

# Supporting document use through interactive visualization of metadata

Mischa Weiss-Lijn, Janet T. McDonnell

Department of Computer Science,  
University of College London,  
London WC1E 6BT, UK

Email - {m.weiss-lijn, j.mcdonnell}@cs.ucl.ac.uk

Leslie James

Strategy Research Team,  
J Sainsbury plc, London

*Digital Libraries (DLs) currently focus on delivering documents. Since information needs are often satisfied at the sub-document level, digital libraries should explore ways to support document use as well as retrieval. This paper describes the design and initial evaluation of a technology being developed for document use. It uses interactive visualisation of paragraph level metadata to allow rapid goal-directed search and navigation within documents. An experimental evaluation of a prototypes performance on representative work tasks is described. Qualitative analysis of the data suggests that there is substantial room for performance improvements, and has inspired design changes to realise this potential. Quantitative analysis proved inadequate to detect a numerical advantage due to the sample size restrictions inherent when doing academic research in a corporate setting. The problems of empirical evaluation in a corporate setting are discussed. A new approach to robust quantitative evaluation is proposed.*

## INTRODUCTION

Although, in principle digital libraries (DL) are constructed to support the information needs of their community of users[2], they currently focus on the task of document delivery. When users needs are examined [10] [1] it turns out that information needs are met by sub-document units of text. Consequently, as well as enabling people to get to relevant documents, digital libraries should support users in getting to the *parts* of documents they need. We describe a new technology being developed to enable information seeking *within* documents on a corporate intranet. The resulting tool, named Gridvis, uses interactive visualization of metadata to support goal-directed document use. An initial evaluation is described and the prospects for progress and further evaluation are discussed.

### Why support information seeking within documents

By regarding a set of relevant documents as the final product of Digital library (DL) use, DLs may be stopping short of their ultimate mission.

Although DLs generally focus on providing documents, the goal of a DL is actually to support the information needs of its community of users [2]. In a study of the needs of digital library users Van House [10] states that: "Workplace users [...] want to retrieve information rather than documents per

se.". The work of Adler et al [1] looked at the reading tasks performed by people in a variety of work settings and found that rapid goal-directed types were very prominent. In goal-directed reading, users only make use of small portions of documents. Hence, it seems that in both document and digital library use, information needs seem to often be met at the sub-documents level. Yet, apart from a few interesting exceptions (e.g.[5] [6]), DLs and the visualization tools designed for them, provide no support for information seeking within documents.

## SYSTEM DESCRIPTION

The development of this technology was motivated by Sainsbury's, a large UK retail organization that sponsors this research. As such, the ideas behind the technology and the resulting application have been developed for a specific corporate environment, in an iterative fashion with the close involvement of potential users.

### The documents and their metadata

Gridvis was envisaged for use with a small but important portion of a diverse corporate document landscape, which can be characterized as being information rich and having a relatively long life. A set of documents with these characteristics was selected (as described in the evaluation section) to serve as a basis for developing a metadata taxonomy and conducting evaluation.

Gridvis parses a document marked up with metadata and one containing a metadata taxonomy. It uses the result to produce visualization and customized HTML views of the document text. The metadata taxonomy is necessary for several reasons, only some of which can be fully described here. For the users of Gridvis undertaking a goal-directed search, it is essential that the metadata has some meaningful organization. With a good taxonomy the user can rapidly locate a small subset of the metadata relevant to their information need, rather than undertaking an exhaustive linear search through all the metadata tags describing the document's contents. Other benefits of the metadata taxonomy are the advantages it has for metadata authoring (discussed below), and the extra power it allows in automatic searches.

The metadata and the metadata taxonomy were constructed by adapting informal ontology construction methodologies [9]. The metadata was manually produced for each paragraph of the documents prepared for Gridvis, using the taxonomy as a guide to encourage complete and consistent tagging. Each instance of metadata was given an

'applicability' attribute which recorded the degree to which the metadata tag applied to a paragraph (1 - 'is principally about', 2 'partly about' or, 3 'mentions in passing').

