Ruddle, R. A. (2010). INSPIRE: A new method of mapping information spaces. *Proceedings of the 14th International Conference on Information Visualisation (IV'10)*, 273-279. DOI= http://dx.doi.org/10.1109/IV.2010.48.

INSPIRE: A new Method of Mapping Information Spaces

Roy A. Ruddle
School of Computing, University of Leeds, UK
{r.a.ruddle@leeds.ac.uk}

Abstract

Information spaces such the WWW are the most challenging type of space that many people navigate during everyday life. Unlike the real world, there are no effective maps of information spaces, so people are forced to rely on search engines which are only suited to some types of retrieval task. This paper describes a new method for creating maps of information spaces, called INSPIRE. The INSPIRE engine is a tree drawing algorithm that uses a city metaphor, comprised of streets and buildings, and generates maps entirely automatically from webcrawl data. A technical evaluation was carried out using data from 112 universities, which had up to 485,775 pages on their websites. Although they take longer to compute than radial layouts (e.g., the Bubble Tree), INSPIRE maps are much more compact. INSPIRE maps also have desirable aesthetic properties of being orthogonal, preserving symmetry between identical subtrees and being planar.

1 Introduction

Twenty years on from its inception, the WWW is a central part of many people's everyday life. Search engines such as Google are remarkably effective for helping us find pieces of information, but not in all situations. For example, there are often occasions when people find it difficult to express in words what they are looking for, so they are unable to formulate a suitable query [18], and knowledge workers waste 15% of their time failing to find information that they know exists [17].

A key limitation of search engines is that they treat websites like a black box, "parachuting" a person to a particular place when they click on an item in the search results, from where browsing rarely involves more than three or four more clicks [13]. As a result, people often find it difficult to recall where a given piece of information is (the problem of keeping found things found; [7]), and have little knowledge of where different items of information are with respect to each other.

In the real world, by contrast, a map is one of the most useful things you can give a visitor when they arrive in a new city. For example, Athens (population 750,000) looks like an impenetrable mass of buildings when the city is viewed from a distance (Figure 1). However, with the help of a map, a visitor can quickly learn to find their way around and remember where places are.

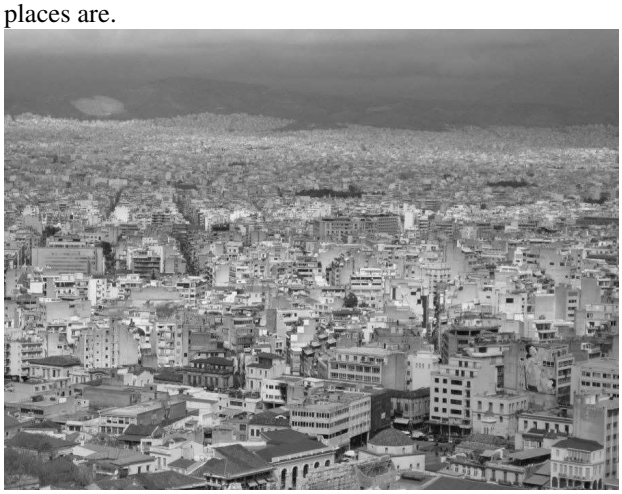

Figure 1. A view of Athens from the Acropolis, situated on a hill, above the centre of the city.

If we can create useful maps of cities, then surely we can create maps that help people to navigate large websites. People could potentially use such maps to visualize the distribution of search results, browse without unintentionally travelling in circles [14], and easily return to information [7].

Creating useful maps of websites is a major challenge for visualization, as can be seen by the fact that most current sitemaps are in the form of a textual index and none of the many graphical techniques that have been investigated are in widespread use [3] [8]. This paper describes a novel method for creating website maps. Our INSPIRE maps are: (1) generated automatically from web crawls, (2) are more compact than radial node-link diagrams (the most common way in which websites are visualized), and (3) still communicate the underlying structure of a site, unlike the Treemap space-filling approach. Our maps are fundamentally different to previous visualization methods that have used geographic metaphors (e.g., BEAD [2] and Euler diagrams [15].

The acronym INSPIRE was chosen because the method is a new way of mapping INformation SPaces for Information REtrieval. This paper describes the algorithm used to generate the maps, and a technical evaluation that used data from webcrawls of 112 university websites. User evaluations of information retrieval are left to future work, as is using INSPIRE to draw other types of tree data.

