
$W^3 + \text{Structure} = \text{Knowledge}$

Why Web Printing is a Difficult Problem

URL for this document: <http://itd.hpl.hp.com/~beretta/ Acrobat/Soapbox/Structure.pdf>

Giordano Beretta

**Hewlett-Packard Laboratories
Imaging Technology Department
Distributed Imaging Systems**

Daniel Lee

Ho John Lee

Dev Chen

Konstantinos Konstantinides

Andrew H. Mutz

Place holder for ITD slide

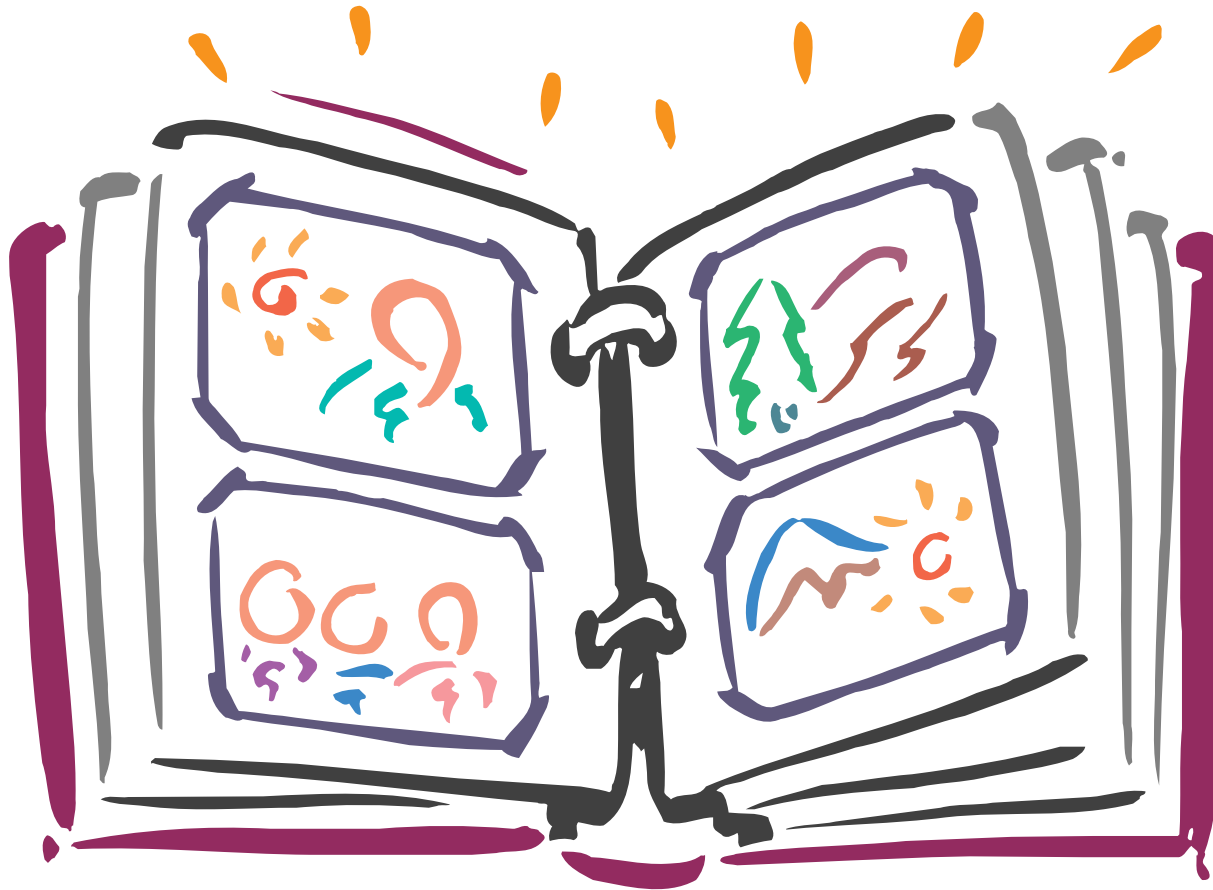
The Bottom Line

- The World Wide Web is the new publishing channel
- Paper is the best medium to present written information
- To own the digital printing market we have to be the best in printing information off the Web



New Emerging Markets

Scanning and digital photography for the web



Purpose of This Work

**Understanding the fundamentals
and identifying the best fulcrum on
which to apply our lever**



- Web printing is hard because Web pages are linked but disconnected: poor structure
- Web printing is hard because in traditional printing
 - *the author decides contents, structure, and appearance*
- ... while on the web
 - *the author decides contents and structure*
 - *the reader decides the appearance*

Desktop publishing:

- WYSIWYG — ability to see during document creation the formatted page as it will be printed

Web publishing:

- a multi-dimensional multimedia communications means
- Holy Grail of Web printing is to create the hard copy *facet* for printed output

Disclaimer:

- We are not talking about Internet printing

Example: Printing a Manual

... we are in the middle of a session ...

