

MAPA: a system for inducing and visualizing hierarchy in websites

David Durand and Paul Kahn

Dynamic Diagrams, Inc.
12 Bassett Street

Providence, RI 02903, USA

Tel: 1-401-331-2014

E-mail: david, paul@dynamicdiagrams.com

ABSTRACT

The MAPA system provides improved navigation facility for large web sites. It extracts a hierarchical structure from an arbitrary web site, with some minimal user assistance, and creates an interactive map of that site that can be used for orientation and navigation. MAPA is designed and most useful for large web sites of from 500 to 50,000 pages. We present an overview of the mapping problem, with a list of 10 important user facilities that maps can offer. Then we describe how the MAPA system analyzes the link structure of a site, and provides effective aids for the navigation of large hypertexts. We also compare MAPA with a number of other web-mapping systems, and conclude with a review of how MAPA stands with respect to our wish-list of map features.

KEYWORDS: hypertext interfaces, structural analysis, hierarchical organization, WWW, Web mapping, data visualization.

INTRODUCTION

Maps are tools to help us recognize where we are, plan where we want to go, and tell us when we have arrived at our destination. While a hypertext is not the world, it presents similar problems of orientation and wayfinding, and a graphical summary of its structure can serve many purposes. We revisit the question of how to map hypertexts, and describe an integrated approach to mapping sets of documents on the WWW that encompasses extraction of useful orientation information from web pages and its clear presentation in a novel visualization.

REQUIREMENTS FOR MAPPING WEB SITES

There is significant precedent in hypertext research for visualizing networks of linked notes. Empirical studies as well as common user experience have shown such visualizations to be both useful and important.

We offer a list of requirements for a web site map. These requirements are a wish-list, answering the question: "What should the map of a web site do?" We distinguish between the requirements of two audiences: the web site visitor and the web site administrator.

Motivation for mapping web sites

The basic concept of hypertext is free connections of associated material. "Free" indicates a non-structured set of connections, links that are not limited by a pre-determined pattern. Bush's imaginary memex device was to be capable

of creating permanent trails among published documents at the reader's command. These trails were intended to record the reader's personal associations, regardless of the location of a story in some editor's or librarian's classification scheme [4]. Similarly, the motivation for [13], where the term hypertext was introduced, was to propose a more flexible file system in which documents could be reached from many locations by links.

The designers of the World Wide Web had this tradition in mind when they created the link mechanism on which the web is based. As Berners-Lee stated in an unpublished talk at Hypertext '93 promoting the World Wide Web as a universal communications medium, "it is a web not a tree". From the very beginning of the Hypertext Transfer Protocol (HTTP) and the Hypertext Markup Language (HTML), any document on the web could be linked to any other document on the web, irrespective of the document's location in a file system. In this sense, the web today is the logical outcome of basic hypertext.

However, while these hypertext features have been successful at making an unprecedented number of documents available on the Internet, web sites suffer from the long-standing visualization problems of hypertext systems: lack of physical context and an organizational paradigm.

Lack of physical context. The covers of a printed document provide a physical context to what it contains. The viewer can see and feel boundaries and sense the part-to-whole relationship of each portion of the page being viewed. In contrast, a web site is viewed one page at a time. While there is a sense of where to start (e.g. the convention of the Home Page), there is no sense of a finish, and often no clear part-to-whole relationship for any page being viewed.

Lack of organizational paradigm. A printed document has a clear order of presentation. On a web site anything can be connected to anything, and the intended order(s) are more difficult to convey. [10] argues for the distinction of local and global navigation structures in web page design. While this is now common practice, there is still great dissimilarity among the organizational paradigms being used.

Need for orientation. The invisibility of the web and the requirement to view its contents one page at a time creates a strong need for orientation cues. Viewers' understanding is

reinforced by knowing where they have been and where they can go at each point in their reading experience. Just as important, they may need to know where they are.

Need for link summaries. Some other differences between the web and previous research hypertext systems affect user orientation. Web browsers do not provide a concise summary of the outgoing links from a page. This was a feature of Intermedia's "tracking map" [20] that helped users to plan their navigation, and even to follow a link before finding its context in the document. A related problem is the unidirectionality of WWW links. There is no way for incoming links to be detected, since there is no central link database to consult. This is a problem for both readers and authors of web pages. While the nature of the web protocols and markup standards makes it impossible to construct a universal link database, it can be solved for well-defined subsets of the web. Knowing what documents (or document parts) refer to a web page can give additional contextual information and navigational options. A solution to this problem is interesting for both readers and authors of web pages and administrators of web sites.

We would argue that it is not additional cognitive overhead that makes navigating in hypertext so difficult, as claimed in [7]. It is the lack of physical context and a clear organizational paradigm in the presentation of the information on the screen that contributes to the reader's disorientation. [6] summarizes data showing that graphical maps improve the efficiency of using hypermedia. We concur and assert that making a map of a web site available to the reader can address the failing noted above and significantly improve navigation.

