The C-ODA Project - On-Line Access to Electronic Journals

Peter Kirstein <kirstein@cs.ucl.ac.uk>
Goli Montasser-Kohsari <gmontass@cs.ucl.ac.uk>

Abstract

In this paper, we discuss the UCL C-ODA project, working with a large database of journal articles of chemical journals in several compound document forms (text/image). As part of the project, we had to set up a database comprising approximately 500,000 pages of technical papers in a mixture of SGML, ODA and bit-map representations; our experiences in, setting up this database and organising its searching are described.

We provide a number of interfaces to access that data over various forms of network - LAN, the Internet and the ISDN in particular. Our users' experiences are presented in working with the data.

1. Introduction

Electronic access to bodies of information like journal articles is a comparatively new facility; it raises many different areas of concern. Some examples are the following:

- The technology used in data preparation;
- The formats used in data storage;
- The technology used in data searching;
- The technology used in data retrieval;
- The facilities the users desire or require;
- The legitimate concerns of the publishers and the service providers to maintain an acceptable revenue stream;
- The legal aspects of the provision of appropriate electronic systems.

It is possible to discuss these problems at length from an abstract viewpoint. It is easy to mount small experiments with small sets of user and collections of documents; these are invaluable to establish the size of database which might arise from a large collection, the legibility of different forms of data presentation, or even the ease with which users can absorb information electronically. However, to make a real impact into the potential and utility of electronic access to journal articles, it is essential to do pilot projects with large bodies of information. There is a considerable body of recent work in this area; mostly this is concerned with specialised document collections, which have been put together for other purposes or even specifically for studying the problems involved. An example of this is the Dienst system [1] from Cornell U, Department of Computer Science. The Dienst project is part of a whole programme funded by ARPA; a number of universities are involved, and most of the system is based around documents for which there are no copyright problems, and where the collection was built up in digital form. That project is designed to access a distributed document collection, and is a logical follow-on from the project we describe here. That project is developing well the distributed search facilities, and makes heavy use of the World Wide Web (WWW) [2]. This has the advantage of not incurring the problems we describe below in building up the database; it cannot give, however, the sort of experience of access to conventional journals that we are describing in this paper. There are other projects attempting to provide access to journal collections; these have normally been restricted to work with scanned images of the journals (e.g. the Red Sage project [3], [4]). This has been partly because this is the only form in which the data could be assembled, and partly because that is the only form that the publishers were willing to supply.

In this paper, we discuss a quite difference sort of activity. The C-ODA project is a much older project than Dienst, having started in 1990. It built up, and experimented with, a large journal collection - providing access to a large body of recent articles of the American Chemical Society. In Section 2 we provide an overview of the C-ODA project, and of a similar US project called CORE; we also discuss some of the differences between the two projects. In Section 3, we summarise the publication chain, and present a very simplistic picture of the different steps which arise in accessing documents. Fundamental to such projects is

the data used; in Section 4 we discuss the different data sets required, and mention the retrieval methods used. An important aspect of the interchange of documents between different systems is the choice of data representation; two popular formats are the Standard General Mark-up Language (SGML [5]) and the Office Document Architecture (ODA [6]). SGML is a tagging language, which is used to mark up documents according to a set of Document Type Definition (DTD); these DTDs can be defined by each publisher, though there is an attempt to use common DTDs in different areas. SGML has no concept of presentation; this is provided by various other mechanisms. ODA has an intrinsic structure and layout syntax and semantics. Both these formats are used in the C-ODA project; in Section 5, we give a brief introduction to SGML and ODA, discuss the different data formats used, describe how and why they were employed, present some of the differences between them, and comment on the problems in converting from one to the other. While we tried to use as much of the work undertaken in the CORE project as possible, there were many activities required to set up the databases and provide the appropriate access to them; this activity is described in Section 6.

Even with the modest quantity of data used in the C-ODA project, there are still tens of GB involved; Section 7 discusses some of the considerations in the storage of the data, its compression and speed of access. Many different User Interfaces could be provided; those used in the C-ODA project are described in Section 8; here we consider also the mechanisms which are used for information retrieval. In Section 9 we present some user experiences; these could only be derived from such a full size pilot. During the preparations for the pilot, we realised the need to provide security features in order to persuade the publishers to allow us to use their data in large-scale pilots. The security facilities were developed, as is described in Section 10, though they were not deployed in the pilots. Finally, in Section 11, we draw conclusions both from our experience in setting up the pilots and the users' reactions to it.

2. Overview of the C-ODA Project

The American Chemical Society (ACS), Bellcore, Chemical Abstracts Service (CAS), Cornell University and OCLC collaborated in the CORE project [7] to deliver electronic information from primary publications to end-user chemists. Although this project originated from the same institution as the Dienst project [1], they came from different parts of Cornell U, and, as far as we can judge, had no relationship with each other. The period of the journals covered by both the CORE and the C-ODA pilots was largely dictated by the availability of data in electronic form. Subsequently to 1980, the ACS had increasingly moved over to producing their text electronically - though mainly in a proprietary format; the ACS made available their typesetting tapes to Bellcore for the purposes of the pilot. Prior to 1995, the diagrammatic material had been included in the production process by "cut and paste"; thus that data was not available electronically from the ACS. As part of this experiment Bellcore scanned most of the pages of ACS journals published between 1989 and 1994 - some several times. They processed the typesetting tapes from the same journal issues into a SGML [5] format so that it may be indexed and used in the pilot. By the end of the project, they were providing electronic access to a large electronic database containing approximately 60,000 articles, representing 400,000 pages of journal articles of the American Chemical Society (ACS) for the period 1989-94. The data was held at the Cornell U Mann Library for access by Cornell chemists.

In the Computer Science department of University College London (UCL-CS), we were involved with members of the CORE project since 1988. This activity relied heavily on the work of Bellcore, and used the data provided by the ACS. It was supported by the British Library Research and Development Department (BLRDD). While we provided facilities similar to the CORE project, we were also interested in applications for the data which the CORE project had not been focused towards or in a position to support. The UCL activity is referred to as the C-ODA project, and covers also areas such as applications of ISO standards, and usage of relatively low-bandwidth networks such as the ISDN. The paper discusses the way the database was set up - which involved conversion from a SGML representation into an ODA one [6], the methods of indexing, the access methods provided, and our user experience. We discuss also the motivation of many of our implementation choices, and the lessons to be learned from our experiences..

The ACS consented to allow the data to be used for these projects, with certain restrictions on distribution - mainly that the data will not be available outside Cornell U for the CORE project, and outside

the University of London for C-ODA. There were differences between the CORE and C-ODA projects [8] but these are not discussed in this paper.

