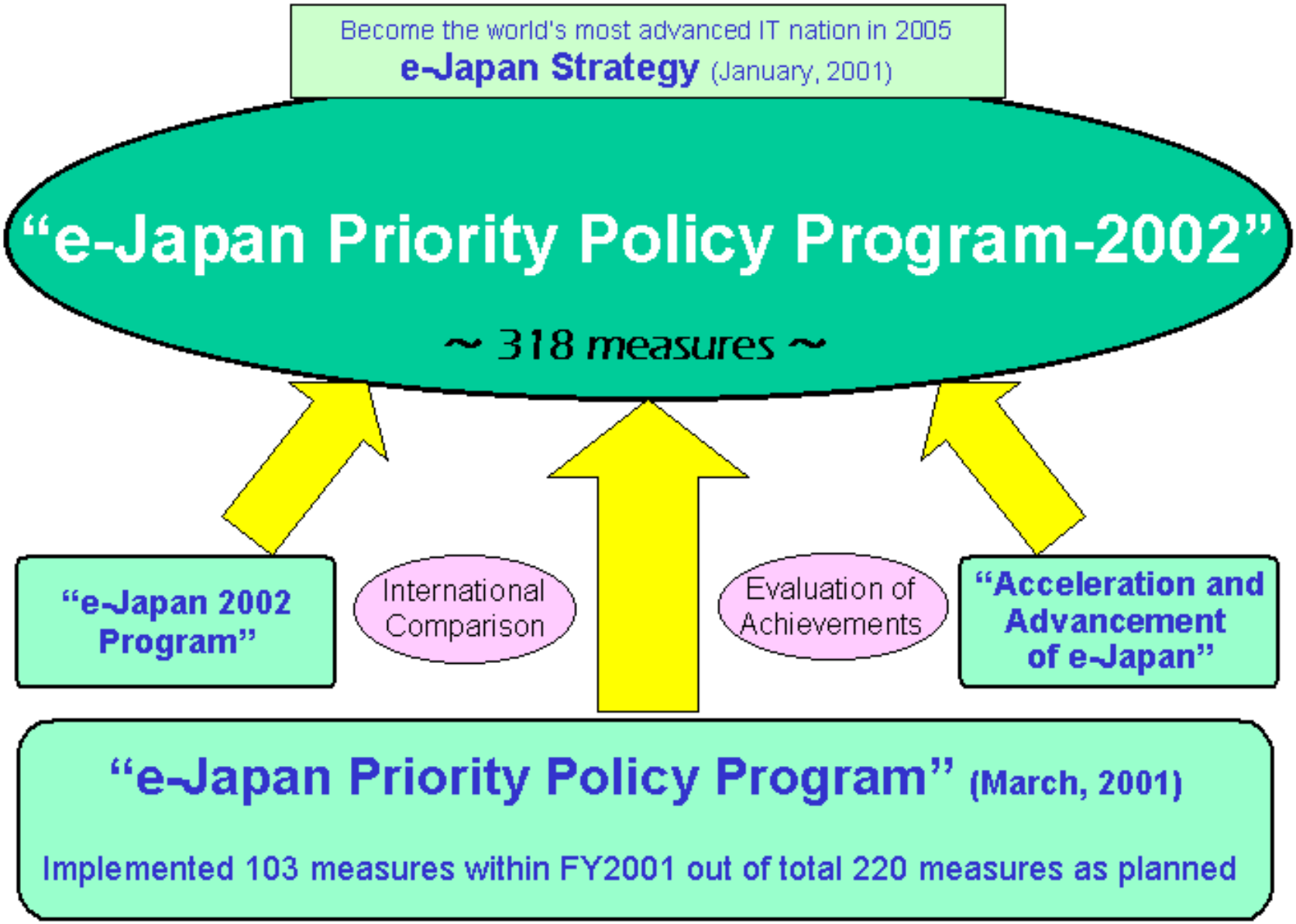
Enhancing e-Content Production:

across institutional, cultural, national borders

- Thematic areas
  - developing new strategies, partnerships, and solutions for designing and producing e-contents and services
  - focusing on e-contents and their multilingual and multicultural interfaces and the associated user/customer services
  - leveraging local, national, and global resources and expertise
- Three content communities as stakeholders
  - “commercial” content community (in place)
  - “corporate” content community (private and public sector, e.g, local or federal government)
  - “Public” content community, including public-private partnerships for a wider deployment of public e-contents
- Localization and internationalization at the same time

