Preserving Readership of Electronic Works

Dr. Bob Jansen* & Trevor Davis†

* Principal Research Scientist,
CSIRO Division of Information Technology,
Locked Bag 17, North Ryde, NSW 2113, AUSTRALIA
Ph: +61-2-325 3133
Fax: +61-2-325 3131
Email: Bob.Jansen@dit.csiro.au

† Executive Director,
National Printing Industry Training Council, NSW Division
PO Box 750, Hurstville Business Centre, Hurstville, NSW, AUSTRALIA
Ph: +61-2-580 1300
Fax: +61-2-580 4488

Abstract

This paper discusses some of the issues associated with the long-term preservation of electronic, multimedia, documents. It begins by asking why we preserve information, presents some 'current' approaches, introduces the new challenges and areas of concern, discusses some possible strategies, and concludes with a discussion of implications for training needs for multimedia-skilled people.

1. Introduction

Much activity and effort is occurring on the preservation of information using electronic media, such as CR-ROM, databases, etc. Although noteworthy in its aims, this effort seems to miss an important point. Without preserving readership, ie., the ability to 'read' the preserved information, this effort is likely to be completely wasted. In the past, this has not been a problem because the tools used to read the information, our eyes, have not changed; the medium used to preserve the information, printing words and images on long-lasting substrate, eg., 100% rag acid-free paper, is the same method used to distribute the information in the first place and hence kept pace with all new developments. In other words, the technology used in creating, distributing and providing access to information was harnessed in its preservation.

Recently, that simple mechanism has changed considerably. As the amount of information has grown to exceed the capabilities of organisations to store and process the paper-based archives, new methods were brought to bear on the problem, photography, microfilm, and, more recently, information technology; methods that have all demonstrated the difficulties with long-term preservation where those methods required mediation by technology.

The multimedia revolution now poses even greater challenges.

Multimedia technology is evolving at such a pace that it is already impossible to access recently archived information because the tools used for reading in the past no longer work on the new range of equipment; the data structures used to store the information are no longer understood; or the amount of information already archived exceeds realistic expectations for finding the chunk[1] required. Preservation of information is a doomed activity unless provisions are established now to cater for future readers. After all, if the future cannot read the information, why preserve it.

This paper discusses some of the issues associated with the long-term preservation of electronic, multimedia, documents. It begins by asking why we preserve information, presents some 'current' approaches, introduces the new challenges and areas of concern, discusses some possible strategies, and concludes with a discussion of implications for training needs for multimedia-skilled people.

2. Archiving

Thibodeau (1994) defines archiving as aiming to identify, preserve, and provide access to that small portion of records that have sufficient enduring value to be retained beyond the margins of business purposes for which those records are created. This is the process of retaining 'important' information for future use. This information is retained to record, amongst other things, significant decisions and events in our time.

The *raison d'être* for archiving is to ensure that people in the future can access significant documents from our era and understand them in the context in which they were created. This captures the nature of our era, its idiosyncrasies,

[1]The term 'chunk' is used in this paper to mean an atomic fragment or collection of material relating to a particular situation or topic.

its turmoil, daily life, and the development of our society, etc. It should be possible to reconstruct the context of a chunk of the archived information so as to enable the reader to appreciate events surrounding the focus of the chunk.

3. Existing Approaches

Existing approaches to archiving all have two things in common: they are predicated on the storage of the information on long lasting stable substrate; and they only work for static information. This approach works well for specific chunks of information, see for example the plethora of old books and manuscripts. The substrate does break down, but, baring catastrophic events, so slowly, that its content can be transcribed before irretrievable loss occurs. In addition, there are now a deluge of techniques to retard the breakdown and conserve the content of old documents, manuscripts, works of art, etc. Some items, however, are not transcribed as they may loose their value (eg., works of art) and hence we aim to conserve these, in their original format, at all cost.

Having recorded the information onto the stable substrate, it is now placed into collections, museums, archives, etc., and remains there for controlled use[2].

This strategy begins to break down when considering the vast quantities of information required to be archived. Experience indicates the amount of space required to store the vast quantities of information quickly exceeds available space, a situation that recurs cyclically unless changes are made to archiving strategies.

Having built huge collections of archived information, it becomes necessary to embark on a process of classification and indexing to ensure timely access to the collection. This becomes a never ending process as the collection is continually enlarged. Classification and indexing, in themselves, are difficult problems as they are immensely context-sensitive and many strategies adopted for their denouement.

These approaches to archiving are predicated on, amongst other things, the owner of the information relinquishing control of their information and minimal evolution in technologies for recording and reading the information.

The information is stored in warehouses, and thus the owner must relinquish control, at least in the physical sense, but also in general, in the logical sense. This relinquishing of ownership, or the transferral of the ownership to the archive, represents an important event in the document life cycle. Ownership is transferred to ensure the document can be physically stored in the archive, remains unchanged and continues to accurately record its context. The document now becomes a 'public' document: the original owner has lost control over it.

The technologies used for archiving and reading must show minimal evolution. Rapid evolution, using existing technologies, would require constant transcription, a process that would be impossible to maintain given available resources. Reading technologies place mutual dependencies on the storage

---

[2] There are some interesting examples of collection strategies in specific archives, see for example the Pitt Rivers Museum in Oxford, United Kingdom, whose collection strategy is based on the notion espoused by General Pitt Rivers (1827-1900) that there could be a systematic classification and arrangement of material objects that would reveal the course of evolution of human culture (Pitt Rivers, 1994).

technologies, and so any change in one usually requires an accordant change in the other. Luckily, our eyes have changed relatively little in the course of human evolution, and so accordant evolution in storage technologies has not been required. Any change to storage technology has been driven by advancing the lifetime of the storage media.

Microfilm technology, although enabling the recording of vaster quantities in small space, complies with the above characteristics, but places new technology between the reader and the information. This mediation negatively effects some attributes of the information, eg., ease of reading, compatibility, etc.

4. New Challenges

With the introduction and evolution of information technology, there are now new challenges to be addressed, including the rapid convergence and integration of previously disparate technologies, context and the change in the nature of information. These challenges are testing the boundaries of our capabilities in archiving and subsequent retrieval and reading.