This project started in 1991, when UCL-CS was heavily involved with ESPRIT PODA projects (e.g. [9]) in the use of ODA. Originally, the CORE project used no standard language for the representation of the text, so that ODA was a natural choice for the C-ODA project. Later the ACS textual material became available in SGML form. The relative advantages of the two forms is discussed in Section 6.

3. The Publishing and Access Chain

3.1 The Chain Itself

While work with the ACS databases as processed by Bellcore were the main activity in the project, we obtained a good insight on how the publishing chain should proceed for this type of activity. The fact that it did not always do so, only made our task harder. The conventional publishing and access chain for journal articles in science and engineering is as follows:

a) Journals articles are submitted in a number of forms by the authors.

The chosen format by authors seems to be predominantly TeX or LaTeX and Postscript, but this is not always the case. Many authors like to use their favourite word processing system, and would prefer WORD or WordPerfect.

b) The article is registered by the publisher.

The article is now in the publisher's system. A lengthy pre-printing chain is then initiated. This should involve minimal reprocessing by the publishers, since it is not yet clear that the article will be published

c) The article is submitted to reviewers.

d) The reviews are returned to the publishers. Parts may be passed on to the authors also for subsequent action in revising the article.

e) The author provides a revised manuscript, in the same form as in (a).

Once the cycle of (e), (c) and (d) has been completed, the article is ready for printing.

f) The articles can be translated into a mark-up language (e.g. SGML with a specific DTD) for typesetting, and printing.

In the way the ACS has produced its journals up to the end of 1994, the diagrams were then stuck onto the Masters before printing. This meant that the typesetting tapes did not include the diagrams; they did include equations and tables. For a full electronic form, the figures must also be provided electronically. Currently the ACS scan such material at high density from the source material, and combine the material electronically with the textual material. This was not done with the data used in the pilot.

In most cases, the article is then returned to the author for proof-reading. This may well involve hand-written revisions, which must be included by the publisher. This cycle is avoided if the submission is in *camera-ready form*. The article is now ready for combining with other articles and being published.

g) The article is combined with other articles, and an issue is prepared. This includes provision of a Table of Contents.

h) For normal publishing, the article is then printed. For an electronic publishing chain, instead of being printed, the data is converted into a form which is suitable as a distribution format, and then sent to the 'electronic library' organisations.

i) The issue, and its component articles, are registered in secondary publications. These may include only articles name and reference; they may also include abstract and keywords.

The form suitable for electronic distribution may be quite different from that desired by the publishers for the article preparation phase of (a) - (f). This is discussed further in Section 5. Once the article has been published, it is necessary to enter the reader access phase.

j) The researcher searches for a specific article, or for the an article about the subject. The search may be by looking at the Table of Contents of the journal itself, or by querying one of the secondary publications of (i). It may even be possible to search electronically the journal article itself.

k) The search suggests that a particular article, or part of an articles, be consulted. This article may then be requested (electronically or by some other means).

l) The article is delivered to the researcher - again either in hard-copy form or electronically.

After perusing the article, the researcher is either satisfied, or continues the cycle of (j) - (l).

3.2 The Different Parties' Interests

An organisation like the ACS is particularly concerned with the automation of the process in the preparation of the article; that means the steps (b) - (g). The end-user is most concerned with the steps (j) - (l). The actual publishing itself is step (h); clearly all aspects of this step concern the publisher. For publishing via paper or CD-ROM, it is then necessarily to continue also the whole distribution chain. Some publishers, e.g. the ACS, produced also the secondary publications of (i); in the case of the ACS this is Chemical Abstracts. For electronic publishing, it may be that the electronic document store is actually provided by another *Value Added Service Provider*, though many publishers plan to provide this service themselves.

In the CORE and C-ODA projects, there was no question of steps (a) - (e); the articles had been published already. This meant that steps (g) - (i) had been completed in some form. However, the form of the original publishing had been on paper. It was thus necessary to completely redo step (f), to provide a new form of the complete issues in electronic form. The Table of Contents of (g) was part of the new step (f). For some versions of the activity, OCLC or Chemical Abstracts provided the secondary publications (i); for others, we did searching on the original articles, so that a formal secondary publication was not required.

In the CODA and CORE projects, the main role of the ACS was to provide the typesetting tapes and the journals themselves - and to permit it to be used for the Pilots. Normally the provision of the pre-printed form of the data (step f) would be done by the publishers; in these projects this was done mainly by Bellcore from the combination of scanning the original journals, extracting the material not on the typesetting tapes, and combining it with the typesetting information. The actual document stores themselves were held in the user premises - Cornell U and UCL. This in itself may be different in future such projects, if the publishers provided the Value Added Services.. Both OCLC and Chemical Abstracts (a division of the ACS) provided some of the secondary services: the Chemical Abstracts data was provided, and the OCLC Newton search engine [10] was used in the CORE project by Cornell U - though not by UCL.

4. The Different Data Sets and their Retrieval

4.1 The Source Data.

The ACS has been preserving the typesetting tapes of all their journals for the past 10 years, and for up to 17 years for some particular journals. The typesetting tapes contain all the textual information of the journals, including highlighting, equations, and tables, and also a large amount of contextual information. This contextual information includes what we may describe *as Document Management Attributes* (DMAs), and also some of the structural information of the articles. The historical typesetting tapes do not contain, however, any of the graphical images, or any layout or presentation information. Bellcore derived the graphic images by scanning the microfilm copies of the published journals and using custom OCR techniques to identify page components such as figures, tables and schemas (captionless figures) since no other record of the images is available [11]. Moreover the formats even of the text material has changed in

the last 15 years; for this reason, while some of the text data does indeed go back to 1980, that before 1989 is incomplete - only 3500 of the articles prior to that date are in electronic form.

The initial format of the typesetting tapes was not SGML, but a proprietary scheme encoded in an IBM database format. This was converted into SGML (SI-D) by Bellcore as part of the CORE project, and we ignore the existence of this earlier form of data in the rest of this paper. They passed the SGML versions of the documents on to us (along with the scanned image components), with the permission of the ACS. This is in a special DTD used only for this data, but based on the American Association of Physics (AAP) DTD. We gratefully acknowledge this assistance from Bellcore and the ACS. The size of the full SGML database for all the journals of the ACS from 1988 - 1994 is about 6 GB.

In practice, the tables and equations were not translated from the typesetting tapes. Instead the graphics, tables and equations were derived from the scanned page images in bit-mapped form (EI-D). The image extraction was done originally by Bellcore. The image extraction was very error-prone, and when the algorithms had been refined, the final version of EI-D was derived by Cornell U When this process had been completed, there were two data sets - the one representing the text in SGML, and the one representing figures, tables and equations. The extracted graphics activity was quite error-prone; a 95% success rate at finding figures was considered reasonably good. There has been a lot of problems in extracting the graphics, and the processing had to be done several times. We have received mainly the 1989-94 data; the size of the database for the extracted graphics for this period is 2 GB.

