Towards the development of a full-text, searchable database: implications from a study of journal usage

Andrew Dillon, John Richardson and Cliff McKnight
HUSAT Research Institute
Elms Grove
Loughborough
Leics. LE11 1RG
England

Abstract

The present paper reports on a study of journal usage amongst professional researchers. The aim of the study was to shed light on how journals are used with a view to making recommendations about the development of a full-text, searchable database that would support such usage. The results indicate that levels of usage vary over time, the range of journals covered is small and readers overlook a large proportion of the contents of articles. Furthermore, three reading strategies are observed which indicate that the structure of journal articles is not ideally suited to their uses. The implications of these findings for developing suitable computer-based applications are discussed.

Introduction

As advances in technology continue to reduce the cost of hardware, computer applications to support the storage, retrieval and presentation of information are now emerging which offer the ability to interact with published material in previously impossible ways. Hypertext is probably the most obvious and discussed example (e.g., McKnight, Richardson and Dillon [1]) but no less deserving of attention from researchers is the possibility of mass local electronic storage of standard format texts that can be presented immediately to interested readers.

To date much work has focussed on possible differences between reading from paper and reading from screen and early research suggested a speed and accuracy decrement for screen reading (Wright and Lickorish [2]). Recent research by Gould et al.[3] has indicated that under optimal conditions —high resolution, black text on white background using anti-aliased characters— such performance deficits could be overcome. Therefore it appears that there is no inherent technological restriction to presenting text on screen (for a detailed review of these issues see Dillon, McKnight and Richardson [4]).

Important as this may appear, it is not enough that image quality is improved. Given a straight choice between screen and paper based texts, readers will almost certainly choose the latter. It is important therefore that screen based presentation exploits the capabilities of technology to support the reader and offer him facilities that paper never could. To do this requires a detailed understanding of the nature of the reader's task and their requirements of printed material.

To date relatively little work has been carried out on these issues. The British Library funded Birmingham and Loughborough Electronic Network Development (BLEND) project (Pullinger [5]) studied the problems of setting up an experimental electronic journal. Although the main aims of the project concerned the use of an electronic communication network as an aid to writing, submitting and refereeing papers, some work was done on the reading interface of an electronic journal. Among the findings was the indentification of three 'strategies' adopted by readers of journals: general filtering through the article, preliminary filter of the title and abstract followed by a request for a photocopy, and skimming through several articles in a journal looking for new ideas.

Such descriptions, though hardly what can be described as strategies, suggest the possible importance of reading style on the likely acceptance or rejection of electronic text. For example, readers who like to skim rapidly through several articles in one session will reject any presentation medium that requires specification of each article to be read in advance or requires the user to wait for several seconds between articles. Readers who only want to look at the abstract to decide if a photocopy is necessary will not want to load complete articles to view on screen, preferring to access numerous abstracts rapidly.

Background to the present study.

Current work at HUSAT involves the development of a full text, searchable database for use by staff. Given the nature of our work, staff are comparatively frequent users of journals. An in-house library exists to serve immediate needs and the main Loughborough University library on campus acts as a supplement. Being professional researchers, the staff afford a useful test base for investigating how journal material is used and the issues involved in developing acceptable electronic versions. The proposed database will contain scanned contents of several years' issues from relevant journals. The present study was motivated by the need to identify pertinent reader characteristics that would affect acceptance or rejection of such a system. It was hoped that by examining how readers interact with journals we would be able to make informed decisions about the type of system required by typical readers.

Method

Subjects

15 subjects were selected to take part in the study. No elaborate sampling procedure was employed but all were practising human factors researchers currently engaged in a variety of projects. Background disciplines were primarily psychology and ergonomics. As a sample they represent individuals likely to be heavy users of journal literature.

Procedure

Subjects were interviewed individually. The first part of the trial consisted of a brief interview which tapped information relating to frequency of use, range of journals covered, requirements of journals, levels of photocopying etc.. When the interview ended subjects were presented with a selection of journals from their specified range of interest, asked to imagine this was the first time they had seen these editions and to interact with the journal as normal, articulating what they were attending to as they did so.