What the map of a web site should do

The map is the most basic visual orientation tool: a diagram of relationships in physical space that describes objects in the physical world. Maps have to satisfy two requirements to be useful in orientation. First, the viewer must be able to comprehend what the diagram represents. Second, the viewer must be able to perceive his own location within the diagram. We can build up a set of requirements for the map of a web site by reflecting on the problems faced by the web site reader, author and administrator.

1. Show high-level view of web site structure. The map should provide an overview of the important parts of the web site. Such a high-level view requires a filtering of major destinations from all possible destinations that can be reached.

2. Show where I am. The map should distinguish the page that represents the viewer's current location from other pages. The location of this page in the web site structure should be visible.

3. Show where I have been. The map should distinguish for the reader the portions of the web site that have been visited and the portion that remains unexplored. Visited pages should not be limited to the current session.

4. Show where I can go from here. The map should show the reader the path or paths that can be followed from the current page. This would be a summary of link relations for a page.

5. Distinguish peer group relationships. The map should show the reader the pages on the web site that are in the same group as the current page. This is often a matter of filtering out links to pages that are within the same peer group as the current page from the summary of link relations.

6. Show how I got here. The map should show the reader the path followed to arrive at the current page. This feature should not be confused with a representation of the reader's general browsing history, which will require a more general view of the web as a whole, rather than a specific view of a reader's path within a single web site.

7. Show how many readers have been here. A web site is a public space, visited simultaneously and asynchronously by many readers. The map should show how many other visitors have viewed the current page. Displaying quantities of visitors per page over large sets of pages on a web site gives the reader the same kind of context as physical wear on the page of a magazine or an often-read library book. Such a map also helps an administrator understand the distribution of traffic visiting a web site.

8. Distinguish interesting pages. The map should show the pages on the web site whose contents would be of interest to the reader. This feature is simple to understand and very difficult to achieve. First, it requires a model of the reader's identity, something not needed for accomplishing any of the previous features. Second, it requires a classification of the reader's interests. Third, it requires a classification of the content of pages on the web site in terms that can be correlated to the reader's interests.

9. Display all this information in a very small space. The map should be legible in the restricted amount of space available on the reader's screen. A restriction of 600 pixels wide by 400 pixels high is a realistic target. Any map that requires a larger display area will be unavailable or difficult to see for readers with a 640 by 480 pixel display. Zooming or panning around in a larger image should be avoided wherever possible.

10. Display the map in a web browser without requiring additional software or lengthy download time. The map should appear quickly and work easily in a standard web browser. Having to retrieve an external program or browser plug-in may or may not be acceptable. This is very important to the reader who occasionally visits a web site and less important to an author or administrator who effectively "live" there.

This list covers a broad range of features. In our work we have focused on features 1-5 and features 9-10. The other requirements offer interesting areas for further research.

TAXONOMY OF VISUALIZATION TECHNIQUES FOR HYPERTEXT STRUCTURES

In order to provide context for the visualization solution we created in MAPA, we will offer a taxonomy of possible techniques for visualizing hypertext structures. Since our solution shares some attributes with 3D visualization, we offer a summary of and comparison to these techniques. Finally, we note two examples from information graphics that have influenced the MAPA visualization.

Types of Maps for Hypertexts and Web sites

There are a number of structures that can be used to create visual maps of hypertexts in general and Web sites in specific. Taxonomically, we can distinguish several common forms: graph-based structures such as webs, hierarchies, and acyclic graphs; and spatial structures such as neighborhoods and abstract spatial metrics. Each of these is described below.

The graph-based structures that follow are “raw” hypertext models, in some sense, with navigation points grouped by the presence of link relationships. While these are frequently explicit hypertext links or some subset of the links, this is not essential to the structuring approach. Some systems (Gloor paper) actually generate relations between document to create overview maps. In general, any link-induction or link filtering technology can produce a graph-based structure for mapping purposes.

Web structures are in some sense the most general model possible for a hypertext map. Unfortunately, as the Intermedia project first demonstrated, and many systems since have discovered, a global web view usually shows too much information for a reader to easily assimilate. Web structures often suffer from a lack of meaningful topology. The distance between and relative position of nodes in the web are arbitrary. Thus the visual shapes created in a global web view have little or no real meaning.



Figure 1: Merzscope Visual Web Map of Systemcorp.com site

An interesting example of a web site map that takes the form of a web structure is the Merzscope Visual Web Map (<http://www.merzcom.com/>). This product builds a visualization of a selected set of links connected a selected set of pages on a web site (Fig. 1). In this map of the Systemcorp.com web site, pages are arranged in clusters around several hub nodes. While the detection of pages and links is done by a program, the selection and placement of the pages and links displayed in this map is done manually.

Hierarchical structures are a restriction on the form of the underlying graph. The advantage of hierarchies is that they support a strong notion of place: documents have clear superior/inferior relationships, sometimes augmented with linear precedence relationships between nodes. Hierarchies are also well known from their applications in many domains, such as classification systems, personnel structures, outlines, and so forth. The rigidity of hierarchical structures aids comprehension, but also creates a certain amount of inflexibility. Some structures, such as several orthogonal, equally important relationships, cannot be well represented as a single hierarchy.