Increasing Dynamism of the e-Content Business

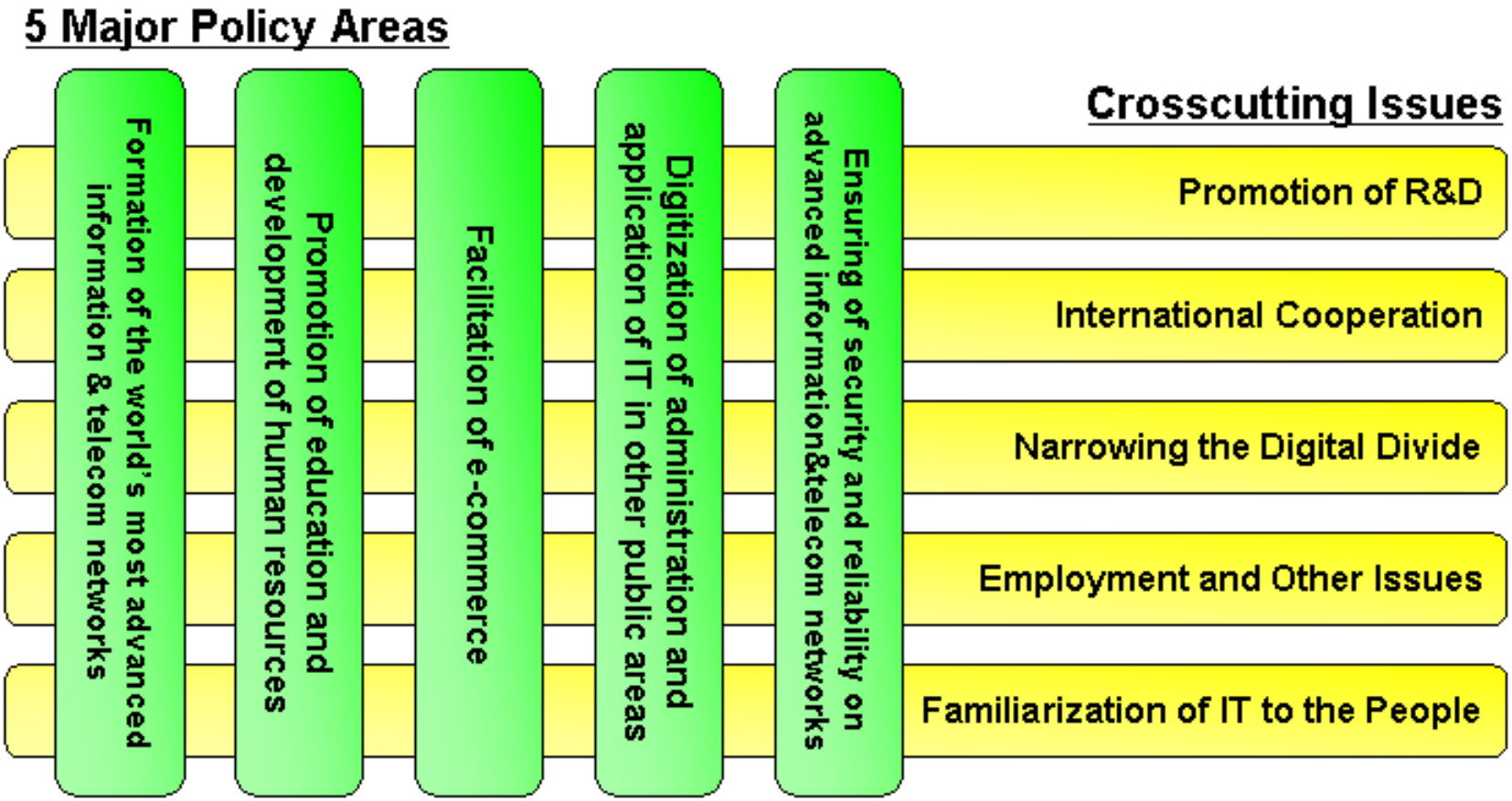
- Bridging the gap between the e-Content business and the capital market
  - Providing different channels to increase access to capital resources by various players
  - Making players aware of available business and tools services
  - Addressing the intellectual property rights and rights trading between e-Content players
- e-Europe may be more ambitious, but e-Japan may get there first
  - i-mode: Successful business model for private e-Content
  - Advantageous IPR policy (www.wtec.org/pdf/dio.pdf)



Source: e-Japan program, Office of the Prime Minister of Japan

Structure of e-Japan Priority Policy Program-2002

- Clearly state ministry in charge and goal of each measure



Source: e-Japan program, Office of the Prime Minister of Japan

4. Digitization of the administration and application of IT in other public areas

Evaluation

- Foundations of electronic government have been steadily constructed.
- Regarding the IT application in public areas, such as healthcare, ITS and GIS, its direction was clarified, and its implementation is expected from now.