2 Related work

Previous research has focused on how the structure of websites should be simplified and methods of visualization. These are discussed in the following sections.

2.1 Structure

Clearly, any graph that shows all of the connections in a website will be highly non-planar. As a result, all methods for visualizing websites simplify the structure involved. Typically, a website's structure is divided into a primary part, which is used to calculate the visualization layout, and supplementary links that may be superimposed to show other relationships.

Most website visualizations use a tree for the primary structure. Sometimes the tree is derived from the hyperlinks that a person followed [19] [12], but usually it is derived from the website's link structure using algorithms based on shortest paths or the site's directory structure.

2.2 Visualization methods

The method used to visualize websites may be classified using the following factors: (1) layout style (orthogonal vs. radial), (2) the projection (linear vs. non-linear), (3) how connections are indicated (lines vs. space-filling), and (4) the dimensions that are used (2D vs. 2.5D vs. 3D). Examples of the methods are shown in Table 1.

If only a small number of pages need to be shown then a traditional orthogonal layout may be used (e.g., PowerMapper's Electrum view). However, radial layouts are the most common method used to visualize websites, because of their ability to maintain a balanced aspect ratio even when some nodes have a large number of children.

A given style of layout may be rendered using either a linear or a non-linear (hyperbolic/fisheye) projection. Taking radial layouts as an example, linear projections (e.g., Astra Site Manager) are faster to render because fewer graphical transformations are required, but non-linear projections (e.g., Inxight Site Lens) were developed so that both detail and an overview may be seen in a continuous display. The disadvantage of non-linear projections is that objects move relative to each other when a visualization is navigated, which severely impedes participants' ability to remember the location of nodes [16]. Linear projections, on the other hand, can be supplemented with a thumbnail to provide an overview of the visualization.

All of the above layouts are line-based, meaning that the connections between pages are shown explicitly as lines. An alternative is a space-filling layout (e.g., Visual Sitemap or Cartia Themescape), which implicitly shows connections either from the way pages abut or are contained within each other. The result is more compact than line-based layouts, but the penalty is that space-filling layouts are poor at communicating the structure of a website.

3D layouts sometimes have aesthetic appeal, but there is little evidence that they are more useful than 2D layouts. This is because navigational movement is much more complex in a 3D space, and it is easier for users to become disoriented.

On the other hand, 2.5D layouts (e.g., the MAPA Z-diagram) use 2D methods for the layout and the third dimension to provide supplementary information and distinguishing features. In this way, 2.5D layouts mimic the cities we live in, where streets are laid out on a surface and tall buildings are very prominent. Manually-created Z-diagrams that contained a few hundred pages have proved useful for the design of major websites [8].

3 Algorithm

The inspiration for our maps comes from the map of London that Richard Horwood created in 1799. Horwood created his map so that even the smallest house could be labelled with its number (in approximately a 10 point font), and the result was the last map to show every single building in London on a single (wall-sized; 4 x 2.5m) display. To be fully appreciated, the map has to be seen for real (e.g., in the British Library), but online images give a useful impression (http://oldlondonmaps.com/horwoodpages/horwoodmain.html). London's population at the time was 1 million, so if that beautiful map could be drawn 200 years ago then surely modern technology should allow us to create useful maps of 1 million page websites?

Style	Dimensions	Orthogonal	Radial	
		Linear	Linear	Non-linear
Lines	2D	Electrum PowerMapper (Electrum view)	Astra Site Manager	Inxight Site Lens
	2.5D	Electrum PowerMapper (Isometric view)	Webspace	-
	3D	-	VR-VIBE	H3
Space-filling	2D	Visual Sitemap & INSPIRE	Cartia Themescape	-
	2.5D	MAPA	-	-

Table 1. A classification of website visualization methods. All of the examples except INSPIRE are taken from [8] [3]. There are no non-linear orthogonal layouts, or 3D space-filling layouts.

3.1 High level concept

Our INSPIRE map adopts the visual metaphor of a city. The map presents a website's primary (tree) structure as streets and buildings. The links between the nodes are the street junctions and where each building touches a street or is connected to another building (in effect, a building's door).

