3-16-2004

Reports of the DAS02 Working Groups

Elisa Barney Smith
Boise State University

David Monn
Center for Communications Research

Harsha Veeramachaneni
Rensselaer Polytechnic Institute

Koichi Kise
Osaka Prefecture University

Alessio Malizia
University "La Sapienza" of Rome

See next page for additional authors

Document Analysis: DAS'02 Working Group on Digital Libraries and Document Image Analysis of Antique Documents Report

Elisa Barney Smith[1] and David Monn[2]

[1] Department of Electrical and Computer Engineering
Boise State University, Boise, Idaho 83725-2075 USA
EBarneySmith@boisestate.edu,
[2] Center for Communications Research
Princeton, NJ, USA
monn@idaccr.org

This report summarizes the discussions of the Working Group on the Digital Libraries and Document Analysis of Antique Documents of the IAPR Workshop on Document Analysis Systems, Princeton, NJ, 19-21 August 2002. Eight researchers from two countries participated: B. Agüera y Arcas, H. Baird, E. Barney Smith, A. Dengel, D. Lopresti, D. Monn, J. Uchill, L. Vincent. The participants represented a mixture of both private industry and universities. David Monn moderated the discussion, and Elisa Barney Smith served as scribe.

The working group recognized that there are already a number of well-known digital librarires available online today, including the Making of America Collection [2], the U.S. Library of Congress [4], and other specialized collections [3]. Still, despite the obvious potential synergies between document analysis research and digital libraries, there has not been much interaction between the two communities. Digital libraries are typically built using off-the-shelf commercial OCR systems, oblivious to the more advanced document analysis techniques under development in our field. On the other hand, most document analysis researchers are not aware of the special problems that arise when building digital libraries, nor do they regard the vast collections of scanned document images now accessible on the Web as a resource that could be invaluable in their work.

The group first identified what it felt were the challenges facing Digital Libraries. We then discussed several of the features we thought would be good for Digital Libraries to have. The remaining time was spent discussing what our community could provide libraries and institutions who are trying to create digital libraries.

1 Challenges of Digital Libraries

Digital libraries are emerging as a supplement to traditional libraries. Still their growth is in its beginning stages. Two goals, in particular, are of concern in constructing libraries of textual material. The first is providing digital images of sufficient quality for use by those who wish to view the documents in their

original form, whether it be for reading, or for examining features such as the printing style, the text layout, marginalia, or non-textual elements such as pictures and graphs. The second is providing an accurate transcription of the text, not only for searching, but also for ease of reading and printing when content is the main focus, and not necessarily how it originally appeared. This second goal usually depends on the first to the extent that the scanned images must be of sufficient quality to achieve high acuracy in the transcription by an OCR engine.

In order for libraries to meet these goals efficiently when converting significant portions of their collections to digital form, the current processes could likely benefit from increased automation. For example, the scanning process alone currently requires a significant amount of clerical support, such as identifying poorly scanned pages, and rotating upside-down text. Certain documents, especially those that are very old, may require specialized imaging techniques.

Since it is often desirable to be able to read the transcription of a document, there are several issues that must be addressed beyond simply having very high accuracy. Determining the proper reading order, for both columns and footnotes poses an enormous challenge. Understanding where in the text each footnote is referenced may require the recognition of special characters (daggers, etc.) that may not even be part of the OCR character set. Some documents may also contain other special characters, such as diacritical markers or section and paragraph symbols (pilcrows). Once one has an understanding of the footnotes and references, and where they occur, there is the issue of how to present them to the reader. Perhaps web-based libraries of the future will be able to provide hyperlinked references and pop-up windows containing the footnotes.

The OCR of mathematical equations continues to be a problem, both in recognizing the symbols and interpreting the equations for proper viewing and printing. The same is true for understanding tables. In technical documents equations, tables, and figures are often numbered and referenced elsewhere. Recognizing these labels and matching them up with the references requires a special understanding.

A lack of funding is the most commonly cited reason why there aren't more Digital Library projects under way currently. New books that are published through digital technology aren't immediately contributed to Digital Libraries, even though the major portion of the cost, which is doing the digitization, would not be incurred. Some of the reason for this was attributed to copyright issues, and worries that publishing online would decrease hardcopy sales.

