Networking Full Text - The Quartet Project

Bill Tuck, PhD
Senior Research Fellow, Department of Computer Science
Paper presented at the CD-ROM Conference, York, 18th and 19th September 1989

1. Background

The origins of this paper lie in a series of experiments that were carried out over the past two years for the British Library under the QUARTET project [1 ,2]. These were based on a very large document image database that was produced by the ADONIS consortium of publishers and consisted of two years' output of approximately 200 biomedical journals [3]. The pages were scanned at a resolution of 300*150 dots per inch (or 300*300 dpi for half-tones) and stored on CD-ROMs. Each disk contained around 6000 pages and a new disk was filled approximately every week. The disk also contained an author/title index which was used to increment the cumulative index held on a separate magnetic store. The total database amounted to some 600,000 pages held on nearly 100 CD-ROMs together with an index database of more than 50 Megabytes.

The objective for ADONIS was to provide an experimental means of document delivery based on a standalone workstation without direct communications links. Requests for documents were keyed in to the workstation by an operator and the retrieved document despatched by post to the client. This fits the mode of working of the twelve or so library-based document supply centres participating in the experiment.

The objective for (QUARTET was to automate the whole procedure by linking the ADONIS database into a communications network (or networks). In this way the end-user could search the index, request the document and have it delivered, essentially without any operator intervention. In many respects, such a system serves as a prototype for many document image databases. The technology used, the database contents and the needs of the users may vary, but the general structure will carry over to a very wide range of applications.

In the course of this work considerable insight was gained into the many and varied practical problems that arise in setting up very large databases of image material. These, together with some of the solutions we devised, form the background to the present paper.

Background

Before elaborating on the work of QUARTET in carrying out this research I would like to remind you of the historical background to the general problem.

At the end of the last war, in 1946, the Royal Society organised a conference to bring together the scientific research interests of all countries of the British Empire, as it was then called. This began in Senate House on June 17 with a Royal Address and progressed in sedate and stately way via the chambers of the Royal Society, to Cambridge University, then Oxford, to end up with the closing sessions at University College London on July 8. (They certainly don't make them like that any more!) The proceedings of this mammoth event were passed on to me during a recent visit to Australia. They make very interesting reading. Among the half dozen principal topics discussed was "information dissemination among research scientists", to which several sessions seem to have been devoted.

The formal title was:

"Discussion of Measures for Improving Scientific Information Services within the Empire. Subjects to be discussed include Indexing, Abstracting, Special Libraries and Microfilms."

The intention of the conference was undeniably idealistic, seeking nothing less than a complete revolution in the way in which scientific research was pursued throughout the Empire. In this respect it was well-matched to the ethos of the times.

Among the ten or so papers on information dissemination (including contributions from ASLIB and the forrunners of the British Library) was one by Professor J D Bernal, then at UCL [4]. The problem he posed was straightforward: how best to manage the information explosion:

"There can be no doubt that the growth of scientific effort in the world has made the task of proper distribution of scientific information a critical one, in that, whereas the annual increment of new knowledge in the whole field of science and in any particular field is rapidly increasing, the capacity for assimilating knowledge of each individual research worker, is absolutely limited."

"The changes required are those which should provide for the workers in science the maximum of information relevant to their work and the minimum of irrelevant information; that is it should aim at efficiency and economy. This can only be done by the better organisation of the production and distribution of the basic unit of scientific publication - the individual paper. This can be achieved without any interference with the autonomy and function of scientific publishing bodies, such as scientific societies, by the formation of an adequate distributing agency using modern methods of reproduction and distribution."

The solution proposed was a sort of central agency that would collect, from the scientific societies, all papers to be published. This agency would then be responsible for redistributing them to individuals, either by standing block order or on request. In some respects, the suggested operation parallels that of the present services of the British Library's Document Supply Centre (DSC), and it may well have been one of the instigators of this development. More generally, however, in spite of the considerable enthusiasm for a coordinated publishing centre, it does not appear to have been carried any futher. No simple technological solution was found to the information explosion, despite high hopes for the new wonder medium of microfilm.

In fact the solution, if indeed there was one, lay in a completely different direction. It is remarkable that in the hundred or so pages of the report devoted to this topic the role of commercial publishing organisations, as opposed to scientific societies, is never mentioned. With hindsight it is now possible to see how publishers, working for commercial gain and not just intellectual idealism, were able to fill the gap. Technology made publishing easier and this, coupled with clever marketing, allowed the needs of individual research workers to be met without necessarily swamping them with irrelevant material. The growth of specialised journals replaced the portmanteau volumes of rather amorphous learned societies.