Acyclic structures are a generalization from hierarchies. They allow a single node to be inferior to more than one other node, representing multiple inheritance. There is a certain naturalness to this kind of diagram for a hypertext, and it can be useful for small sets or restricted views within a larger set. However, it becomes very complex as the hypertext grows, quickly approaching the tangle of the web structure.

Spatial structures allow mapping by explicitly using a coordinate system to relate navigation points. They are the most straightforward way to create a purely spatial structure: some number of significant axes are chosen, and all navigation points are positioned in the space according to their projections on the significant axes. A map based on such a structure is a meaningful display of the spatial relationships thus created, typically by directly representing significant objects in positions according to their underlying coordinates. High-dimensionality coordinate systems present special problems and work is required to provide effective interpretation of such information [1].

Neighborhood structures are an interesting sub-case of spatial maps, since they lack a meaningful coordinate system. Some forms of cluster analysis algorithms like the *self-organizing map* [11, 12] position items within a metric space such that closely related items (under some measure) end up “close” to each other in the final map. The actual spatial coordinates are inherently arbitrary, and are only significant because closer coordinates tend to have more related items in them.

We chose hierarchical structure as the most appropriate basis for our visualization. We felt that the visualization should be based on the link relationships among the pages on the web site, rather than other content attributes of the pages and wanted our visualization to work independently from any particular authoring system. While acyclic

structures would provide more flexibility in expression, we felt that organizing link relationships into a single hierarchy offers a better opportunity to present a clear visualization for very large collections of pages.

We considered the techniques that have been developed for 3D Information Visualization Systems. Placing nodes in 3D space is a promising direction for mapping a complex structure such as a large web site. [14] provides three classifications of systems in his survey of 3D information visualization: mappings, presentation techniques, and dynamic techniques. He includes in this the work on cone trees [16], a technique for placing hierarchy (tree) structures into 3D space to create denser information presentation.

All these systems need a metric for placing nodes in space. The simple example is visualization of a file system hierarchy. Many of the systems Young surveys organize nodes according to some measure of content similarity, similar to the neighborhood structures described above. An interesting example of this is the Bead system described by [5] which creates document “landscapes” by grouping nodes according to information retrieval characteristics.

The tendency in these systems is to use computational methods to place solid objects in 3D space. The user can then “fly through” the model to view and understand it. Other examples of applying this 3D fly-through technique to web site hypertext structures include Apple HotSauce and Perspecta SmartContent viewer.

HotSauce was available from Apple Computer from September 1996 through November 1997 (Fig. 2). The product came from Project X, a visualization technique developed at Apple’s Research Lab for displaying hierarchies contained in a file expressed in Meta Content Format (MCF). The visualization represented pages as color-coded text labels suspended in a 3D space. Labels higher in the hierarchy appeared larger, and overlapped lower labels. The reader could “fly” into the labels, select and reposition them, or click on them to navigate to a page.

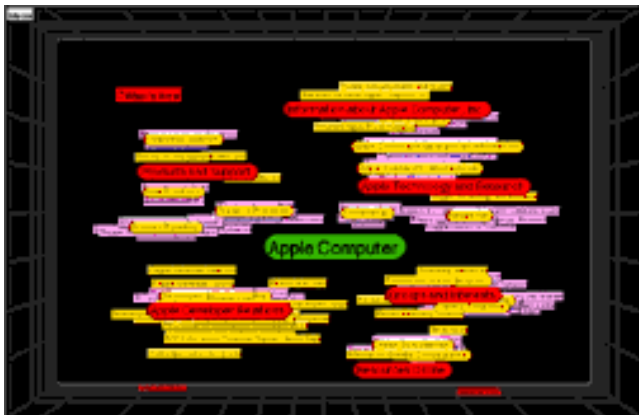


Figure 2: HotSauce view of the Apple Computer site



Figure 3: Perspecta SmartContent viewer

The *Perspecta SmartContent* system is a technique for organizing, visualizing and browsing knowledge structures, based on work described in [19] and [15]. Textual labels are placed in 3D space and visually grouped and/or connected by lines to express relationships (Fig. 3). The relative size of the text grows and shrinks as the reader “flies” into the information space.

Predecessors from the Information Graphic Tradition

The tradition of information graphics goes far beyond the recent work in computer-aided visualization. Since a large part of the problem in mapping web sites is simply fitting a lot of information in restricted space (requirement #10), we also reviewed a range of information graphic techniques used in printed maps. Two examples provided important lessons for the MAPA visualization: the Turgot Plan of Paris and the Beck diagram of the London Underground.