4.1. The Rapid Convergence and Integration of Technologies

Over the last few years, there has been a rapid convergence and integration of previously disparate technologies. For example, in the past, computers were normally operated in stand alone mode. Advances in telecommunication technologies have commenced the linking of computers so that now it is sometimes difficult to distinguish where the computers end and the telecommunications service begins. This convergence has been made possible by advances in computer technology and the desire to manipulate, in real time, the various types of information processed in our daily lives. One of the driving forces behind it all is of course the money to be made by the computer suppliers: the technologists.

This convergence, apart from enabling the manipulation of the real types of information processed daily, places new demands on archives. Archives are now required to process new types of information, and more importantly, information that is not necessarily static. It is now feasible, some say necessary, to support movies, sound, text, graphics, etc. all within a single framework. More importantly, it is now possible to 'read' these types of information in an integrated fashion, and so this mechanism must be supported by the archive.

This move to *multimedia* information brings with it capabilities for interaction, with accompanying effects unforseen by the creators of archiving paradigms.

4.2. Context Depends on Users' Interaction

As the amount of information and its complexity increases, more sophisticated techniques are required to make sense of what is available. Context is one of the most important tools in this task. Without context, intelligent processing is impossible (Jansen, 1993; Jansen & Bray, 1993).

Once a static notion captured in part within the content of a document, context is now is something that depends also on the user's interaction with the information space. Once dimensionally constrained, context is now dimensionally unconstrained. This has effects on the processes of classification

and indexing, effects that beg the question from the classifiers and indexers, 'Where do we stop?' and from users, 'Where do we begin?'.

4.3. The Change in the Nature of Information

Further changes are discernible in the nature of information. Whereas previously, the information was easily identifiable through a physical manifestation, it is now possible to identify information that has no physical counterpart. Electronic documents[3], derived information, virtual documents, etc, are new information types that now are regular contenders on the information stage.

5. Areas of Concern

The discussion so far has addressed the nature of the archiving process and new challenges facing archivists. This section introduces some areas of concern: concerns that need to be considered when planning the archiving of electronic information or documents.

5.1. Access/Reading

The prime concern, in our mind, is that of access and reading. Unless we can ensure the future can read our archived information there is no need to archive it. This is the major paradigm shift to cater for the new information types and technology given the rapid evolution in this new technology.

Unlike the present day, the archiving process must now also account for the different technologies used in creating and subsequent reading of the information. One approach is to archive the relevant programs and hardware that were used in the creation of the information because it is the programs (software) that enable us to make sense of the binary form used to store the information.

5.2. Data

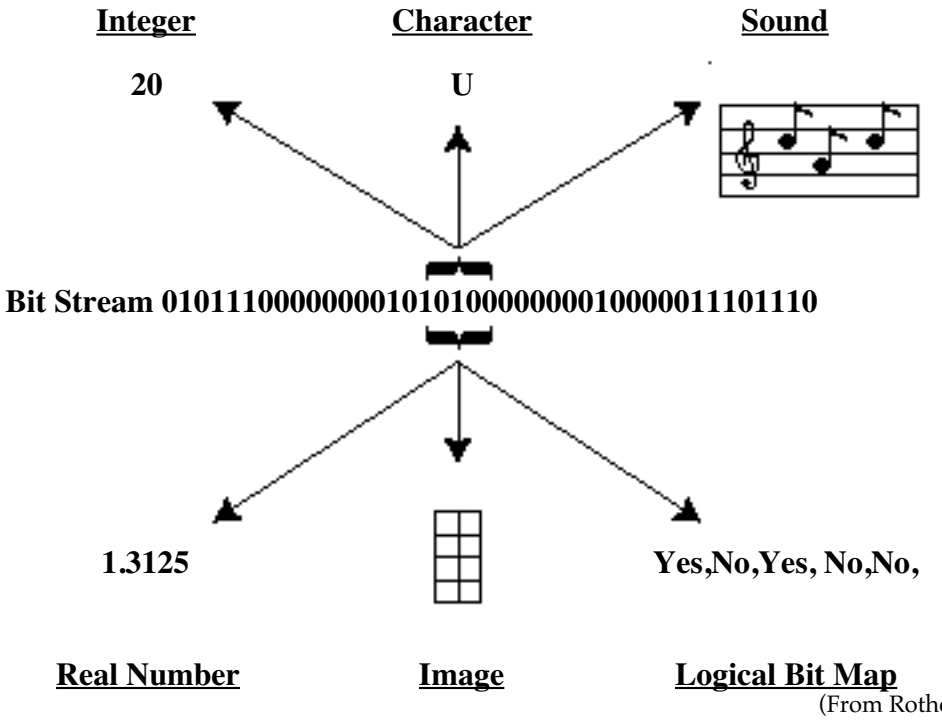
In terms of the actual stored data, there are two main concerns: the meaning of the data stored in a binary format; and the method for storage of that data.

5.2.1. Meaning

A problem that exists even today is the meaning of a binary data stream read from a storage disc. Consider migrating a data file from a word processor program to a painting program. Obviously, intelligent humans, knowing the context and content of the word processor data would exclaim, why that will never work! But consider two hundred years hence, when a file is retrieved from an archive believing it to be a painting program file. What will the painting program interpret that data stream as? It would almost certainly fail to interpret the file in any way, and so, to all intents and purposes, its contents are lost. Figure 1 shows the problem, a binary data stream can be interpreted in a multitude of ways with the interpretation depending on the original purpose of the program doing the interpreting.

---

[3] Currently, electronic documents are parodies of their paper-based counterparts, but this situation is changing rapidly. See for example Smith (1984) for a discussion on book paradigms.