When a map is created to show the full detail of a website (i.e., showing every page separately), every branch node is portrayed as a street and every leaf node is a building. When a map shows the site at a reduced level of detail, pages that are linked and indistinguishable at the map's level of detail (e.g., in the same domain, if just the domain level of detail is being shown) will be consolidated together and portrayed as buildings.

Cities are usually modelled by placing a node at each junction and making each street a link. INSPIRE maps are the opposite way around (street junctions represent the links between webpages), an approach that is also used by the architectural theory of Space Syntax, which is used to predict where people will travel in urban landscapes [5].

A consequence of our city metaphor is that an INSPIRE map only implicitly shows links, instead of explicitly drawing links as lines, as is done with most website visualization. In other words, INSPIRE uses a space-filling approach, but the space between streets and buildings means that a websites structure is shown more clearly than with other space-filling methods such as a Tree Map.

A city metaphor has a number of important advantages. First, it should be easy to comprehend for anyone who has used a street map to navigate in everyday life, whereas the radial layouts that have been used for most previous website mapping are unfamiliar to the general public. Second, the city metaphor means that principles for the design of intelligible cities [10] could be directly applied. For example, districts could be shaded differently to identify the major regions of a website, paths could show the most common navigation sequences, and landmarks could act as anchors for people's mental models to aid revisitation. Third, the city metaphor allows us to draw on hundreds of years of experience that people have gained in cartographic design, for example, arranging labels, the use of color, and level of detail generation for semantic zooming [11].

The following sections explain how INSPIRE maps are generated. This covers extraction of a website's structure, rules that govern a map's layout, and making use of the third dimension.

3.2 Layout rules

The INSPIRE algorithm generates a layout from the bottom (leaf nodes) up, which guarantees symmetry between identical sub-trees. Other factors that assist comprehension are that the map generates an orthogonal layout, which is constructed from a small number of basic shapes. The layouts are not optimal in terms of the space occupied (that would be an NP-complete problem), but the layouts are compact and sufficiently efficient in computational terms to allow the algorithm to scale to sites that contain hundreds of thousands of pages (see Technical Evaluation).

By default, each page may either be a square of nominal size (1 x 1 unit) or a rectangle sized to accommodate the page's label (Figure 2). With the former the label can be presented using a mouseover, whereas the latter writes the label within the space that is occupied by the page in a manner that's similar to link labels in the GreenArrow system [20]. Dynamic, scrolling labels could also be implemented, as in that system.

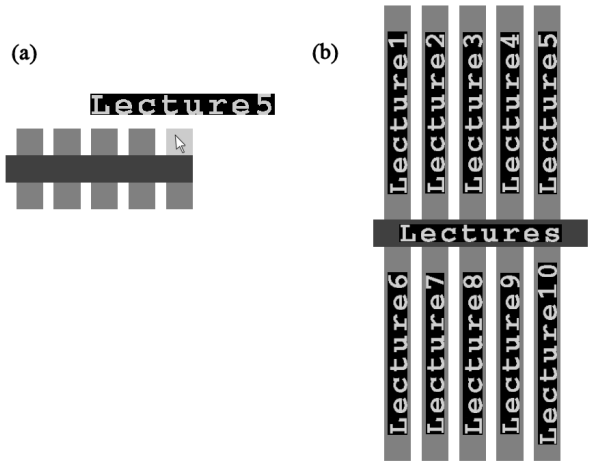

Figure 2. An 11 page extract from a website, showing: (a) Pages drawn at the default size with one label displayed via a mouseover, and (b) Pages sized so that their labels may be drawn inside each page. In this and the other figures, each level of the site is shown in a different shade of grey.

When a set of pages is added to their parent, it is stretched from its original length to whatever length is necessary to accommodate all its descendents. A useful byproduct of this is that the higher up the graph hierarchy you go, the longer and more prominent a page becomes.

To guarantee that pages will not overlap once the whole map has been generated, children are added to their parent so all descendents lie in the $X \geq 0$ region of the parent's coordinate space. As well as the above, maps are generated according to the following layout rules.

3.2.1 Number of children added to each side

If children were divided equally between the two sides of a parent then more space than necessary would often be used. Therefore, INSPIRE uses descendents' bounding boxes to calculate how many of the children should be on each side, so that the space they occupied by is balanced (Figure 3).

