The Web's Missing Links:
The Search Engine & Portal Industry

Thomas Haigh
The Haigh Group &
University of Wisconsin, Milwaukee
thaigh@computer.org

ETH, Informatik Lunch, 18 June 2007

About Me

- Dual training
  - B.Sc. & M.Eng in Computer Science (Manchester)
  - Ph.D. in History & Sociology of Science (Pennsylvania)
- Main interest in history of IT use in US business
- Published papers on history of
  - Management Information Systems concept
  - Early data processing
  - DBMS concept
  - Word Processing
  - Packaged software industry
  - Sources for ACM history
- Chair SHOT SIG on Computers, Information, Society
- Involved with IEEE, SIAM & ACM projects

Background of Project

- Two chapters in MIT Press edited book, "The Internet & American Business," Aspray & Ceruzzi
  - Software infrastructure chapter – web, email, protocols
  - Search and portals ("Web navigation business")
- Contemporary history, somewhat journalistic
  - Recounting of basic events from secondary sources
  - Focus on interplay between technology and business models

Aims

1. Situate web with respect to other electronic publishing technologies
   - And earlier Internet story
2. Tie together
   - Web publishing economics
   - Web navigation economics
   - Technical choices built into web design
3. Write analytical history from journalistic sources

Social Construction of Technology

Two key concepts established since 1980s

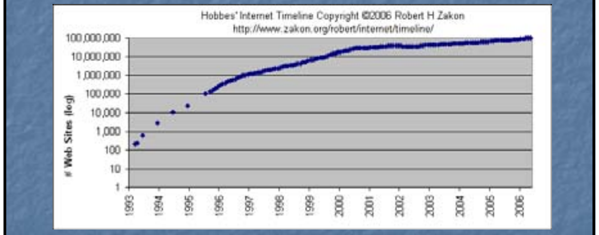
- 1: Mutual shaping of technologies and society
  - Influence of social factors on technological design choices
- 2: Power of technological SYSTEMS
  - Combine users, firms, standards, technologies
  - Lock-in effects of dominant systems as "Technological Momentum"

Reconstruction of Technology

- Commercialization of Internet infrastructure
- What happens when an already "shaped" technology gets
  - New uses
  - New "relevant social groups"
  - New cultural meanings
- Thoughts at the back of my mind
  - VHS vs Beta, QWERTY vs. Dvorak? –
    - which is the net?

2: Narrative Overview

Web Hosts Growth

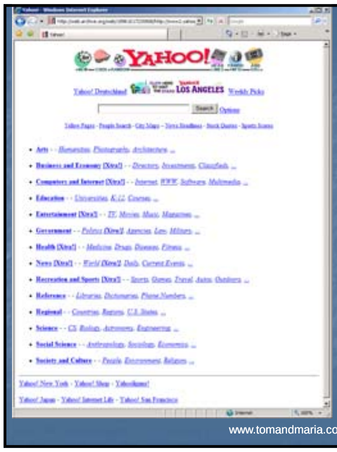


Timeline of Developments

- 1991: Web introduced at CERN
- 1993: Mosaic popularizes the Web
    - 130 servers to 10,000 in 18 months
- 1993: First web crawlers
- 1994: Yahoo directory service founded
- 1995: AltaVista, Lycos, Excite, Infoseek & OpenText index web
- 1995: Netscape IPO
- 1996: Yahoo, Excite, Lycos & Infoseek IPOs
- 1998: Google, Inc. founded
- 1999: Search firms converge on Portal model
- 2000: Dot com crash signals end of easy money
- 2000: Google starts selling AdWords
- 2004: Google IPO.
- Today: Google dominates search, Yahoo is primary U.S. Portal

Web Directories

- The Web As Its Own Catalog
    - Link directories are special-purpose websites
    - Yahoo is most successful
- Humans visit lots of websites
    - Find the best ones on a topic
    - Add them with topic code to a simple database
    - Directory listings are batch generated
- Basically the yellow pages of the Internet
    - Businesses pay for prominent position
    - Firms advertise to reach searchers



Yahoo! Deutschland · Weekly Picks

Options

Yellow Pages - People Search - City Maps - News Headlines - Stock Quotes - Sports Scores