(From Rothenberg, 1995)

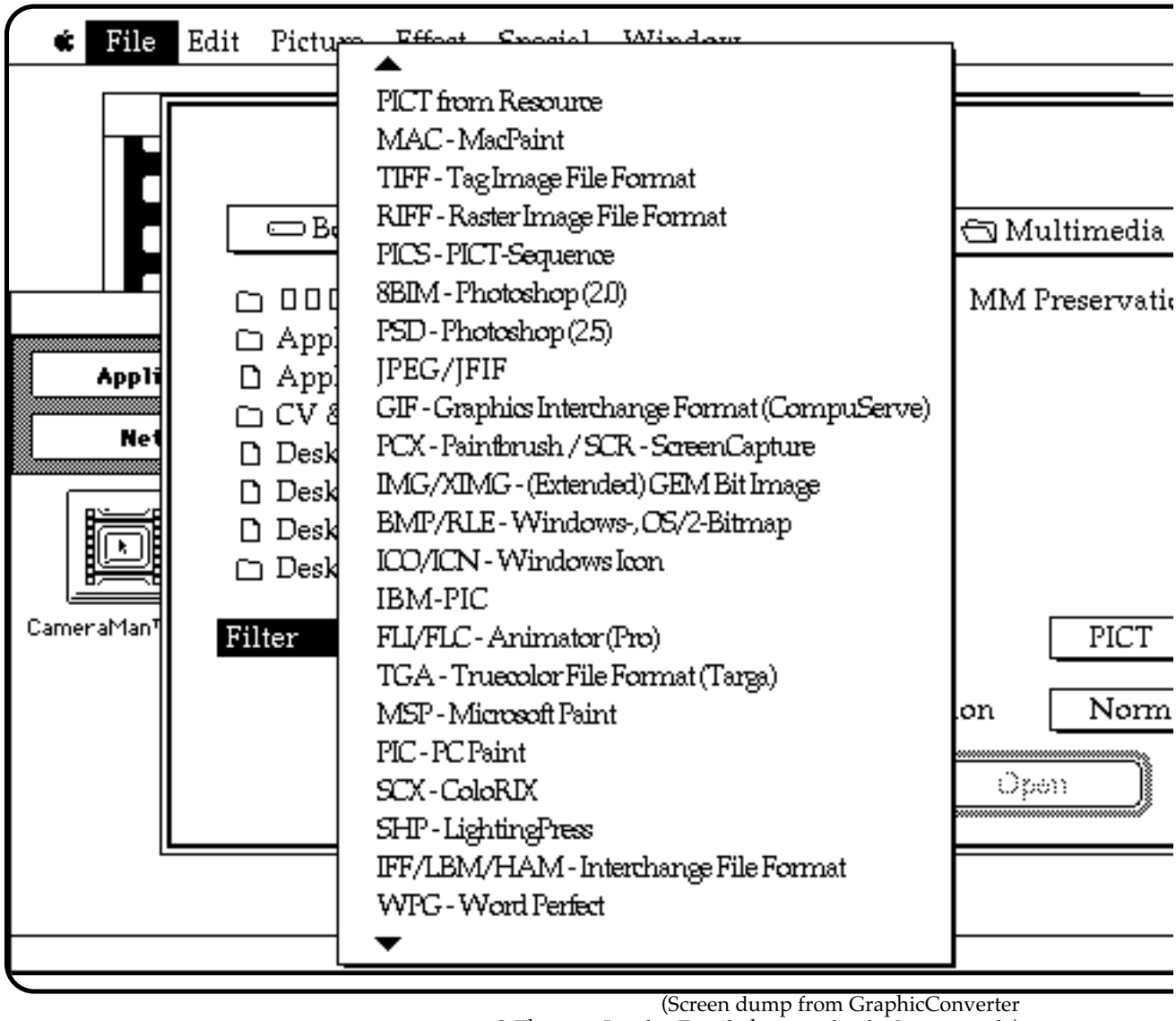
Figure 1 - Multiple interpretations of a binary data stream

Thus, the data file must be accompanied by extra information that specifies what type of information is being recorded, ie., the meaning ***and*** means of interpreting the binary data. Without this crucial information, or meta information, we lose the ability to understand the binary stream.

Interpreting the meta-information is no simple process, for this information is in itself recorded as binary data. Thus the problem, rather than being resolved, has moved to become, 'How can we ensure the meta-information is meaningful in the future?' Answer, record meta-meta-information, *ad nauseam*.

5.2.2. Storage

The number of formats for storing chunks of information is increasing daily with the complexity of the information being stored. Figure 2 shows some of the formats available using GraphicConverter, a program to convert between different formats.



(Screen dump from GraphicConverter
© Thorsten Lemke, Email: thorsten_lemke@pe.maus.de)

Figure 2 - Some available formats for storing graphic information.

The capabilities of individual formats are overlapping but potentially different. The choice of format is dependent on the software being used. Increasingly, individual software programs are capable of manipulating multiple formats, but there still exist requirements for programs like GraphicConverter, for translating between formats.

