WORKING PAPER SERIES

http://www.casa.ucl.ac.uk/working_papers/paper94.pdf

 ISSN: 1467-1298

94
07 | 05

# INFORMATION MAPS: TOOLS FOR DOCUMENT EXPLORATION

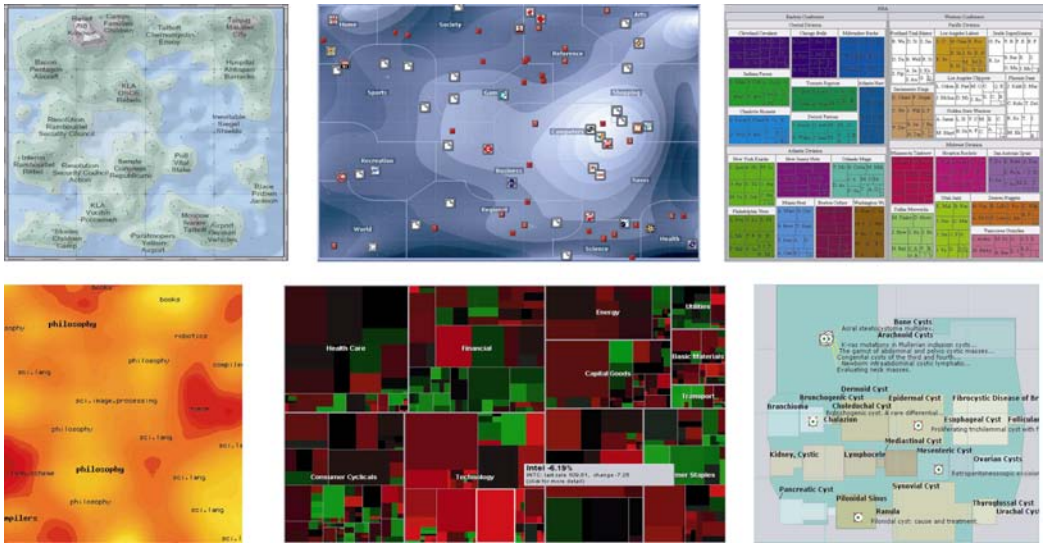
**Martin Dodge**

www.casa.ucl.ac.uk

T +44 (0) 20 7679 1782 • T +44 (0) 20 7679 1813 • F +44 (0) 20 7813 2843 • E casa@ucl.ac.uk

Centre for Advanced Spatial Analysis • University College London • 1 - 19 Torrington Place • Gower St • London • WC1E 7HB

# Information Maps:
# Tools for Document Exploration



## Introduction

> *So much has already been written about everything that you can't find out anything about it.*
>
> **James Thurber (1961)**

This white paper defines information mapping and reviews a range of examples. The focus is on two-dimensional interactive information maps that can be used to summarise large volumes of textual information, provide interfaces to browse the whole corpus and retrieve particular documents of interest. Information maps are being developed to tackle the modern-day challenges of information overload – too much, too fast and too unmitigated [1].

The average person sitting at a networked PC has access to a huge (and hugely growing) array of textual data that can be utilised as a vital information resource in helping to make decisions. The data are in the form of documents on their own hard disk, messages pouring in through email and IM, from databases on intranets and digital libraries on the Internet and, of course, there is the expanse of the World-Wide Web, encompassing several billion static pages and an unknown volume of dynamically generated content. All these data are available at a mouse-click but they are beyond easy comprehension in their original forms, and can more often hinder, rather than help, us in finding answers and making appropriate decisions. To take the case of the Web, we are all familiar with using current Internet search engines to find an answer to a basic query and being overwhelmed with a lengthy list of matching pages. Too often, one retreats from this type of information-seeking with a sense of defeat, having been buried by the sheer volume of potentially relevant data and yet not being able to find answers.

Information mapping is one of the most useful means of assisting information seeking by summarising and visualising the non-visual content, structures and interrelationships of thousands of documents. Crucially, the information map should

[1] *Data Smog: Surviving the Information Glut*, by David Shenk (Harper San Francisco, 1998).

summarise the salient content and context of the documents, unlocking their value, without one actually having to read them all.