The CORE project did not work only with the text/image form of the data; they were also interested in providing the full image data to their users. The size of that database is critically dependent on the resolution of the scanning; at the 300 dpi eventually adopted, the size of this database for the period 1991-94 period is 55 GB.

To the CORE project, The ACS Chemical Abstracts data (CAS) was provided, together with the Newton Search Engine from OCLC. OCLC also indexed the textual data for use with its Newton Search Engine, and provided the indexed database to Cornell U.

4.2 Databases for Electronic Access

There are many forms of database which could be provided for this type of data. Many organisations provide Abstract Services - often searchable electronically; when these are used the actual journal articles can be requested (often electronically) and delivered in paper form or even by facsimile. Some organisations provide facilities for full text search. The ACS has been providing this for some of its journals; the journal articles are then still delivered in paper or facsimile form. Both the CORE and CODA projects wished to provide electronically not only the facilities for search, but also the documents themselves - in text and image form. For this reason, it was necessary to provide three forms of database:

a) The text portion of the journal articles - in a form suitable for user access;

b) The image portion of the journal articles - in a form suitable for user access;

c) Any part of the database representing the original journal articles not contained in (b) - but linked to it, in a form for user access (Examples are notes and experimental results, submitted but not printed);

d) An index database of the text data - in a form suitable for electronic search - and pointing to the article itself.

The format of the typesetting tapes need not be the same as (a) or even (b).

4.3 *The Text/Image Data*

The text portion of the journal articles themselves were provided in several forms - many discussed in [8]. The basic raw text data was received in SGML format from Bellcore (SG-D). The extracted image data (EI-D), mentioned in Section 3.1, were used for the tables, equations and figures. In fact the tables and equations were available on the original ACS typesetting tapes, but none of the parties had the effort to convert the equations and tables into the SGML form - though this is what should really have been done.

For some purposes we intended using the full scan images, together with a an appropriate retrieval mechanism; hence we needed a full database of these images (SI-D), which we received from Bellcore. Finally, there was a proprietary Hypertext/image system called SuperBook [12] from Bellcore, so that all the data for one year was provided also in this form (BK-D). We had the image data for 1988-94. However, we have made available to our users only .the articles from 1991-94 in the full image form in SI-D. Partly this is because we are not particularly interested in pursuing this approach - because we want to provide also remote access (via lower speed networks), and do not like the amount of storage this form of presentation requires. For the same reason, the version of BK-D to which we gave our users access was only xx GB - the data for 1993.

Our main user activity was with the whole database in the form of the Open Document Architecture (ODA) [6] form. The comparisons between ODA and SGML are given in [8] and Section 6; of some importance is that ODA has a unitary data content which includes text and image, so that the image portion with extracted graphics could be encompassed in the same database (ODA-D). On the other hand its representation of text is 8 bit and contains font information; hence it is not suitable for textual search by WAIS. Thus when the ODA representation is used for the journal contents, another form - e.g. SGML - is still required for the textual search.

We deliberately wished to experiment with access of journal data by use of the WWW; hence we developed a simple SGML-HTML converter, and .transformed some of the articles from SG-D into HTML (100 articles, HTML-D). ACS wished to keep tight control of the data, and we did want to put stringent access controls onto our WWW servers, so that we were constrained to put only 100 articles from 1993 in this form.

Finally, we worked with two voluminous databases directly from the CORE project, which have been mentioned above - EI-D and SI-D. Because of their size, we decided to locate them on the Juke box (JB). An additional database was required (IM-DB) giving the location of this data on the JB.

4.4 The Index Databases

Several index databases derived from the source data were used in the C-ODA project. We were deliberately experimenting with various forms of data representation - and hence of indexing data and retrieval mechanisms. Some of these retrieval mechanisms were developed at UCL, some was provided by the CORE partners.

From the SGML form of the data, both we and the CORE project have derived field-indexed versions of the text data which can be searched very fast. UCL decided to provide full text search using the text searching package WAIS [13]. WAIS is the Wide Area Information Server tool developed and placed in the public domain by Thinking Machines Corp, and now being developed further as a commercial product by WAIS Inc. WAIS provides tools for full-text indexing of different types of data, and allowing that index to be queried by a remote machine. It is a classic client-server system with a back-end (the WAIS server) which searches an index based upon queries provided by a front end (WAISQ - WAIS Question). The WAIS server can provide both lists of documents with their 'scores' according to some query, and whole documents when a user selects a document from a list. Xwaisq is an X-based question program which is provided with the WAIS distribution. The size of this index database (Idx-oda-DB), which pointed to the data in ODA-D, is 6.1 GB.

WAIS has the capability to access documents via the WWW using WAISGATE [14]. To exercise this feature, we derived an index of the HTML data in HTML-D, using the WAIS indexing mechanism, which was called Idx-html-DB.

Finally we made available to our users the articles from 1991-94 in the full image form (SI-D) and the articles for 1993 in SuperBook form (BK-D). Both of these required index databases which had to be derived from their relevant retrieval engines - and were provided by Bellcore. The first was Idx--si-db, the second Idx-bk-db.

4.5 *Data Search*

It is usual in the library field to do bibliographic search on attributes such as author, journal and subject - and key words. In the CORE project, with its library orientation, a standard search engine called Newton [10] from OCLC was used; this has facilities for field search with Boolean Operations. This search engine was available only near the end of that project, and we felt it rather limiting in its capabilities. There is an alternate approach in which search on contents could be a free search on any bit strings - though this could be supplemented by restricting the search to specific fields.

The earlier Public Domain versions of WAIS have facilitates for free text search; they were used in the early trials. The commercial version of WAIS from WAIS Inc [14] also has the capability for field search with Boolean operations; this is the version we used on most of our trials, since it met most closely the wishes of our users. In fact there is now also a Public Domain version of this search engine (WAIS-SF) [15] with similar facilities. For our trials, we used the version of [14], which had somewhat more advanced features including the ability to use multiple processors (though only on different databases in the release we were using), and could return a larger number of hits. Using a specific version of WAIS, it was then necessary to construct the Index databases of Section 4.4 to allow the retrieval of the articles in question. The version of WAIS we used then also indicated in the retrieved articles where the search terms occurred. For a search engine like WAIS, it is necessary to match bit patterns. As discussed in Section 5.2, the form of text representation of SGML allowed such a search, that of ODA does not. For this reason, Even when the documents to be retrieved were in the ODA format (ODA-D), the searching was done on the SGML text version (SG-D) - but these were arranged to point to articles in ODA-D.