Suppose a customer needs a hard copy of an oscilloscope manual

- Typically the manual will be stored as a number of HTML pages, one per section
- Typically the user has to perform the following steps:
 - 1 find first section
 - 2 click URL of section
 - 3 click on print button
 - 4 click OK in print dialog
 - 5 click back in browser
 - 6 increment section
 - 7 if endOfDocument exit else goto 2

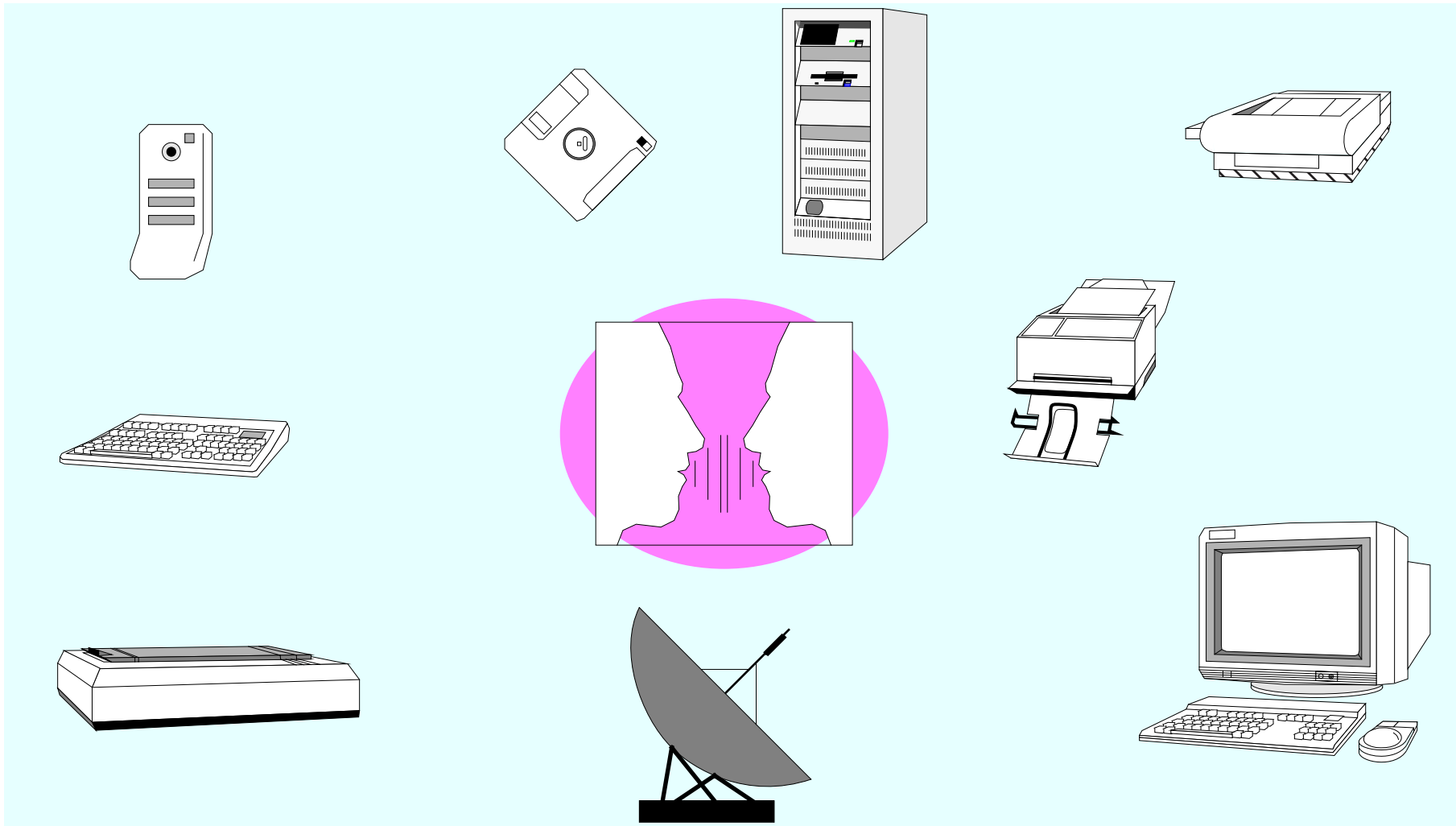
The user is required to perform many tedious and repetitive steps to perform one action