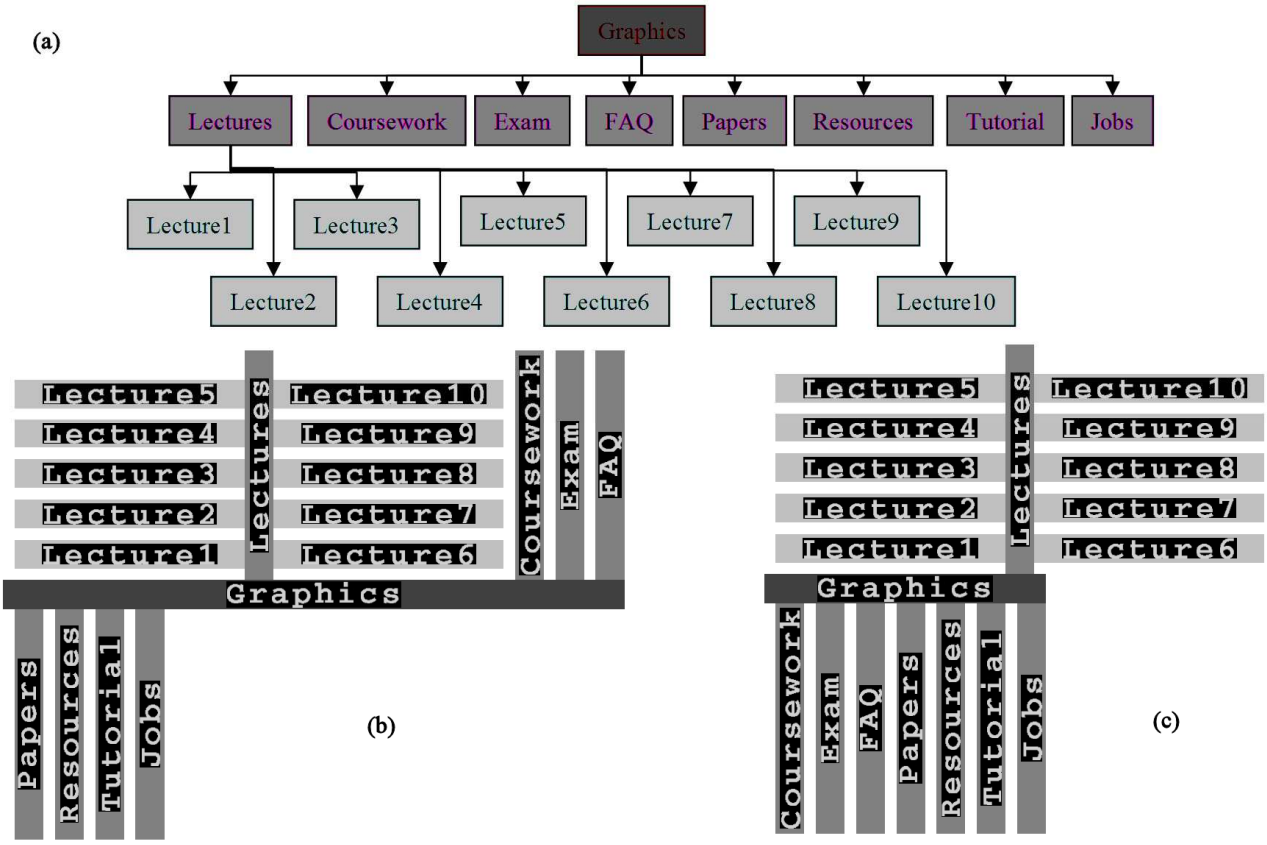

Figure 3. A 19 page extract from a website, showing: (a) A traditional orthogonal tree diagram, (b) An INSPIRE map that places an equal number of children on each side of a parent, and (c) An INSPIRE map that uses bounding box information to balance the space occupied by the children. The bounding box area of (c) is 19% smaller than (b).

3.2.2 Shape grammar

When a series of pages each have a small number of children then spatially inefficient layouts can be generated. To overcome this, the shape grammar concept is used. Before a set of pages is added to their parent, the area that will be occupied by the parent and all its descendents is calculated for *perpendicular* and *parallel* arrangements (Figure 4), and the one with the minimum area is chosen.