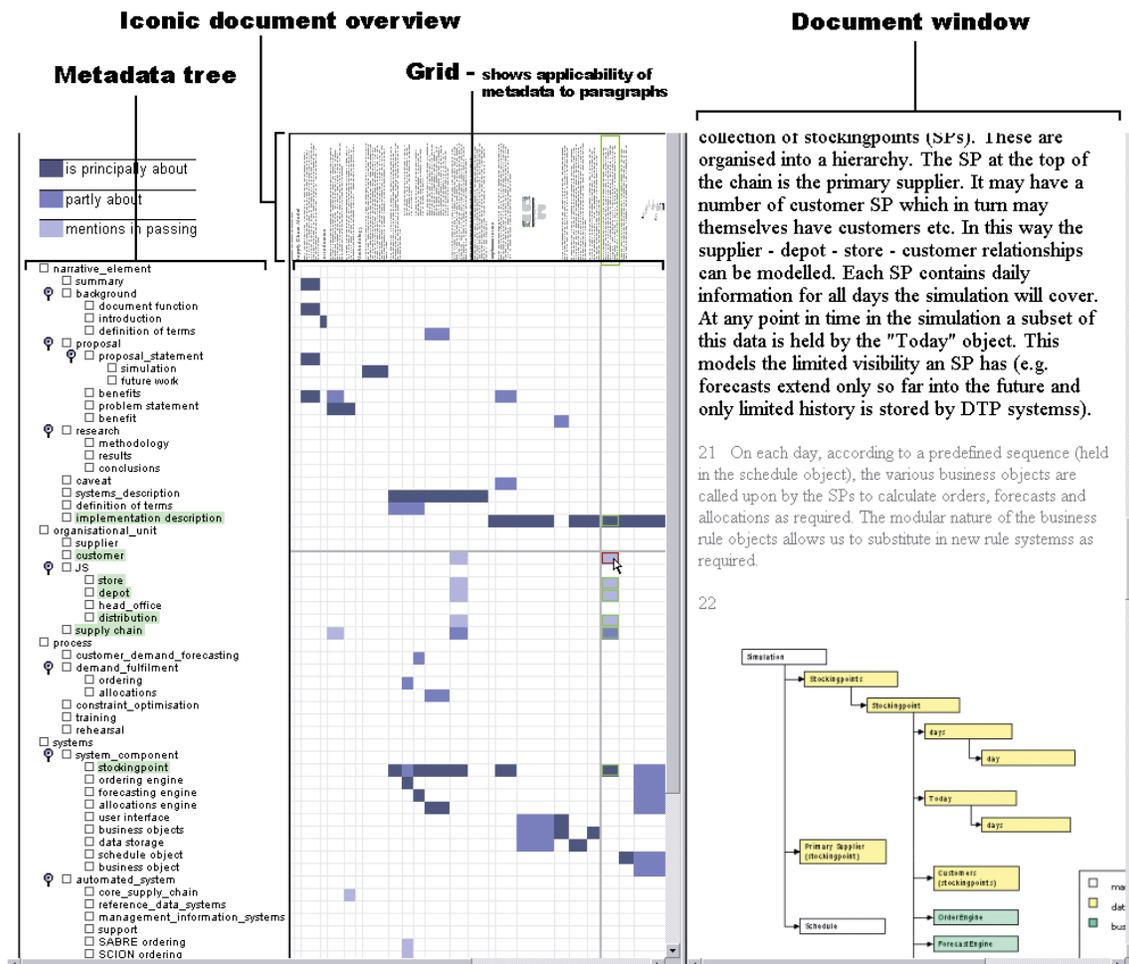
Since Gridvis is designed for use on a corporate intranet in a corporation where software installation for low priority applications is problematic it was given a client-server architecture: the visualization is a client-side Java applet, the queries are answered by a Java servlet using XLS-T to produce customized HTML documents. The documents and taxonomy are encoded as XML. The document structure is described with a standard hierarchical XML structure; the metadata is embedded within this alongside the text it describes.

Existing DL's make extensive use of both metadata and the technologies and standards upon which Gridvis has been built. Gridvis could therefore be smoothly integrated into their functionality offering.