- Ideally the user just clicks on the printer icon and the browser instructs the printer to negotiate with the server printing the manual
- Style sheets do not solve this problem
- A Web site can have many facets
 - *fast or slow link view*
 - *PDA view*
 - *hard copy*
 - *read or browse*
 - *different environment capabilities*
 - *different audiences*
 - ...

Anthropocentric View of Web Publishing

10

Contents is communicated via documents



Business Culture

Two paradigms for conducting business

transaction driven	relation driven
how much does it cost? when can you deliver it?	do you know somebody who can sell me ...?
sign this contract	my bank will talk to your bank
trading people	agrarian
USA, Middle East	Japan, Europe

- Identity management and networking are key for success in the USA

Case Study: Aldus Manutius

12

Venice in 1500

- Gutenberg's printing press
- Brain drain from Byzantine Empire
- Well educated Venetians with disposable income



Key idea # 1:

- Become a publisher, not a printer
- People pay for contents, not for media

Values

Key idea # 2: Make value judgements

- Decide to drop comments
- Segment market and use best language for each segment

Drowning in Data — Starving for Information

WWW Values

- The W^3 is about publishing
- With today's tools, the average American citizen can be at most a poster and a consumer
- Layout and graphic tools of desktop publishing are just details
- Opportunity: help the American citizen to become a publisher

New Holy Grail: Structure

- Tools are needed to help people organize and compile information into knowledge
- Quality of knowledge is measured by how effectively it is communicated
- Effective communication requires clear organization
- Clear organization is achieved by introducing good structures

From Chaos to Order

Plagiarized from 1996 Xerox Annual Report

Help people to

- make W^3 publishing better
- make better W^3 publications
- allow to work better with W^3 publications

- New interpretation of mathematics after 1935
- Relational construct: a set with a relation
- System of axioms represents properties of constructs
- Mathematical creativity: find new constructs by defining maps that preserve the relations
- Two-step approach
 - *find a good system of axioms*
 - *find a good isomorphic construct*

Hypertext System

- W^3 is like HyperCard with the data stored on the Internet instead in a file
- User wanders around by clicking on hot spots representing links
- Several other hypertext systems have been conceived and implemented over the years

Uniform Resource Locator

- protocol — *http*
- username/password — *frank:secret1*
- host name — *www.hp.com*
- file path — *~smith*
- paragraph name — *8765*

<http://frank:secret1@www.hp.com/~smith#8751>

- Text (character sequences)
- Images
- Areas in images (image maps)

Ariadne's String

vs.

- Surfing the net, channels, cruising
- Goal oriented travel
 - *Acquisition of knowledge*
- Many roads can lead to the destination

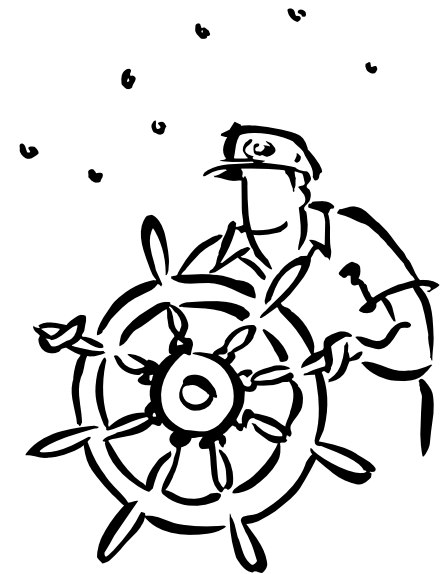


Avoiding to Get Lost

22

Jean Weydert's questions:

- Where am I ?
- What can I do here ?
- How did I get here ?
- Where can I go, and how do I get there ?