If the resulting arrangement has a large aspect ratio then a *concertina* shape is applied to the parent (Figure 5). Even though the concertina occupies more space than a non-concertina shape, the lower aspect ratio reduces the total area required for a whole map.

4 Technical evaluation

The evaluation was divided into three parts. The first investigated the effect of different layout rules on the maps that were generated. The second compared INSPIRE maps with a well known radial algorithm called the Bubble Tree [4], which is implemented in the Tulip graph software (downloadable from sourceforge). Tulip [1] is well known for its ability to handle very large graphs. The third investigated INSPIRE maps showing different levels of detail of websites.

4.1 Method

The INSPIRE algorithm was implemented in C++ application, using the SG, UL and FNT libraries from PLIB (http://plib.sourceforge.net/). The evaluation was run on a 64-bit Linux PC, running Fedora Core 9, with a dual core 2.13 GHz Intel processor and 2Gb RAM.

The data used in the evaluation were crawls of 112 British universities' websites, which were performed in 2006 and are publicly available (http://cybermetrics.wlv.ac.uk/database/). The websites had from 699 (University of Chichester) to 485,775 pages (University of Cambridge), averaging 65,385 pages.

Each of the webcrawls was processed to create a tree graph, by calculating the weighted shortest path (to preserve a website's underlying directory structure, link weights increased with the distance between pages in the directory structure). For each website, output files were created in Tulip's format and a format we had created for INSPIRE.

The INSPIRE application loaded a tree, calculated its map layout, and output the resultant graph and a logfile. Data stored in the logfile included the time taken to compute the layout, and the map's width and height. For the Bubble Tree layouts, an application was created

using version 3.0.2 of Tulip. The application loaded a tree, calculated the Bubble Tree layout, and output the resultant graph and a logfile that contained processing time and layout data.

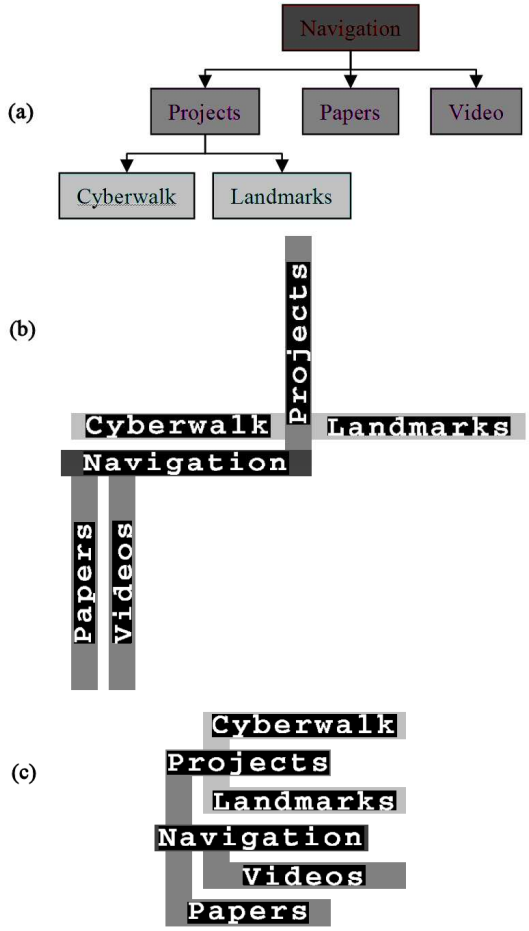

Figure 4. A 6 page extract from a website, showing: (a) A traditional tree diagram, (b) An INSPIRE map that arranges all pages perpendicularly, and (c) An INSPIRE map that chooses a parallel arrangement and occupies 75% less space than (b). The perpendicular and parallel arrangements are applied for all children on a given side of a given page, so most maps will incorporate both arrangements.

4.2 Results

4.2.1 Effect of INSPIRE's layout rules.

To evaluate the effect of different layout rules, maps of each website were generated using three combinations of rules: (a) The basic INSPIRE algorithm (an equal number of children allocated to each side of a page (Figure 3b), and only the perpendicular shape (Figure 4b), (b) Balancing the space occupied by children using the bounding box calculation (Figure 3c), and (c) Using the space balancing rule and shape grammar (Figures 4 & 5). For each combination of rules, one map was generated using a nominal (1 x 1 unit) page size (Figure 2a), and another was generated using a page size of 20 x 1 units, which was chosen to simulate label-sized pages (Figure 2b).