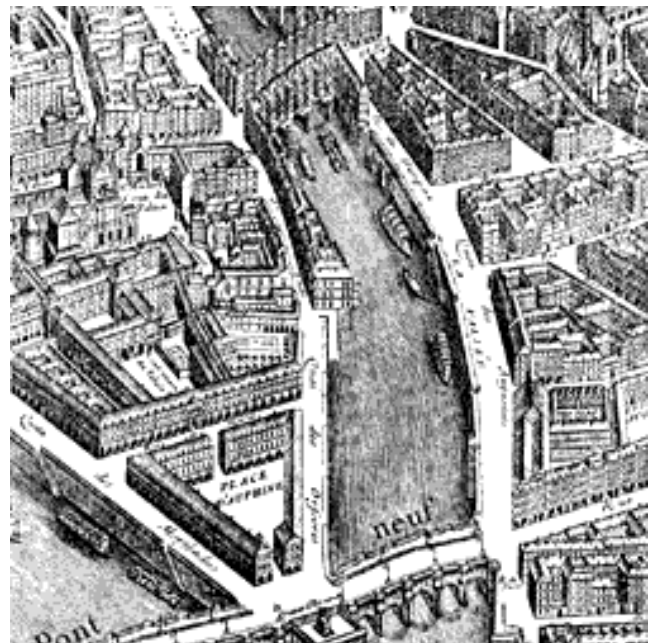


Figure 4: Small section from the Turgot Plan of Paris

The *Turgot Plan of Paris* (Fig. 4), reproduced in [17], is named for the official who commissioned it and was drawn and engraved by Louis Brétez and Claude Lucas between 1734 and 1739. It is one of the most remarkable examples of eighteenth century engraving and design, its twenty sheets offering an elevated view of every building structure in the entire city. The map is often incorrectly called a “perspective plan” [9], but in fact the map uses an orthographic projection with no vanishing point. This technique creates the least amount of visual distortion. By raising the projection’s point of view the engravers can represent the architectural detail and still have space to label the major roads.



Figure 5: Detail of central London from the original Beck diagram

The *London Underground* plan was developed by Harry Beck in 1933. Its development and application are described in [8]. This diagram (Fig. 5) contains three important innovations that we wished to apply in our work. First, Beck stylized the angles of the elements in his plan. The River Thames does not actually flow at 45-degree angles, nor does the Underground run in an absolutely straight line between Notting Hill Gate and Oxford Circus. By placing these line elements on a regular grid, Beck created a simple visual rhythm that is easy to grasp. Second, his plan distorts the distance between stations (nodes) to accommodate a regular placement of text labels. The distance between two stations on a map is far less important than the viewer’s ability to read their names. Third, he used bold and contrasting colors to separate the different “lines” that make up the Underground system. The colors make it easy to grasp the overall pattern of a line and help to visually separate densely packed information.

From the *Turgot Plan* we wanted to apply the lesson of orthographic projection as a method for using the “z-factor” of 3D space without the distortions associated with perspective projections. From the Beck Underground diagram we wanted to apply the lessons of a regular grid, placement that favored legibility of text labels, and use of bold contrasting colors.

A DESCRIPTION OF THE MAPA SYSTEM

The MAPA system induces and builds a hierarchy from the network of hypertext links among pages on a web site, and provides a navigable, interactive visualization of the resulting hierarchy. This visualization, in the terminology described above, is a comprehensive map with an underlying hierarchical structure, presented as interlinked local maps. The MAPA software is independent of the web server and authoring system used to create the site being mapped and works entirely from the analysis of linked HTML documents. MAPA can run anywhere as it accesses the site via HTTP, just as a user would.

MAPA gathers link information from a web site with a *web walker*. This walker determines the boundaries of the site by means of pattern-matching on encountered URLs, and gathers a number of pieces of information that are used in later processing. The walker’s results are processed by the *organizer*, which induces a hierarchy from the walker data, and stores that hierarchy and the link and meta-information gathered by the walker in a central database. This database contains complete link connectivity information for the site as well as the results of the organizer. The *link list* is a CGI-generated page that can show a summary of all the links relating to a particular document. The final component of MAPA proper is a *visualization applet*, written in Java, that communicates with the database and creates dynamic maps of the hierarchy created by the organizer.

The Walker

The walker is responsible for gathering several pieces of information about page content from in-document markup, including HTML `<meta>` tags. This includes some specialized information to adjust the automatic processing of the organizer:

- *An MD5 digest of the retrieved page contents.* This is used to resolve “URL aliasing” problems. The fact that many web pages have more than one absolute URL creates havoc in creating a meaningful web map, since pages will appear more than once, as if they were separate nodes. The MD5 digest provides a practical way to efficiently test for document uniqueness with very low probability of failure.
- *The page title.* This is used to identify pages in the map.
- *Equivalent Pages.* these are URLs that are functionally the same; this information is used to collapse links and references to the primary version of the pages on the map. A typical use of the equivalence markings is to collapse imagemap and text-only versions of a page.
- *Email address of the page’s owner.* MAPA tracks changes to page information, and can send email notifications to the page’s owner when that information changes.
- *Links and link type information.* MAPA uses either HTML `<meta>` tags or the REL attribute to indicate link types that can influence the mapping process. This information is saved along with the destination information for every link.

- *Parent page status.* This records whether the page should be treated as a navigational hub by the organizer. The set of parent pages influences the structure of the resulting map.