- Arts -- Humanities, Photography, Architecture, ...
- Business and Economy [Xtra!] -- Directory, Investments, Classifieds, ...
- Computers and Internet [Xtra!] -- Internet, WWW, Software, Multimedia, ...
- Education -- Universities, K-12, Courses, ...
- Entertainment [Xtra!] -- TV, Movies, Music, Magazines, ...
- Government -- Politics [Xtra!], Agencies, Law, Military, ...
- Health [Xtra!] -- Medicine, Drugs, Diseases, Fitness, ...
- News [Xtra!] -- World [Xtra!], Daily, Current Events, ...
- Recreation and Sports [Xtra!] -- Sports, Games, Travel, Autos, Outdoors, ...
- Reference -- Libraries, Dictionaries, Phone Numbers, ...
- Regional -- Countries, Regions, U.S. States, ...
- Science -- CS, Biology, Astronomy, Engineering, ...
- Social Science -- Anthropology, Sociology, Economics, ...
- Society and Culture -- People, Environment, Religion, ...

Yahoo! New York - Yahoo! Shop - Yahooligans!

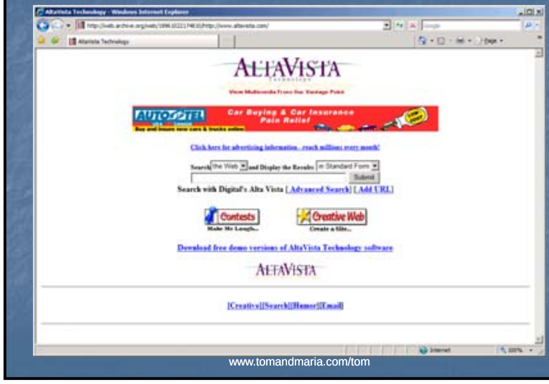
Yahoo! Japan - Yahoo! Internet Life - Yahoo! San Francisco

Yahoo, 1996

Search Engine Model

- Crawlers index the web
    - Technology already developed for ftp sites, gopher headings
    - Keywords entered by users are looked up in index
    - Index & search developed for online services, full text databases like OED
- Hard to do well!
- How to make money?
    - Subscription model fails for Infoseek
        - Standard for online databases like LEXIS
    - Advertising supported
        - Popular keywords sold at a premium from 1995
    - Also sell tech or services to other websites

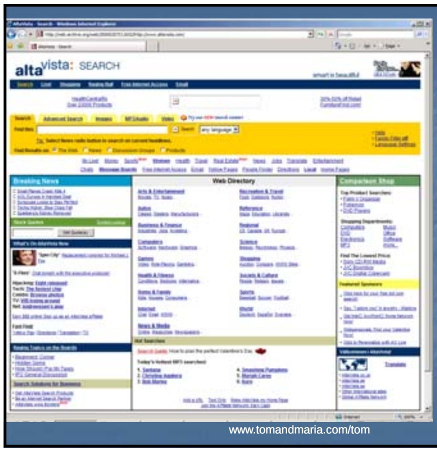
AltaVista 1996


AltaVista Technology - Windows Internet Explorer
AltaVista Technology
ALTAVISTA
Technology
View Multimedia From Our Vantage Point
AUTOBYTEL
Car Buying & Car Insurance Pain Relief
Buy and insure new cars & trucks online
Click here for advertising information - reach millions every month!
Search the Web and Display the Results in Standard Form
Submit
Search with Digital's Alta Vista [Advanced Search] [Add URL]
Contests
Make Me Laugh...
Creative Web
Create a Site...
Download free demo versions of AltaVista Technology software
ALTAVISTA
[Creative][Search][Humor][Email]
Internet
100%

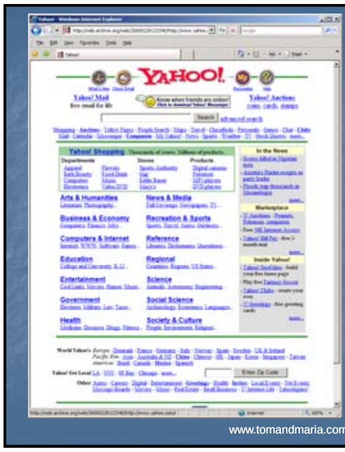
Portals

- Internet navigation firms add content
  - Both Yahoo (directory)
  - And Excite, Lycos & other search firms
- Theory: add "stickiness" – be more like AOL
  - Good search sends users away quickly
  - Keep them around instead
    - News, Weather & Horrorscopes
    - Free email
    - Shopping "malls"
  - They watch more banner advertisements
- But unlike AOL aren't online services