**The visualization**

The Gridvis client-side application contains a visualization

of a document's metadata (see fig 1). This visualization can be considered as three interlinked sections; the metadata tree on the left-hand side, the iconic document overview running along the top, and the grid sitting at the center. The metadata tree consists of the document's metadata tags and the metadata taxonomy. The taxonomy is used to create a tree structure and the metadata tags are added to this as leaf nodes. Each tag corresponds to one row in the central grid. The iconic document overview is a miniature image of the document laid out horizontally along the top of the visualization. Each paragraph in this overview lines up with one of the columns in the grid below. The central grid shows which metadata tags have been applied to which paragraphs in the document; each column represents a paragraph and each row represents a tag. Hence, to show that a tag has been applied to a particular paragraph the cell where the appropriate column and row meet is shaded. The color of the cell is determined by the value of the applicability attribute given to the tag when it was applied to



**Figure 1** A screenshot of Gridvis. By placing their mouse over a cell in the grid the user has highlighted a paragraph in the iconic document overview and all the metadata terms in the metadata tree relevant to this paragraph.

this paragraph; the higher the applicability the darker the color (a legend for this scale is always visible in the top left hand corner). The metadata tags and taxonomy, iconic document overview and grid, are thus brought together to produce a visual overview of the document.

The user can query the visualization with their mouse through a combination of dynamic querying and brushing [3]. When the mouse is placed over a cell in the grid, all the tags in the column are highlighted; hence the paragraph's content is described. By looking along the row of an interesting tag, the user can see where and to what extent that topic comes up in the document. So by moving the mouse in the central grid the user can explore what different parts of the document are about, and where in the document particular topics are covered.

The user can also query the document by clicking on any cell in the grid. Clicking on the grid will produce a query that is sent to the servlet formulated with the metadata and paragraph corresponding to the selected cell. The servlet will produce an HTML version of the document in which every paragraph tagged with the selected metadata is highlighted using a bold font. The browser will display the paragraph selected; thereby offering details on demand [3].

#### **An illustrative scenario**

Scenarios [4] have been found to be effective design and communication tools. Consequently we have included the scenario below to give a better idea of how the functionality described above might be used in practice.

Clair, a Sainsbury's (JS) employee, sits down to look at this week's 'industry update' newsletter, with a view to finding out whether the company plans to sell its US interests. She starts up Gridvis and looks through the metadata to see if any of JS's US subsidiaries are mentioned. She scans down the top-level entries in the metadata taxonomy until she reaches 'organizational unit'. Then she scans the next level of entries under 'organizational unit' until she reaches 'JS'. Under JS she sees an entry for 'Shaws' which is a US subsidiary of JS.

Having found a relevant tag, Clair proceeds to see if any of the paragraphs relevant to this tag will say anything about plans for a sell off. She looks along the tag's row in the grid and sees that two cells have been colored in. Passing her mouse over the first she sees the other colored cell in the same column, and its tag, become highlighted. The other cells tag label reads 'loyalty cards'. Seeing that therefore the paragraph is not relevant she moves her mouse on to the next colored in cell. In this case the 'acquisitions' and 'USA' tags light up. This surprises Clair, since she expected Shaws to be on sale, not making acquisitions; she decides to look at the paragraph and clicks on the cell. The right hand window is refreshed, with the paragraph she wanted appearing at the top of the screen in bold.

#### **EVALUATION**

An experiment was undertaken to see whether Gridvis offered a performance advantage for goal-directed search tasks.

#### **Method**

The documents selected were a set of five reports written by an analytical research team within Sainsbury's. These are

substantial papers, which may inform business choices on specific supply chain issues for the foreseeable future. Five of these documents were used in this experiment; these were tagged with metadata, and paragraph numbers.

Three information needs for each document were elicited from the original audience of the documents. One to two people were interviewed for each document (corporate restructuring meant this was the best that could be done). They were asked what they had been looking for, or would look for now, when they read the document. This process was undertaken to ensure the information needs used would be as representative as possible of what Gridvis would be used for in a real work situation.

The relevance of paragraphs in the document to each information need was established using three judges and an adjudicator. In each case one of the judges was the document's author, and the other two had been subjects in the experiment described below. They were asked to read through the document carefully to locate every relevant paragraph. They were asked to classify the relevance of each paragraph as one of 'contains key information', 'contains relevant information', and 'contains related information'. An adjudication session, including three of the judges and the researcher as an adjudicator, was used to decide the relevance of paragraphs where the original judgments showed large disagreements. Where the differences were small, a mean score was calculated. This process resulted in a single set of relevance judgments with which the retrieval performance could be measured.