Metadata about pages and links can also be separately stored in the database. Such external metadata must be separately managed along with the contents of the page itself, but in some contexts it is a significant convenience to adjust mapping information while viewing the page and map simultaneously. MAPA can send mail to the responsible parties with updated markup to be placed in the document itself when external meta-data is modified.

The Organizer

The organizer program examines the total link topology determined by the walker, and culls those links to create a single hierarchy, using a combination of heuristics and any metadata added by the user. The resulting hierarchy is loaded into the database and serves as the basis for generating the interactive maps. Practical use of the system over more than a year has shown that only a small fraction of pages in the first few levels of a site need to be hand marked to produce a useful map.

MAPA extends previous work on structural analysis of hypertexts, and takes advantage of a critical practical difference between the authoring and reading conventions of the World Wide Web and those of hypertext in the abstract. That key difference is the notion of a “home page” — an author-defined unique starting point for readers of a site. This eliminates, at the outset, one of the more significant problems for previous systems: the determination of a good “root node” for a hierarchy. In part because of the lack of a unique starting point, much previous work on structural analysis of hypertexts [2, 3] is based on relatively computationally intensive algorithms such as the *all-pairs shortest path* problem (with equally weighted links).

The simplest form of hierarchy induction is the one implemented in most commercial web mapping products: traverse the site in depth-first order, starting at the home page. All links are treated the same, and pages show up at a hierarchy depth that is determined by the smallest number of links from the home page. The problem is that fewest-links navigation does not necessarily reflect the author’s intended organization. MAPA’s organizer improves on this by using slightly different algorithms, exploiting the “home page advantage.” Rather than depth first enumeration of nodes, MAPA uses Dijkstra’s algorithm for the *single-source shortest path* problem, with heuristically-determined and user-assigned link weights that influence the shape of the map.

Even in the case where all heuristics fail, and links are equally weighted, the hierarchy produced by MAPA will have a significant difference from other web analysis systems, since duplicate URLs will be collapsed, something that can decrease apparent map size by a significant amount, and that does not confuse users by representing the same information more than once in different places.

We have also changed Dijkstra’s algorithm to support multiple inheritance by maintaining, for each node, a list of incoming links with equal weight. This tells us that for typical sites about 10% of all nodes are ambiguously placed using our heuristics: interestingly, the arbitrariness of the placement of non-parent nodes is only rarely of great concern. Currently, we do not visualize this information, although we do provide a count of multiple ancestry pages as feedback to the map-builder. We conjecture that the extra information gained from visualizing multiple ancestry may not be worth abandoning the simplicity and familiarity of hierarchical organization.

Unlike some of the commercially available manual tools, the map maker does not directly tell MAPA where to put a page. Instead the map maker tells the organizer what the parent/child relationship is between specific pages, and the program uses this information to organize the hierarchy. A real advantage of this kind of control is that it is local to individual pages and links between pages. Changes to the web site may invalidate some relationships, but whatever metadata remains is still useful in building the next map.

The organizer also allows the map maker to assign parent/child and equivalence relationships in the link information, without having to modify documents on the web site. Attributes are assigned directly to the page record in the database table, which then influences the subsequent organization of the map.

The organizer supports loosely-coupled collaboration via email. The author email `<meta>` tag is used to automatically notify both the author of a page and the map maker when modifications to the parent/child or equivalent information for a page are committed.

Link weighting. MAPA currently implements a very small number of simple link-weighting strategies. The organizer is configured with a weights file that lists a series of tests, in a user defined order. Each successful test modifies the link weight, sets the link weight, and can also terminate the weight determination process. While a large number of conditions could be used to adjust link weights, the following are implemented:

- *Explicit child relationship.* If a link is marked as a child link in the metadata, this condition is true.
- *Parent page status.* If both ends of a link are parent pages, this condition is true.
- *Parent non-child status.* This condition is true for links that are not marked as child links, but whose source is marked as a parent page.
- *URL sub-path match.* These conditions test the URL hierarchy of the source and destination, and allow the detection of whether a potential child link points up or down the URL hierarchy.

The usual mode of using MAPA is to give all child links a weight of 0 (to ensure that they will be part of any resulting map). Parent page status reduces the weight, as does URL sub-path matching. The “parent non-child” predicate was an experiment in tighter page placement control that had

parent pages push away nodes not their children. The results proved confusing and unreliable when sites changed, since failure to update the child information would cause newly linked pages to be pushed to surprising places.

While a large number of possible heuristics might be used to weight links, we are most interested in adding a new metadata type we refer to as *child repellent*. Many web sites have pages with ancillary comprehensive indexes of pages. Unfortunately, such index pages are often easily accessible from the home page, and create low weight but inappropriate paths to pages that are more appropriately placed deep in the site hierarchy. A page marked with child repellent would place an effectively infinite weight on all its outgoing links, so that a page would only end up as a child if it was not accessible by any other path.

Also potentially useful would be heuristics to identify candidate parent pages. Botafogo [2] defines statistical metrics for identifying candidate *index nodes*: nodes that link to many other nodes; and candidate *reference nodes*: nodes that are referred to from many places. These measures could be used in identifying MAPA parent nodes, and child repellent nodes.