Defining a Construct Set

- *Site*: a neighborhood in the space of data, consisting of those data items to which the user has direct access at a given moment
- *Mode*: a subset of the set of commands, consisting of those commands that are active at a given moment.
- *Trail*: a feasible time-sequence of pairs <current site, current mode>
- *Trail editor*: supports the conventional help function (inspect and extrapolate the user's current trail) and conventional command macros (trail editing and re-using past trails)
- *Universal commands*: a small set of commands that are always available

Introducing Relations

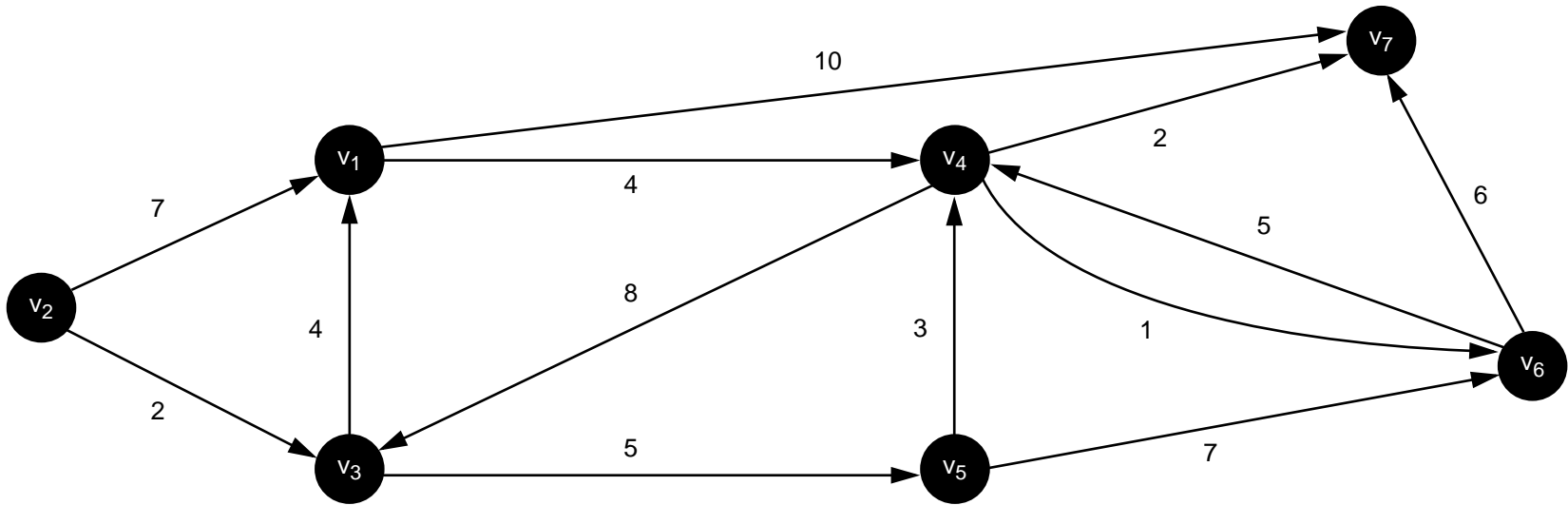
- Not all links are equal
- How long does it take to get there ?
- How many words do I have to read to reach the hot spot ?
- How difficult are the words ?
- ... the sentences ?
- ... the concepts ?

Categorization

The most difficult step in distilling knowledge from information is categorization

- Key words — *image processing, JPEG, rate-control*
- Difficulty — *beginner, intermediate, advanced*
- Audience — *student, scientist, buyer, family*
- Discourse level — *main thread, note, comment, reference, detail, source*
- Presentation medium — *computer monitor, TV set, printer, communications speed*
- Many more ...

Weighted digraph



Plurality of Graphs

There are many possible graphs for a fixed set of nodes

- **Example: One graph for each category or set of categories**
 - *graph by out medium: PC monitor, printer, TV, PDA*
 - *graph by difficulty: pupil, student, scholar*
 - *graph by audience: family, buyer, developer, user*
 - *graph by interest level: curious, information seeker, desperate*
 - *graph by intellectual challenge: tabloid, encyclopedia, treatise*
 - *graph by spin: republican, democrat, green, tory, socialist*
 - *graph by ethnicity: Afro-, Native-, Asian-, Hispano-American*
 - *graph by subculture: jet-set, VC, transcendental*
 - ...