Twelve participants were found who were employees of Sainsbury's that had not read the documents and claimed to be familiar with the concepts in them. Each had the visualization explained to them and were asked to familiarize themselves with it in a training session where they attempted to satisfy three information needs as quickly as possible. The subjects were then asked to repeat this task with four new documents. Two of the documents were presented with Gridvis; the other two were displayed in a conventional browser window. When using Gridvis, the subjects were asked to only make minimal use of the document window's scroll bar and thus to rely on Gridvis for navigation and search.

The detail of users' interactions were preserved through a continuous screen recording of the entire experimental session. Every document query action (i.e. when the participants clicked on a cell in the grid) was automatically recorded in a text file and the relevance of paragraphs found were noted by the subjects. They were asked to record the paragraph numbers of germane paragraphs and to classify each paragraph's relevance. A paragraph could be given three levels of relevance, as described above.

#### **Results and discussion**

A variety of performance measures were calculated from the results. But the large variance observed precluded a statistically significant result with anything but large difference between performance with and without Gridvis. For the medium effect size observed we would require a sample size of 126 to get a significant result[7]. A controlled study of this size would not be tractable in the

corporate environment for which this tool was developed. A qualitative analysis of the data was therefore undertaken to determine why Gridvis did not produce a large performance improvement and what improvements might be possible. This analysis started by looking at what strategies the subjects might be using to make queries with Gridvis. The results of an earlier task analysis of Gridvis use were used to focus the analysis of strategy usage. The task analysis divided document query production into two components. Firstly, the user has to find a relevant metadata tag, then they must select from amongst the paragraphs relevant to this tag. The task analysis also suggested a number of strategies that might be used to make these selections.

This initial set of strategies was used to classify each tag and paragraph selection action of the 12 subjects. The classification was made on the basis of inspecting video and text log files of the session. During analysis of the first few subjects the classification was refined until it could account for 95% of tag selections and 88% of paragraph selections. The strategies used in this classification will only be listed briefly, with only those that are raised in later discussion being explained. The tag selection strategies were: keyword mapping, synonym keyword mapping, related concept keyword mapping, and narrative keyword mapping. The paragraph selection strategies were: 'left-to-right', 'applicability only', 'co-occurrence only' and finally 'applicability and co-occurrence'. An applicability strategy is one that favors paragraphs of high applicability (i.e. the tag is highly descriptive of the paragraph). A co-occurrence strategy is one that favors paragraphs tagged with other relevant tags.

The frequency of usage of different tag selection strategies was examined first. Subjects were found to use a range of strategies, with no one strategy being unduly neglected. The only problem observed here was a tendency for subjects to overlook some relevant tags as evidenced by the use of a mean of 3 out of the 6 tags relevant to each information need. However, there did seem to be problems with the use of paragraph selection strategies to pick paragraphs related to a tag of interest. Here the users were found to seldom use strategies that capitalize on information given by the visualization. Only a third of tags were queried using the 'applicability only' strategy. Additionally, users seemed to make little use of the 'co-occurrence only' strategy. Co-occurrence information is important since it can serve to more fully describe a paragraph's content and disambiguate metadata tags. The analysis revealed that co-occurrence seemed to be used to select paragraphs only for a fifth of the tags inspected. Hence, in the case of paragraph selection, the information offered by the visualization was little used. This apparent under-utilization of the information provided by Gridvis suggests performance improvements would be realized if the information were used. The validity of this suggestion depends on the assumption that making full use of information about applicability and co-occurrence will lead to better performance. An initial test of this assumption was undertaken by exhaustively applying the tag

and selection strategies listed above.

Recall (proportion of all relevant paragraphs, located) was increased to threefold of that achieved by subjects in across conditions by exhaustive application of tag selection strategies. It was also found that the resulting decrease in precision (the proportion of located paragraphs that were relevant) could be offset using the paragraph selection strategies. These strategies were used to rank the set of paragraphs related to the tags identified with the tag selection strategies. The top-ranked half of this set of paragraphs had a recall and precision about 60% higher than the bottom-ranked half. Hence, the tag and paragraph selection strategies can be used to increase precision and recall.