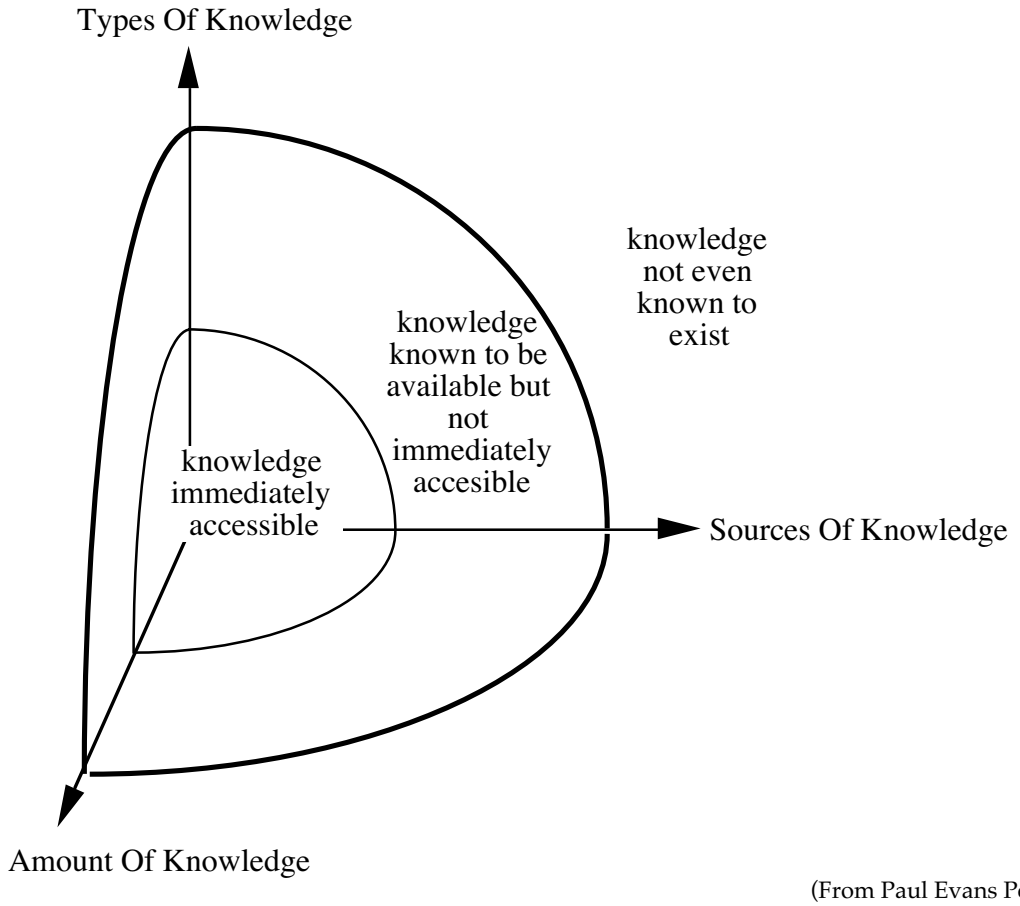
New formats are continually being created as the complexity of information we process increases. The problem of compatibility between different formats thus becomes more and more of a problem, and as new formats are born, old formats die, thus causing severe problems for archived information and its readability.

Storage formats are increasingly seen as revenue earning for their creators: they may be protectable under intellectual property laws. If this continues, users of those formats will have to pay royalty fees to the owners, thus adding an extra burden to the cost of archiving. The most recent example was the GIF format, the graphic format for the World Wide Web, usage of which, it seems, is subject

to protection under IP laws (Unisys, 1994), with concomitant constraints on authors and readers.

5.3. Amount of Information

The amount of information available today is increasing at an exponential rate. Paul Evans Peters (Association of Research Libraries, USA), in a lecture at the University of Technology Sydney, characterised the problem as one involving three dimensions: the types of knowledge available; the sources of knowledge available; and the amount of knowledge available, and three zones: that knowledge immediately available; that knowledge known to exist but not immediately available; and that knowledge not even known to exist. Information technology increases the volumes of the first two zones by reducing the volume of the third zone. The volume of the third zone is increased by research, creation and discovery in general.



(From Paul Evans Peters,
lecture at University of Technology, Sydney, 1994)

Figure 3 - The knowledge space available to a researcher

As the scale of this knowledge space increases, a number of important questions need to be asked for archiving purposes:

- What information is available;
- What information deserves to be retained; and
- Who owns the information?

Answers to these questions will have a major impact on determining what is the 'important' information, whether and how it should be retained, and whether it is feasible to archive it.

5.4. Virtual or Derived Documents

In the current world of the Internet's World Wide Web (WWW) information systems, a trend away from physical documents has commenced. Documents can now be created as *virtual documents*: documents without physical manifestation.

A virtual document is one whose content is wholly, or partly, derived from other sources. This has always been possible by using cut-and-paste techniques, but nowadays, this virtual document may only exist during the time of access. Outside access, the document consists only of a set of pointers to those chunks from other sources from which this document is to be derived and the rules guiding such derivation. A further complication involves chunks of information computed from other information, for example, a spreadsheet table appearing only as a row totals in an executive summary. Using existing technology, the executive summary could, in fact, calculate the latest total every time it is accessed.

This is true for a growing number of business documents. For example, a report may be written in Microsoft Word™ with links to tables in Microsoft Excel™. Archiving the report without the supporting Excel documents effectively renders it useless.

In the above scenario, the question to be addressed is, how can such a document be archived? In a world where information can be derived from derived information which in turn is derived from derived information, and so on, what is the analogue to the traditional physical document?

What is archived? Is it a snap shot of the final document or the total content of all documents from which a virtual document is based? Is there a 'derivation threshold', whereby all documents whose derivation distance to the virtual document lies within the threshold are archived (*viz*. spreading activation semantic nets (Cohen & Kjeldsen, 1987))?