## What is information mapping?

> *Where is the wisdom we have lost in knowledge?*
> *Where is the knowledge that we have lost in information?*
>
> **T.S. Elliot, *The Rock* (1934)**

There are many different ways of creating a useful, high-level representation of a large volume of data – for example: a well-written abstract or executive summary; a bar chart of a long column of numbers; a mathematical model. And, of course, there are maps. Cartography has proved itself to be a particularly useful and effective tool for summarising and representing large volumes of data. For several thousand years, people have created maps of the world around them as a means of understanding, and controlling, that world. Indeed some say map-like pictorial representations were one of the earliest means of recording and transferring information, pre-dating written languages and number systems[2]. Since the Renaissance, the developing science and art of cartography have given us maps that act as powerful information management tools, summarising large volumes of geographical data, usually on a single sheet of paper, so that one can comprehend the whole and yet also retrieve a particular detail (e.g. instructions for navigating between two points) [3].

Information maps are visual tools for the abstraction, summarisation and presentation of large volumes of data which facilitate interactive exploration by users[4]. They usually map the information in a non-geographical, abstract space.

---

[2] *Cognitive origins of graphic conventions*, Barbara Tversky. In Understanding Images, edited by F.T. Marchese (Springer, 1995).

[3] *History of Cartography Project* <http://feature.geography.wisc.edu/histcart/>.

[4] *Map Displays for Information Retrieval*, by Xia Lin. Journal of the American Society for Information Science, 1997, Vol. 48, No. 1, pages 40-54.

Information mapping can be considered part of the larger research field known as *information visualization* [5].

There is, of course, a long history of information presentation using charts, graphics and map-like displays[6], but it is only really since the 1990s that interactive graphics have been possible on affordable desktop PCs, thus enabling dynamic visual representations of large volumes of information accessible to large numbers of people. Previously, the user of an information map typically had access only to a static hardcopy product where the data selection, processing and presentation were predetermined and fixed by the map author.

Perhaps the two most widely cited 'classics' of this type of static information map, from the era before interactive graphic computing, are the London Underground diagram (the 'Tube map') designed single-handedly by Harry Beck[7] in the 1930s and John Snow's map[8] of the cholera outbreak around the Broad Street pump from 1854. Beck's Tube map is an incredibly useful map that enables millions of people to successfully navigate London's complex subway network, but it also provides a very powerful framework for understanding the overall shape and geography of the city above ground. (This can sometimes be problematic as the map has gross and variable distortions of geographic scale.) Many of the spatial design metaphors developed by Beck are still used today on network and subway maps around the world. Snow's map illustrates the potential of mapping (in this cases on a geographic template) to reveal previously unknown patterns which provide genuine insight into the problem at hand, i.e. attributing the cause of cholera to infected water, as demonstrated by the cluster of outbreaks around a particular water pump.

---

[5] *Readings in Information Visualization: Using Vision to Think*, edited by Stuart K. Card, Jock D. Mackinlay and Ben Shneiderman (Morgan Kaufmann Publishers, 1999).

[6] See the exemplary work of Edward Tufte in reviewing and critiquing the field <http://www.edwardtufte.com/>.

[7] The fascinating story behind the Tube map is told in *Mr Beck's Underground Map*, by Ken Garland (Capital Transport Publishing, 1994).
The current Tube map can be consulted here <http://www.thetube.com/content/tubemap/>.

[8] See the John Snow site < http://www.ph.ucla.edu/epi/snow.html>.


Dot map of cholera deaths

The three key advantages of well-designed information maps are:

1. *'a sense of the whole'*: the ability to summarise and meaningfully convey a large amount of information in a limited space (usually a single PC screen).
2. *'revealing hidden connections'*: show the underlying semantic structures of a collection of documents through shared concepts and the similarity between them.
3. *'exploration'*: support users to easily and intuitively browse and forage through the information space.

In a metaphorical sense the map should enable one to get 'above' the information space, hovering a mile or two up in the air, so you see the whole area. The map can then provide a wide view of the lie of the information landscape, what is where, the location of clusters and hotspots, what is related to what. This kind of birds-eye view function has been memorably described by David D. Clark, Senior Research Scientist at MIT's Laboratory for Computer Science, as the missing 'up button' on the Web browser. Most information maps are trying to provide this 'up-button', with varying degrees of success. Ideally, this 'big-picture' all-in-one visual summary needs to fit on a single standard computer screen, without scrolling or flipping between pages.