Lessons from CAI Systems

28

- Navigational information is valuable only if data is structured systematically
- Users get lost in generic graphs
- Cycles make it most difficult to stay on course

Trees

- Hierarchical
- Root
- No ambiguity
 - *exactly one path between two vertices*

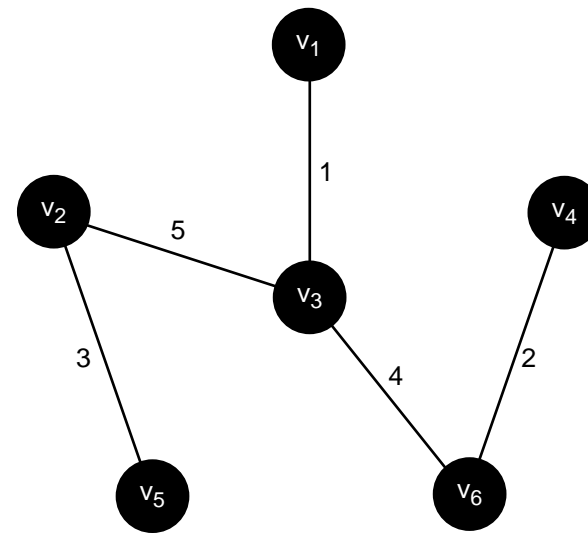
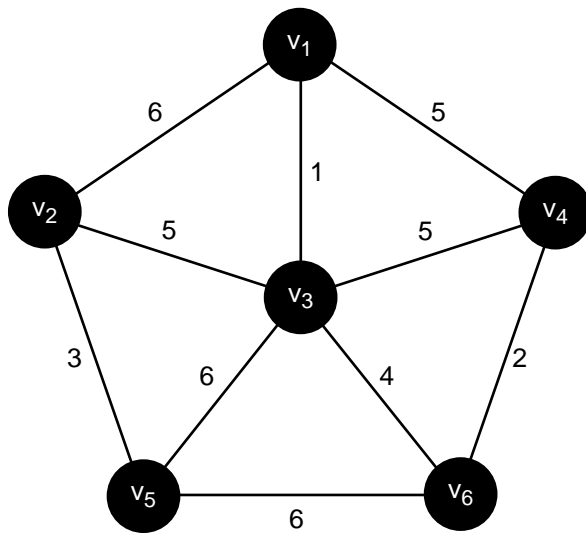
Conclusion:

Each collection of Web pages should be organized as a tree

Note: Each graph represents a collection

Spanning Trees

A tree that is a subgraph of G and contains every vertex of G is called a **spanning tree** of G



- What good properties are known about spanning trees?
- The spanning tree with minimum total edge weight is called a *minimum spanning tree* (MST)
- Good algorithms are readily available to compute an MST
 - *Prim's algorithm*
 - *Kruskal algorithm*

What Have We Done ?

- We have designed a methodology where a user collects information and links it
- With help from a standard lexical analysis program the user can assign weights to the links
- The authoring tool finds the MST
- The MST publishes the knowledge at the lowest intellectual cost
- The method scales well and be applied to sites with dozens of diverse pages

Is it Limiting ?

Have we limited the author's creativity ?

- No !
- There are many graphs for each web site, predicated by the categorization
- Authors can create rich knowledge by interconnecting MSTs
- Rich \leftrightarrow compelling

Role of the Tool

- The authoring tool does not necessarily automatically build a Web site (but it can)
- The tool proposes good structures (e.g., one per category) that a human editor can interweave
- Important property: it scales !
- Example: the comments thrown out by Manutius can be reinstated

Requirement for Browsers

Need ability to disambiguate links by category

- Proposal: encode using color
- Color is related to appearance
- Category to structure

Other proposed syntactic changes:

- Use symbols to tag each paragraph by category
- In each HTML file's header section, store a representation of the current subtree
- For each node of such subtree, list all attributes