Results

Frequency

The frequency with which individuals access journals seems to depend on their purpose. Virtually all subjects distinguished between problem-driven journal usage where work demands require literature reviews or rapid familiarisation with a new area, and personal usage where journals are browsed in order to keep up with latest developments in one's area of expertise or interest, the former being cited more frequently than the latter. Obviously work demands vary and periods of heavy use are matched by times of little requirement for articles. Therefore it is difficult to express a frequency rate for usage of this type over several months or more. Usage of the second variety is somewhat easier to quantify. 60% of subjects reported accessing journals at least once per week for this purpose. The highest levels of usage were twice per week (27%) and the lowest were once every two months (13%).

Full frequency results are presented in table 1:

Rate of usage	Subjects (%)
2+ per week	4 (27%)
1 per week	5 (33%)
2 per month	2 (13%)
1 per month	3 (20%)
1 per 2 months or more	1 (7%)

Table 1: Reported usage rates of journals

It should be emphasised that such rates refer to browsing journals, not reading articles in-depth. This will be discussed further in a later section.

Range of journals

Given the professional interests of the subjects it is not surprising that the most frequently cited journals were International Journal of Man-Machine Studies, Behaviour and Information Technology, Human Factors, and Ergonomics. Applied Ergonomics and Design Studies are both read by more than 2 subjects but noticeably less than the previous four journals which are seen as the primary journals in this area. A variety of other journals ranging in subject matter from expert systems to social psychology are read by individual researchers on a regular basis.

Coverage

Subjects were asked whether they felt that their usage of journals resulted in satisfactory coverage of relevant material and if they were confident that important material was not being overlooked. Interestingly none of the subjects reported that they felt they captured everything and 20% reported that they probably missed a lot of relevant material. However the majority of users (54%) felt they only missed some and a third were satisfied that they covered most of the relevant material in the course of their routine interactions with the literature.

Some interesting comments were passed about coverage that warrant reporting. All subjects work in research teams and the onus of identifying and assimilating relevant articles in journals does not lie with any one individual. All subjects reported that articles are often pointed out to them by colleagues and that collectively, most relevant material will be assimilated by the team if not the individual. Whether or not this is a particular case of diffusion of responsibility remains an open question. Other subjects stated that while their coverage of the literature may be restricted to the primary journals, if important work was being carried out it would eventually surface in one of these, or put another way, if it isn't published here it probably isn't very important.

Access to library

While the in-house library stocks most of the relevant journals in the area, individuals wishing to undertake more detailed research and access a wider range of material are expected to use the campus library. While all subjects are aware that superior facilities exist on campus only one subject regularly visits this library. The full results are presented in table 2:

Visits to the Campus library	No. of Subjects
Regularly	1 (7%)
Occasionally	6 (40%)
Seldom	7 (47%)
Never	1 (7%)

Table 2: Reported frequency of visits to campus library

The main reason given for not utilising the facilities available on campus was the physical distance of the workplace from the library (a distance of two miles). Most subjects claimed they would visit the library more frequently if it was located nearby. A further reason, not unrelated to the first, was lack of time.

5 subjects (33%) use Inter-library loans to obtain material and 7 subjects (47%) have a personal subscription to one or more journals, though one must point out that membership of the Ergonomics Society renders subscription to one journal (Ergonomics) obligatory.

Photocopying journals

Subjects were asked to describe where and when they read journals and the extent to which they made photocopies of articles. None of the subjects reported spending time in the library reading material in detail. All prefer to identify relevant articles in the library and select material that will be read later at home or in the office. Three users expressed strong dislike for the atmosphere of libraries, describing it as "depressing".