Some more useful link type inference rules have been defined and explored in the ParaSite system [18]. While ParaSite attempts to gather and use information about the topics of pages, to produce more meaningful search results, it also needs to look for hierarchical sub-structures of the spaghetti that is the Web. ParaSite uses link data for the same reason that we do: unlike the output of an automated analysis, each link is known to be meaningful — at least to the author of the document. In addition to picking up hierarchy information from the structure of URLs, ParaSite uses the internal markup structure of a document to determine possible hierarchical relationships between links. This would be useful to the organizer.

The MAPA visualization

Splitting the map and the link list makes complete local navigation information available to users without cluttering the map. The graphical map is a *local* map, providing a perspective on the location of a single page, the *focus page*, within the overall structure of a site. Like the maps commonly found in public places such as malls, hotels or museums, MAPA aims to answer the question “where am I?” The following information from the organizer’s hierarchical decomposition of the web is presented in the graphical overview (Fig. 6):

- *The focus page.* The focus page is presented in a central position to clearly highlight its importance to the map.
- *The ancestry of the page.* The chain of parent pages up to the home page is indicated by a series of iconified pages. This is placed on the lower right of the diagram, along the diagonal axis of the isometric projection. This position in the diagram, while offering very little room to spread, allows for a large number of overlapping page glyphs arranged in a stack.

- *The child pages.* All of the children of this page are shown. The child pages are placed along the horizontal diagonal of the isometric projection. This position allows for the largest spread of page glyphs with no overlap.
- *The grandchild pages.* In order to see the underlying scope of the immediate navigational choices (the child pages) the grandchild pages are also shown, indicating not only what navigational choices may be possible, but giving a generally reliable visual estimate of what that scope of different subparts of the site may be. The grandchild pages are placed behind the child pages in parallel rows. This position allows room for the largest number of overlapping page glyphs arranged in stacks. The page glyph overlap always exposes the top and right portion of the glyph, which is enough to provide both a target for viewer interaction and a space for displaying information symbols.

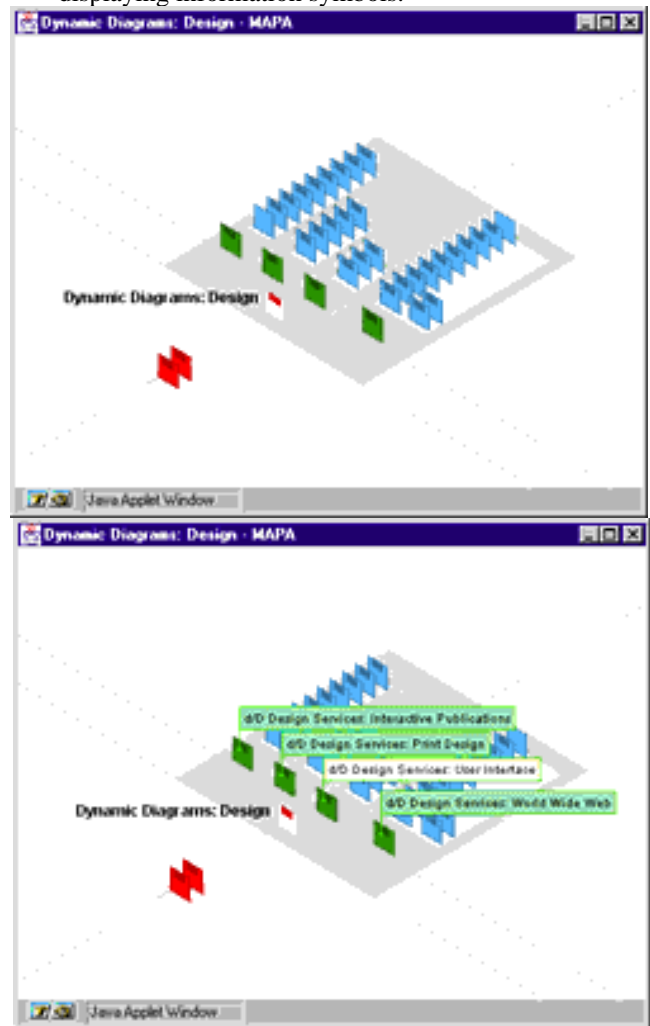


Figure 6: MAPA 2.0 Visualization
Title flags appear when the mouse remains over a page.

- *Page titles.* It is essential to provide textual information about the identities of pages on the map (though web-page titles are not always as carefully chosen as they

should be!) Title information is dynamically displayed for groups of pages: ancestors, children, and groups of grandchildren under a particular child page. (Fig. 6).

The most original feature of this visualization is the use of placement in a virtual space to indicate logical connections between pages. Most visualizations of hypertexts are extremely cluttered with lines representing links. The MAPA visualization displays its set of “important” links purely by front-to back arrangement in space, visual clustering by means of controlled amounts of empty space, and the use of colored “carpets” to reinforce and enhance the spatial groupings created by the empty space.