Benefits for Web Printing

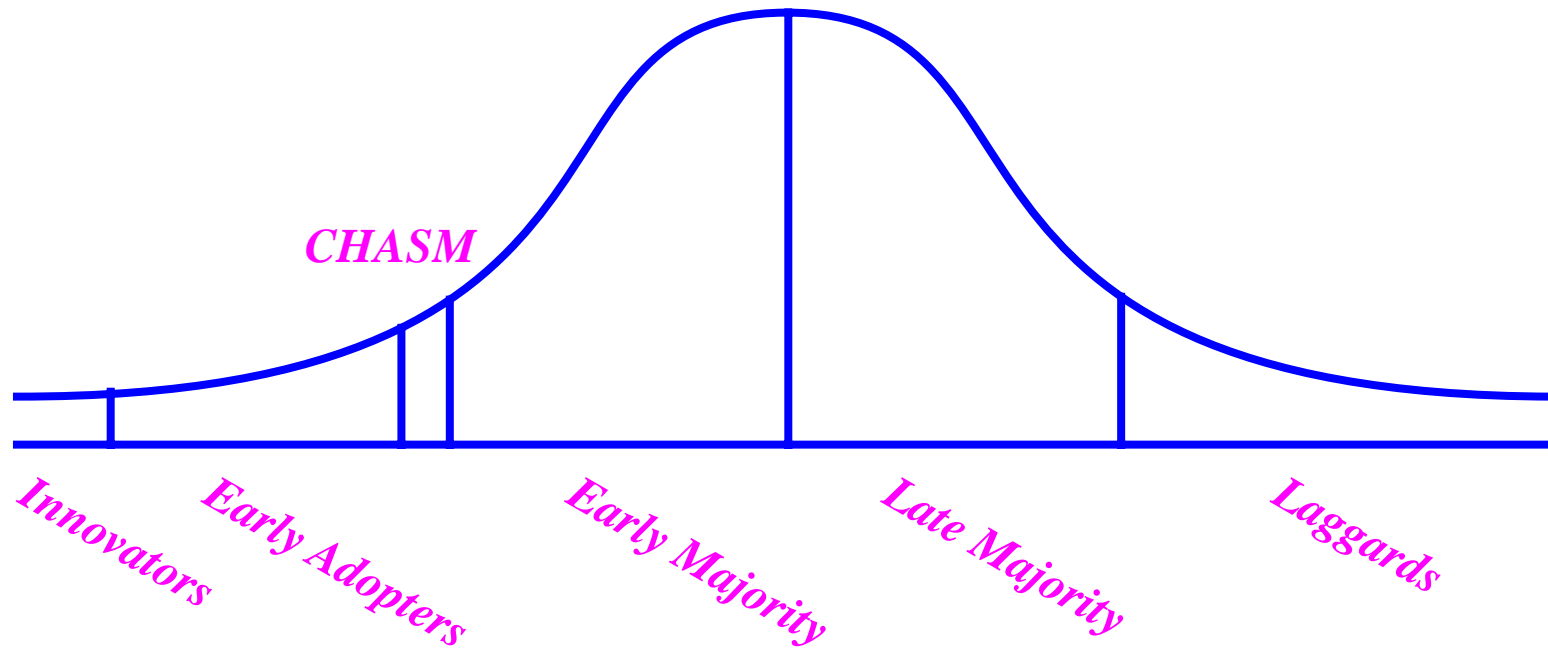
- Tree is easily linearized into a print job, e.g., by traversing it depth-first
- Size of print job can easily be estimated when file size attributes are stored with a category related to file implementation details
- Contents rendition for hard copy can easily be assembled from categories
 - [http://frank:secret1@www.hp.com/~smith#8751\\$facet=print](http://frank:secret1@www.hp.com/~smith#8751$facet=print)
- Same for cellular palmtop devices
- Asynchronous down-loading and printing enabled by subtree availability (look-ahead)

Automatic browse modes

- By tree traversal algorithm
 - *depth-first*
 - conventional
 - *breadth-first*
 - manager's view
 - compare to FlashPix a.k.a. NIF



Technology Adoption Life Cycle



- *Innovators*: use search engines, compile lists of cool sites
- *Early adopters*: use SiteMill or FrontPage etc. to better communicate their knowledge. hard labor to build systems that scale
- *Early majority*: still trying to figure out what to do with the Internet Assistants and PageMill thrown at them

The chasm has not yet been crossed

Manager's Lingo (continued)

41

- Tornado: Internet frenzy
- Chasm: Cross it with this scalable methodology
- Bowling pin: Reason for doing this despite HP not being a software company

Evaluating Alternatives

- Search engines: good for finding information, not for delivering knowledge

to be or not to be

- Java: a possible delivery medium for an HP authoring tool (bowling pin)
- Other tools or targets:
 - *HTML editors: for people who just desire W^3 presence (1–3 pages)*
 - *Site editors: small personal sites (< 12 pages)*
 - *Databases: publish repetitive information*
 - *No commercial turn-key tools for complex knowledge*
- Issue: how well does it scale ?

Opportunities

Hypertext is old and a large body of knowledge is available for exploitation. Examples:

- Vannevar Bush: Memex
- Doug Engelbart: NLS Journal
- Ted Nelson: Xanadu
- Jay Nievergelt: XS-1

DIMS Involvement

Distributed Imaging Systems Project

- Ho John Lee (PM) and Andrew H. Mutz involved in IETF and W³C
- Working in conjunction with INS and Bristol Labs
- Thank you for attending our Symposium