Some publishers have found that simultaneously publishing books in both paper and electronic form has actually lead to an increase in the sales of the paper version. Amazon.com has many sample pages of books available on the web now as a tool to increase their sales. Many workshop and conference proceedings, including the proceedings of this conference are also available electronically [1].

There are several very nice Digitial Library projects currently under way, but they are not interconnected. Should they be? What effect would a large centralized Digital Library have on smaller Digital Library projects or smaller paper libraries? What should the architecture of a global digital library be?

2 What features would the DIA community like to see?

There were many additions that were felt would increase the value of a Digital Library. It seems possible that advanced document analysis techniques could open up new options for the delivery of content to users, thereby increasing the perceived value of the information. Decisions on how to best deliver the content to the users need to be made. Configuring digital libraries so that the contents can be easily viewed on a PDA was the first feature working group members suggested be added to future Digital Libraries. Having multiple layers in the document to represent the image, the OCRed text, hyperlinks, highlights, notes, etc. would also add value to the library.

To make the content of greater use, the libraries must be easily searchable. The search engine can focus on the text data, the interpreted content represented by that text or the structure of the document. Members agreed that integrating these would enable Digital Libraries to go beyond a simple search.

It was also discussed that when a digital library is created it would benefit our community and other users if the meta information on the collection was included. It was felt that most DIA researchers were not aware of all existing Digital Library corpuses, or how much of what types of data is in each one. This information could make the digital libraries a useful source of data for DIA research.

The concept of personal Digital Libraries arose. Are tools are out there or would it be reasonable to develop tools such that people could create their own personal Digital Library from their own resources? Scanner hardware that was less intrusive to personal users such that people could digitize a document page by page as they read it, perhaps a digital camera mounted on eyeglass frames was one suggested idea. The availability of scanning hardware that was less harmful to antique books was another issue brought forth.

3 What could the DIA community provide?

Members agreed that the DIA community can use its experience to make recommendations to libraries as they begin a digitization project. One topic of interest to members was whether libraries were using the best scanning resolution when doing their digitization. The resolution need depends on the type of input document, particularly where antique documents are concerned. Older documents shouldn't be rescanned often so starting with the correct resolution image is important. For these documents the details about the printing and paper is often as important as the textual content making higer resolution scans more important than for a recent publication. Some guidelines were suggested by the group that documents printed before 1600 should be scanned at 2000 dpi, documents from 1600-1800 should be scanned at 600 dpi and documents printed more recently than 1800 should be scanned at 400 dpi. These were all heuristics, and developing some better reasonings for these guidelines is a possible direction for growth. The decision on whether to scan in color versus grey scale versus bilevel was

also touched upon as an area where our community can help guide the libraries involved in these projects.

When the digitized document is viewed it needs to be decided whether to keep the original digital image or just the converted OCRed text. Enhancement of the original images might be necessary for improved legibility, or to improve OCR. These two criteria are not always equivalent. It was agreed that when enhancement is done, the original image should still be saved for future use.

It was mentioned that the value of the library could be increased by expanding it through annotations, corrections of the OCR layer, addition of knowledge, etc. and that the users of a digital library could contribute to this if the proper framework was developed.

The discussion continued on to what software our community can provide to libraries. Should we encourage our software to be embedded on their Digital Library site or make the software external to the site in the same vein as a search engine, like Google, is external to other websites, but can be used to search a local site?

4 Summary

The working group concluded that Digital Libraries were of interest to DIA researchers and members looked forward to the increase in size and number of Digital Libraries. The ending thoughts were: What DIA technology is applicable to what aspects of the problem? We have developed many useful tools as parts of our research, but integrating them and making them available to other researchers and implementers of Digital Libraries could be a place for improvement. What fundamental technologies should we be focusing on to help Digital Libraries expand?

References

1. 2002 ICPR Workshop on Document Analysis Systems, August 2002. http://link.springer.de/link/service/series/0558/tocs/t2423.htm.
2. Cornell University Library: the Making of America Collection. http://moa.cit.cornell.edu/.
3. Princeton University Library Papyrus Home Page. http://www.princeton.edu/papyrus/.
4. U.S. Library of Congress: Digital Library Initiatives. http://memory.loc.gov/ammem/dli2/index.html.