The reduction in visual clutter allowed by the removal of explicit link representations enables the display and assimilation of information about a very large number of pages: MAPA displays of more than 500 pages easily fit in their entirety on a typical 17 inch monitor (though a complete view of so many pages will not fit in a 600 x 400 window without scrolling). MAPA maps are extremely compact compared to most variations of linked node displays. The only other display that fits a comparable number of items in a single overview diagram is the cone tree visualization [16]. While the flying and rotation abilities of the cone tree make it more flexible for visualization, MAPA maps don’t require the user to control 3-dimensional rotations, nor can objects block one another. The only way information can be hidden in MAPA is due to window size, and this can be handled by the common user interface gestures of resizing the window, or dragging the view in two dimensions.



Figure 7: InXight Visualization

When we compare the display techniques such as the fish-eye view hyperbolic tree used in InXight and the 3D perspective placement of text labels in Perspecta to the

MAPA display, we see that only MAPA manages to present the three levels of context within a single display.

The majority of the display area in the InXight view (Fig. 7) is taken up with lines connecting text labels. We are given a great deal of information about connections but little information about what is being connected. The labels themselves are truncated by the layout in such a way that their meaning is difficult to comprehend. Because of the combination of fish-eye and parabola distortion it is impossible to see labels more than one link away from the focus node.

Perspecta (Fig. 3) has a very clear presentation of text labels, using good typographic forms and color contrast. This system shrinks the labels at the next level to represent depth in the display. Labels grow smoothly as the viewer goes deeper into the display, reinforcing the sense of depth. However, this shrinking and growing of the text leaves the viewer able to see at most one group of labels at a time, in addition to their ancestor.

Navigation in MAPA. The MAPA visualization also supports two forms of direct navigation:

- *Map navigation* enables the map to be refocused on any page visible on the display, given a new map centered on a different focus page. Thus a user can use MAPA as a local search tool for navigation of a complete hierarchy.
- *Site navigation* is also possible, since a MAPA map can be used to request display of a mapped page easily. Thus, if the answer to the question “Where am I?” is “I’d rather be *here*,” a MAPA map allows the user to quickly go to that information.

The interface for map refocusing takes advantage of object constancy [16] in a unique way. Instead of using a realistic set of spatial transformations, an animation is used to move any pages that are common to the starting map and the new map into their new positions. First, pages that will not be on the new map vanish (by “sinking into the ground”). Next, the pages that will be in the new map move to their final positions. Finally, the new pages appear (by “rising up from the ground”). The animation allows the user to see the conceptual relations between the old and new maps, and also allows them to see where any landmarks are in the new map.

See <http://www.dynamicdiagrams.net/minimapa.html>

The relatively unconventional use of an orthographic projection also has some useful properties for an interactive visualization. Since no foreshortening occurs, any size difference visible on the screen must be interpreted as a matter of visual emphasis. It also makes the display of page titles possible, since the titles can be placed on a regular grid (Fig. 6). The lack of a vanishing point also means that two-dimensional dragging of the view is visually as good as three-dimensional navigation – making for simpler implementation as well as simplified, and consistent user control.

Showing the title of only one page or no pages is too little information for an overview of a site. Showing every title creates a display consisting almost entirely of overlapping and partially illegible titles: too much information. Fortunately the conceptual groupings provided by the hierarchical model provide logical sets of pages for which title information can be displayed. The visualization allows sufficient space so that a legible, regularly laid-out display of such groups is possible.

Link List

The link list (Fig. 9) is the other major visualization feature of MAPA. Its purpose is to address, for navigation within a single site, the lack of a central database of link connectivity information. The link list presents summary information about other pages that are linked to or from a focus page, and sorts and labels these links as to type. The following types are distinguished:

- *Ancestor link.* This is a page that is in my chain of ancestors back to the root.
- *Parent page.* If the link comes from a page that has children in the hierarchy, then it is has a bar.
- *Child parent page.* This is an incoming link from a page that is a child of this page, and that has children itself.
- *Child leaf page.* This is an incoming link from a childless page that is a child of this page.
- *Other non-parent page.* This is a hierarchically unrelated page that is linked to this page and has no children.
- *Other parent page.* This is a hierarchically unrelated page that is linked to this page and has children.
- *External page (for outgoing links only)* This is a page that is not part of the site being mapped but that is linked to from this page.

This typology of links helps the user to scan and interpret what would otherwise be a long and undifferentiated list of potential destinations.

The provision of incoming links is intended to answer the question “How *else* could I have gotten here?” Especially in the case of interpreting the context of pages found as a result of a search, access to incoming links may provide a useful access point to related information.

Summary and Future work

We believe that MAPA is a useful solution to some of the mapping problems we listed in the introduction. We provide support for 7 of the 10 items on our wish list. Aside from the problems of usage tracking, and distinction of pages based on user interest, which we decided not to address, we fail only to directly provide peer information about the current page; the peers of a page are only shown for children and grandchildren of a focus page. To view the peers of a page, that page’s parent must be made the focus.

One of the most successful aspects of the MAPA visualization is the use of the orthographic layout. Indicating structural relationships without lots of graphical objects to show links frees needed screen real estate to

display additional orientation information, at the same time reducing visual clutter that makes the map harder to interpret.