The SuperBook version [12] of the system had its information in a Hypertext format (BK-D), with its own search and retrieval engine; the indexing therefore had to be done independently, leading to an index database of Idx-bk-DB. In the case of the page image format (SI-D), it was again necessary to search via the SGML text database (SG-D), and a new index database was required (Idx-si-DB).

4.6 *Summary of Databases Used*

In summary the following databases are received, derived or indexed:

Data Received:

SG-D	Raw text data in SGML format from Bellcore
EI-D:	Extracted Image data in TIFF format for figures, tables and equations from Cornell U
SI-D:	Scan Page Image data in TIFF format from Bellcore
BK-D:	SuperBook text data in SuperBook format from Bellcore (only 1993 data)

Data derived:

ODA-D:	Combined text and image data in ODA format converted from SG-D to ODA from UCL
HTML-D:	HTML text data converted from SG-D to HTML from UCL (only 100 documents)

Indexed database:

Idx-oda-DB:	Indexed SG-D for retrieval of articles in ODA-D by WAIS Client from UCL (6.1GB)
Idx-html-DB:	Indexed HTML-D for retrieval of articles in HTML format by WWW client from UCL
Idx-bk-DB:	Index BK-D for retrieval of articles in SuperBook format by SuperBook from Bellcore
Idx-si-DB:	Main Index file for retrieval of Scan Images of pages by PixLook from Bellcore
IM-DB:	Database giving the position of SI-D records on the Juke Box by UCL

5. Document Distribution Formats

5.1 Document Distribution Needs

A distribution format should have the following properties as a minimum:

- **Presentation** It should contain presentation information sufficient to generate a pleasing image for the reader. For example It should enable titles and headers to be in larger font, and allow for typographical effects such as italicising and boldness.
- **Content** It should contain the words of the article (or possible the front matter of the article) in order to facilitate searching.
- **Viewing** Tools should be available for readers to view the system on-screen, and possibly generate hard-copy as well. These tools must be friendly, reliable, and well-supported.

Electronic Journal (EJ) delivery involves a publisher generating documents and distributing the electronic form to organisations which will pass these on to the users. For the sake of argument we will call these organisations 'electronic libraries', even though they may not be what are currently recognised as libraries. The reader of these documents will require them in one of two ways. Either they will be receiving a new issue of the EJ, in which case they will wish to inspect the table of contents, browse the articles, and/or read a number of articles in-depth. Alternatively, they will wish to search against a collection of journals, using some kind of query mechanism, and then browse or read the articles that were found. However, it is also possible that a reader may wish to browse old journals, or search in a new issue, and the user should be able to do both.

When viewing the EJ, the reader will expect that the articles be clear and contain formatting suitable for supporting the document structure. Moreover, all readers and screens are not equal, and so some method of changing the size of fonts and of diagrams in the documents would be advantageous.

Having an on-line database of scientific journals offers many advantages over the conventional paper-based journals; many of these advantages fall into the areas of search and access. It is much easier, and more productive, to search texts for information electronically, using a computer system, than manually. In our environment all the journals are indexed so that, despite the size of the database, searches are very fast. Electronic access provides additional advantages:

- It is non-exclusive - any number of people can access the same journal simultaneously.
- It is distributed, so it is not necessary to be in close proximity to the database in order to access its information.
- It can be integrated with the users' facilities, so that it is possible to extract information for other purposes - always subject, of course, to consideration of copyright and other constraints.

5.2 *The Main Characteristics of SGML and ODA*

5.2.1 The Characteristics of SGML

SGML [5] is a system of specifying intellectual mark-up for documents. The point of intellectual mark-up is that one denotes what an element represents, rather than what it looks like. The mark-up should describe a documents structure and other attributes rather than specify processing that is to be performed on it, as descriptive Mark-up need be done only once and will suffice for all future processing. For example, one would mark the title of an article with the tag <title>, rather than say 'Centred, Bold, 16pt Times Roman'. The description of <title> is then contained in a "Document Type Definition" or "DTD".

SGML uses an ASCII-based representation which has certain in-built limitations and advantages. It is not possible to embed arbitrary binary data within an SGML document, since elements are terminated by a special character sequence - and clearly that sequence is possible in arbitrary binary data. It is possible to

circumvent this using escape sequences, but there is no defined way to do this within the ISO SGML standard. The accepted method is to refer to external entities for such items.

One of the reasons SGML is well-used is because it is easy to generate and transmit the ASCII representation. On receiving an SGML file, it is possible to scrutinise it effectively using just a standard text editor. This allows it also to be searched by a text-based search engine like WAIS. Another advantage is that it is so flexible that it is possible to express the styles of the publisher by providing a specific DTD or set of DTDs.

5.2.2 The Characteristics of ODA

ODA has quite different properties from SGML. The ODA standard [6] describes an abstract view of an office document and a document processing model as well as an interchange format. A document consists of components such as the document profile, generic structures (logical and layout object classes) and specific structures (logical and layout objects), styles (layout and presentation styles) and content portions. These components give two views of the document; a logical structure which represents a logical view of the document such that a letter header consists of a date, an addressee, a subject, etc., and a layout structure which represents a layout view of the document such that a letter header page consists of a logo frame, a date frame, etc. ODA incorporate data management attributes such as author, title, date etc as Document Management Attributes

The ODA standard defines three kinds of document form; a processable form with logical structures is created after an editing process, a formatted form with layout structures is produced by a formatted process, and formatted processable form with both logical and layout structures is also produced by a formatted process. An imaging process takes a formatted or formatted processable form and produces a final document in an interchange format called ODIF (Open Document Interchange Format). All these document forms are encoded into the ODIF stream in which all components are represented as sets of attributes.

The ODA structure is defined using an ASN.1 syntax. This allows any 8-bit sequence to be used. However, font information is tied into the character representations, and binary sequences are located inside the document. Therefore it is not possible to use a standard text search engine like WAIS to identify specific words in an ODA representation.