All users reported that they made or read photocopies regularly. Since readers prefer to work outside of the library and only bound journals are released for short periods, this is not surprising. However journals that are stored in-house and which can be borrowed are also photocopied regularly and subjects expressed a number of reasons for this. Firstly, they claimed rather altruistically, that they did not want to remove the journal in case other people wanted it so they made a photocopy of the article that interested them. Secondly, making a photocopy provides readers with a personal copy that can be read in detail at their convenience. Thirdly, they can mark, highlight and write on photocopies, and finally, photocopies can be stored and retrieved for later use with minimum inconvenience.

While these reasons seem plausible, 8 readers (54%) admitted that they sometimes photocopied articles that were not read and there were distinct differences between subjects in terms of their likelihood of photocopying material. 6 subjects (40%) claimed they only photocopied articles when they were sure such material was important or directly relevant to their needs, while 8 subjects (54%) admitted being less discriminate and making copies of anything that appeared to have some interest to them or their team members.

Proposed Database

The proposed development of a full-text, searchable database of articles from leading journals was outlined and subjects were asked if this would be of use to them. All users felt that it would be of some use, though more as a supplement to their current journal usage than a replacement. Subjects varied in terms of how far back the contents of the database should go, suggestions ranged from 3 to 8 years (anything over 5 years old is reckoned to be considerably out of date in this area).

Reading style

Subjects were given unbound issues of several journals typical of those they read, asked to imagine this was the first time they had seen these editions and to interact with them as they would normally. Comments about where they were looking and what decisions they were taking about articles were elicited throughout this activity.

It is possible to identify a general pattern to these interactions by tracing the sequence of activities typically performed and the decisions made by readers as they used journals.

Firstly all users look at the table of contents of the issue. A preference was expressed for contents printed on the front or back page which made location of relevant articles possible without opening the journal. At this point readers tend to scan the contents by looking primarily at the titles of papers or the authors. 6 readers reported looking at titles only, 4 reported looking at all the titles and only then referring to authors and 4 readers reported looking at all the authors first and then looking at the titles. Only 1 reader reported looking at both the title and author of each paper. Titles were considered to describe the contents of the paper though several readers remarked that they were often misleading. Authors were primarily scanned to check if certain individuals familiar to the reader had papers in the issue. Lack of familiarity with authors was reported as the major reason for not attending to authors when checking the contents.

If the reader fails to identify anything of interest at this point the journal is put aside and, depending on the circumstances, further journals may be accessed and their contents viewed as above. When an article of interest is identified then the reader opens the journal at the start of the relevant paper. Title and author of paper tend to be ignored at this point except to confirm that the right article has been accessed. 3 readers reported attending to the author's address in order to gain an impression e.g. nationality, academic or industrial background etc. and it seems likely that early judgements about the probable quality of the paper are occasionally made on this basis.

The abstract is usually attended to next. However, many of the present sample were very critical of abstracts, describing them as "misleading" or the author's attempt to "sell" the article to the reader. 2 readers actually ignore the abstract and start at the introduction. Only 5 readers actually read the abstract fully, the remaining 8 all scanned it quickly. At this point a decision may be made about the suitability of the article for the reader's purposes. However, given the poor estimation of abstracts, most readers tended to at least view other parts of the text before fully rejecting an article.

The next phase of reading tends to be a quick scan of the rest of the article. Most readers reported browsing the start of the introduction at this point before flicking through the article to get a better impression of the contents. At this stage the readers in this study reported attending to the section headings, scanning the diagrams and tables, noting both the level of mathematical content and the length of the article. Browsing the conclusions seems to be a common method of extracting central ideas from the article and deciding on its worth. 10 subjects reported reading or browsing the conclusions at this point.

References are browsed by some readers in order to further their impressions of the article. This may involve browsing the actual reference list or noting names as they appear in the text. In all, 6 subjects reported attending to the references at this stage. Doing so was seen as a way of identifying the theoretical perspective of the author and appreciating the relevance of the work to the reader's needs. Therefore even though the author may be using terminology that appears relevant, he may actually be approaching the problem from a very different perspective and the references may indicate this. 2 readers remarked that if they were familiar with most of the names cited in the article then it was unlikely that the article was saying anything very new.