In addition, of course, one should not neglect the importance of the indexing and abstracting services, coupled later with the technological advances of online databases -- though here again it was often the commercialisation of the service (through the efforts of the distributing agents, such as Dialog) that made it work, not the technology *per se*.

More recently, however, a new problem has emerged, or rather the old problem has re-emerged in a new guise. The explosion of periodical titles, rising subscription costs and falling real incomes means that libraries are finding it increasingly difficult to cope.

The extent of this problem and its implications for research have been presented in a recent study carried out for the British Library [5] . One of the principal findings of this report is that reduced library finances are beginning to have significant effects on research. Expenditure on periodicals, whilst it increased by over 59% between 1981/82 and 1985/86, nevertheless lagged behind a recorded increase of 72% in the Blackwell's periodicals index over this same period.

The usual response is to cut periodical subscriptions, often by as much as 25% in one go (University of Guelph, Canada). At the same time the number of periodical titles being published is increasing remorselessly. The net consequence to the reader is that there is a decreasing likelihood of finding any particular article on the library shelves. This has lead to:

- increased use of the inter-library loan system
- increased use of online search facilities
- less rigorous researches
- increased tendency to bypass libraries
- inability to place as much reliance upon browsing for meeting information needs as formerly

It is unlikely that the financial resources of libraries themselves will ever be increased sufficiently to overcome this problem. What is needed is a joint effort between the publishers and the technologists to harness the two sides to provide the right balance between a market-led and a technology-led approach. It is for this reason that the ADONIS project is so exciting. Unlike many earlier attempts at online document storage [6], it is not primarily concerned just with the technology, but with finding new ways to deliver the product -- in this case, science research papers -- to the market.

Optical storage, particularly in the form of CD-ROM, thus appears to offer the possibility of a

new solution to the age-old problem of supplying information on demand. But is this really the case or is it just another illusion on the road to a perhaps unattainable goal? Will it, too, go the way of microform and be absorbed by the system without fundamentally altering our basic way of doing things?

There are good reasons for believing that it is qualitatively different, and these centre on the fact that storage is no longer in analogue form but digital. This difference is fundamental: Every time you make a copy from an analogue medium you lose resolution/quality; but the digital domain makes possible infinite replication with no loss of quality -- and at an infinitely declining cost.

2. Introduction -- the general problem

Any information system can be decomposed into just three basic components: the storage facility, the display facility and the part that links the two together, the communications.

At one level, the only difference between any standard online database service (such as Prestel, Reuters, etc.) and a document image service, is in the kind of material that is being dealt with.

The main characteristics of document image databases such as ADONIS (and of image databases in general) is that they are very large (both in total volume and in the data size of each "unit" -- in this case the page) and the print is often very small (needing high resolution displays). These two factors create the main technical problems: how to squeeze these files into a small enough space for storage and how to drive screen displays with sufficient resolution and control for the document to be readable. The constraint in both cases is cost: anything can be done if you are willing to spend enough money on it, but that isn't the point -- costs must be commensurate with the value of the application.

If the first two problems can be solved then the third, the communications problem, becomes relatively easy. In broad terms, this is simply about getting the data from the storage unit to the screen (or printer) in reasonable time and at reasonable cost.

The first -- and most obvious -- observation is that it would be a good rule to perform all communications with the data in compressed form. Since image data must be compressed to solve the storage problem, it should stay in the same form for transmission and only decompressed at the display or delivery end.

The second observation is that the system is almost invariably going to be asymmetrical, with many more display systems than storage systems. For this reason, the allowable cost of the storage facility can be very much higher than that of the display systems. Likewise, the "cost" of compression (in both processing time or hardware cost terms) can therefore be much higher than that of decompression (which must be "cheap" and very fast). This can influence the choice of algorithm or method used for compression/decompression.

Thirdly, one should never store anything unnecessarily. Storage costs are high (the example of Telecom Gold is a useful reminder of the true costs of data storage: 20p per kilobyte per month corresponds to 10 pounds per page image per month). This means that as little as possible should be stored on the display end, and nothing should ever be stored within the network (ie., in between the document store and the display -- unlike electronic mail, which operates on a "store-and-forward" basis). A consequence of this is that the transmission between storage end and display end should be as fast and efficient as possible, so that users will not be tempted to cache stuff at the display end.

3. The technology options

The list above establishes the parameters within which the system must work. Here we look at some of the technology options available to solve these problems. This basically boils down to answering two questions:

- What data encoding method to use?
- What communications link?