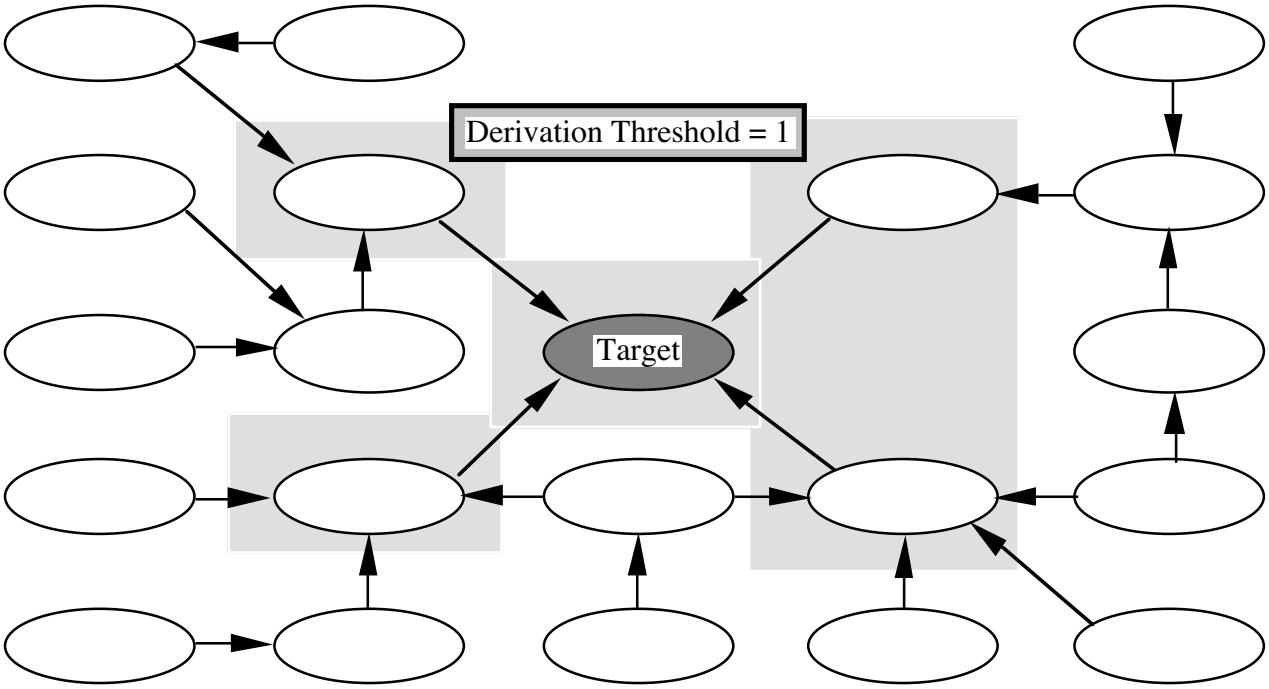


Figure 4 - Only chunks within the derivation threshold of 1, those chunks a single derivation distance from the target chunk, are archived

Issues introduced above question the very nature of documents. Jansen *et al* (1992) describes work on the integration of different media to support explanations in an electronic information systems (Figure 5). An outcome of this project was the blurring of the distinction between a document and computer program, where both describe the same knowledge. This concept led to a theory of *knowledge types*, subsequently implemented in *IntelliText* for representing and reading of electronic documents which was trialed on the Finnegans Wake archive (Jansen and Ferrer, 1994).

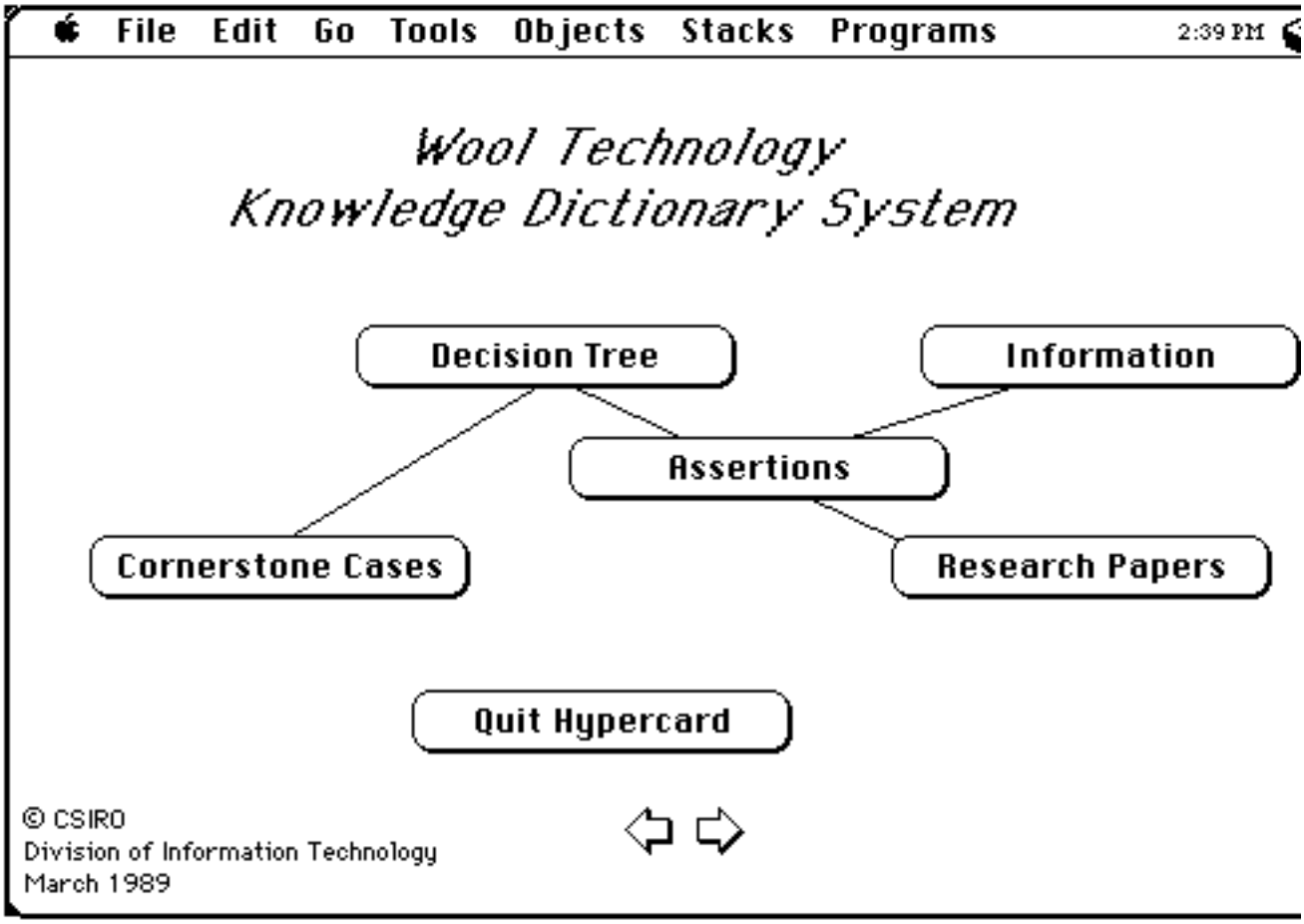


Figure 5 - Architecture of the Dark Fibre Risk Information System

6. Possible Strategies

Many strategies are being advocated for addressing some of the issues mentioned above (see for example, Garrett & Waters (1995)). Needless to say, no single strategy addresses all aspects and countries may have specific requirements. This section will discuss some possibilities for the considerations raised in the previous section.

6.1. Access/Reading

Preservation, and hence the problem of subsequent access and reading, was introduced as a result of the duplicative powers of print (Eisenstein, 1993, pp113).

As the effects of current information revolution may be as far reaching as the introduction of the printing press by Johann Gutenberg, it is too early, at this stage, to suggest any strategies for ensuring access and readability of electronic information. The best available strategy is to get on the learning curve and be prepared to make mistakes as investment and training for the future.