Implemented Policies

Digitization of the administration

- Introduction of electronic tendering and bid-opening for public works
- Formulation of a basic plan toward the "single window" for import/export and harbor-related procedures
- Submission to the Diet of the laws aiming at enabling all administrative services available online

Application of IT in other public areas

- Formulation of a strategic grand design for digitization in the healthcare field
- Revision of Road Traffic Law to enable private servicers to provide the data of road and traffic information

Future Policies

Digitization of the administration jointly promoted by central and local governments

- Formulation of action plans for electronic filing of all governmental procedures by each ministry [FY2002]
- Introduction of electronic tendering and bid-opening for all projects of public works under ministerial jurisdiction [by FY2003]
- Establishment of government structures for the promotion of e-government [FY2002]

Support to local governments

- Presentation to local governments of standard procedures for online transactions of major services such as passport issuance [by FY2003]
- Promotion of the use of ASP for the operation of common systems of e-local governments [from FY2002]

Application of IT in other public areas

- Formulation of a roadmap toward the world's most advanced Intelligent Transport Systems [FY2002]
- Promotion of digital archiving of cultural assets and artworks [by FY2005]
- Enhancement of information provision services on reliability of food [from FY2003]

Source: e-Japan program, Office of the Prime Minister of Japan

Digital Government (DG)

An example of applying Digital Libraries technology

- Components of Investment
  - Vision: The PITAC report www.ccic.gov/pubs/pitac/index.html
  - Research: Linkage to DLI programs; DG Research initiative by NSF www.cise.nsf.gov/eia/dg
  - Implementation: All government levels, led by the Federal agencies www.firstgov.gov/
- Dimensions of System Design
  - Architectural relationship they have with their clients
  - Types of services they can provide to their clients

Unique Aspects of Government Information Services

- Security, privacy, and integrity as prime architectural and design criteria
- Scale: Instead of a core business, government is in every business
- All citizens and organizations as its equal customers
- Government as a huge customer for information technology: leverage and limitations
- Diversity of systems and applications

Service Levels for a Digital Government System

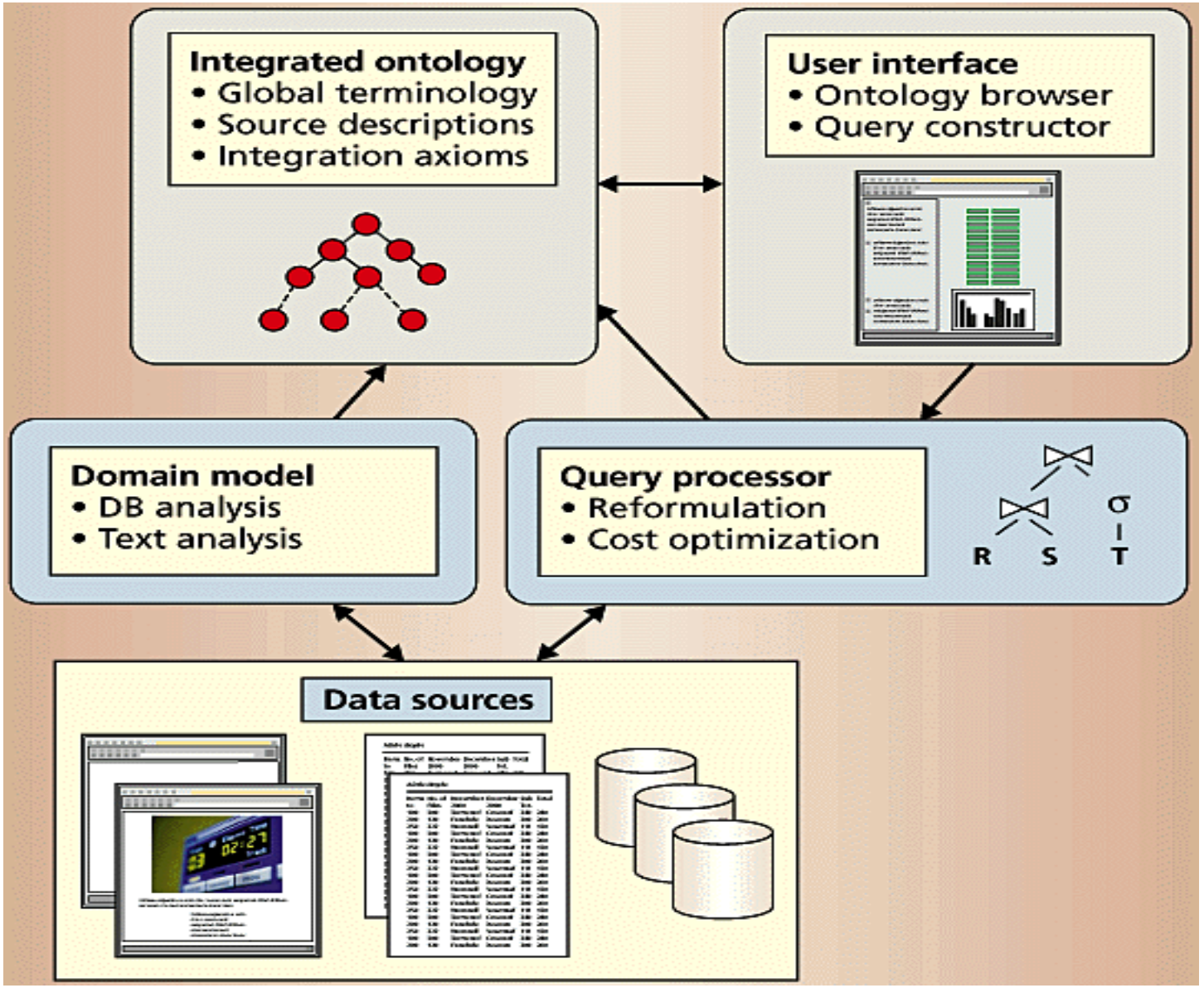
Level	*Key functions and uses*	*e-Contents and management*
First (low)	Provide one-way communication for displaying information about a given agency or aspect of government	Usually fixed type, limited to a single domain, one medium, simple data structure
Second	Provide simple two-way communication capabilities, usually for uncomplicated types of data collection such as registering comments	Similar to level 1, but may need more complex data structure and management
Third	Facilitate complex transactions that may involve interagency workflows and legally binding procedures. Examples are health and welfare services	Usually involves multiple databases and ontologies; need collaboration and coordination among agencies and with private sector, e.g., service providers
Fourth (high)	Integrate a wide range of services across a whole government administration and possibly several governments, domestic and international. Examples are crisis management and immigration & custom services.	Usually requires a hierarchy of ontologies and database structures; extensive coordination and collaboration among agencies; partnerships w/ private sector in content development and management

