Electronic annotation for researchers: a feasibility study
Item 4: Detailed Project Description

1 Structure of this document

This description has the following form: first we discuss some overall issues; we then describe the study; finally we look at some more detailed issues, together with the outputs from the project.

2 Background and hypotheses

Annotation of paper documents has been an invaluable aid to researchers across all disciplines. *Our hypothesis* is that electronic annotation systems have the potential to do the same for documents read from a computer screen. Electronic annotation systems are currently relatively crude and weak. We believe there is potential for much better systems, and particularly for systems that exploit the electronic medium to do what cannot be done with paper annotations.

We hope there are parallels with the success of hypertext. Hypertext has been a computer-based technology that has had a huge impact, primarily via the World Wide Web, on almost all computer users. We believe that electronic annotation has similar potential. Indeed previous studies by Brown[2][3] have shown great similarities between hypertext and annotation: each involves linking one document to another (where a single annotation is regarded as a 'document' linked to the document that is annotated).

Our *secondary hypothesis* is that, although annotation practices differ across disciplines, there are likely to be common underlying principles. Thus there are opportunities for general electronic annotation systems, rather than ones geared to particular types of usage.

To test our primary hypothesis we will perform a feasibility study that covers a range of disciplines, and looks at researchers' current practices and future needs. As a result of this study, we hope to identify underlying principles, to support our second hypothesis. We will then formulate the nature of a possible general electronic annotation system, based on these principles, and usable across disciplines.

3 Concepts

Our project is concerned with documents. With electronic media we take an all-encompassing view of what a document is. Thus it could be a web page, a Word document, a spreadsheet, or the results of a database query. Documents may be textual or multimedia.

We also take a wide view of what an annotation is. The purpose of an annotation is to encapsulate the annotated document and the user's views on it. Some existing software mechanisms can be regarded as a special case of annotation: for example adding a document to a Favourites list is equivalent to adding to the document the annotation "I wish to read this often". We hope that taking a broad view will enable us to understand underlying concepts better.

We regard each annotation as itself a document: ideally an annotation is a 'first-class citizen' that can itself use any feature normally found in electronic documents, e.g. hypertext linking. An

annotation may be produced by the user (currently the normal case) or automatically produced or enhanced by the computer[9].

Ideally an annotation could also be attached to a set of operations rather than to a document. Thus if the user firstly selected menu item A, secondly menu item B, and thirdly typed something, then this set of three operations might be given the annotation "How to reset page size".

4 Individual and shared annotation

A lot of the existing work on electronic annotation, such as Annotea[5] from the WWW consortium, has been aimed towards shared annotations: many users simultaneously attaching annotations to the same document. In its extreme this form of annotation becomes a discussion group anchored on the annotated document. We are somewhat sceptical of such systems: the success-rate of collaborative systems, such as collaborative authoring systems, has been poor[7], mainly because such systems are heavyweight in that they require special servers, control, extra security, synchronisation mechanisms, etc. Marshall [6], in a famous paper, has found that shared annotations are quite different from personal ones. For example our personal annotations are related just to us, can be in a notation only we understand, can be done informally and quickly, and can record our true opinions ('This is rubbish').

The focus of our project is on individual annotation. There is also a focus on long-term use of annotation: a researcher may annotate a paper, and may wish to look again at the annotated paper many years in the future.

5 The study

We will now describe the form of the feasibility study to identify annotation needs. We will start by outlining the nature of the participants.

Clearly a university is an ideal environment to investigate the habits and needs of researchers. Nevertheless plenty of researchers operate in non-university environments, and perhaps have different needs (e.g. our past links with Pfizer showed a need for recording formally every document that has been read, in order to safeguard against future court cases that claimed negligence in not looking at past literature on possible side-effects). Our study will cover university work, in the University of Exeter, plus outside organisations.

We have done some exploratory work within the University of Exeter, and have identified a number of subject areas that are enthusiastic about participating. We mention three areas here. The first is Pallas, a team for Humanities Computing. A particularly interesting project for our study is their electronic archive of Dartmoor life, which involves the community as well as academic researchers. Thus it would be useful for members of the community to add annotations that shed light on materials within the archive. A second area is bioinformatics, in which Prof. Narayanan is a leader: here researchers look at massive amounts of diverse data/documents, and often need to go back later to re-look at information that previously appeared to be of limited interest. Clearly annotation can help in tracing information, and capturing the researcher's view at the time it is read. Bioinformatics also has relevance to the Social Sciences: the University has a 2.2 million pound ESRC grant to fund a multi-disciplinary centre for Genomics in Society. This Centre is in fact part of our third area, which is the School of Historical, Political and Sociological Studies (SHiPPS). SHiPPS has several projects that span History and the Social Sciences.