6.2. Meaning

The meaning of meaning (Ogden & Richards, 1946) is a complex problem that dogs many areas in the move from the printed to the binary word. Two issues stand out from the rest: context and standards.

6.2.1. Context

The issue of context, a major component in meaning, is largely addressed in printed formats through linguistics but is further complicated in electronic information systems where context is derived not only from content but also from the reader's complex interaction with the information space (eg. virtual documents). To-date, only static strategies exist for coping with the context of information chunks, eg. Herbrand universe (Herbrand, 1967) although research is proceeding on understanding the nature of context and its intelligent use in information systems.

6.2.2. Standards

Standards are an attempt at providing a common understandable format across software systems. The IT industry, after several years of proprietary systems, is now adopting the standards approach as a method of addressing the real needs of the end users. The adoption of a standards-based approach to archiving is predicated on the hope that standards are indeed standard across all systems and system versions.

The problem that has been experienced since the move to standards, is the adoption of those standards by the hardware and software manufacturers. In the IT industry to-date, individuality, facilitating specialist functionality, guaranteed survival. In the age of standards, most developers will adopt a superset approach, namely, implement the standard, the core standard, adding to it a superset to guarantee extra, specialist functionality. Although not dubious in its own right, implementors and users come to rely on these extra functionalities, from which the transfer to another system becomes extremely difficult for all involved.

Contrary to belief, standards continually change. Programming languages have seen their standards, their specifications, suffixed with the year of adoption, eg., Fortran 88. Multimedia standards, being new, are in a constant state of turmoil as new facilities are added: some required, some desired.

As with hardware and operating systems, software standards are not guaranteed to be backward compatible. Thus the adoption of a new version of a standard can lead to significant maintenance activity as content needs to be re-marked-up to conform to the new requirements of the standard.

An example of this is the draft standard for HTML[4], version 3.0. The draft documentation elicits the following statement related to anchors:

> "Note: the NAME attribute has been superseded by the ID attribute. User agents should include support for NAME to ensure backwards compatibility with legacy documents produced during previous versions of HTML" (Raggett, 1995)

Thus, if an archive had marked-up their content using the NAME attribute for anchors, an extensive maintenance exercise has been indicated to alter that markup to the ID attribute. That maintenance need not be undertaken immediately, but can an HTML reader's functionality be guaranteed to recognise the NAME attribute, and if so for how long?

---

[4] HyperText Markup Language, the page layout scheme currently in vogue for World Wide Web information systems

6.3. Storage Formats

The problem of storage formats is most tractable and can be addressed by:

- storing relevant hardware and software for the lifetime of the information; or
- implementing policies whereby the information is continually transcribed, or re-represented, without loss to a suitable format and software environment (i.e., becomes forward compatible); or
- designing and implementing a proprietary storage format and massaging all information to be archived into that format.

6.4. Amount of Information

The ever-increasing amount of information is essentially an unsolvable problem. It seems most probable that the amount of information will continue to increase with little if any restraints. Thus archiving strategies will continually have to address this issue.

6.5. Information is Media Independent

At a panel session, The World Scene[5], Kenneth Thibodeau claims "... we first of all start out by eliminating the issue of media altogether. Media is simply the something you use to preserve stuff on... If the medium has sufficient market penetration, there should be a lot of expertise out there and that expertise should be available to you over the period of time that the media is cost effective." (Thibodeau, 1994, pp98). This begs the question, what if an archive had chosen 5 $^{1}/_{2}$" floppy discs as storage media? Would it still be cost effective and where is the expertise today? From an archiving perspective, consider that many of the media available today have already failed to remain readable for even one hundredth of the time that the Rosetta stone has been readable. (Rothenberg, 1995, pp25)

Hence we need to find the analogue of archival paper: a long term, stable, electronic storage medium. It is impossible to maintain all media types, either for an organisation or society. It must be remembered that some versions of software are hardware configuration dependent and thus, in following this approach, it would be necessary to archive all relevant configurations and manuals of all hardware and operating systems for perpetuity. And then what about redundancy? How many copies of each piece of hardware must be kept to ensure, for example, paper tape readers remains useable one hundred years from now?

Since, as discussed above, the ability to read and make sense of a binary data stream is storage form and software program dependent, it may be profitable to establish a proprietary format for archiving. The adoption of this strategy is not without concerns, including: is the proprietary format semantically powerful enough to cater for changes in requirements; what is the cost of maintaining the knowledge of the format in-house; is it extensible to cope with future needs; etc.

---

[5] Australian Archives 'Playing for Keeps' Conference, 1994

6.6. Use Information Technology

An alternative strategy is to use advances in information technology as they become available. Issues that must be addressed include the rapid product life cycle of this technology, version control and backwards compatibility, the appearance of 'virtual' documents and control of documents.

6.6.1. Rapid Evolution in Information Technology

Currently information technology product lifecycles are, in some instances, now measured in terms of months. Hardware on the shelf is already out of date and keeping software up-to-date involves continual upgrades at significant cost. The issue for concern is what policies can be put into place, given limited funds, to keep up-to-date with latest trends? One argument is that this is not a problem, since by the time an archive receives a chunk of information, its creation environment is already out of date. It is possible to adopt a policy whereby electronic information will not be accepted for archiving without its creation, storage, access and reading environment.

More to the point, new versions are not necessarily backwards compatible and may not support legacy software systems. This is a greater problem in operating systems. For example, Windows'95™ will not run many of the existing DOS™ and Windows programs and hence information archived using DOS or Windows is lost to Windows'95 unless new versions of the software programs are released. This may mean that the archived information becomes lost to the users of state of the art technology. Data migration, where old data is migrated to new formats, is one possible strategy, albeit an expensive one.

6.6.2. Document Control and Central Archives

As stated previously, current approaches to archiving are predicated on the transfer of control of the document to the archive. In a virtual environment, where some information may be calculated at run time, who owns the document? It is possible, by doing away with a central location for the information, ownership ceases to be a problem. There would be requirements, however, for ensuring information owners are aware of their legal requirements and have the environment, competency and skill to meet those requirements.

