Systematic approaches to long term digital resource management

Edmund Balnaves
School of Information Technologies
University of Sydney
Sydney NSW 2007
ejb@it.usyd.edu.au

Abstract

Digital only subscription is increasingly popular as a means of journal and book delivery among our major libraries. The advantages of digital delivery are apparent, but unlike traditional publications, digital subscriptions are commonly not housed within our national boundaries. With an increasing large proportion of book and journal subscriptions being digital only this presents an as yet unquantified risk to the collections of the major research and state libraries. At present very little attention is directed to the continuity of access to increasingly important research resources through periods of economic, social or military instability. This is a metaphor for long-term resource management on the Internet generally.

A model for managing the risks associated with these new directions must address both business risks of digital collection continuity and systems issues of content discovery, sharing and reuse. Escrow contracts are an established method to guarantee continuity of business when licensing business-critical software applications. The paper explores the example of low cost community driven resource sharing networks (the Gratisnet case study). It also discusses new approaches for Content Reuse developed by the author at University of Sydney in systematic methods for Content Reuse and Content Management Systems. A case for the establishment of digital escrow database at the community level is presented with an architecture that embraces both the business and systems issues of long-term management of the digital resource supply.

Keywords

Legal and policy issues; Content Reuse; Escrow Management of Digital Resources

COLLECTIONS IN TRANSITION

The rapid growth in digital publishing on the Internet is accompanied by an equal growth in digital publishing of academic journals. Online delivery of digital research resources has heralded a new era of research opportunities emerging from enhanced capabilities for information discovery and resource delivery. The strong user acceptance of digital journals, increasing e-publishing activity, and static library budgets, are all drivers for the adoption of digital-only delivery of library resources. Professional associations such as IEEE have established their own presence in the direct delivery of large document collections through their digital library. The consolidation of substantial collections with direct delivery has seen the gradual attrition of subscriptions to the traditional print format (Fox & Marchionini, 1998; Weiderhold, 1995).

Journal collections in our major research, state and national libraries form an important national research asset. Even outside major institutional research libraries, smaller research collections provide an important service - for instance, research libraries in hospitals provide important day-to-day support for medical professionals and research in health delivery.

Libraries themselves have historically served a dual role of supplying the immediate information needs of clients and the preservation of intellectual resources over extended periods of time. The transition to digital-only subscription has attendant risks that are substantially different from those of print asset management. This paper addresses those risks in the context of strategies for better long-term management of digital resources.

In medieval times, there was an obvious relationship between the many days of labour that went into copying a manuscript and the value of its long-term preservation. Historically, printed collections have survived not through the efforts of a single library collection but through the distributed collection building across countries and continents, assuring the preservation of intellectual resources over periods of time greater than a single decade or century. Copying of digital resources is quick, accomplished easily and results in a functionally identical copy of the original, to the dismay of traditional publishers. However, the ease with which digital content can be copied belies the long-term archival difficulties surrounding the digital content.

Digital content suppliers enhance the value of the resource itself by providing rich database resources backed by document delivery. These databases are characteristically located outside our national boundaries.

Where the library retains ownership of the printed journal, electronic resources can be on the licence terms of an annual subscription.

RISKS ASSOCIATED WITH DIGITAL RESOURCES

The move to subscription-based digital library collection building has associated with it inherent risks of technological availability and obsolescence. With digital assets already in outdated formats (for example the 8" floppy disk) libraries are often in no position financially to undertake the necessary fundamental research in areas of technological migration from different versions encoding, media storage or content delivery platform (Ekman, 2000; Phillips, 1998). However, the budgetary and research benefits for moving to digital collection building and digital subscriptions are such that small and large library services are in the process of making or have already made key strategic decisions to subscribe to journal publications in digital form only (Fox & Marchionini, 1998; Weiderhold, 1995).

Digital resources present both short and long-term risks to institutions. To understand these risks it is important to understand the transition from management of a physical resource to the management of an Information System. Business Continuity Planning (BCP) is a set of business strategies to provide assurance of continuing operation for the multi-faceted logistics of Information Systems operation. BCP comprises a risk analysis that lays the foundation for a Disaster Recovery Plan (DRP).