Research Areas for Digital Government Initiative

Topical Areas	*Research Description*	*Illustrative Examples*
Intelligent Information Integration	Shared ontologies; Mediation of multimedia data; Collaboration tools	Content searching for government data; Information systems for crisis management
Very Large-scale Data Acquisition and Management	Technologies to acquire, integrate, view, and assure the integrity of geographic, biological, environmental, and economic data and metadata	Access to linked statistical data sources in the 70+ agencies; A master U.S. data center for Crisis and emergency management
Advanced Analytics for Large Data Collections	Infrastructure to broadcast range of data analysis techniques; Visualization of large and complex data sets	Data mining facilities and computing services for citizens; Information-on-demand services for emergency management
Electronic Transaction and e-Commerce Techniques	Common transaction media between government and citizens; Data integrity and authentication techniques; Migration strategies from batch transaction to online systems	Electronic services delivered via WWW; Distributed kiosks at public sites for any-time transaction; Demonstrate capability of public key technology in multiple domains
Information Services for ordinary Citizens/Customers	Enhanced human-computer interactions, visualization and presentation technologies	Kiosk-based access for multiple services; Universal access for citizens with varied physical capabilities
Applications of IT to Law, Regulation, and other Mission Domains	Research on information, store, access, and management specific to mission agencies	Archiving, record keeping, and preservation; Systems in support of law enforcement and regulatory process with citizen inputs
Information Services for Large-scale Government R&D Projects	Engineering software and other computing services for large national projects in dedicated missions or across agencies	NASA launch monitoring and control; Bureau of Census integrated data services; Information services linking Social Security Administration and Health Services