In the electronic environment, determining ownership is complicated. The electronic document has many components: the core data, the algorithm used to derive the document or information, the code implementing the algorithm and the display system. Each of these components may have an IP owner thus complicating the establishment of final ownership

The archive, in this situation, needs only contain pointers to the information itself. However, as with the existing World Wide Web, there are problems with owners of information being unaware of its connectivity and changing the documents with serious complications to the logical nature of the navigation web[6].

™ Windows'95 and DOS are registered trademarks of Microsoft Corporation.

[6] In the existing World Wide Web, an owner of a document can determine the *fan-out* of the document, ie., the pointers in this document pointing to other documents, but is unable to determine the *fan-in*, ie., those documents having pointers to this document. The fan-in is a major contender in determining the logical place of the document in the document space. Fan-out and fan-in are terms used in electronic chip design.

6.7. Make Information Creators Responsible

The recent report from the US (Garrett & Waters, 1995) advocates, as one option, making information creators responsible for the archiving of their information. Although at first glance attractive, several concerns exist.

Most creators of information have no notion of long-term archiving requirements or techniques. Thus to invoke this option necessitates extensive education and legislation which would be difficult, if not impossible to enforce. After all, a breach of legislative requirements would only be noticed on attempting to access non-existent information, by which time it would be impossible to re-create.

Organisations, hosting information creators, are extremely conscious of costs and would see any imposition of long-term archiving responsibility as a further cost to bear, a cost having no benefit for the organisation. As stated previously, any defaulter would be practically impossible to detect.

The case of virtual and derived documents adds further complexity. Who is the creator of this type of document? The reader, in which case the document might be archived as multiple instances; the creator of each individual chunk, in the case of a virtual document; the creator of the algorithm in a derived document? An organisation might argue that although its employee has created the virtual document through a set of interactions with the information space, it is not the creator of the individual chunks and thus should not carry the burden of archiving.

7. Conclusions

The multimedia revolution has only just begun. The consequences of this revolution are still unknown, however, experience to-date indicates the situation is evolving rapidly - maybe too rapidly. It is difficult, in the current situation, to realistically address long-term issues.

So, the message is clear, do not make any decisions today unless absolutely necessary: any decision made today will be out of date tomorrow. Try to cope with what is required, and use the experiences gained as training and investment for the future.

Complacency, however, is also not the right choice. It is time to get on the learning curve (DG XIII/E, 1993), make the mistakes and learn the lessons, but do so in a carefully controlled manner. Recognise that over the next few years, new technologies will evolve that will necessitate existing data to be transformed into new formats, processes that will cost money.

HTML/SGML/Acrobat/ODA/proprietary format - which should I choose? Does it matter? Yes, it matters, some are easier than others to adopt, each has its advantages and disadvantages. Adobe Acrobat is good for storing data as in the printed form, pages are kept and viewed as pages, however, if the screen is too small, then cropping occurs necessitating annoying page manipulation. HTML browsers, on the other hand, automatically flow the text and figures to fit the screen window, but support for 'real' reading is poor. In addition, the language is still unsophisticated and semantically limited compared to the printed word. SGML is complex, some say too complex. The Text Encoding Initiative (Sperberg-McQueen & Burnard, 1990) is a masterful attempt at structuring

information, but is an extremely complex model to understand. SGML editors are still not too common and not cross-platform. Office Document Architecture (ODA) is challenging SGML to be the de-facto standard, challenges based primarily on geographic lines (Europe vs USA). And then we have Hytime, Hot Java, etc (see also the paper by McLean in this conference). The important thing right now is to choose the best format for the job: get the experience in handling information in an electronic format, understand its advantages and limitations.

An important attitude to cultivate is: don't forget the reader. It is easy in this day and age to be beguiled by the increasing number of authoring tools becoming available. However, remember that authoring is different from reading: although it is possible to conceive Tolstoy authoring War and Peace with a word processor, it is highly unlikely that you will read it that way.

8. Training: The Way Forward

Training represents the way forward: get on the learning curve (DG XIII/E, 1993), make the mistakes and learn the lessons, as an investment in the future.

At the outset it must be said that traditional themes, programs and delivery of training generally are of little support and consequence to the specific requirements of people involved in the rapidly evolving markets of multimedia and electronic publishing. In general, vast training resources are developed on the presumption that it is either near to generic or can be adapted to meet emerging needs.

Clearly, this is a new world requiring a back to basics approach with the creation of training being synonymous with the creation of the technology for which it will be designed. It is clear that no single existing industry sector will dominate the new multimedia market. Instead, convergence and emergence are the key: convergence of existing industries in concert with the emergence of new industries. Multimedia workers will require a mixed bag of skills, some existing within traditional areas, eg. film, audio, computing, graphics, etc, and others which are still to be recognised. Thus cross-skilling will play a major part in the development of requisite skills.

Although an unknown quantity in terms of specific details, it is clear that organisations entering this new technology environment must mount the learning curve by building and experimenting in the use of products (DG XIII/E (1993), *op cit*). It is important also to recognise that whatever is built today may require rebuilding at a later stage as the proficiency of the organisation's staff improves with concomitant increase in expertise.

Strategic alliances should be investigated where opportune. For example, the printing industry in Australia is a participant in traditional trades based training but has rapidly become a player in the development of specific training for the electronic and multimedia arenas. This industry through its training council formed an association with the CSIRO Division of Information Technology early in 1994 with a view to exploring the future technological requirements of the printing and graphic arts industries.

This association lead to a mutual involvement through membership of the 'Interact' Co-operative Multimedia Centre Consortium, being an applicant for funding under the Commonwealth's Creative Nation initiative. The alliances

formed through this consortium provide a forum for exploring technological advances and the commensurate training needs for the future of this industry.

Brandt (1994), commissioned by the training council, further provoked the industry into realisation of the role to be performed in order to meet the needs, particularly of pre-press operators, in the industry.

What therefore is needed?

The first stage must be an educational process designed to acquaint and also commit industry with the realities and potential within the electronic publishing and multimedia world.

This would be a team effort perhaps coordinated by the industry training board but to which every stakeholder in the industry, together with the CSIRO as a key external partner, would commit.