ODA supports the functionality a 'Document Class', and also allows presentation information to be bound to the document elements. ODA has been designed primarily as a self-contained interchange format for documents. ODA is supported by commercial word-processor manufacturers, and converters are available for interchange between ODA and commercial word-processor formats. As a result, it is possible to interchange documents blindly between different word processors

5.3 The Use of SGML and ODA

ODA and SGML are suitable format for document distribution, but have different properties:

- A single ODA file can encapsulate a compound document; its distribution as ODA only requires a single file to be passed, whereas a compound document in SGML is likely to consist of a number of separate files.
- The ODA file contains enough information to render the file on screen or paper in a pleasing and meaningful manner. SGML requires that the DTD and a translation specification file be sent also.
- The viewing tools for both ODA and SGML data are of similar quality. However, the SGML viewing tools have different types of translation specification file; such a file would be needed for each viewing tool which end-users intended to use. The take-up of DSSSL [16] will remove this difficulty, but for the next year, possibly two, this will be the problem.
- Both SGML and ODA can be readily converted into a wide range of commonly-used word-processing formats; however often these converters lose the structure available in the original form. There are converters available which convert ODA or SGML to WordPerfect, Microsoft Word, Microsoft Word

for Windows, IBM DisplayWrite, DCA-RFT, and DecWrite formats. It is possible for a system which holds documents in ODA or SGML to deliver them to users in a format which they can view on their normal equipment. Moreover they can edit these documents, annotate them, or extract parts into their own documents; all within their normal document processing environment.

- The ODA format is remarkably compact. The format supports geometric graphics, and bitmaps are compressed using the Group 4 fax algorithm - an excellent lossless compression scheme or Group 3 fax algorithm or Bitmap.
- The ODA format does not suffer the ASCII-related problems with which SGML files must contend. The ODA files do not need altering when files are transferred between ASCII and EBCDIC-based machines, or between machines with different byte orders, or between ASCII-based machines with different line break characters (for example between DOS and UNIX). However, the ODA format does need a proper browser to be read.
- Both ODA and SGML allow editing and modification of the source; this is not the case with a page description language like Postscript.
- Converters between ODA and SGML are straightforward to implement [8].
- An SGML text file can be searched directly with a string-based search engine; to search an ODA representation for the occurrence of text strings, it is not possible to index the ODA file itself; it is necessary to index an auxiliary file such as an SGML version.

Our general impression is that SGML is an excellent authoring format, due to its more sophisticated data-modelling potential, SGML is an excellent search format, and ODA a good distribution format. The concept of authoring in SGML and distribution in ODA brings together the best of both worlds. It is comparatively straightforward to move between the two forms; the really important step is to have the distribution mechanism using one such format, rather than something like Postscript. While Postscript is suitable for printing, it is impossible to transform back into a revisable form.

6. Activities in the C-ODA Project

The C-ODA project had two main strands - replicating the work undertaken by Bellcore and its partners in the USA [17], and also extending the work into using more efficient storage, lower data transmission, and more generalised document searching tools The starting point for all is the work of Michael Lesk at Bellcore who built a number of tools to convert the original ACS data from their typesetting form into the SGML format [5]. This data is augmented with scanned images of the journals and diagrams to form a rich enough base of information to build the database upon - again an activity undertaken by Lesk [11].

It was our intention to provide a document database which could be queried in a convenient manner, and allowed the users to browse their results on-screen using a number of different tools. We have provided facilities for a number of end-user chemists to access the database at various locations within the University of London - both at UCL with Local Area Network (LAN) access, and via the University of London Wide Area facilities (WAN) which include the Internet and the ISDN. At the beginning a portion of the data was provided originally in the same form as in the CORE project; now, the database is supplemented by transforming the whole data which we have into the ODA/ODIF format (9 GB), and making it available to the University of London (UL) chemists in that form (ODA-D). This data is much more compact, so that the ISDN access is feasible; in the scanned image format, it is only practical to access the database at speeds at least equal to that on Local Area Networks (LANs). At present we provide a number of interfaces to access that data, including WAIS [13], SuperBook [12] and PixLook (also from Bellcore[18]). We are also evaluating how SuperBook, can be extended to give intelligent Hypertext guidance to users [19].

At the time we started the C-ODA project, due to the size of the dataset, the most sensible device for storing the documents was an Optical Juke Box (JB) - with the more rapid reduction of the cost of magnetic storage than the magneto-optical, this may no longer be the case. We have developed a JB interface library which virtualises the JB as a single large storage device, so that the application programs do not need to track the locations of files among the discs in the JB - to which a high speed storage server, with some 18 GB of disk space is attached as front end. A reverse index of all the document text is held in the disc

storage. For the whole of ten years of data this contains about 6.1 GB. All searching of document contents is done from the disc storage; the retrieval of the documents themselves is from the JB which holds the documents in all forms.

The work we have undertaken in this project is as follows:

- Develop a converter from SGML into ODA. This is a flexible converter which can be used with any DTD. The converter uses a translation specification which determines how the SGML elements are to be converted into ODA entities. While the development of this tool represented a major part of the UCL work, it has been well described in [8], [20];
- Due to the size of the dataset, the most sensible device for storing the documents is an Optical JB; we have developed a JB interface library which virtualises the JB as a single large storage device, so that application programs do not need to track the locations of files among the discs in the JB. Again, while this was a very time-consuming task, modern jukebox software now provides these facilities - though the software can be expensive;
- Replicate the Bellcore/OCLC work at UCL, and extend the interface tools to use the ODA representation. The Bellcore tools do not adequately deal with the problem of text and graphics on the same page, whereas the ODA-based viewers provide a much more natural presentation of such material.
- Interface the various databases with the WAIS search and retrieval system. Various other access mechanisms were also considered - but these have been described elsewhere.
- Provide remote access to the database over Basic Rate ISDN and the University of London WAN.
- Work with users and provide facilities as they require it.
- Finally to assuage the worries of publisher, we felt it essential to add various forms of access control, integrity control, authentication and audit trails.

7. Storing Data

7.1 The Use of an Optical JukeBox

We have installed a large document store, consisting of a Hewlett Packard optical JB with 4 Sony drives, a Sun SparcStation (Sparc-5 with 96 MB of primary store) as a dedicated server, and 18 GB of Magnetic storage. The main storage consists of 144 magnetic optical platters each with 600 MB of data; this allows 90 GB of rewritable storage. Access to arbitrary data is slow - 15 seconds. However it is possible to stage the data into the disc storage.

At the time of writing the paper UCL had received, but not yet put into operation, integrated software for storing the data in the JB; some recent JB software allows an application running on a workstation to access transparently any disk in an optical JB via standard Unix functions. It treats the whole JB as an integrated disc store - while still giving us some control on what to cache in the magnetic store. We are still investigating the advantages of that type of software.

We store all the index databases (Idx-xxx-DB) on magnetic storage. This allows content searching to be done relatively fast. The available textual data requires approximately 6 GB of storage for the whole such data - which ever access system is used. In practice we have supported only relatively small periods of data for any sets other than the main Idx-oda-DB/ODA-D which we are mainly supporting. Nevertheless, the one year of SI-D requires 30 GB of data, and the extracted graphics (EI-D) requires 45 GB; these data sets are kept on the JB. We have moved the search function via WAIS to a 4-processor Sparcserver 1000 - which is not the JB controller. This is because we wish to use WAIS with many other applications, to search some of our WWW data, and to support simultaneous access by many researchers simultaneously.