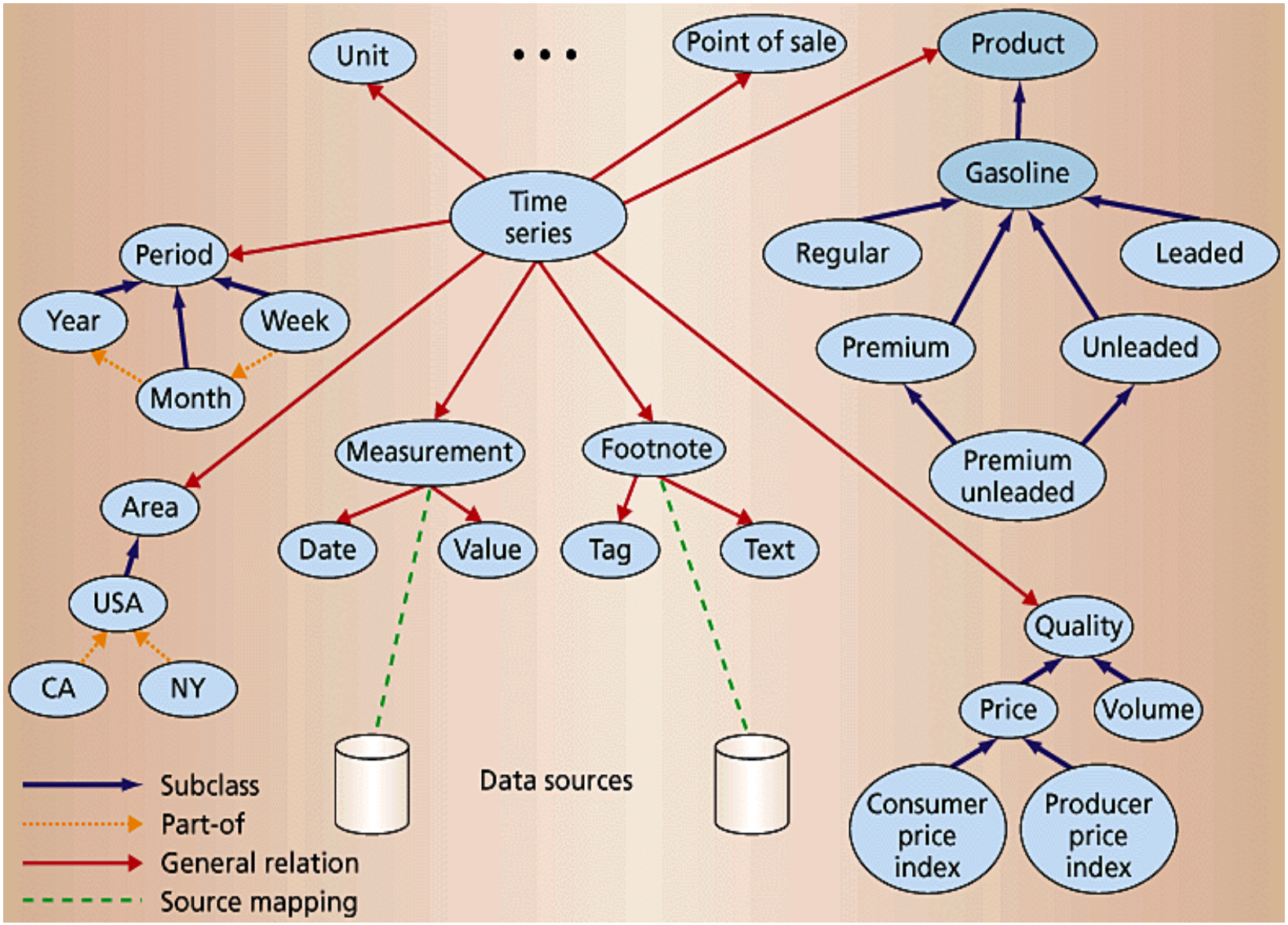
S

The Energy Data Collection (EDC) Project: System Architecture



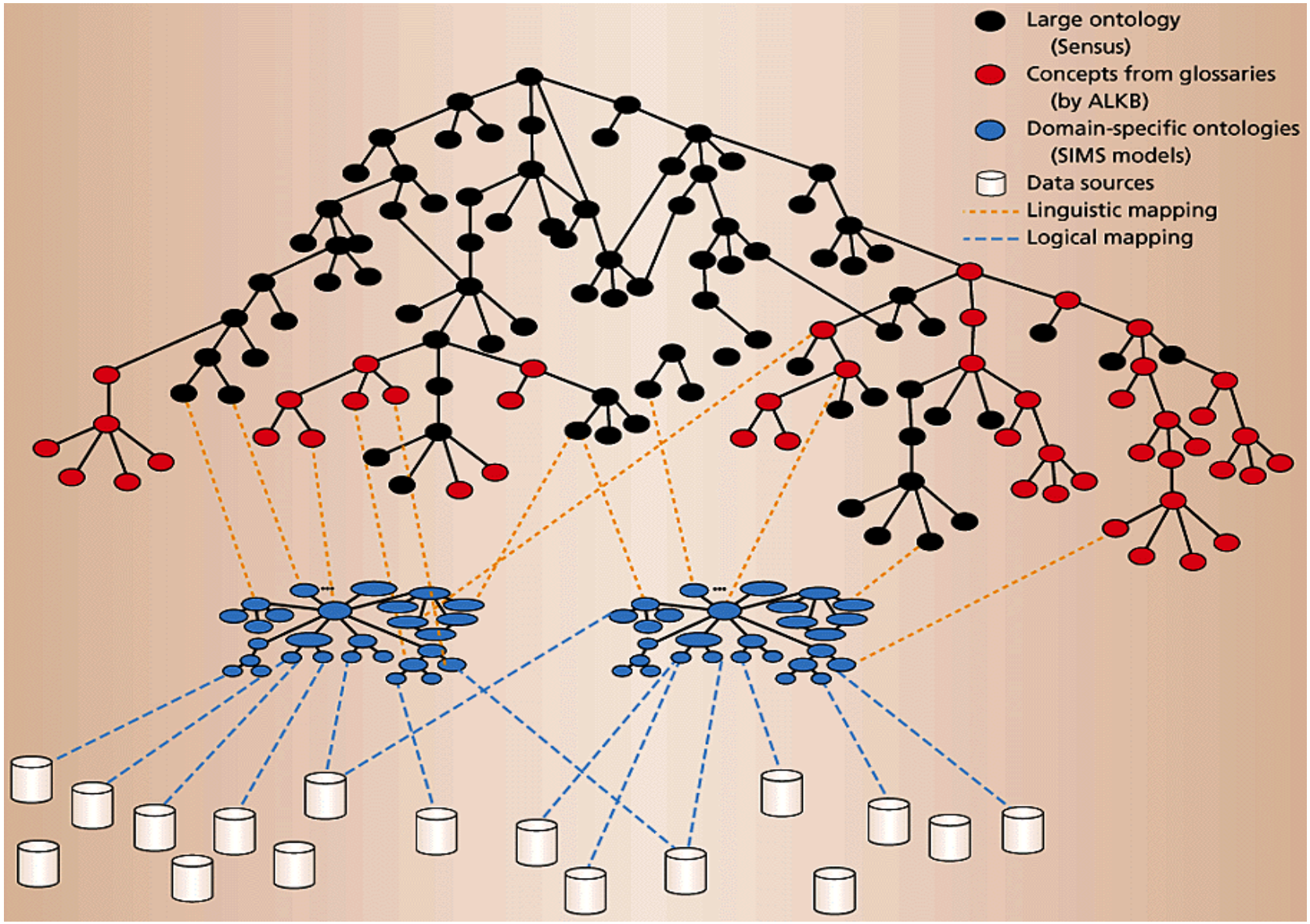
Source: NSF DG Pilot Project at USC/ISI

Fragment of an EDC domain model



Source: NSF DG Pilot project at USC/ISI

EDC Ontology and Domain Models



Source: NSF Pilot DG project at USC/ISI

DL Cross-cutting Issues

- Architectural levels
  - Applications, User services, Domain Knowledge Management, Collection Management, Data Handling, Storage
- Distributed Repositories
  - standards, tools, scalability, sustainability
- Integration and Interoperability
  - local, regional, global collections
  - data, access, service levels

Core business is DL middleware

Creating the Core Business

- Metadata providing information about the unlimited resources on the Web (e.g., the W3C semantic web activity, the Dublin Core Initiative, Resource Framework, etc.)
- Automated processing of Web information by software agents, including new concepts of search engines (next Google?)
- Facilitating applications that require open and public rather than constrained and proprietary contents
- Internetworking between applications: e.g., merging contents from multiple applications to create new information
- To do for the applications contents what the Web has done for hypertext: to allow contents to be processed outside the environment in which they were created at the Internet scale

The Anatomy of a Large-Scale Hypertextual Web Search Engine: Google

Sergey Brin and Lawrence Page

{sergey, page}@cs.stanford.edu

Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date.

Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

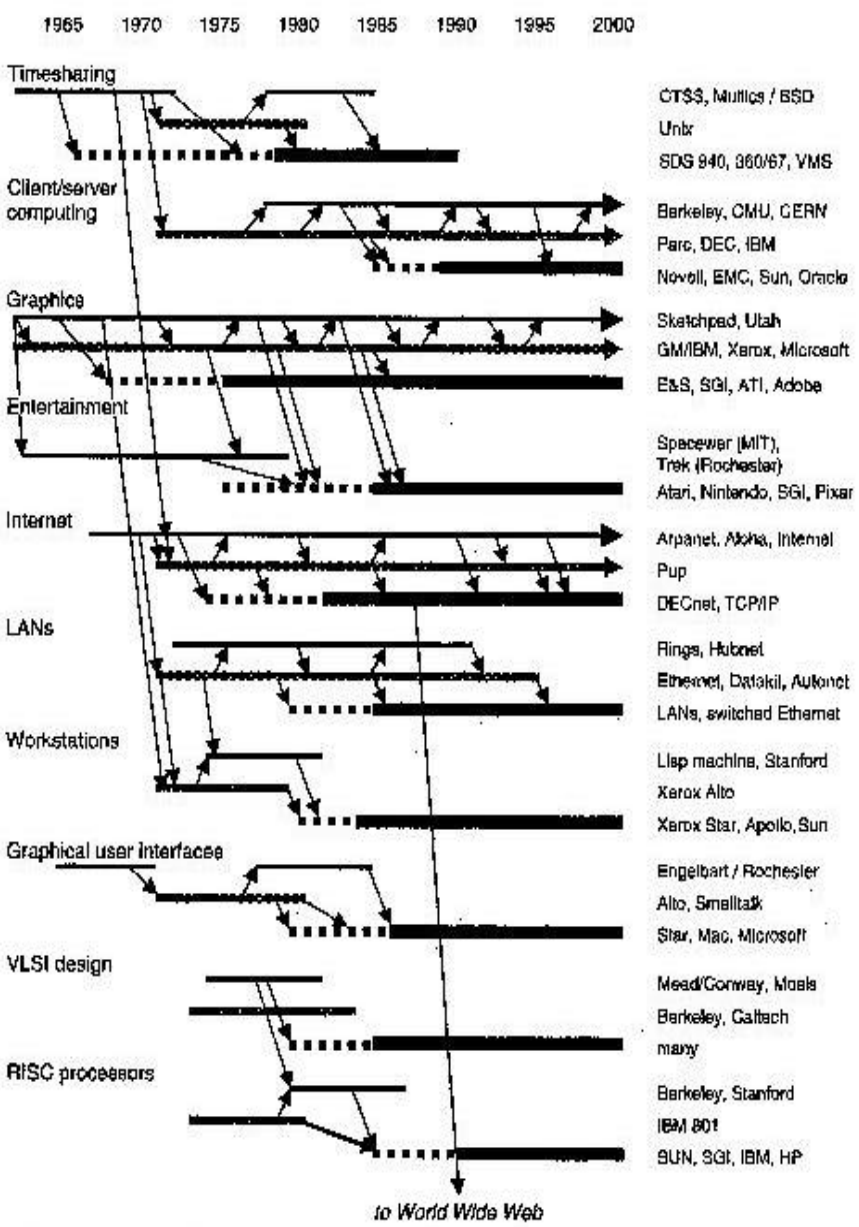
Keywords: World Wide Web, Search Engines, Information Retrieval, PageRank, Google

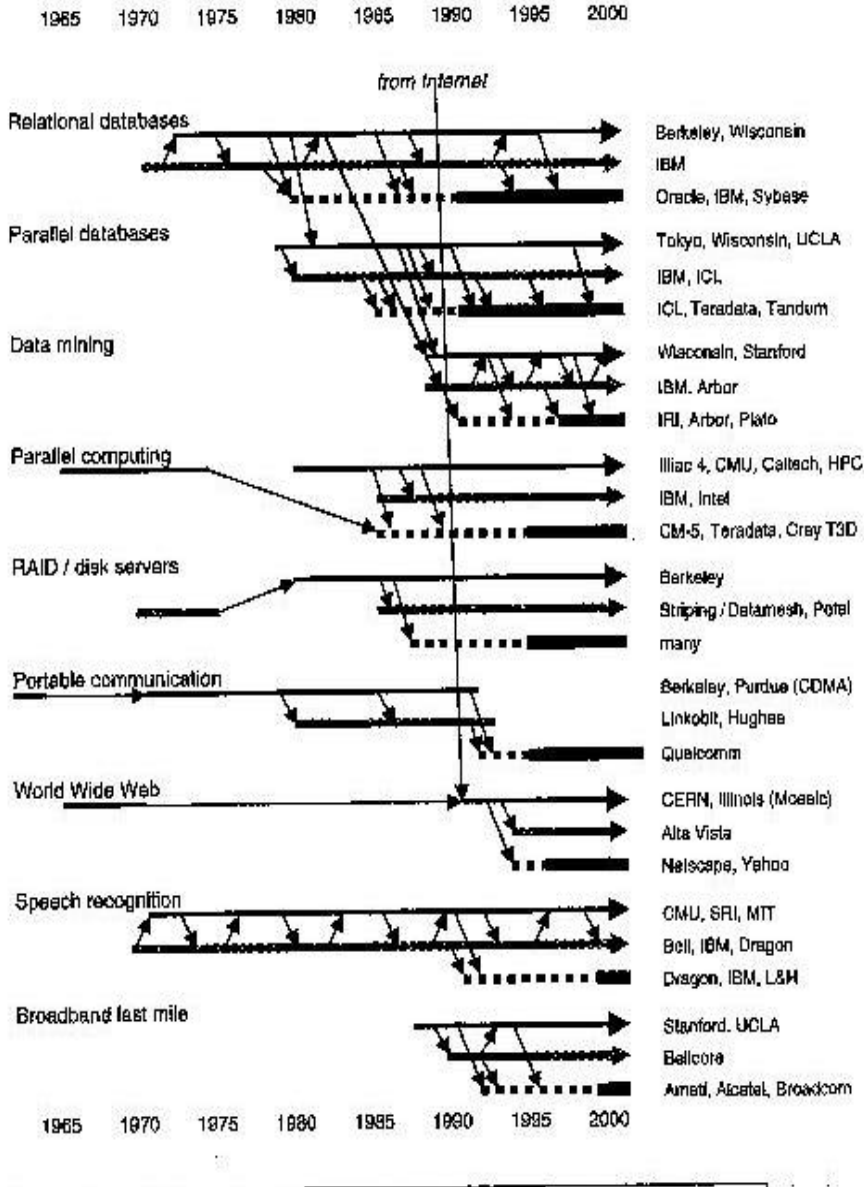
Google, Inc.: from university research to business