The 'big-picture' overview is particularly important given the nature of much of information seeking is via unstructured and poorly formulated browsing and foraging.

We often have a vague notion of what we are looking for, and likely places to start looking and we can usually tell when we have found an answer. An approximate answer is all that is needed most times (and sometimes *'the'* answer does not exist). As Robert Spence, Professor of Information Engineering at Imperial College, London, succinctly notes,

> "… a user may be unable to say exactly what they are looking for in a collection of documents because they may not *know* exactly what they are looking for. They may want to discover *roughly* what is available in the collection and then, by exploration, gradually refine their inquiry." [9]

Information maps are potentially much better at supporting this type of browsing, than conventional key-word driven search engines. Browsing is a dynamic process of exploration and the information map design and interface must support this fluid interaction. Exploration often gives rise to unexpected perceptual inferences and insights and one should never dismiss the importance of serendipity in finding answers. Dynamic exploration is also very often a haphazard and iterative process – we wander off in different directions based on what we have seen, reformulating what we are actually looking for. Part of the power of the Web has been that it supports this type of exploration - with hyperlinking enabling the user to seamlessly jump to new sites and to follow new pathways at a single click. But the major problem with current Web representations is that we are stumbling along largely blind as we have no maps.

Information maps must also be able to show, in an intuitive and meaningful fashion, the structures of the data space in terms of direct relationships between documents (via citation or hyperlinks, for example) but also similarity in terms of shared themes and semantic connections. These structures and relationships are usually completely hidden in the presentation of conventional search engines. And yet this is often where we find insight and answers in the assimilation of some sense of how documents fit together to provide a mosaic of understanding. As Jacque Bertin reminds us,

> "Items of data do not supply the information necessary for decision-making. What must be seen are the relationships which emerge from consideration of the

[9] *Information Visualization*, by Robert Spence (Addison-Wesley, 2001), page 179.

entire set of data. In decision-making, the useful information is drawn from the overall relationships of the entire set." [10]

Maps and cognitive models

The role of the map is to provide a tangible external representation that aids people in forming an internal mental model or cognitive map of a space [11]. The map in this sense is simply a tool to aid understanding. The understanding actually happens in the brain. A good map can provide the vital 'Ah-ha' factor when one sees structure, patterns or find new connections that had not been apparent without the map. This is due in part to the fact that the human brain has excellent spatial processing capabilities (we navigate through a complex three-dimensional world without conscious effort) and powerful spatial memory (which enables us to remember where a multitude of things are, and how to find them again). As Professor Robert Spence notes,

"The relevance of spatial memory to visualization is clear. Since visualization is the creation and enhancement of an internal model, it is useful to know that spatiality as a basis appears to lead to a robust internal model, and therefore that phenomena having no inherent location could perhaps be assigned an artificial location – consistently used throughout – in order to enhance the likelihood of a robust internal model." [12]

The cognitive processes at work in the acquisition of this insight from the map are still not fully understood by scientist, being the subject of ongoing research in cognitive psychology and behavioural geography. However, insights clearly do arise from mapping information, as many of us can testify from personal experience.

[10] *Graphics and Graphic Information Processing*, by Jacque Bertin (trans. by Walter de Gruyter, 1981), page 64.

[11] *How maps work: representation, visualization, and design*, by Alan M. MacEachren (Guildford, 1995).

[12] *Information Visualization*, by Robert Spence (Addison-Wesley, 2001), page 60.

Map design

Information maps use a two dimensional display to represent data content, structures and relations via a range several spatial properties. The most important properties are:

- Area
- Position
- Proximity
- Scale

The *area* of graphic objects in the map is perhaps the most important representational tool as it is used to show the size or volume of different groups of data. The larger the area on the map, the more data it represents. The *position* of graphic objects can also have meaning - for example, they can be arranged into hierarchies, or ordered by time, size, importance/relevance or simply alphabetically. *Proximity* means that the closer together graphic objects are drawn on the map, the more alike they are. Distance on the map equates to some metric of similarity between documents. This gives rise to local neighbourhoods and regions in the map that are likely to contain documents with similar content and themes. The ability to create maps of a data-space at different *scales* is also vital as it allows different level of detail to be shown at appropriate times. Large-scale maps can provide a good overview by showing a large extent at low detail and then when required the user can use a smaller-scale map to reveal rich local details. Moving between scales can be by simply switching display or a smooth transition via some kind of a zooming function. The application of these spatial properties to summarise and represent information has been term *spatialization*[13]. Note, not all information maps make use of all these spatial properties.