We have also had initial discussions with three of the provisional participants of our study outside the University. The first is the Devon Record Office, which is largely paper-based and is much used by "amateur" researchers, both young and old. The Devon Record Office offers a further dimension in that it looks after the Exeter Cathedral Archive, which attracts different researchers. The second is the Met. Office, and its associated large research institute concerned with climate change. The Met. Office has a extensive body of experience on the habits of researchers, covering

a wide range of disciplines. The third is an overseas institution, the University of Würzberg, with which Derek Lewis has links. The University has particular expertise in parallel corpora — an area that adds a new dimension to annotation. We have no prior reason to think that overseas researchers have different habits from UK ones – indeed research culture is international. Nevertheless we think it useful to involve an overseas institution in the study, particularly as (a) its work is of relevance to us, and (b) no excessive travel costs are needed.

Finally we plan to exploit existing links with Southampton University, where Professor Brown is Visiting Professor, particularly during the later stages of our project when conclusions are being formulated.

5.1 Methodology of the study

The first month of the project will be spent defining the methodology for the study. Our ideas at the moment are that the study should *not* be based round a fixed questionnaire (yielding results such as "74% said XXX"). Instead when we interview researchers we should find out their current practices: what sort of documents they deal with, how they annotate and organise paper documents; how they currently annotate electronic documents (if they do); how they currently do tasks that electronic annotation could perform; what future aids might help their use of electronic documents; how long their projects are and the extent of collaboration. We will be particularly interested in the unusual, such as the use of extremely dense or nested annotations in certain disciplines. The results of each interview will initially be recorded as a set of notes, and these notes will subsequently be collated in summary documents.

During the first month we will create a proforma, based on the above ideas, that records issues that interviews should cover. We will also conduct some initial interviews to see how well the proforma works, and will revise where necessary. It is also vital that data protection issues are resolved in this first month.

The research assistant will start in the second month, and will conduct most of the subsequent interviews, with help from Prof. Brown. As the interviews proceed, we will gradually compile the final report, which captures what we have learned.

6 Detailed discussion

In the following sections we give a detailed discussion of some of the technical issues relating to annotation. Clearly we hope our study will throw more light on these issues.

6.1 Chronology

Chronology is particularly important in Humanities subjects and in Law. In our case, the issues are chronology of underlying documents and chronology of annotations. The former is known to be a difficult issue. As a web page containing a Jane Austen extract illustrates, the key date is *not* when the web page was produced, but when the underlying material was written. Recording chronology of annotations is, however, an easier problem, but we need to identify the exact needs of users as regards both recording and presentation.

6.2 Future hardware solutions

An ideal is to integrate the paper and electronic worlds so that everything done on paper is automatically captured electronically. The increased cheapness of cameras and projectors might bring this ideal closer. In fact, Professor Brown worked at Xerox Cambridge when the Digital Desk[10] was pioneered: this allows the user to work exactly as they would with paper, but captures electronically everything the user does, and integrates it with electronic versions of the

paper material. A more recent and cheaper development is the Anoto Pen[1], which allows the user to annotate (specially prepared) paper, and again offers digital capture.

Let us hope these hardware developments eventually lead to universally applicable devices. If so, some of the software we propose will be overtaken by events: however the underlying principles will, we trust, not be. Clearly these are questions for our final report.

6.3 Retrieval and catering for change

Retrieval is at the core of our project, as there is no point in preserving annotations unless they can be effectively retrieved. If a researcher uses an electronic annotation system for all the documents they read, then there may be tens of thousands of documents and annotations to be retrieved from, even for a single researcher.

There must, therefore, be effective ways of retrieving past annotations even when the required material represents a needle in a haystack. We need to look at researchers' retrieval needs: is retrieval by content likely to be adequate, or does each annotation need to be augmented by extra material (metadata) to aid subsequent retrieval? If metadata is needed, could this be in the form of a data type, chosen by the user, to record the nature of each annotation, or could some metadata be captured automatically (see Use of context, below)?

We also need to identify users' requirements on how annotated material, when it is retrieved, is displayed. For example one possible approach is Fluid Annotations[11], whereby the body of an annotation can be inserted into the document in-line, thus making the document expand. Is this better, and more suited to electronic media, than putting annotations in the margin?

Catering for change is a central issue in any IT project, and applies ever more strongly to retrieval as material ages. An extreme, but highly desirable, software system that might arise from our project might be support for 'Annotation for life': a researcher captures annotations throughout their working life and can retrieve any of them at any time. Clearly such software would be threatened by change in the content or URL of web pages, by incompatibilities between succeeding software versions, by licensing issues, by obsolete software/hardware, etc. Some good past research has made annotations more robust over change[8], but this is just a palliative.

Our study needs to look at how long users want/need to preserve their annotations, and whether their expectations are realistic.

6.4 Usability

Usability is a potential Achilles heel of electronic annotation systems. This is particularly true of the *creation* of annotations, especially when such annotations are short (e.g. underlining a phrase and putting a mark in the margin). Short annotations are incredible easy to create on paper documents, and *can be done so quickly that the user's flow in reading the document is not interrupted*. In the electronic world annotations will always be harder and slower to create.