A second stage would see applications for funding through appropriate government sources to identify, develop and distribute such materials as would be required to develop a training needs survey for analysis.

Finally, with the inclusion of the public and private training providers there would be a need to develop training resources, packages and programs linked specifically to industry and manufacturing processes.

Within this latter step we would need to link the identified key objective, namely the 'Preservation / Readership of Electronic Works' as a priority to the overall training carried out.

To summarise the printing, telecommunications and information technology industries need to embrace this major objective. They would then form an alliance with the CSIRO and such other partners as may be identified, to commit to a major project in the immediate future. They would jointly undertake this initiative which underpins both Australian industry, the advancement of technology as well as the training and education activities of the future.

Acknowledgments

We acknowledge the work of Phillip Brandt in compiling the report on the Australian Printing Industry on behalf of the National Printing Industry Training Council, NSW Division. This report and many subsequent discussions were the basis for understanding the requirements for education and training for this industry sector in accommodating the multimedia revolution.

We would also like to thank Peter Outteridge, currently at CSIRO DIT, for many hours of interesting discussions related to this important topic and advanced information systems in general.

References

Brandt P, *Desktop Publishing: Developments and Training Implications for the Printing Industry*, A National Printing Industry Training Council NSW Division Survey, May 1994

Cohen PR & Kjeldsen R, *Information Retrieval by Constrained Spreading Activation in Semantic Networks*, Information Processing & Management, Vol. 23, No. 4, 1987, pp255-268

DG XIII/E, *New Opportunities for Publishers in the Information Services Market*, Commission of the European Communities DG XIII/E Report No. 14926 EN, February 1993

Eisenstein EL, *The Printing Press as an Agent of Change*, Cambridge University Press, 1993 edition

Garrett J & Waters D (eds), *Preserving Digital Information: Draft Report of the Task Force on Archiving of Digital Information*, The Commission on Preservation and Access & The Research Libraries Group, 1995

Herbrand J, *Investigations in Proof Theory: The Properties of True Propositions,* in From Frege to Godel: a Source Book in Mathematical Logic 1879-1931, J van Heijenoort (ed), Harvard University Press, 1967, pp525-581

Jansen B & Ferrer D, *IntelliText: An Environment for Electronic Manuscripts* , presented at the 4R's Conference, Canberra, Australia, March 1994

Jansen B, *Context: A Real Problem for Large and Shareable Knowledge Bases*, Proceedings of the International Conference on Building and Sharing of Very Large-Scale Knowledge Bases (KB&KS'93), Tokyo, Japan, 1993

Jansen B & Bray G, *Context & Knowledge Types Vs Serendipity*, in Proceedings of the 13th International Conference Avignon'93, France, 1993, pp85-95

Jansen B, Colomb B, Henderson-Sellers A, Gallagher J, Robertson J & Bray G, *Printed Material as a Knowledge Representation*, in proceedings of HyperOZ'92 (invited presentation), Adelaide, 1992

Ogden CK & Richards IA, *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*, Routledge & Kegan Paul Ltd.(pub) London, 1946

Raggett D, *HTML Markup Specification Version 3.0*,
URL = http://www.w3.org/hypertext/WWW/Markup/Markup.html/, 1995, pp41

Rothenberg J, *Ensuring the Longevity of Digital Documents*, Scientific American, January 1995, pp24-29

Smith KA, *Structure of the Visual Book*, Keith A Smith (pub), New York, Book No. 95, 1984

Sperberg-McQueen CM & Burnard L (eds), *Text Encoding Initiative: Guidelines for the Encoding and Interchange of Machine-Readable Texts*, Document Number TEI P1, Text Encoding Initiative, Chicago, Oxford, Draft Version 1.1, 1990

The Pitt Rivers Museum: A Souvenir Guide to the Collections, Pitt Rivers Museum Publication, United Kingdom, 1994

Thibodeau K, *Panel Session The World Scene*, Playing for Keeps Conference, Australian Archives, 1994, pp97-109

Unisys/CompuServe GIF Controversy,
URL=http://www.lpf.org/Patents/Gif/Gif.html, 1994

The Authors

Dr. Bob Jansen is a Principal research Scientist at the CSIRO Division of Information Technology specialising in electronic publishing and the reading of electronic information. He has many years experience in information retrieval having worked on many important projects including the prototype electronic Hansard at the Federal Parliamentary Library, the West Australian Department of Consumer Affairs, and the Victorian Ministry of Housing. Since joining CSIRO, Dr. Jansen has focussed on the application of artificial intelligence techniques to improve information indexing and retrieval. Dr. Jansen has designed IntelliText, a software system supporting the reading of electronic documents which he has applied, in conjunction with the CNRS Institut des Textes et Manuscrits Moderns in France, to representing the archive of James Joyce's Finnegans Wake. He organised and convened the Summit on Electronic Publishing in Sydney, 1994, is a founder member Interact CMC proposal and is currently a member of the Australian Vice-Chancellors' Standing Committee in Information Resources on Electronic Publishing and the Towards Federation 2001 Management of Material in Electronic Format Working Party. He remains an honorary member and guest researcher at CNRS ITEM.

Trevor Davis is the Executive Director of the National Printing Industry Training Council (New South Wales Division). He has spent 30 years in private enterprise organisations in Australia in positions with responsibility for human resources and in particular in roles specialising in education, training and development. Since 1988 he has had primary responsibility for the identification of training needs for all sectors of the printing and graphic arts industry which includes pulp and paper, publishing, stationery, binding, fibre board and the associated processes and technology. During this period he has witnessed one of the most significant technological changes to what was known as traditional printing resulting in the virtual elimination of some manufacturing processes which have been replaced by fully computerised pre-press designs and methods. The printing industry under his management have formed a strong association with the CSIRO's Division of Information Technology for the purposes of membership in a major consortium bidding for funding to establish a Co-operative Multimedia Centre and in the development of an application to the Board of Vocational Education and Training to establish a Printing and Publishing Centre in New South Wales upon which the industry can focus it's technological and resourcing advances. Trevor Davis' qualifications are in the area of adult education continuing his studies towards his master degree in business (human resource major).