AltaVista 2000



Yahoo, 2000



Influence of .com Boom

- Portals copy AOL with "strategic partnerships" with doomed startups
  - E.g. "Exclusive CD retailer on Yahoo"
  - Excite@home pays $780 million for online greeting card company
  - Companies valued on number of visitors
- Institutional Ismophism – companies copying each other
- Need rising numbers to justify valuation
  - YHOO stock rises 100 times in 4 years from IPO
  - Lycos (#3 portal) sold for $12.5 billion in 2000

Portals Largely Wiped Out

- Had deemphasized search
  - Full of advertising & paid results
  - Swamped by search engine spam
  - Little investment in improvements
- Crippled when easy money dries up in 2001
- By 2003 Yahoo is only significant non-ISP portal
  - AOL and MSN retain online service portals

3: Special Features of the Web

Why Was the Web Special?

- Web is the first functional
  - Very large scale
  - Highly distributed (no index or catalog)
  - Hypertext
  - Electronic publishing system
- So, how was it different from other electronic publishing systems?
  - And how did this influence the web navigation industry?

Web Navigation Business

- Unlike earlier electronic publishing, the web has no search or index built in
  - Makes publishing very easy, retrieving very hard
  - Hypertext seen as alternative to searching and indexing
- Unlike earlier electronic publishing systems
  - Navigation and indexing content is a separate business from publishing content
- Creates huge business opportunity. 2 models
  - Web Directory (Yahoo, Magellan)
  - Web Search (Excite, Lycos, AltaVista)

The Early Web

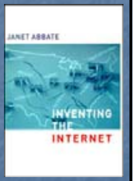
- Leverages existing Internet technologies
  - TCP/IP, FTP, news, Gopher, SGML, SMTP etc
  - New elements: HTML, HTTP, URL
- Simple design
  - elegantly tackles immediate needs
- Fundamental problems ignored
  - Searching
  - Hyperlink issues
- Follows cultural traditions of Internet

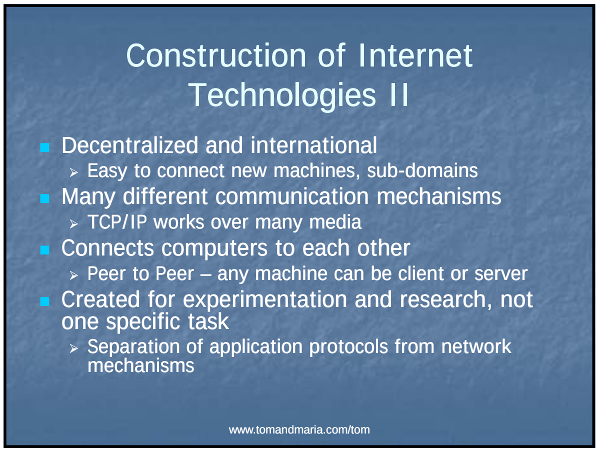
Layering of Protocols

FTP Client	Mail client	Web browser	Many others….
FTP (File transfer)	SMTP (Mail transfer)	HTTP (Web)	Video, chat, news, P2P, instant messaging
Socket API			
TCP/IP (also DNS shared by applications)			
Ethernet	SLIP/ PPP	Satellite	Fiber Optic, Etc.

Construction of Internet Technologies (1970s-80s)

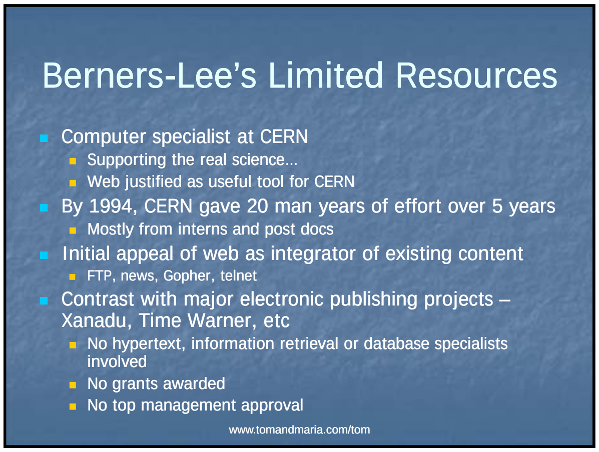
- Closed, homogenous, small academic population
  - Results: Rely on social mechanisms for security, elimination of troublemakers
- Practical, working network
  - Rather have it next week than perfect
- Non-commercial
  - No mechanisms to bill for use of resources
- Support for many machine types
  - Compatibility through standards, not code