Such matters will not be a primary focus in our study of existing use of annotation — people cannot really judge the usability of some hypothetical electronic annotation software to replace existing approaches. These matters must, however, be covered in our report on potential software solutions.

6.5 Use of context — looking beyond current practices

The work of Rhodes and Maes[9] on *Just-in Time* (JIT) retrieval has pioneered new thinking in two respects.

- firstly, it introduces the idea of annotations automatically produced by the computer to help the user with what they are currently reading or writing. In essence the computer acts as a annotating agent that automatically personalises the user's current document. JIT annotations take account of the user's *context*. In [9] this was mostly the user's computing

context, e.g. their past e-mails, their bibliographies, their history of web browsing. Thus a JIT annotation might provide a link between (1) a name in the current document, and (2) citations of the named person's work or e-mails received from that person. More generally the context could include physical aspects such as the user's current location: then a JIT annotation on a list of restaurants might be 'This is the nearest one'.

- secondly, it brings together the world of context-aware retrieval with the world of document manipulation. This was especially pleasing to Prof. Brown since these two areas are his main research interests – they were previously thought disparate. In particular his work on context-aware retrieval (for which he is most grateful for past Leverhulme support) is relevant to this project.

Thus the strengths of JIT are (1) that it offers new types of annotation, only possible in an electronic environment rather than a paper one, and (2) that, by introducing context, it offers new potential for retrieving annotations (e.g. 'Retrieve the annotations about statins that I made while working at the Glaxo Laboratory').

These two examples relate to an overall aim of the current project: *to find new and valuable annotation methods that become available in an electronic environment.*

7 Analysis and evaluation

As the project proceeds we will analyse the results of the interviews, and will hope to identify common principles. Hopefully these principles will be sufficiently comprehensive that we can call the whole an *annotation theory.* We will then sketch the design of software that is an electronic annotation aid to all researchers. Obviously we hope that our research will spur the building of this software, or, at first, limited versions of it. During our design work we will look as existing commercial systems such as iMarkup[4]. Indeed we hope to influence the future development of these systems.

Following on from this software design, we will create 'proof of concept' demonstrators. Given the limited time available, these may well be limited to simulated screen shots. We will get feedback on these demonstrators by re-visiting some of our interviewees, and asking for their evaluation. This evaluation will doubtless lead to some revision of our ideas.

Our *final report* will cover (a) the results of the interviews; (b) the software design; (c) an evaluation of the original hypotheses.

8 Publication

Initial dissemination will be via a web site we will create for the project. We will encourage outside participation in this site. We already have experience, on previous projects, of building participatory web sites.

We plan to publish journal/conference papers on our work. A number of our contacts have indicated interest well beyond just participating in interviews: we greatly hope that this will lead to joint papers in journals, e.g. in journals devoted to computing and the humanities.

References

[1] *The Anoto pen*, `http://www.anoto.com`, 2005.

[2] Brown, P. J. and Brown, H., 'Is annotation a form of hypertext linking?, `http://www.dcs.ex.ac.uk/~pjbrown/annotation/doceng.pdf`, 2002.

[3] Brown, P. J.. 'From information retrieval to hypertext linking'. *New Review of Hypermedia and Multimedia*, Vol. 8, pp. 231-255, 2002.

[4] *iMarkup: annotate, organize and collaborate on the web*, `http://www.imarkup.com/products/annotate_page.asp`, 2003.

[5] Kahan, J., Koivunen, M-R., Prud'Hommeaux, E. and Swick, R. R., 'Annotea: an open RDF infrastructure for shared web annotations', *Proc. WWW10 International Conference*, Hong Kong, pp. 623-632, May 2001.

[6] Marshall, C., 'Toward an ecology of hypertext annotation', *Proc. ACM Hypertext '98*, Pittsburgh, Pa., pp. 40-49, 1998.

[7] Nöel, S. and Robert, J-M., 'Empirical study on collaborative writing: what do authors do, use, and like', *Computer Supported Collaborative Work*, **4**, 1, pp. 63-89, 2004.

[8] Röscheisen, M., Morgensen, C. and Winograd, T., 'Interactive design for shared World-Wide Web annotations', *Proc. CHI '95*, Denver, Co., volume 2, pp. 328-329, 1995.

[9] Rhodes, B.J. and Maes, P., 'Just-in-time information retrieval agents', *IBM Systems Journal*, **39**, 4, pp. 685-704, 2000.

[10] Wellner, P., 'Interacting with the digital desk', *Comm. ACM*, **36**, 9, pp. 87-96, 1993.

[11] Zellweger, P., Bouvin, N. O., Jehoej, H. and Mackinlay, J. D., 'Fluid annotations in an open world', *Proc. 12th ACM Hypertext Conference*, pp. 9-18, 2001.