The maps were compared by calculating the area occupied by each map (Table 2) and its aspect ratio. For maps generated using a nominal page size, removing shape grammar caused the maps to double in size, and removing shape grammar and space-balancing caused the maps to treble in size. Label-sized maps were, of course, larger, with shape grammar and space-balancing providing even greater benefit. However, the type of label had little effect on the maps' aspect ratios. With space balancing and shape grammar the average aspect ratio was 1.8, compared with 2.0 for the other combinations of rules.

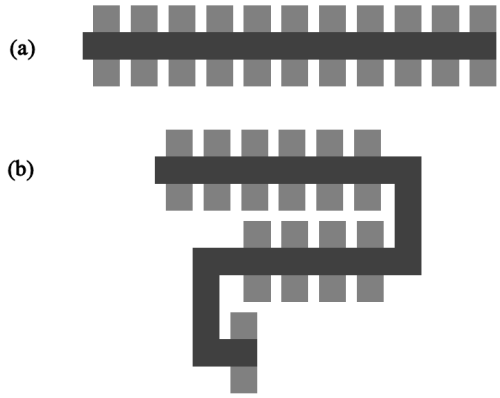

Figure 5. A page with 22 children drawn: (a) In the default shape, and (b) With the parent concertinaed. Although the bounding area of (b) is double that of (a), the aspect ratio is much smaller (1.0 vs. 5.1).

Layout	Rule	Nominal page size	Label-size page
INSPIRE	Basic	3.2 (3.9)	82.7 (35.1)
	Space-balanced children	2.3 (3.7)	49.6 (17.3)
	Space-balanced & shape grammar	1.0 (0.0)	7.6 (1.7)
Bubble Tree	-	23.9 (16.5)	3410.6 (2701.2)

Table 2. Relative area required for the INSPIRE and Bubble Tree layouts. The data are the mean (standard deviation) for the 112 university websites. For each website, the area of the full (space-balanced & shape grammar) nominal page size INSPIRE layout rules was set to one.

4.2.2 Bubble tree comparison

Radial layouts are the most common method used to draw website layouts [3; 8]. The Bubble Tree is a recent radial algorithm [4] that our tests show generates much more compact layouts than the Radial Tree [6].

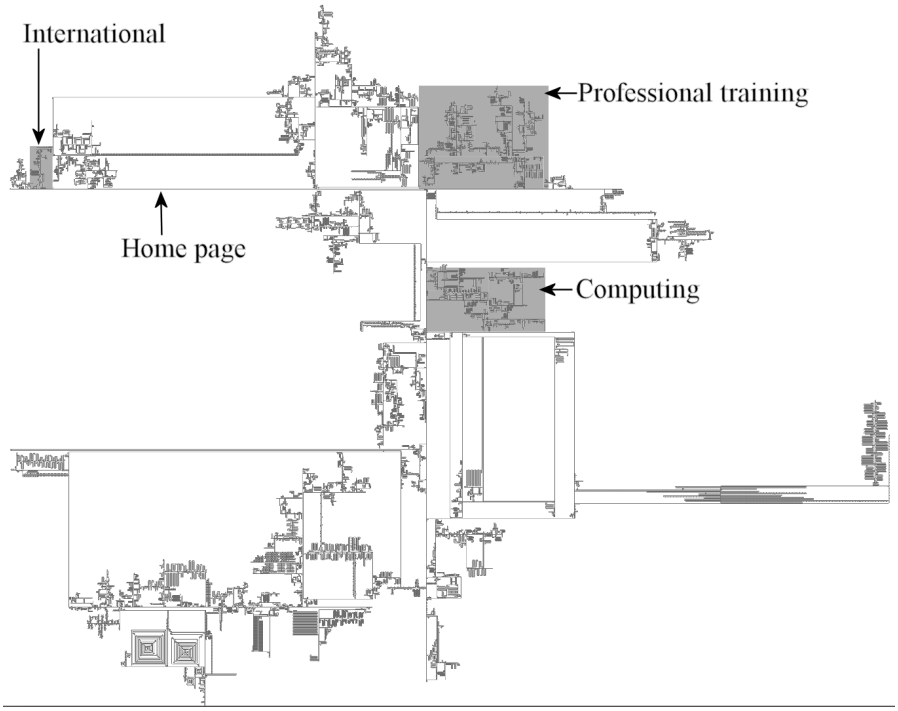

Figure 6. INSPIRE map of the University of Surrey's website. Three regions of the site are highlighted. Most of the rest are undergraduate and postgraduate pages (below & above Home page, respectively).