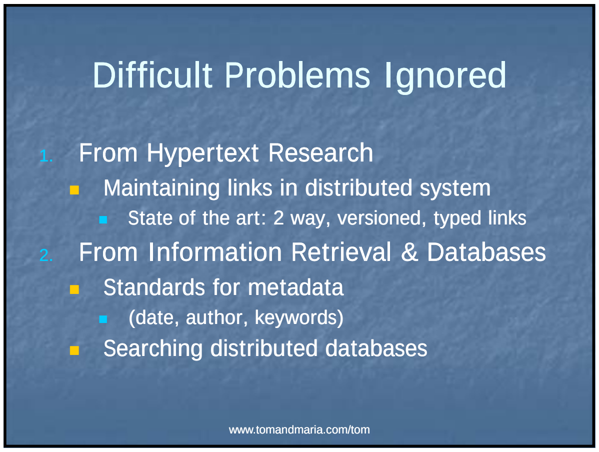
Construction of Internet Technologies II

- Decentralized and international
  - Easy to connect new machines, sub-domains
- Many different communication mechanisms
  - TCP/IP works over many media
- Connects computers to each other
  - Peer to Peer – any machine can be client or server
- Created for experimentation and research, not one specific task
  - Separation of application protocols from network mechanisms



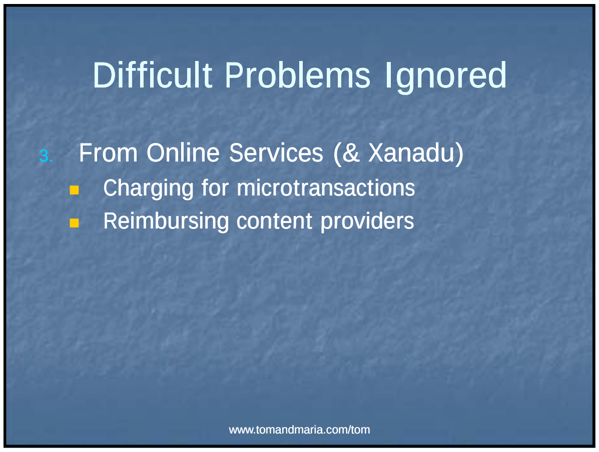
Berners-Lee's Limited Resources

- Computer specialist at CERN
  - Supporting the real science…
  - Web justified as useful tool for CERN
- By 1994, CERN gave 20 man years of effort over 5 years
  - Mostly from interns and post docs
- Initial appeal of web as integrator of existing content
  - FTP, news, Gopher, telnet
- Contrast with major electronic publishing projects – Xanadu, Time Warner, etc
  - No hypertext, information retrieval or database specialists involved
  - No grants awarded
  - No top management approval



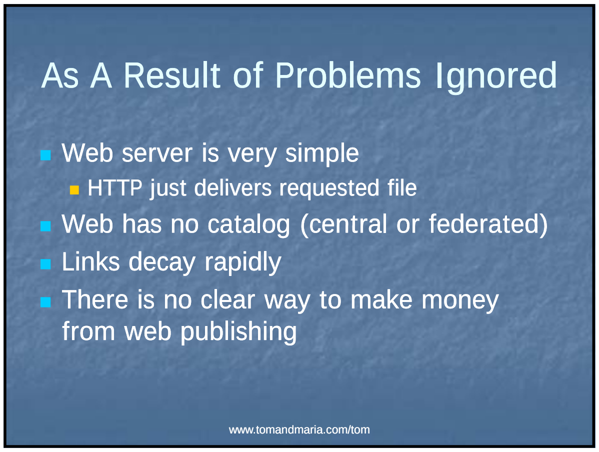
Difficult Problems Ignored

1. From Hypertext Research
   - Maintaining links in distributed system
     - State of the art: 2 way, versioned, typed links
2. From Information Retrieval & Databases
   - Standards for metadata
     - (date, author, keywords)
   - Searching distributed databases