Additional graphic properties such as colour, shape and labelling, for example, can encode further details of the data in the map. For example, colour is often used to show magnitude or change. Labels can be very useful in identification of specific items or regions of the map. And the shape of graphic objects in the display can

[13] *Spatial Metaphors for Browsing Large Data Archives*, Sara Fabrikant (Unpublished Phd dissertation, Department of Geography, University of Colorado-Boulder, 2000) <http://www.geog.ucsb.edu/~sara/html/research/diss/spatialization.html>.

convey useful subsidiary attributes of documents. For example, the particular aspect ratios of rectangular tile symbols can encode data in terms of the width and height values. So, for example, long, skinny rectangles would be different to short, fat squares.

## Reviewing Information Maps

*The map is a help provided to the imagination through the eyes.*
**Henri Abraham Chatelain, *Atlas historique* (1705)**

In the 1990s a range of different information maps has been developed. In this section of the white paper we review the significant examples:

Research prototypes:

- Treemaps (Human Computer Interaction Lab, University of Maryland)
- ET-Map (Artificial Intelligence Lab, University of Arizona)
- WEBSOM (Neural Networks Research Centre, Helsinki University of Technology)

Commercial products:

- ThemeScape (Cartia, Inc.)
- Map of the Market (SmartMoney.com)
- Visual Net (Antarct.ca Systems, Inc.)
- WebMap (WebMap Technologies, Inc.)

## ThemeScape


*Company*: Cartia, Inc.

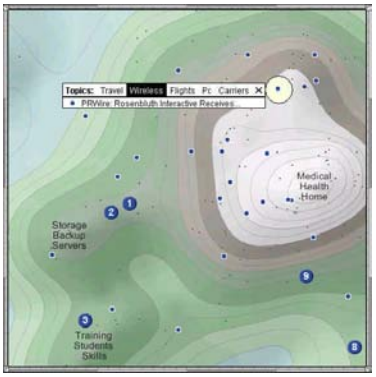
*Address*: Bellevue, Washington State, USA.
No current website.

*Status*: Initial research at PNNL between 1994-96. ThemeScape launched in 1998. Cartia taken over by Aurigin Systems in 2000. Further development of ThemeScape is unclear. No online demonstration available.

*Key developers*: James A. Wise, David Lantrip and Marc Pottier.

*Further reading*: *Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents*, by J.A. Wise, et al., Proceedings of InfoViz'95, IEEE, pp. 51-58.

*NewsMaps: Topographic Mapping of Information*, Mappa.Mundi Magazine, September 2000.



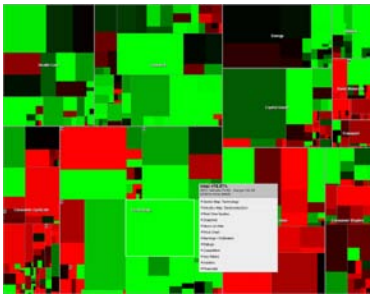

ThemeScape information analysis and mapping technology is one of the most sophisticated information maps developed. It produced 'thematic terrains' showing the principle topics or themes of a large collection of documents in a single overview map as well as providing easy interactive browsing of the map to locate specific documents.

The ThemeScape map uses a realistic looking landscape of mountains, hills and valleys created via shading and contours. The elevation of the hills and mountains in the terrain reveals the relative prevalence of different themes. The valleys are the natural transitions between one topic and another. The concept of spatial proximity is used in that the closer together two hills are on the map, the more similar their information content. The actual location of individual documents used to construct the map is indicated by small black dots.

Themescape maps are fully interactive, with the user able to pass the mouse cursor over an area of interest to display the top five topics within a small radius in a pop-up window. Clicking once on the terrain will cause a pop-up list of available documents in the area to be displayed, from which the full document can be opened in a new browser window. Users can zoom in on a region of the map to see greater local detail and also do a key word search for documents of interest or select documents from a topic list. The results of which are shown prominently by large blue dots on the map, numbered according to their relevancy ranking. It is also possible to stick small red marker flags into the terrain to identify documents of interest for future reference and to zoom in and pan around the map display to reveal more detail.