With a nominal page size, Bubble Tree layouts were from 4 to 129 times larger than INSPIRE maps (mean (*M*) = 23.9). With label-size pages the difference between the Bubble Tree and INSPIRE was even larger (*M* = 446.4), so much so that the label-size (20 x 1) INSPIRE maps were 3 times smaller than Bubble Tree layouts that only had a nominal (1 x 1) page size (Table 2).

The aspect ratio data showed that Bubble Tree layouts tended to be squarer than the INSPIRE maps (*M* = 1.3 vs. 1.8), and the Bubble Trees were very fast to compute (the algorithm is linear; maximum time = 4.2 s). Time data showed that, in terms of the number of pages in a website, compute time for INSPIRE was $n^{1.8}$. Actual compute times were < 1 second for all the websites with less than 9000 pages, and < 60 seconds for all the websites with less than 80,000 pages. For a small number of websites the compute time was lengthy, but none took longer than 536 seconds except the University of Cambridge's website (1554 seconds; 485,775 pages).

To illustrate INSPIRE, a map of the University of Surrey's website has been chosen, because it was the one that was the closest to "average"' in terms of number of pages (65,385) and the area of the layouts that were generated. The INSPIRE map of all the website's pages (1688 x 1345 units; Figure 6) occupied an area that was 26 times smaller than the Bubble Tree layout (6876 x 8464 units).

4.2.3 Levels of detail

All of the above results are for maps that showed every individual page on a website, but INSPIRE is also well suited to automatically generating maps at lower levels of detail. This section provides data about Level 1 and Level 2 maps, where Level 1 only considered the domain of a page (e.g., http://abc.org) and Level 2 considered domain and the first subdirectory (e.g., http://abc.org/def/). Each subtree of the full structure that comprised a set of pages that were all in the same domain (Level 1) or subdirectory (Level 2) was reduced to a single item. These items were rendered as buildings, meaning that the map showed a set of buildings and streets that were connected.

A building's area was made equal to the number of individual pages that the building represented, which meant that the sum of the minimum possible area of all the streets and buildings was the same in every map (number of pages x nominal page size), irrespective of its level of detail. Despite this, Level 1 & 2 maps occupied 11.5 and 5.4 times less area, respectively, than maps that showed every individual page.

5 Conclusions

This paper describes INSPIRE, a novel method for creating maps of large information spaces. The method generates a map from the primary (tree) structure of a space and was designed for use with websites, but could also generate maps of other types of information space such as a filesystem or on-line help in a computer application.

INSPIRE maps use a city metaphor, and are generated entirely automatically from webcrawls. The maps represent an information space as streets and buildings, which are generated by a small number of rules that each play an important part in generating a compact layout. If every individual page in a space is shown then leaves and branches in the tree are represented as buildings and streets, respectively.

However, overview maps consolidate pages together into other buildings, if those pages are indistinguishable at the map's level of detail (e.g., pages that are in the same domain), while still retaining the tree's structure.

Compared with other methods for portraying information spaces, INSPIRE has several advantages. First, INSPIRE generates layouts that, on average, were 23.9 times more compact than the Bubble Tree, which is a recent example of the commonly used radial layout. Second, the area occupied by an INSPIRE map only increases moderately if long page (node) labels are displayed, unlike the Bubble Tree. Third, unlike traditional space-filling approaches like the Tree Map, INSPIRE's street network explicitly shows the structure of an information space. Fourth, the INSPIRE algorithm guarantees that the maps are orthogonal, preserves symmetry between identical subtrees, and are planar (nothing overlaps), all of which are desirable aesthetic properties of graphs. Fifth, the city metaphor should help to make INSPIRE maps easy to comprehend for anyone who has used a street map to navigate in the real world.