The service requirement may be local (within the same building or campus) or it may be global (literally world-wide) -- or it may be anything in between. The particular application will determine which of these it fits (or if it is a mixture of several). In each case the data encoding and, more importantly, the communications options must be chosen to match the requirement.

3.1 Data encoding

At first sight, the data encoding method would seem to have little relevance to the communications problem. As mentioned earlier, the overall objective of encoding is to compress the document image as much as possible within the twin constraints of system cost and processing time. On these grounds there are strong arguments for going with the CCITT Group 4 protocols for compression.

Standardisation and wide availability of compression hardware and software means that implementation costs should be relatively low. Processing time is likewise reasonable (but will depend, of course, on the method of implementation). The codes chosen were optimised over a representative ample of standard documents, so compression is reasonably near optimal (at around 100K per page at 300 dpi -- a reduction to 20% of the original).

There are alternative methods of compression that may have advantages in certain circumstances. Arithmetic compression algorithms, TIFF, SGML, and other OCR-based methods, may all be possible, and may produce very much greater compression. The main constraint is the time and cost of performing the encoding. The virtue of fax is its simplicity and low cost.

For documents that can be archived at the point of origination (usually the publishers) some form of marked-up ASCII text may be a preferable form of encoding. Standardised mark-up (in the form of SGML) has many advantages in addition to those of spatial economy [7].

In addition, the requirement of being able to annotate the database is another factor that may need to be considered. Thus the original documents might be held as fax images, with annotations (possibly linked to specific page positions) held in ASCII form. Annotations might include diagrams, tables, spreadsheets, etc., as well as text. The ODA ("Open Document Architecture") specification is a CCITT/ISO standard being developed for complex structures of this kind. Hypertext is a further generalisation.

Although it is often best to think of the encoding and communications problems as quite separate, there are factors that link the two. Choosing Group 4 fax (CCITT T.6) for the encoding leads logically to using the CCITT T.73 protocol for communications. But this is not essential and many systems operate on proprietary file transfer protocols.

Adopting ODA, however, poses a more serious possibility. ODA is intended to define the message content type for X.400 electronic mail. X.400 itself provides a very elaborate communications protocol for handling ODA documents which could, at least in principle, be used to manage the transactions upon an image database. The problem is cost and inefficiency -- X.400 is a 'store-and-forward' system, so its use would contradict one of the basic principles for image systems. In the case of a public or wide-area network the costs may be very apparent and prohibitive. Within a local area network, costs for an X.400 based service may be less easy to determine.

3.2 Data storage options

The data storage options are fairly limited and again are primarily determined by cost and function. CD-ROM seems ideal for archival use, but we still need a good way of handling hundreds, if not thousands, of discs (200 periodicals generate 50 CD-ROMs per year, 50,000 -- the size of the BL-DSC serials holdings -- will generate over 12,000 discs per year). This may well prove to be a crucial limiting factor on the technology.

Laser cards, writable optical discs, erasable optical discs and even magnetic storage will all find a role in the business of creating a digital archive.

In many ways, however, the question of storage medium is of less interest than the data encoding format.

3.3 Communications options

At the lowest level, the communications technology options are very broad and include at least the following:

- Local area: ethernet, token ring, PBX or data switch, fibre optic (FDDI)
- Metropolitana area: fibre optic (FDDI), cable tv technology, ISDN, QPSX (IEEE 802.6)
- Wide area: ISDN, packet-switched networks, satellite

The best option will again be determined by the characteristics of the application. The performance ratings of the technologies vary widely (as does the cost!). The choice of technology must depend on the level of service required, likely network loading, and constraints of cost. These will be different for different applications.

The best I can do here is illustrate the arguments from the particular example outlined in the first section.

The potential client base for the document delivery service from the ADONIS database was of two kinds: those local to the university library where the database was held (ie., located on the university campus) and those at a considerable distance (anywhere in the UK, say).

3.3.1 Local connections

Local connections potentially available between database and client included a number of different networks:

- Local packet switch
- Data PBX
- ISPBX ("Integrated Services" PBX)
- Ethernet LAN
- FDDI ring

The problem with the first two is that the connections are far too slow for image data to be transferred in realistic time if any degree of interaction is required. At 9.6K bps, a single page (assumed around 100K bytes) will take approximately 100 secs to transfer (allowing for some overhead). This would possibly be adequate for an overnight document delivery service. Doubling the data rate to 19.2K or using a synchronous link would give only a marginal improvement in performance -- still inadequate for real-time applications.

The possibility of using the 64K bps lines of a digital PBX is an interesting alternative as it would significantly improve on the 9.6K lines of the standard link through the data switch. This option will be explored further when considering ISDN links below.