- 1994: DLI-1 initiative began; Stanford U Consortium funded for its Infobus project
- 1995: Grad students Larry Page and Sergey Brin developed a search technology called “BackRub”
- 1997: Research paper by Brin and Page, “The anatomy of a search engine Google”, published
- 1998: Page and Brin launched Google, Inc.; Search engine answered 10,000 queries per day
- 2002: www.google.com/corporate/facts.html.
  - Answers more than 150 million queries daily
  - Searches more than 2 billion web pages
  - Has 55+ million unique users per month
  - Global reach: More than 50 percent of traffic is from outside the US; search covers some 80 languages

Why a business model? Adding a DL entry to the innovation pipeline

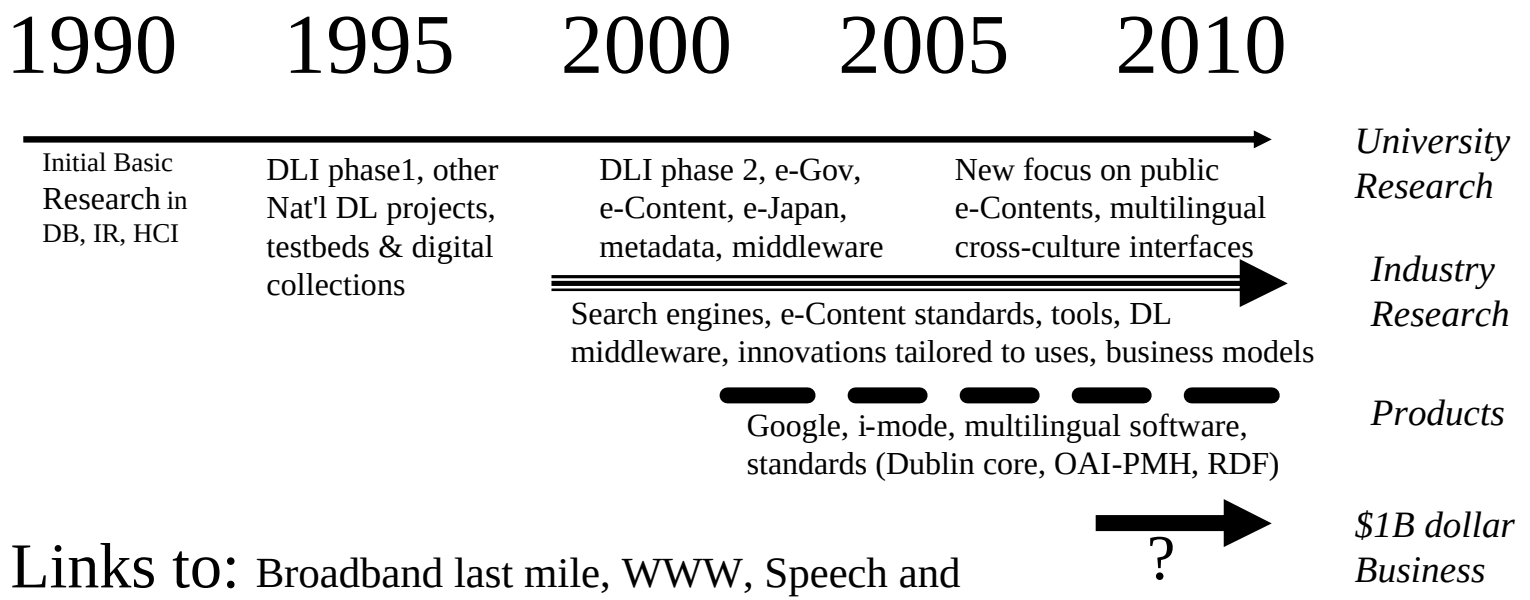
Source: NRC report on IT Research and Innovation





DL Middleware

Milestones for a New Entry in the R&D pipeline



Links to: Broadband last mile, WWW, Speech and language technology, Portable communications, Data mining, Relational databases