At this point readers seem to have completed a cycle of interaction with the article and decide whether or not to proceed with it. A number of factors may lead to the reader rejecting the article. The main reason is obviously content. The reader by now has a strong impression of the type of material contained in the paper and will be able to make an informed decision on the relevance of it to his needs. How accurate this impression is remains an empirical question. If the article is heavily mathematical it tends to be rejected by the readers in this sample. Poor sectioning, large method and results sections, small discussions and large size in terms of number of pages were all cited as factors that would influence a reader's decision on whether or not to reject an article. Whether or not to photocopy the article is often decided at this point too.

The concept of article size is interesting. Large articles obviously require a significant time-investment which is often seen as a disincentive. Perceptions of what constituted a large or small article varied. Large articles were described as being anything from 6 to more than 30 pages long, medium length articles as being between 5 and 20 pages long and small articles being between 3 and 20 pages long. In other words what one individual rates as large, another may rate as small. Median responses suggest that articles more than 20 pages long are large and those articles that are about 5 pages long are small. Approximately 10 pages is considered to be medium length.

If the article is accepted (or photocopied) for reading it is likely to be subjected to two types of reading strategy. The majority of readers (10) will scan read the article in a non-serial fashion to rapidly extract relevant information. This will involve reading some sections fully and only skimming or even skipping other sections. Typically the method and results sections of experimental papers are skim read while the introduction or introductory sections and the discussion/conclusions are read fully. Readers may highlight points or make notes at this stage.

The second reading strategy is a serial detailed read from start to finish. This was seen as "studying" the article's contents and though not carried out for each article that is selected, 11 readers claimed that they usually read selected articles at this level of detail eventually. 2 readers stated that they rarely did this unless their scanning of the article failed to provide them with the requisite information or the discussion brought up points that required closer reading of the method or results section to fully appreciate. Some readers (3) expressed a preference for this reading strategy from the outset over scanning though acknowledging it to be less than optimal.

While individual preferences for either strategy were reported most readers seem to use both strategies depending on the task or purpose for reading the article, time available and the content of the article. Original and interesting work is more likely to be read fully than dull or routine papers. Reading to keep up with the literature requires less "studying" of articles than attempting to understand a new area. If reading the article with a view to citing it in one of their own papers, readers expressed a stronger tendency to read the article fully. However, even when reading at this level of detail some readers still reported skimming particular sections that were not intrinsically relevant to their particular needs at that time. One user reported reading serially from the introduction until interest waned whereupon she read the conclusion to decide if the paper was worth reading at this level. If the conclusion led her to believe that it was, she returned to the point where she had been and continued her detailed start-to-finish read. This may happen several times in the course of reading one article, she claimed. Since actual reading of the article at this level was not part of the study it is impossible to comment meaningfully on these reports.

Summary

Readers select text on the basis of their current purpose. Articles overlooked for work purposes may be selected at other times for personal use and vice-versa. The range of journals used on a regular basis is restricted.

Information gathering tends to be problem-driven but is not particularly systematic. There were few mentions of citation indexes, abstracts or on-line searches. Few subjects seemed aware of the range of services offered by the main library.

A large proportion of the material in articles is overlooked by readers. Despite claims that they usually read articles in detail eventually it seems as if this is not the norm. Coupled with the fact that articles are sometimes photocopied and never read it seems as if the introductory sections and conclusions provide most users with their information about an article.

Although many respondents admitted that they probably "missed" some relevant material this did not concern them unduly. Most respondents claim to select articles on the basis of titles, yet these are rated as often misleading. One wonders how often potentially useful articles are ignored.

Contributory factors to article rejection are a high proportion of mathematical content, very detailed method and results section, large size, poor sectioning, poor or patronising style, and overly-technical language.

It is possible to distinguish between three types of reading that articles are subjected to:

1. Rapid scanning of abstract and/or introduction, section headings and occasional paragraphs within sections, figures and tables, and discussion.

2. Detailed scan of relevant sections: usually the introduction and discussion but may include any part of the text that the reader chooses. May include all text, not necessarily serial.