Standard ethernets operate at a nominal 10M bps. In principle, this should permit a page to be transferred in 0.1 sec, but the actual rate will depend on the number of users sharing the network and the load generated. Random (but somewhat unsystematic) tests on an ethernet at UCL indicated that the time to retrieve 100K from host file server to workstation was between 1 and 5 sec, with an average around 2 secs. (Although the network was quite heavily loaded, most transactions were much smaller than the 100K of a page image -- so it is difficult to extrapolate from this to predict the effect on a network used extensively for image transactions.) Decoding of a T.6 Group 4 fax image can be performed in less than 1 sec on a Sun 3/50 workstation. (Hardware decoding on a PC, or software decoding on a fast/386 machine would be expected to give comparable times.) For fully interactive work, performance times will be critical and it may be questionable whether current 10M bps LANs are able to deliver adequate bandwidth. (There are indications, however, that a more efficient use of the 10M bps ethernet is possible by judicious modification of the protocols and that tuning it for a particular type of traffic could increase the actual throughput to within 70% of the maximum.)

The ADONIS database was located in the library at UCL and could potentially be linked to a local ethernet. Most clients, however, would not be on that ethernet but on some other, departmental, ethernet. The intention is to link these through an FDDI backbone network (operating at a nominal 100M bps). The resulting local network would thus consist of many small departmental networks operating at 10M bps, linked to a 100M bps backbone. This is likely to become a standard configuration for local 'campus' networks. Evaluation of such composite networks for handling image databases is still under consideration.

3.3.2 Wide area connections

There were also several options for linking remote clients into the document delivery service, including:

- JANET/PSS
- PSTN and Group 3 fax
- IDA/ISDN with group 4 fax
- IDA/ISDN with PC-PC file transfer

Again it is cost and performance that determine the essential differences between these transmission channels. Packet switched networks, such as PSS and JANET, typically deliver 2400 or 1200 bps terminal connections. Higher rates are available but at a price -- 48K lines may be in excess of £8000 per year rental.

Group 3 fax on the telephone network (PSTN) operates at 9600 bps. The availability of cheap receivers (in the form of standard fax machines) makes this an attractive possibility for certain kinds of document delivery from an image database. The main problems are the poor quality of output (at 200 dpi) and the relative slowness.

Group 4 fax operates at either 400 dpi or 300 dpi, but requires a 64K bps channel for transmission. This would ideally be over ISDN, although current versions use the IDA (Integrated Digital Access -- British Telecom's prototype ISDN) network. Equipment cost is high at £15,000 for a Group 4 fax terminal. IDA line rental is £500 per year, and the usage charge is the same as telephone -- which works out at around 2p per page (this is dwarfed, however, by the cost of printing, line rental, machine depreciation and maintenance, which combine to give a charge of around 15p per page on the assumption of a 50,000 pages per year demand rate to each receiving site). The transmission time is roughly 10 secs per page.

A cheaper alternative for wide area distribution would be PC-to-PC file transfer across IDA. On the same assumptions of throughput, the cost in this case works out at roughly 10p per page.

Whether to take up leased lines or to use a public network will again depend on the characteristics of the application. High and/or continuous traffic load indicates a leased line; infrequent "bursts" of traffic indicates a public network.

3.3.3 Higher level protocols

The network channel -- whether ethernet, fibre optic, or satellite -- only provides the basic path along which the communications takes place. On top of this comes the consideration of what protocol to use for the data transfer. Options include: CCITT Group 4 (T.73), ISO/OSI FTAM, X.400, or some proprietary protocol. As usual, the choice will depend largely upon the application.

An additional factor is the communications links between the "head-end" components themselves. Databases consisting of several million pages will require more than one processor to handle the storage elements (CD-ROM or WORM jukeboxes) efficiently. Index systems on magnetic media and request management databases (ie., records of requests held for billing) will also need to be managed. A distributed system of this kind demands a sophisticated "network operating system" for its effective control. The QUARTET/ADONIS experiments at UCL used a facility based on "remote procedure calls" for this purpose. Proprietary systems (such as Sun Microsystems' "NFS") might also be adequate for this function.

4. Integrated systems

There is more to a document delivery system than simple document delivery: Indexes must be searched, requests need to be transmitted to the database server, error messages (if any) or requests for payment will need to be communicated back to the user, and so on. All will require some method of communication. And the channel chosen to transmit the actual documents may not be appropriate for these other types of information -- and vice versa.