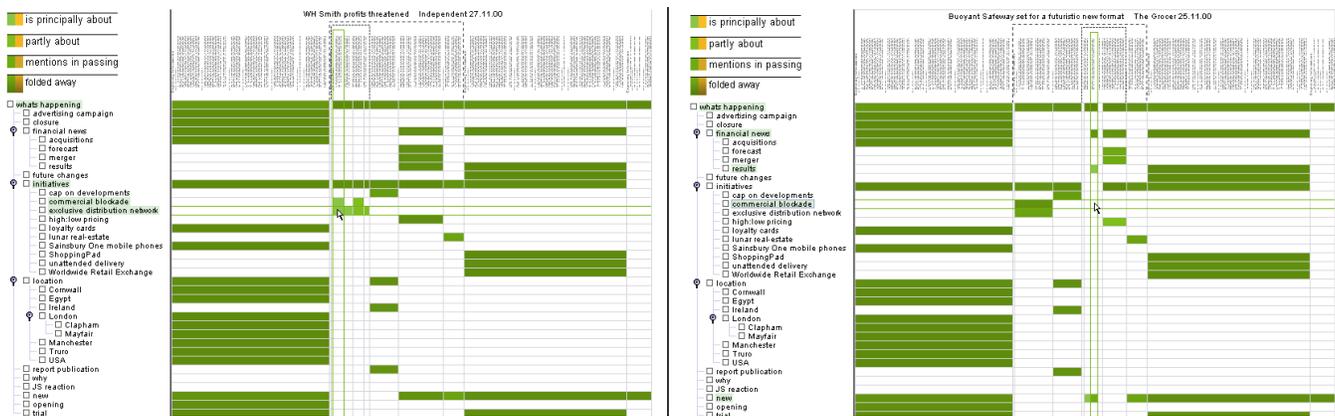
If making more use of the information given by the visualization would have improved performance, why did users not do so? Lack of familiarity and expertise might be one reason. A more important problem is that the cost of using this information might be too high. This seems particularly true for co-occurrence information. To use this, the user must read the metadata tags applied to each paragraph. The tags for each paragraph must then be held in memory while their relative relevance is assessed. Moreover, the application design makes using co-occurrence information particularly difficult and time consuming in some circumstances. For example, if there is a set of interesting tags, the user might want to find paragraphs relevant to them all. This requires the user to scan along the row of each interesting tag to find colored cells. Upon arriving at each of these, the user must then scan vertically to check if the cells relating to the other tags of interest are colored in.

Another possible factor inhibiting the use of applicability and co-occurrence information is Gridvis' voracious appetite for screen real estate. With longer documents the users have to scroll horizontally to see the whole grid. This makes it impossible to see every paragraph relevant to a tag and therefore confounds attempts to select the most appropriate paragraph.

### **1.1.1 Design improvements**

New functionality has been added to Gridvis, to address these problems. The co-occurrence strategy being made easier through partial automation; if the user selects tags of interest, Gridvis indicates the number of tags that co-occur on each paragraph. The screen real estate problem is being addressed by a new implementation that dynamically redraws the grid at varying levels of detail to only show in full the parts of the document relevant to the users current activity (see fig. 2).

Design improvements have not all resulted from the evaluation described above. Many have arisen during informal user testing and discussion with the potential users at Sainsbury's who are sponsoring the research. For example, they complained that it was difficult to keep their eye on the same row when looking along a row in the grid to see which cells were filled. To ease this task it is now possible to select a tag in the metadata tree and have its paragraphs automatically highlighted



**Figure 2** The two screenshots of Gridvis illustrate the level of detail management feature. As the mouse is moved from one paragraph to another, only the paragraphs in the section the mouse is currently over are shown, the rest are shown at the section level.

### 1.1.2 Future evaluation

Ideally, Gridvis should be evaluated on precisely the tasks for which it was designed, i.e. goal-directed search of corporate documents with the goals of corporate users. Moreover, the evaluation should be able to answer a sizeable portion of the multitude of questions needed to inform application design and improvement.

Robust quantitative empirical evaluation using controlled experimentation often depends, as it has with Gridvis, on having a reasonably large sample size. This presents an insurmountable problem to the academic researcher working in a corporate environment.