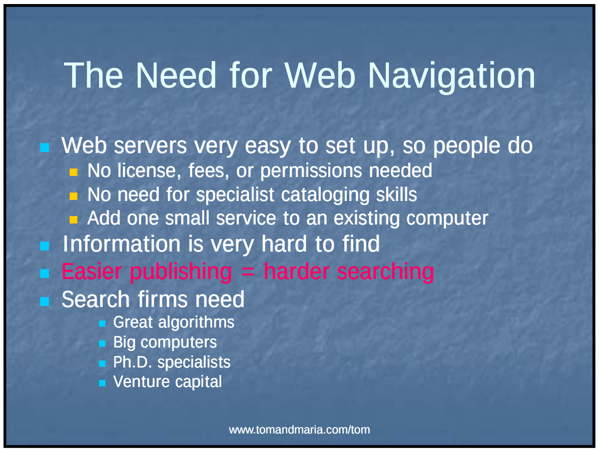
Difficult Problems Ignored

3. From Online Services (& Xanadu)
   - Charging for microtransactions
   - Reimbursing content providers



As A Result of Problems Ignored

- Web server is very simple
  - HTTP just delivers requested file
- Web has no catalog (central or federated)
- Links decay rapidly
- There is no clear way to make money from web publishing

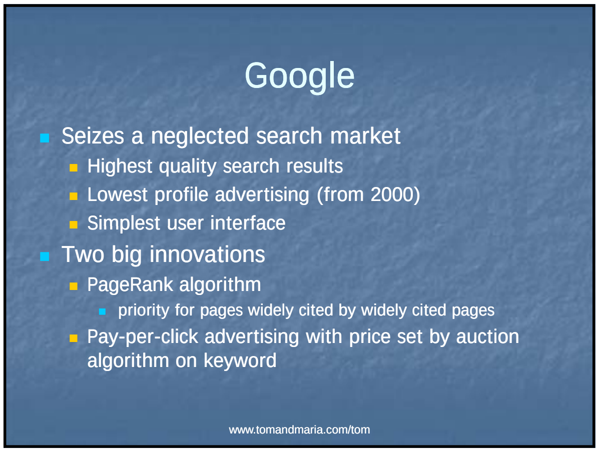


The Need for Web Navigation

- Web servers very easy to set up, so people do
  - No license, fees, or permissions needed
  - No need for specialist cataloging skills
  - Add one small service to an existing computer
- Information is very hard to find
- Easier publishing = harder searching
- Search firms need
  - Great algorithms
  - Big computers
  - Ph.D. specialists
  - Venture capital

4: The Triumph of Google

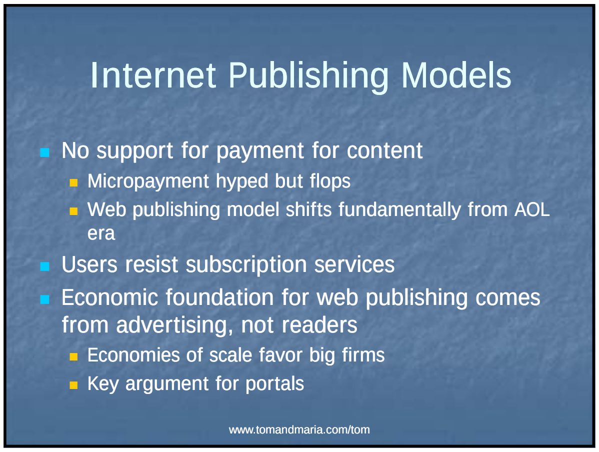
www.tomandmaria.com/tom



Google

- Seizes a neglected search market
  - Highest quality search results
  - Lowest profile advertising (from 2000)
  - Simplest user interface
- Two big innovations
  - PageRank algorithm
    - priority for pages widely cited by widely cited pages
  - Pay-per-click advertising with price set by auction algorithm on keyword

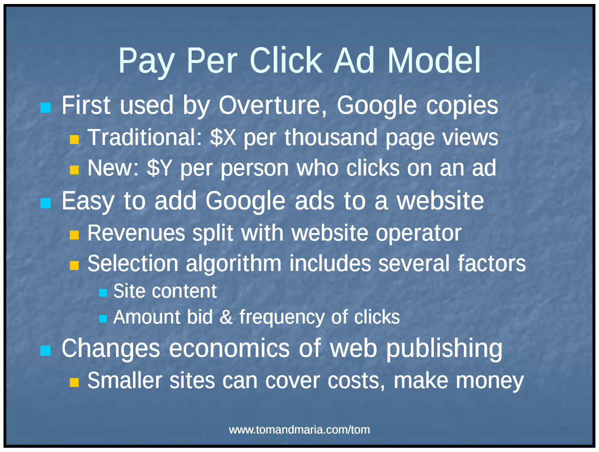
www.tomandmaria.com/tom



Internet Publishing Models

- No support for payment for content
  - Micropayment hyped but flops
  - Web publishing model shifts fundamentally from AOL era
- Users resist subscription services
- Economic foundation for web publishing comes from advertising, not readers
  - Economies of scale favor big firms
  - Key argument for portals