This is best illustrated by again referring to the example of document delivery from the ADONIS database. In such a system, the basic activities are online searching of bibliographic indexes and transmission of requests to the supply centre. These have very different characteristics and requirements from that of image delivery. Online searching, for example, is highly interactive but does not require a very high channel capacity. Packet-switched networks are traditionally the most cost-effective way of providing this kind of service. More recently, the CD-ROM index held locally (or accessible via a local area network) is beginning to replace this method. For document requests (and error messages or payment requests) the requirement is a channel able to transmit numerous small messages in machine readable form; they need to be processed automatically, without having to be re-keyed by an operator. Here, interactivity is much less important. Electronic mail is an ideal medium for this kind of communication.

In contrast, the primary requirement for the document delivery path itself is cheap high volume data transfer capabilities. Time (within limits) may not always be an overriding consideration.

The implication of this asymmetry between index search, document request and document delivery is that the network services required for an integrated system are likely to be heterogeneous. Fax is not appropriate for everything -- and nor is email.

5. Trends in the marketplace that may influence full-text networking

Two areas of development that could be an important influence on the communications strategies of the future are:

- developments in the supporting technology
  eg., satellites, personal communications
- developments in the information marketplace
  eg., resale of information, facilities management
- international standards and regulatory policies

Early experiments (particularly, ESA's Apollo project) verified the practicality of using high bandwidth satellite channels for document delivery from a document image database, but the costs were unacceptable. The emergence of a mass market for DBS receivers, however, could have the same effect on the use of satellites for data transmission as the compact audio disc had on data storage. Coupled with relatively low bandwidth personal communications networks this could provide an ideal low cost delivery system for many applications.

Changes in the way information is marketed are also likely to have a profound impact on image database services. Currently, most effort is being spent on developing local or "departmental" facilities. Future systems, however, may well need to be better integrated with a wide area network to capture a potentially much wider market -- even if still only within the company as a whole.

Regulatory policies and international standards are likely to have a marked influence on the pattern of

development within certain sectors (such as government networks). Relative tariffs on ISDN and X.400 will be critical for determining the viability of these channels as carriers for ODA or fax.

6. Conclusions

Expressed in its simplest form, the underlying motive for the development of document image systems is to increase productivity by getting rid of all the paper; to create a "just-in-time" information system in a manner analogous to the efficient operation of inventory management and parts supply in a manufacturing process. It has long been realised that maintaining large local inventories is inefficient and ties up capital. Just-in-time, on the other hand, demands much closer links between supplier and manufacturer. By the same token, efficient information management requires good communications between client and document supplier.

The objective -- higher productivity -- is, of course, quite reasonable. On the other hand, information technology has not always been able to deliver the benefits it has so often promised and it is important to weigh the advantages of any such system against the costs of its operation. This is particularly true in the relatively new area of document databases. There is little doubt that the benefits can be substantial, the only question is whether the cost and performance of current systems can match the requirements.

This brings us back to the original problem Bernal posed to the Royal Society in 1946: "...to provide for the workers in science the maximum of information relevant to their work and the minimum of irrelevant information; that is, it should aim at efficiency and economy."

With an automatic document delivery system from an electronic archive the efficiency and economy of the operation may well be increased -- at least from the library's point of view. Reduced storage and binding costs, reduced labour and increased throughput are all likely consequences of large-scale CD-ROM deployment.

But from the client's side, the gains are perhaps less obvious. Maximising the relevant seems a fine ideal, but is it really the library's function to help minimize the irrelevant? Or are we in danger, in going along this path, of preempting too many creative possibilities? At least the CD-ROM, with its "more is better" convictions, has no such pretensions!

Acknowledgment

The work on which this paper was based was funded by the British Library Research and Development Department as part of the Quartet project and we acknowledge with gratitude their support.

References

1. Tuck B, McKnight C, Archer D and Hayet M (eds): Project Quartet Final Report. British Library, (to be published).

2. Tuck B: New directions for document delivery: Quartet's experiments with Adonis, Interlending and Document Supply, Vol 17 no 3, July 1989, pp 94-100.

3. Stern B: "Publishing on CD-ROM in mixed mode", in Proceedings of 10th Online Conference, London 986, pp 23-31

4. Bernal J D: "The form and distribution of scientific papers", in Proceedings of the Royal Society Empire Scientific Conference, pp 698-699, London 1948.

5. Pocklington K and Finch H: "Research Collections Under Constraint." British Library Research Paper 36, 1987.

6. Mastroddi F A (ed): Electronic Publishing: the New Way to Communicate, Kogan Page, London 1987.

7. Coombs J H, Renear A H, and DeRose S J: "Markup systems and the future of scholarly text processing", Communications of the ACM, Vol 30 No 11, November 1987, pp 933-947.