An alternative approach, perhaps more suited to evaluation in a corporate setting is performance modeling. We are currently using quantitative optimization models of information seeking developed at Xerox PARC [8] to predict the best performance achievable with the interface given a set of well-defined strategies. This ‘best case’ can then be compared to a control condition where people perform the same task with paper documents, where they are already expert performers. Such comparisons needs much smaller sample sizes of around 25 subjects [4]. Another advantage is that many different task parameters (e.g. document length, query difficulty) and design possibilities (e.g. improving the speed of certain operations) can be examined more cheaply and more powerfully.

Perhaps what is most important about this approach to evaluation is that it will make it possible to concretely assess whether the application is potentially useful. If the ‘best case’ performance for Gridvis is worse than, or the same as, that for paper, one can be sure that Gridvis will never prove an effective tool. In that case, future researchers attempting to support document use can productively avoid making a futile attempt to achieve this with the interactive visualization of metadata. On the other hand, if the ‘best case’ performance exceeds that of paper we can be confident that the application will be useful given the user interface and/or user training is good enough.

### CONCLUSION

The work a DL does not stop once useful documents are

retrieved. DL research should explore ways in which to support the ‘last mile’ of digital library use; information seeking *within* documents. This paper has described Gridvis, a tool which attempts to accomplish this. It enables rapid goal-directed retrieval through an interactive visualization of paragraph level metadata.

Empirical evaluation was used to examine whether Gridvis increases the efficiency of information location. Subjects were asked to satisfy a set of information needs using online documents with, and without, the aid of Gridvis. Quantitative analysis did not reveal a significant numerical improvement in performance when using Gridvis. However, qualitative analysis suggested that subjects were not making full use of the metadata. This result helped to identify several problems with the Gridvis interface, which have been tackled with a set of design improvements.

Problems with the evaluation approach use are discussed and a model-based approach is suggested. The proposed evaluation methodology will be able to assess the ultimate value of Gridvis within the constraints imposed by working in a corporate context.

### ACKNOWLEDGMENTS

This research was sponsored by J Sainsbury plc and undertaken within the Postgraduate Training Partnership established between Sira Ltd and University College London. Postgraduate training partnerships are a joint initiative of the DTI and the EPSRC. The first author would also like to thank Dr John Gilby for his wise counsel and Anne Pomery for her spiritual support.

### REFERENCES

- [1] A. Adler, A. Gujar, B. L. Harrison, K. O'Hara and A. Sellen, A diary study of work-related reading: design implications for digital reading devices; Conference proceedings on Human factors in computing systems , 1998, Pages 241 - 248
- [2] Borgman C. L. Final Report, UCLA-NSF Social Aspects of Digital Libraries Workshop, '96
- [3] S. K.Card, J. D. Mackinlay, & B. Shneiderman

(Eds) (1999) Information visualisation: Using vision to think, Morgan Kaufmann, p 223

[4] Carroll, J. M. (2000) Five reasons for scenario based design, *Interacting with computers* 13 (2000) 43-60

[5] CIMI. Cultural Heritage Information Online (CHIO)  
[http://www.cimi.org/documents/tagging\\_guide/tg1.htm](http://www.cimi.org/documents/tagging_guide/tg1.htm)

[6] Hearst M A., and Plaunt C (1993) *Subtopic structuring for full-length document access*; Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval , 1993, Page 59

[7] D. C. Howell, *Statistical Methods for Psychology* (4<sup>th</sup> edition) Duxbury 1997

[8] P. Pirolli and S. K. Card (1999). Information Foraging. *Psychological Review* 106(4): 643-675.

[9] M. Uschold (1996) *Building Ontologies: Towards a Unified Methodology*, <http://www.aiai.ed.ac.uk/~enterprise/enterprise/ontology.html#papers> - abridged version appears in Proceedings of the Workshop on Ontological Engineering in conjunction with ECAI 96

[10] N. A. Van House, User Needs Assessment and Evaluation for the UC Berkeley Electronic Environmental Library Project, *Digital Libraries '95: The 2<sup>nd</sup> Conf. on the Theory and Practice of Dig. Lib.*, June '95, Austin, US