It is an important aspect of the C-ODA project that the JB uses magneto-optical re-writeable storage. The CORE project used Write Once Read Many (WORM) storage; as a result, CORE was very concerned about getting the data right before it is put onto the JB. Since we have found that it requires many passes

through the whole data in practise, this has had the impact of making all their data manipulation a very long-winded process; CORE has usually worked for a longer time with smaller databases on disc store, and been very hesitant to commit to using the JB.

7.2 Database Sizes and Access Times

We now have considerable experience on the size of the data, and on the access times [8]. We have the text components of the database for most of 1988-1994, and the bitmap form for much of 1989-1994. For recent years, there have typically been 15-16K articles p.a. in the database; of these we have typically made available 8K articles p.a. in bit-map form also.

The full data for 1989-94, including the SGML and the extracted images, requires some 50 GB of image; this we have loaded onto the JB. From the above it is clear that the actual data management of these large collections, when they pass through so many stages of processing, is difficult.

We treat each year as a separate database, and the search for any particular word combination is done on each database. Thus, for example, searching for any single word (e.g. Robb), would take less than a second on each database; in one such search, 847 documents were found. It is also possible to do a field search on the same data; if the same database was searched in a field sense (e.g. author = Robb), then the search time was little changed, but the number of documents retrieved was more manageable and precise - only 23 documents. Finally, the current version of C-WAIS has limited facilities for parallel work in a multiprocessor system; it can operate on several databases in parallel - one to each processor, but not with more than one processor per database. Thus our multiprocessor WAIS server, which has four processors, can operate significantly faster than a single processor one - even for single searches - because of the way the DB is organised..

8. User Interfaces

Having an on-line database of scientific journals offers many advantages over the conventional paper-based journals; and many of these advantages fall into the areas of search and access. Much of the UCL-CS interest in the project is in providing different means of search and access, and gauging the comparative value of the different methods. Electronic searching texts for information is much easier and more productive than manual. We support full-text retrieval - every single word in the document is indexed so that the searches go beyond any keywords that the author/classifier has deemed appropriate. Again, search responses are virtually instantaneous - with the limited number of users we currently support..

Electronic access provides additional advantages. Access is non-exclusive - any number of people can access the same journal simultaneously. Access is distributed - it is not necessary to be in close proximity to the database in order to access its information. Access can be integrated with the users' facilities, allowing extraction of information for other purposes.

Most search requests are based upon some type of word-based search, the system looking for occurrences of the words in its document base. Searches may be restricted to certain kinds of data in the documents such as titles, author names, or abstracts - or may be applied to the whole of the text in the document. One of the interfaces (WAIS) will support relevance feedback - this mechanism allows the user to mark one or more documents in the database as being relevant to the query and the search algorithms will favour similar/related documents in subsequent searches. Algebraic text searching allows greater control over the text queries if more than one word is to be searched for in the document database; it allows the user to specify rules about how those documents are to be searched. Say a search is looking for the words "petroleum" and "refinement". The number of documents containing both words could be quite high, although there is no guarantee that a document containing both words may be about the refinement of petroleum - the occurrences could have been on separate pages. However, if the search were to look for "petroleum" and "refinement" in the same paragraph, then one would expect a higher "hit-rate" of appropriate documents. Some of the interfaces will allow some degree of algebraic searching.

Browsing is another type of searching - just looking through documents for contents of interest - much as one would skim a book. For browsing to be effective, it is essential that page update be quick.

At UCL-CS we are particularly interested in widening the scope of the project to include remote access to the document database; this involves relatively low-bandwidth communications - for example, Basic-Rate ISDN lines operating at 64 Kbps. At this speed a typical page in bitmap form, occupying 100 KB, takes at least 12 seconds to deliver. However delivery of the document form is nearer 1 second per page, or perhaps three or four seconds if images were also transmitted. Our technology provides access to the database outside the high-bandwidth LAN at UCL - though the ACS constraints do not allow us to offer such a service outside the University of London. We enforce our constraints by the use of security techniques (cf Section 9. We expect to introduce at a later stage other document stores, which have less constraints on their usage than the current ACS ones.

A fuller discussion of the various User interfaces available in the C-ODA and CORE projects is given in [8], [20].

9. User Experiences

We had had considerable feed back from the users' on what was needed to make them interested in participating in accessing journals electronically, or what they require once they have started to participate. Their views can be categorised into availability, accessibility, presentation and facilities. Each is considered in turn below.

9.1 *Availability of databases*

It is a significant effort to gear up to use a system such as an electronic database. It is easy to organise students to access journals experimentally - and even to feed back their reactions. There is little enthusiasm, however, to have academic chemists make serious use of the system: unless the following is guaranteed at a minimum:

- *There is enough data available to make the effort worthwhile.* One year was considered of marginal value; the last three or four years were considered worth taking the trouble.
- *The databases should be available at the time serious use starts.* It was found counter-productive to talk much about the system unless the data was actually available.
- *The continuity of recent data should be guaranteed for a reasonable time.* We believe that something like a year ahead is the minimum that justifies a new user taking the trouble to gear up to use a new method. We should not underestimate the time it takes chemists to get really familiar with new systems, and to change their method of working.
- *Good well-supported access facilities are available.* The take-up of such a system depends far more on the enthusiasm of individual protagonists and the provision of good advisory and technical facilities than propinquity. King's College, some distance from UCL but with excellent computer facilities inside the chemistry department, were much keener users than those in UCL.
- *There is a significant range of data available.* Researchers do not wish to use many different systems. In one experiment, we had just all editions of one journal available; the chemists were quite unwilling to bother to tool up to learn a new methodology just to access one journal. By contrast, the ACS journals covered enough of their interests, that there was enthusiasm to use that database.
- *The data facilities are stable* The chemists had no tolerance of frequent changes in facilities.

9.2 *Accessibility*

- *Immediacy of access is more important than quality of access.* Although the chemists are prepared to travel to the Computer Science department in order to take advantage of the workstation screens, they would still like more immediate access to the data on cheaper workstations via lower-bandwidth lines - and most have only such workstations available. For use from their own desks or from home, researchers were prepared to sacrifice a lot of speed of access.

- *Integration of access facilities* There was considerable interest when we showed the chemists that we could make access available via the WWW and private pages. Many are already using the WWW for other purposes, and this will certainly ease the problems of chemists getting started.
- *Appropriate culture in the department.* Our universities are not yet necessarily set up well to support this sort of access to outside the university from inside non-CS departments; some departments are more open to such usage than others.