In terms of application scenarios, search results for a given website could be superimposed on an INSPIRE map, for example, to improve the saliency of clusters that weren't ranked in the top 10. A similar approach could be applied to general Web searches, superimposing the results on a map that was generated from domains or topics in the full set of results. If a user's browsing path was superimposed on a map then they would be more likely to find information efficiently than circuitously [14], and spatial cues provided by the map should aid revisitation [7]. In addition, information providers could use INSPIRE maps to design and analyse the usage of websites.

Finally, the following future work is planned. First, enhancements to the layout rules will incorporate established principles for the design of navigable environments [10], and apply space constraints [9] so that INSPIRE maps evolve from one level of detail to the next and are resistant to local changes in a website's structure. Second, 2.5D versions of the maps will be created using an isometric projection, meaning that buildings stand out more from each other, can be textured with thumbnail images and become salient landmarks. Third, proof of concept applications for the above scenarios will be developed and evaluated.

6 Acknowledgements

This research was partly carried out at the Max Planck Institute for Biological Cybernetics (Tübingen), supported by an Alexander von Humboldt Foundation Fellowship for Experienced Researchers to RAR.

7 References

[1] D. Auber. Tulip. In *Lecture Notes in Computer Science* (Vol. 2265), Springer, 335-337. 2001.

[2] M. Chalmers. Using a landscape metaphor to represent a corpus of documents. In *Lecture Notes in Computer Science* (Vol. 716), Springer, 377-390. 1993.

[3] M. Dodge and R. Kitchin. *Atlas of cyberspace*. Addison-Wesley. 2001.

[4] S. Grivet and D. Auber and J. P. Domenger and G. Melancon. Bubble Tree drawing algorithm. In *Computer Vision and Graphics* (Vol. 32), 633-641. 2006.

[5] B. Hillier and A. Penn and J. Hanson and T. Grajewski and J. Xu. Natural movement: or configuration and attraction in urban pedestrian movement. *Environment and Planning B: Planning and Design, 20*, 29-66. 1993.

[6] T. J. Jankun-Kelly and K. Ma. MoireGraphs: Radial focus+context visualization and interaction for graphs with visual nodes. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'03)*, IEEE, 59-66. 2003.

[7] W. Jones and H. Bruce and S. Dumais. Keeping found things found on the web. In *Proceedings of the 10th ACM Conference on Conference on Information and Knowledge Management*, ACM, 119-126. 2001.

[8] P. Kahn and K. Lenk. *Mapping web sites*. Rotovision. 2001.

[9] G. Liotta and H. Meijer. Voronoi drawings of trees. In *Lecture Notes in Computer Science* (Vol. 1731), Springer, 369-378. 1999.

[10] K. Lynch. *The image of the city*. MIT. 1960.

[11] A. M. MacEachren. *How maps work: Representation, visualization, and design*. Guilford. 1995.

[12] S. Mukherjea and J. Foley and S. Hudson. Visualizing complex hypermedia networks through multiple hierarchical views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 331-337. 1995.

[13] P. Pirolli and J. Pitkow. Distribution of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web, 2*, 29-45. 1999.

[14] R. A. Ruddle. How do people find information on a familiar website? In *Proceedings of the 23rd BCS Conference on Human-Computer Interaction (HCI'09)*, 262-268. 2009.

[15] P. Simonetto and D. Auber and D. Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum, 28*, 967-974. 2009.

[16] A. Skopik and C. Gutwin. Improving revisitation in fisheye views with visit wear. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 771-780. 2005.

[17] J. Teevan and E. Adar and R. Jones and M. A. S. Potts. Information re-retrieval: Repeat queries in Yahoo's logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 151-158. 2007.

[18] J. Teevan and C. Alvarado and M. Ackerman and D. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 415-422. 2004.

[19] A. Wexelblat and P. Maes. Footprints: History-rich tools for information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 270-277. 1999.

[20] P. K. Wong and P. Mackey and K. Perrine and J. Eagan and H. Foote and J. Thomas. Dynamic visualization of graphs with extended labels. In *Proceedings of the IEEE Symposium on Information Visualization*, IEEE, 73-80. 2005.