Figure 9: MAPA Link List

The organizer not only provides one solution to the problem of finding a hierarchical arrangement of a web site. It provides a flexible and fully-general solution, based on the use of an optimization algorithm driven by any relevant link weight metric. We see several possible directions for improvement. The addition of child repellent, while simple, would increase the usability of MAPA considerably. Taking advantage of the ability to add more heuristics (particularly content similarity tests) for site organization is a clear avenue for exploration, as is creating a proper interface to enable control of the weighting function. We intend to implement and test a new strategy for mapping web sites that use frames (which destroy the unique URL->document mapping). More distant possibilities would be an editing interfac, so that meta-information could be managed directly in terms of the map itself. This requires faster, incremental algorithms for re-organizing the map, such as (Find proper citation for Ramalingam and Reps JA article).

One of the most interesting future prospects for MAPA is to explore the possibilities to overlay additional information

on the maps. Many distinctions of color, shape, size and decoration are available for the display of such additional information as search results, web traffic, topical relevance, and other data *about* a web site.

ACKNOWLEDGMENTS

The authors would like to thank the following people involved in the development of MAPA. Design: Krzysztof Lenk, John Shepherd; Software Engineering: Matt Ayers, Andrew Gilmartin; Documentation and Testing: Magdalena Kasman, Michael Roy, Elaine Froehlich, Joe Quackenbush; Consulting Support: Paul Mende, Elisabeth Bayle, Brook Conner, Patrick Chan. We would also like to thank Carol Moore, Ed Costello, and Alex Wright of the IBM Internet Program, and Rick Levine and Nicole Yankelovich of Sun Microsystems for their support in making the development of MAPA possible.

REFERENCES

- [1] Bershers, C. and S. Feiner. "Generating Efficient Virtual Worlds for Visualization Using Partial Evaluation and Dynamic Compilation" ACML SIGPLAN Symposium on Partial Evaluation and Semantics-Based Program Manipulation. 107-115, 1997.
- [2] Botafogo, R. A., E. Rivlin and B. Schneiderman. "Structural analysis of Hypertexts: Identifying hierarchies and Useful Metrics" TOIS. **10**(2): 142-180, 1992.
- [3] Botafogo, R. A. and B. Schneiderman. "Identifying Aggregates in Hypertext Structures" Hypertext '91. 63-74, 1991.
- [4] Bush, V. As We May Think. Atlantic Monthly. 101-108, 1945.
- [5] Chalmers, M. "Design perspectives in visualizing complex information" Proc of IFIP 3rd Visual Database Conf. 1995.
- [6] Chen, C. and R. Rada. "Interacting with hypermedia: a meta-analysis of experimental studies" HCI. **11**: 125-156., 1996.
- [7] Conklin, J. "Hypertext: An Introduction and Survey" IEEE Computer. (September): 17-41, 1987.
- [8] Garland, K. Mr. Beck's Underground Map. 1994 Capital Transport Publishing. Middlesex.
- [9] Hodgkiss, A. Understanding Maps, A systematic history of their use and development. 1981 Dawson. London.
- [10] Kahn, P. "Global and Local Hypermedia Design in the Encyclopaedia Africana" Hypermedia Design. Fraïsse, Garzotto, Isakowitz, Nanard and Nanard ed. 1995 Springer. London.
- [11] Kohonen, T. Self-Organizing Maps. Springer Series in Information Sciences. 1995
- [12] Lin, X. "Map Displays for Information Retrieval" Journal of the American Society for Information Science (JASIS). **48**(1): 40-54, 1997.
- [13] Nelson, T. H. "A File Structure for the Complex, the Changing and the Indeterminate" ACM Proceedings of the 20th National Conference. 84-100, 1965.
- [14] P. Young, C. T. and Report, Dept. of CS, Univ. of Durham, UK. Three Dimensional Information Visualization. CS Department, Univ. of Durham.: Report 12/96. 1996. URL: <http://www.dur.ac.uk/~dcs3py/pages/work/documents/3d-survey/IV-Survey/index.html>
- [15] Rennison, E., L. Strausfeld and D. Allport. "Issues of Gestural Navigation in Abstract Information Spaces" CHI. 1995.
- [16] Robertson, G. G., J. MacKinlay and S. Card. "Cone Trees: Animated 3D visualizations of hierarchical information" SIGCHI'91. 189-194, 1991.
- [17] Rouleau, B. Le Plan de Paris de Louis Bretez dit Plan de Turgot. 1989.
- [18] Spertus, E. "ParaSite: Mining Structural Information on the Web" World Wide Web Conference 6 Hyperproceedings. <http://www6.nttlabs.com/HyperNews/get/PAPER206.html>, 1996.
- [19] Strausfeld, L. "Financial Viewpoints: Using Point-of-view to Enable Understanding of Information" CHI. 1995.
- [20] Utting, K. and N. Yankelovich. "Context and Orientation in Hypermedia Networks" TOIS. **7**(1): 58-84, 1989.