ThemeScape was originally developed by Cartia which in turn was a spin-off formed by information visualization researchers at Pacific Northwest National Laboratory (PNNL) in the mid 1990s. The research concepts developed were called SPIRE - Spatial Paradigm for Information Retrieval and Exploration.

## Map of the Market



*Company*: SmartMoney.com

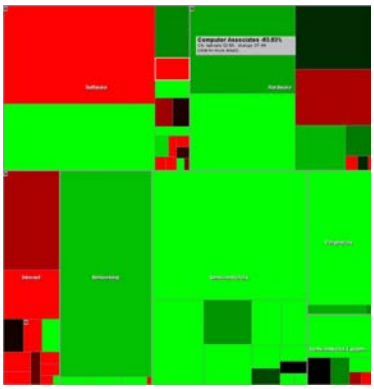
*Address*: 1755 Broadway, New York City, NY, USA..
www.smartmoney.com/maps/

*Key developer*:
Martin Wattenberg

*Status*: Launched at the end of 1998. Extended to include IPO and Mutual Fund maps as well stand alone Mapstation and Excel plug-in. Full product is online.

*Further reading*: *Visualizing The Stock Market*, by Martin Wattenberg. Proceedings of ACM CHI'99 Conference, 1999, pp. 188-189.

*Show me the money: The Map of the Market*, Mappa.Mundi Magazine, August 2001.

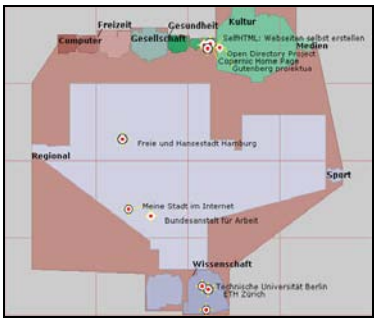


Map of the Market aims to answer a simple, but vital question for investor, 'how is the market doing today?'. It does this by mapping the performance of the stock market in a single visual snapshot showing the capitalization and changing stock price of over 500 publicly traded US companies. On one single map one can quickly gain a sense of the overall market conditions, yet still see many hundreds of individual data elements. This overall picture is very difficult to comprehend from more conventional listings of stock prices.

Each tile in the map represent one company, with size, colour and position encoding key attributes. Firstly, the tile size is scaled to the market capitalization of the company, so the bigger the tile, the more valuable the company in terms of stock market valuation. The colour of the tile represents the percentage movement in company's the stock price. The brighter the colour the bigger the change (either positive or negative) in stock price. One can interactively choose the time period over which the stock price change is calculated (from the last market close to six months or a year). Company tiles are grouped into familiar hierarchies based on classifications of industries (e.g., software, networking, semiconductors) which in turn form broader sectors (e.g., technology, energy, financial). Within these groups the spatial arrangement of individual tiles are arranged using a neighbourhood technique that places similar companies near to each other, where similarity is a metric based on historically similar stock price performance. It is fully interactive, allowing the user to access a great deal of information, statistics and news on the companies by clicking on their tiles. Users can also zoom into the map, to focus on a particular sector or industry of interest.

Map of the Market uses a visualisation technique called treemaps to generate the space-filling nested, regular tiles spatialisations. (see page *xx*). It is probably the most success of the current range of information maps in terms of popularity and usability.

## Visual Net




*Company*: Antarcti.ca Systems Inc.

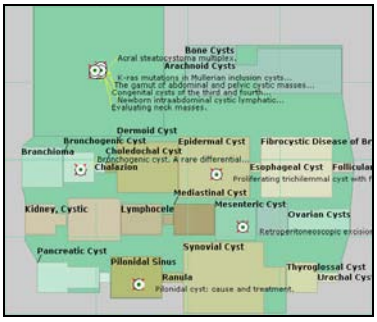
*Address*: 1198 Homer Street, Vancouver, BC, Canada. http://antarcti.ca

*Status*: Company formed in 1999 and Visual Net and Map.net launched in November 2000. Online demonstrations available mapping the ODP Web directory, the PubMed National Library of Medicine's database, and Canadian VC activity.

*Key developers*: Tim Bray, Dave Ashworth and Glen McMillan.