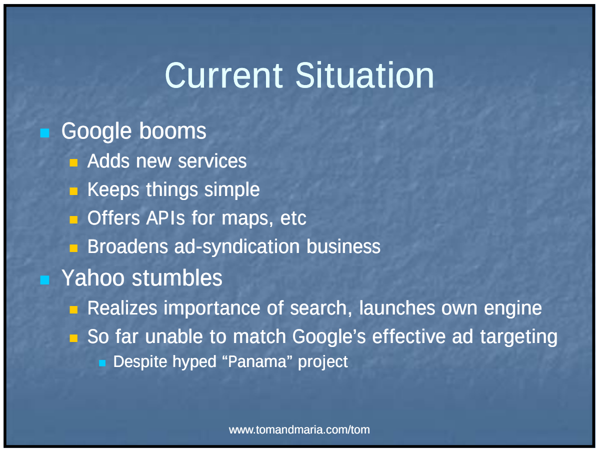
www.tomandmaria.com/tom



Pay Per Click Ad Model

- First used by Overture, Google copies
  - Traditional: $X per thousand page views
  - New: $Y per person who clicks on an ad
- Easy to add Google ads to a website
  - Revenues split with website operator
  - Selection algorithm includes several factors
    - Site content
    - Amount bid & frequency of clicks
- Changes economics of web publishing
  - Smaller sites can cover costs, make money

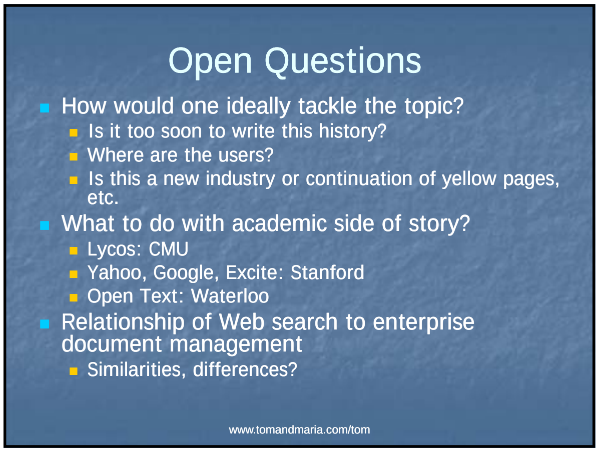
www.tomandmaria.com/tom



Current Situation

- Google booms
  - Adds new services
  - Keeps things simple
  - Offers APIs for maps, etc
  - Broadens ad-syndication business
- Yahoo stumbles
  - Realizes importance of search, launches own engine
  - So far unable to match Google’s effective ad targeting
    - Despite hyped “Panama” project

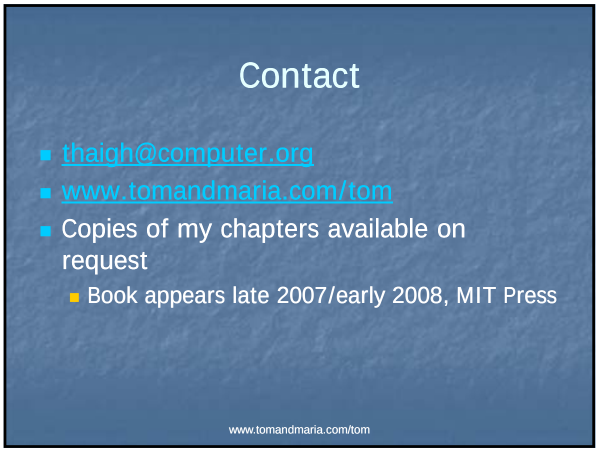
www.tomandmaria.com/tom



Open Questions

- How would one ideally tackle the topic?
  - Is it too soon to write this history?
  - Where are the users?
  - Is this a new industry or continuation of yellow pages, etc.
- What to do with academic side of story?
  - Lycos: CMU
  - Yahoo, Google, Excite: Stanford
  - Open Text: Waterloo
- Relationship of Web search to enterprise document management
  - Similarities, differences?

www.tomandmaria.com/tom



Contact

- thaigh@computer.org
- www.tomandmaria.com/tom
- Copies of my chapters available on request
  - Book appears late 2007/early 2008, MIT Press

www.tomandmaria.com/tom