9.3 *Presentation*

- *Print is still essential.* While many chemists were eager to look at articles on the screen, most wanted to take away hard-copy of those articles that particularly interested them. Paper is still considered to be the best form for reading a Journal article in depth. Users did not feel that they would be happy to absorb a journal from the screen.
- *Scroll Bars.* The lack of scroll bars on the right hand side of windows was considered an important omission.
- *Scan Density.* Users like the 100 dpi size for browsing, but considered it inappropriate for reading. Similarly speed is considered good for these images. However, when shown the re-generated text from typesetting tapes, they thought this was a major improvement. ;Some of the pages even at 300 dpi have unusable pictures.
- *Legibility of presentation.* In general it is difficult to read scanned images of whole pages; the text/image form, where the image can be expanded or superposed, and the text can be reformatted, is considered much more usable.
- *Fonts and Character Sets.* In this environment, the exact font used was not important. However it was essential to have a full range of character sets, and to have both suffixes and superscripts. Artificial synthesis of character sets was found very disturbing. This has serious ramifications for the provision of such facilities through the WWW, until the software using the next generation of HTML has been released.
- *Accuracy of Data.* There is about a 5% error rate in locating figures in articles when the mixed-mode ODA style was used; this was due to errors in the automatic generation of the extracted image data from the scanned images. This was not registered as causing serious problems by the users. Of course this problem will disappear with the new ACS production method, which will obviate the need for scanning the complete pages..

9.4 *Facilities*

- *Hypertext access to references..* One aspect which greatly excites the users, and differentiates this system from paper-based ones, is the automatic following of references.
- *Scrolling through search lists.* The ability to scroll a highlight through a search list is important, because this automatically tracks the place in a list of documents.
- *Augmented Search.* The need to view, edit and augment previous searches was considered to be very important. The lack of such a feature discourages casual browsing;
- *Processable retrieved data.* Many users would like to re-process the retrieved data - e.g. re-use diagrams, references, etc. This is impractical with page images data.
- *Registry numbers.* Some of the chemists who are familiar with on-line databases are keen to use registry numbers.
- *Access to chemical structures.* Many chemists would have liked access to the chemical structure files generated by the ACS.
- *Access to Chemical Abstracts.* There was a mixed response to the need to have access to Chemical Abstracts; some chemists wished us to be able to provide it, many did not care.

- *Access to Additional Data.* Many chemists would have liked to have access to the auxiliary data which was available in the ACS tapes, but was not printed in the journals.
- *Switched access.* We demonstrated the feasibility of access via ISDN, but did not offer the service. We had the impression that UL chemists would have been greatly concerned at the need to pay variable communication costs - but this would not have been a serious consideration for chemists in industry..

10. Database Security Features

Restricting document access is important in two ways. First, publishers are going to make document access available only if it can be constrained and charged. C-WAIS can restrict access only to workstations with specific IP numbers; this allows already restricted access only to workstations in the University of London if this is desired. Second, we have added a public key system to the data access mechanism; this allows non-repudiable access to be achieved - which is important for later billing. Another use of such techniques, is that it is straightforward to add a digital signature to each article, which indicates the source of the document (e.g. the UCL-CS datastore), and the copyright holder (e.g. ACS); this could even be augmented by the person who accessed the store. Such additions could be used in several ways:

- To indicate the integrity of the document by signing with the secret key of the document provider
- To indicate the source of the document; documents where the signature had been removed could be regarded as *a priori receivers of stolen property*. (of course any such interpretation would require relevant changes to the legal framework.

The implementation of this technology has been described in [21]; it depends on the OSISEC [22] security package developed at UCL which implements the services described within the X.509 Authentication Framework, viz.: *Data Confidentiality, Data Integrity, Origin Authenticity, and Non-Repudiation of Data Origin.*

11. Conclusions

Database Construction

- As usual all underestimated the work required to put together such a large and complex database. The text portion was more difficult than expected because of the fonts included; in addition, the librarians were very concerned with fonts and spacing being followed very exactly. The equations were complex because of the absence of standards for equations in some of the systems used (in particular ODA and SGML); as a result even in some systems of compound documents, the equations were displayed in image form. the figures were hard to extract accurately by automated means from the scanned images; it was often difficult to distinguish figures from equations, or to differentiate between one and two figures across a page.
- The use of a small database was invaluable in exercising the technology, learning to understand its limitations and gauging the extensions needed.
- The database construction will be greatly simplified for future issues because the ACS has gone over to a fully digital production process - including for the image data, and has adopted SGML internally.
- While all the production chain of (a) - (i) of Section 2.1 will be eased, the distribution process will be only marginally addressed. Several related DTDs are defined for the different phases of the production process, and there will be some production of CD Roms, and even access to one or two journals. None of the on-line access or the CD-Roms will be in compound document form at present.

User Access

- The use of small document databases was invaluable also to get subjective feedback on what user facilities were required, and the relative advantages of the different types of user access.
- For access to documents with mixed mode (e.g. SuperBook or WAIS/ODA), the ISDN gives quite respectable performance. pre-fetching the complete paper improves this performance.

- For remote usage, the provision of small versions of diagrams, with the ability to request larger ones if desired, is very useful.
- Colour workstations are important in highlighting aspects of the searches; they are easier to use than monochrome ones.
- For user interest, the guaranteed availability of a substantial database for a long period is of paramount importance.

Document Formats

- Only the compound document format (e.g. the one discussed in this paper) could realistically deliver the whole document remotely; the bit-map forms were rather voluminous for extensive on-line perusal from outside a LAN (until a fast network like SuperJanet is available!).
- It is inconvenient that we cannot store one form of database, and allow access by three different methods. Although we had several different access methods, each required a different form of database.
- The ODA form of document was the most convenient to incorporate into other documents. It was the only one in which the management aspects of the document are incorporated into the same database as the information itself. It is also the only one in which security features have been standardised.
- The SGML format is clearly the most appropriate for the publishers and can well incorporate full house styles; ODA is more suitable for blind reading of a number of different databases. The lack of agreement on SGML DTDs is still a considerable nuisance - as we discovered in trying to go between the C-ODA software and that of the Institute of Physics with another journal.
- It was relatively easy to 'layout' the SGML into ODA once we ignored the problem of retaining the SGML structure for a subsequent conversion back into SGML. ODA is as good a choice for a presentation form as any other.
- Storing data in an ODIF form does not limit the user choice of tools. It can be used by any other editors which can read ODA documents. At the moment plenty such editors are available in the Market.