*Further reading*: *Visual Net: Technical White Paper*, version 1.0, March 2001, Anarcti.ca Systems.

*Measuring the Web*, by Tim Bray. Fifth International Conference World *Wide Web*, 6-10th May 1996, Paris.



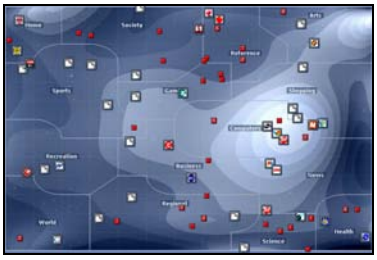

Visual Net: Antarcti.ca's self-stated aim is to *'transform networks into places'* using their Visual Net technology. This provide tools for giving documents a spatial x,y location on a map along with suitable graphic representation.

A key technology showcase for Visual Net is the Map.net site. This is a free, online use map over 2 million web sites from the Open Directory. It is a two-dimensional, multi-level map using a tiling of irregular shaped polygons onto the land surface of continent of Antarctica. They are grouped in a hierarchy of categories which are represented visually on the map display as nested polygons. The category polygon are colour coded and their size is in proportion to the number of websites they contain. The categorization is human derived rather than through automatic classification of websites. The tiling arrangement is not based on similarity of content, but is simply laid out in alphabetical order from top-left to bottom-right. The maps are fully interactive, allowing the user to move up and down the hierarchy of maps. Simply clicking on a tile of interest will usually provide a more detailed map of relevant sub-category.

The position and characteristics of individual websites within categories are also shown on the maps. These are represented by the distinctive circular targets and clicking on one will open the site in a new browser window. By default, the top twenty most visible websites in the map region are shown. Visibility is an important metric computed by Map.net and is an overall score to identify the 'best' websites in the vast milieu of the Web.

The Map.net demonstrator has further advanced navigational features to aid browsing of the visual directory. These include bookmarking, teleportation, real-time chat to other Map.net users in the same neighbourhoods. There is also a fully three-dimensional information landscape available which shows local area of the made as a cityscape with individual websites as different types of buildings.

## WebMap



*Company*: WebMap Technologies, Inc.

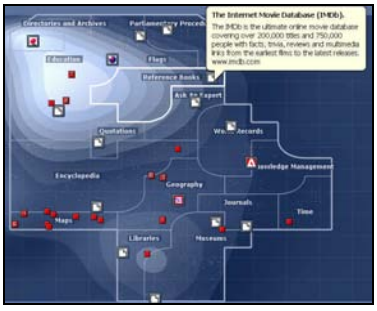
*Address*: 175 Portland Street, Boston, MA, USA. www.webmap.com

*Status*: Formed in November 1999 and initial map application launched in spring 2001. Development is ongoing focusing on travel, pharmaceutical and financial services. Free download of demonstration InternetMap.

*Key developers*: Michael Iron, Ohad Ranen and Roi Neustadt.

*Further reading*: WebMap Technologies, Inc. White Paper, April 2001.

*Visualizing the Web*, by Paul Heltzel. Technology Review, April 2001.



WebMap Technologies is developing an interactive, multi-level information mapping system. Their self-stated aim of their Information Mapping Technology is "*to make interaction with the Internet, intranets, extranets and wireless environments more useful and efficient through a visual rather textual format.*" The map comprises three layered components – categories, ratings and document icons. Their flagship mapping technology demonstrator, InternetMap, launched in the spring of 2001, maps more than 2.5 million websites classified by the Open Directory Project. The top-level map comprises the standard high-level categories. Through an interactive zoom you can explore more than 350,000 categories, as well undertake key word searches with the location of matching websites shown by small red icons. Your browser favorites are shown as icons on the map.

The system uses a well implemented client-side browser plug-in to deliver an interactive map of a collection of documents. Documents are grouped together into categories based on similarity of topic. These are represented on the map as irregular shaped regions. A terrain style map underlies the category regions and acts as a kind of 'topography of popularity'. The higher the terrain (shown by the lighter shading contours) the more highly rated the documents are in that region. Various icon symbols are used to mark the location of particular documents, such as book marked documents or results from a conventional key word search. The map is multi-level an clicking on a region will smoothly zoom the view an set the display to this region and change scale to reveal more detailed categories.