3. Detailed serial read from start to finish.

It is unlikely that any of these are rigidly adhered to. Detailed serial reading is likely to involve some degree of jumping for example and rapid scanning may actually include some detailed scanning of sections.

Implications for the design of a full-text database

It is unlikely that merely reproducing the style of the paper version will be of benefit to the users of any technology designed to support reading. In order to encourage use of such facilities it is important that users can perceive clear advantages for investing time and effort in learning a new system. Obviously, if the database provides rapid access to material not immediately available in the library then it may encourage use, however our experiences with users of the innovative ADONIS system show that this may be a necessary but not sufficient condition.

If we note the reading styles outlined above several performance characteristics emerge that would appear important for design purposes. Firstly all users attend to the Contents page of journals and prefer these to be easily accessible. It would seem therefore that a facility to scan lists of titles and authors would be desirable. These should be grouped as they are on paper i.e., in "issues", but the ability to scan continually should be available.

Secondly, since the full contents of the paper are not attended to at this point it is better that users are given brief information about the paper and offered the chance of jumping around to various sections of the text. The default mode of article presentation should not be the same as the paper equivalent. A likely presentation style based on the present findings might be: the title of the paper, the author(s), the abstract, a list of section headings that are selectable and the references cited. Further information about the size of the article might also be useful.

Thirdly, rapid browsing facilities are vital. At the initial stage of article browsing fast page-turning is common as readers jump back and forth through the article. The electronic version must support this activity by allowing both serial scrolling of the next page/previous page variety and rapid jumping to particular sections e.g., from the introduction to the method or the discussion. It might be desirable to facilitate jumping to "landmarks" in the text such as tables or figures too. We have implemented a reading interface that will facilitate such interactions and investigations of user performance with and ratings of such a system are currently underway.

Fourthly, the ability to print the article at any point would be desirable as obtaining hardcopies of selected articles is a major concern of most journal readers. Keeping a record of interesting articles which can be batch printed later may be desirable. Given the observed reading styles of the present sample it might be useful to offer the facility to print sections rather than the full article. For example, readers might choose to print the introduction and discussion sections only. This would have the advantage of reducing costs of obtaining hardcopies and save on unnecessary use of paper.

Obviously these are relatively general considerations. It is not possible to generate specific principles for interface design on the basis of the present study. The optimal structure for electronically presented articles and the most usable way of navigating through this structure are empirical issues which require much detailed research. We are currently addressing these questions in our further research. However it is clear that typical article structures in paper are wasteful and not particularly usable for anything but a detailed serial read of the complete text.the introduction and discussion sections only.

Physical access remains a major stumbling block. This concurs with the findings of the BLEND project. Virtually all subjects stated that they would like the database on their desk. Making the effort to use facilities available elsewhere (even if it is only someone else's office in the same building) is seen as a deterrent to use. It is unlikely that such a facility will be available on every desk for the foreseeable future but some attempt should be made to increase physical access by careful location of the proposed system.

Acknowledgements

This research was funded by the British Library Research and Development Division and was undertaken as part of Project Quartet.

References

[1] McKnight, C., Richardson, J. and Dillon, A. The authoring of hypertext documents. In R. McAleese (ed.) Hypertext Volume 1, Norwood, N.J.: Ablex (in press)

[2] Wright, P. and Lickorish, A. Proof-reading texts on screen and paper. Behaviour and Information Technology 2 (3) 1983, 227-235.

[3] Gould, J.D., Alfaro, L., Finn, R., Haupt, B. and Minuto, A. Reading from CRT displays can be as fast as reading from paper. Human Factors 26 (5) 1987, 497-517.

[4] Dillon, A., McKnight, C. and Richardson, J. (1988) Reading from paper versus reading from screens. The Computer Journal (in press).

[5] Pullinger, D. NOTEPAD teleconferencing for BLEND 'electronic journal'. Behaviour and Information Technology 3 (1) 1984, 13-24.