User Interest and Facilities

- Users are much more interested in viewing documents from workstations in their vicinity than going any distance to a workstation. For the UCL Chemistry users this meant that at the least we needed to install Unix workstations locally. They would have preferred to use their own PCs or MACs from their offices.
- They are more comfortable in reading papers they really want on paper; we have not yet installed convenient printing facilities, but they are vital.
- A 100,000 document database was the minimum size to interest chemists in using the system - and even then their interest was limited. The principal bar to use was the limited number of years - and of journals - in the database. Unless there is a reasonable chance of the chemist finding the references wanted, there is little motivation to use the system.
- The ability to highlight through a search list is important. Viewing, editing and augmenting previous searches is important.
- There was considerable interest in the possibility of using the system to search automatically through references. This type of usage probably requires the full database.
- Chemists who are familiar with on-line databases are keen on Registry Numbers.
- In the image database, the use of the cruder 100 dpi size for browsing is convenient - but considered inappropriate for reading: proper text was considered better than image versions of it. Speed for images is important. Even 300 dpi was considered unusable for some pictures. With the WAIS/ODA

version, software limitations in the UCL software only permit 80 dpi for the diagrams and equations - but the picture has been converted, and then does not cause any complaints.

The technical feasibility and utility of electronic access to these documents have been demonstrated. We are now discussing with the ACS and various British bodies whether, and if so under what terms, a large scale National testbed of this data might be mounted.

Acknowledgements

We acknowledge the help given to the project by a number of People. David Gold did much of the work described here while he was leading the project; Mike Lesk (Bellcore) has been a major driving force both to the CORE and C-ODA projects; Lorrin Garson (ACS) has kindly allowed us to use the ACS data and discussed their new production process; Chemistry users have been important in the trials; Peter Williams (Sterling Software and UCL) and Sammy Sameshima (now Hitachi Software) have been instrumental in the implementation of the security facilities..

The substantial support of the British Library Research and Development Department through a substantial period under several different grants is gratefully acknowledged.

References

[1] Davis, RD and C. Lagoze, "A protocol and server for a distributed digital technical report library, TR 94-1418, Dept of Computer Science, Cornell U, 1994.

[2] Handley, M and J Crowcroft, The World Wide Web, UCL Press, 1995.

[3] Lucier, R.E.: "Red Sage Electronic Journal Project", http://www.library.ucsf.edu/Projects/RedSage

[4] Story, GA et al.:" The RightPages image-based electronic library for alerting and browsing", Computer, 25, 9, 17-26, 1992.

[5] ISO, "Information processing -- Text and office systems --Standard Generalised Mark-up Language (SGML)",ISO, IS 8879, 1986.

[6] ISO, "Office Document Architecture (ODA) and Interchange Format", ISO, IS 8613, 1988.

[7] M. Lesk, "The CORE Electronic Chemistry Library", Proc. ACM SIG Information Retrieval Conference, Chicago, 1991.

[8] P. Kirstein, and A. Montaser-Kohsari, "The C-ODA project - experience and tools", to be published in Comp. J, 1996.

[9] S. Golkar, P. Kirstein and A. Montasser-Kohsari, "ODA activities at University College London", Comp. Netw. and ISDN Syst., 21, 187-196, 1991.

[10] Newton Search Engine, OCLC proprietary search engine, //http://www.oclc.org:5046/oclc/research/research.html

[11] M. Lesk, "Images in document retrieval: extraction of figures from pages". Proc. Anglo-French-US Conf. Image Storage in Libraries and Museums. York, June 25-26, 1990.

[12] J. Remde, L. Gomez, and T. Landauer, "SuperBook: an automatic tool for information exploration - Hypertext?", Proc. Hypertext '87, Chapel Hill, N.C., pp 175-188, 1987.

[13] B. Kahle, "Wide Area Information Server Concepts", Tech. Rep., TM Limited, 1989.

[14] WAIS Server and WAIS Workstation for Unix Administrator Manual, Release 2.0 WAIS Inc., Menlo Park, CA, USA. http://www.wais.com

[15] Pfeifer, U: "Searching structured documents with the enhanced retrieval functionality of free-WAIS sf and Sfgate, U Dortmund Lehrstuhl Informatik VI, D-4221, Dortmund, http:www.fhg.de/www/www95/paper/47/fwsf/fwsf.html

[16] Document Style Semantics Specification Language (DSSSL) Lite Standard, ISO/IEC DIS1017.92, ISO, Paris, 1995.

[17] Making a Digital Library: The Contents of The Core Project, http://community.bellcore.com/lesk/chem94/chtx.html

[18] Lesk M. (1994), Electronic Chemical Journals, Analytical Chemistry 66 (14), pp 747A-755A,

[19] M. Hu. "An Intelligent Hypertext System", Ph.D thesis, University College London, UK, 1994.

[20] G. Montasser-Kohsari and P. Kirstein, "On-Line Access to Multimedia Documents", BLRDD R&D Report 6139, London, 1994.

[21] J. Sameshima and P. Kirstein, "Secure Document Interchange - a Secure User Agent", to be published in Proc JENC '95, Terena, 1995.

[22] P. Kirstein and P. Williams, "Preparing to Pilot OSI Authentication and Security Services on a Medium-scale", Proc.4th JENC, pp 50-54, 1993.

ABBREVIATIONS

ACS	American Chemical Society
ASN	Abstract Syntax Notation
BLRDD	British Library Research and Development Department
CAS	Chemical Abstracts Service
DSSL	Document Style Semantic and Specification Language
DTD	Document Type Defintion
EJ	Electronic Journal
ESIS	Intermediate Language
ESPRIT	European Strategic Programme for research and Dvelopment in Information Technology
HTML	HyperText Mark-up Language
LAN	Local Area Network
OCLC	On-line Computer Library Consortium.
ODA	Open Document Architecture
ODIF	Open Document Interchange Format
OSI	Open System Interconnection
PODA-SAX	ESPRIT project 5320 Piloting ODA Extensions and their Applications in systems
SGML	Standard Generalised Mark-up Language
UL	University of London
UCL	University College London
WAIS	Wide Area Information Servers
WAN	Wide Area Network
WWW	Word Wide Web

The Authors

Peter Kirstein is a Professor in the Department of Computer Science at University College London. He has been leading research projects in computer communications, computer networks, telematic services and related activities for over 20 years. Amongst recent projects which he has led are the ESPRIT PODA project on ODA, VALUE PASSWORD project on security, and a BLRDD one on accessing electronic documents.

Goli Montasser-Kohsari is a Senior Research Fellow in the Department of Computer Science at University College London. She has a PhD from Newcastle U in Computer Science. She was responsible for the UCL-CS activity on PODA-SAX, and had the technical leadership of the C-ODA project. Goli has been responsible for all the recent UCL-CS activity on ODA implementation, ODA-SGML conversion, and C-ODA piloting.