The server side of the system can analyze documents, grouping them in categories based on content and positioning them using similarity proximity. The actual analytical techniques and clustering algorithms employed by WebMap are proprietary. The system also features a degree of map customization including the ability for the user to change categories of a document by dragging and dropping the icon and also a personalization engine that "*floats the links most often accessed by the user to the top levels of the map, while sinking the least often used links down into the lower levels of the map.*"

## Treemaps

*Research center*: Human Computer Interaction Lab (HCIL)

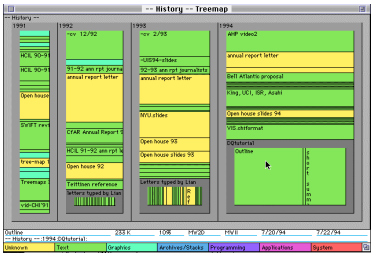
*Address*: University of Maryland, College Park, MD, USA. www.cs.umd.edu/hcil/

*Status*: Initial technique developed by Shneiderman in 1990. Developed through 90s and latest release of java-based Treemap 3.0 is available for commercial licensing. Research into treemaps is ongoing at HCIL. Free online demonstration is available.

*Key developers*: Ben Shneiderman, Brian Johnson and David Turo.

*Further reading*: *Treemaps for space-constrained visualization of Hierarchies*, by Ben Shneiderman, November 2000.

*Treemaps: a space-filling approach to the visualization of hierarchical information structures*, by Brian Johnson & Ben Shneiderman, HCIL Paper HCIL-91-06, 1991.

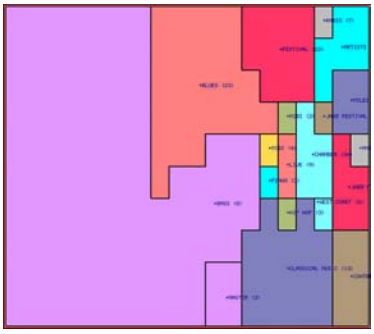


Treemaps are defined as a space-constrained visualization of hierarchies and they are one of the most well developed information mapping techniques having begun life back in 1990 and been refined and extended throughout the decade by researchers at HCIL. Treemaps show hierarchical data as a set of nested rectangular tiles that fill the whole display. Each data element is represented on the treemap as a single tile and key attributes of the data are represented by the size and colour of the tiles. Treemaps enable a large volume of information to be usefully displayed on a single map screen. In certain task treemaps have been shown to facilitate users in comparing data elements, gaining a sense of the whole data space and identify patterns and anomalies. The inherent hierarchical structure of the input dataset is represented in the treemap display by the nesting which is visually enhanced with border around each level in the hierarchy.

Shneiderman originally developed the treemap concept and algorithm in 1990 as a visual aid to the file management tasks on congested hard disk of shared PC. The hierarchical data was the directory structure of files. Each tile in the treemap display was an individual file on the disk, with the area of the tile proportional to the file size and the colour indicating different file types. Treemap therefore mapped the content of the disk and revealed which files (and directories) were taking up precious disk space. Shneiderman comments, "*I think treemaps are a convenient representation that has unmatched utility for certain tasks. The capacity to see tens of thousands of nodes in a fixed space and find large areas of duplicate directories is very powerful …. It does take some learning for novices to grasp the tree structure layout in treemaps, but the benefits are great.*"

Treemap algorithms have been refined to improve layouts and produce more legible displays. While a number of new implementations have been released throughout the 1990s. Treemaps have been applied to a diverse set of information domains, including financial data, basketball statistics, decision-making processes and network management.

## ET-MAP



*Research center*: Artificial Intelligence Lab, University of Arizona.

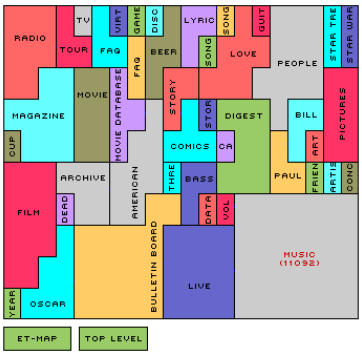
*Address*: Management Information Systems Department, University of Arizona Tucson, Arizona, USA.
http://ai.bpa.arizona.edu

*Status*: Research prototype developed in the mid 1990s. Development has ceased as efforts of the center are applied in other project in intelligent information classification and retrieval.

*Key developer*: Hsinchun Chen

*Further reading:* *Internet Categorization and Search: A Self-Organizing Approach*, by Hsinchun Chen, Chris Schuffels, and Rich Orwig, 1996.

*A Map of Yahoo!*, Mappa.Mundi Magazine, February 2000.



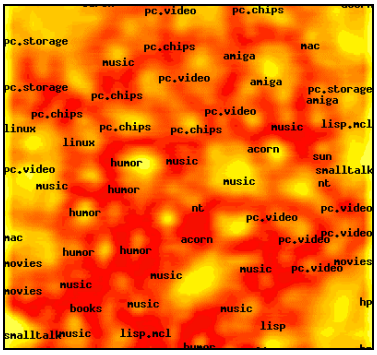

ET-Map is a prototype Internet homepage categorization system developed by the AI Lab. It is attempt at a meaningful multi-layered information map based on part of the Yahoo! directory.

They analyzed and automatically classified the content of 110,000+ entertainment-related homepages producing a multi-layered map with different subject regions. Larger subject regions occupy a bigger space on the map and conceptually related subjects are often grouped in close proximity. Regions which contain 100 URLs or more produce another map; regions which contain fewer than 100 URLs produce a ranked list of URL summaries.

ET-Map starts with a the top-level map showing forty odd broad entertainment 'subject regions' represented by regularly shaped tiles. Each tile is a visual summary of a group of Web pages with similar content. These tiles are shaded different colours to differentiate them, while labels identify the subject of the tile and the number in brackets telling you how many individual Web page links it contains. ET-Map can be browsed interactively, explored and queried, using the familiar point-and-click navigation style of the Web to find information of interest.

ET-Map was created using a sophisticated AI technique called Kohonen self-organizing map (SOM). See next project.

## WEBSOM



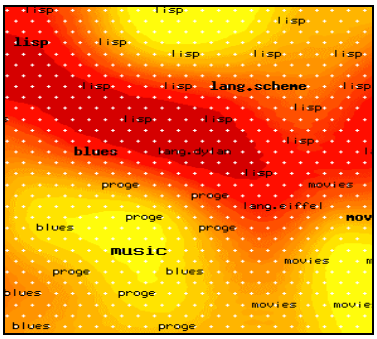

*Research Center*: Neural Networks Research Centre, Helsinki University of Technology.

*Address*: Helsinki University of Technology, Espoo, Finland http://websom.hut.fi/websom/

*Status*: WEBSOM research began in the mid 1990s and is ongoing. Online demonstrations are available.

*Key developers*: Teuvo Kohonen, Krista Lagus, Timo Honkela, Samuel Kaski.

*Further reading*: *Self Organization of a Massive Document Collection*, by Teuvo Kohonen et al. IEEE Transactions on Neural Networks, vol. 11, no 3, pp. 574-585.




WEBSOM uses a well developed neural network technique called self-organizing maps (SOM) to create interactive, browseable information maps for exploration of large collections of text documents. SOM's sophisticated algorithms automatically organizes the documents onto a two-dimensional map grid so similar documents lie near each other on the map. This spatial proximity by content similarity helps in finding related documents once any interesting document is found. It is also important to realize that SOM operates automatically, without direct human categorization / classification of documents so it can scale to cope with tens or hundreds of thousands of documents. SOM techniques have been developed by Teuvo Kohonen over the past twenty years or so, while the interactive WEBSOM maps have been developed by Kohonen and colleagues at the Neural Networks Research Centre, Helsinki University of Technology.

WEBSOM maps have multiple scale levels and fully interactive to facility user 'point-and-click' exploration of the document space. The maps use a continuous representation of topographic shading that provides an information terrain. The shading denotes the density or the clustering tendency of the documents across the map. The light coloured areas are the densest clusters (the hill and mountains in the terrain), while the darker shaded areas are empty space (the valleys). Different areas of the terrain are also labeled with the key topic of documents in that region. Clicking on any area on the map brings loads a smaller-scale map which reveal more detailed classification. Each white dot on them map marks a document location. Clicking on the dot will provide full details and access to the actual document content.

Example WEBSOM maps include Internet Usenet newsgroups, news bulletins and a database of 7 million patents abstracts.