The short-term risks typically include the problems associated with application stability and change management, network access and operational continuity of the client and server technology infrastructure. Long-term risks are systemic to Information Technology, not the least of which is long-term certainty in resource delivery. The record of long-term persistence of Web-based resources is generally poor even over durations as short as five years (Lawrence *et al.*, 2001). With e-journal publishers experimenting in different content formats for content delivery, subscription comes with often unquantified attendant technology risks.

Risks associated with external suppliers can include business failure, failure to deliver supplier in an effective and timely manner, and discontinuation of service supply. While many elements of Information Systems architecture are interchangeable over time, some critical systems may have a single supplier, and service risks associated with these assets are typically managed through some form of escrow arrangement. The escrow is a contract between the supplier and the client giving assurance based on specified conditions of access to resources to allow continued service delivery independent of the original supplier. Traditional DRP strategies for business rarely need to address service continuity beyond time required for system replacement. The additional responsibility faced by libraries for long-term collection building present a different cost dynamic to DRP Traditionally in journal publishing, neither the author nor the publisher has borne the burden of archival management of digital resources. Arguably, the publisher is now principle archival source for their digital resources. However, where the publisher is also the distributor, libraries face precisely the risk of continuity of service supply from a single source. Publishers and distributors are not immune over the long term from economic hardship or disruption due to social factors beyond their control. In this context libraries face the risk of substantial collection loss.

SUMMARY OF RISKS

The risk management of traditional print collections revolves around:

a) The physical risks to the collection itself (fire, and other forms of destruction)
b) The risks of media deterioration over time.

The Figure 1 classifies the risks as they pertain in the short, medium and long term in the persistence and archival management of digital resources.

In the short term all Information Systems face the standard system continuity issues relating to hardware, networking and Operating Systems interruption. In addition to these risks, Digital subscriptions are increasing placed through suppliers of heterogenous collections of journals, with the resulting financial risk of duplicated journal subscriptions. In the medium term, issues of supplier continuity become prevalent, with the fragility of license contracts, risks of fee escalation, and through business transitions such as business takeovers, failures and bankruptcy administration. The ongoing management of digital collections may also require continuing accommodation to new media storage formats, content encoding methods, potentially with the costs of retrospective conversion of collection resources (where copyright allows).

In the long term, issues of media obsolescence and vulnerability increase in importance. Unlike print media, even partial damage to digital storage systems can compromise the entirety of the digital contents. This risk is ameliorated by the ease with which digital content can be copied (through backup systems and distributed databases). The obsolescence of software architectures and changes in software versions across platforms can

present problems in reviewing content that might be scarcely a decade old. Changes in encoded media formats and shifts in technological architectures for delivery can be profound. There is simply no experience in the management of digital resources for time spans exceeding 60 years.

The delivery of information systems infrastructure is increasingly dependant on many overlapping systems, vulnerable to failure in many ways not the least of which now is disruption and attention from would-be-terrorists(Longstaff, Chittister, Pethia, & Haimes, 2001).

SHORT TERM (0-5 years)	MEDIUM TERM (6-15 years)	LONG TERM (> 15 years)
System continuity. Duplication of subscriptions with different aggregate database providers. Heterogeneity of architectures to support Hardware, Networking and Operating System interruption. Business failure or ownership changes in key service and subscription suppliers.	*Short Term Risks plus:* Business failure of publishers (affecting subscription suppliers). Changes to media storage and content encoding. The cost of retrospective conversion of content into current delivery methods Licensing changes in content supply.	*Medium and Short Term Risks plus:* Inability to sustain subscription charges for retrospective material. National boycotts Proliferation of online resources Long-term content location and identification Long-term persistence of access Media longevity for archival digital content.

Figure 1 Risks associated with Digital Resources

For libraries, digital collection integrity depends on the long-term continuity of access to the suppliers Information Systems. The failure, for whatever, reason (networking failure, business failure, war, natural disaster) of the supplier to provide continuity of service can therefore compromise retrospectively the established collection.

Few libraries have significant resources for new software development, let alone retrospective archival content conversion. Some national efforts have seen the establishment of organizations specifically focused on the preservation issues, such as the Archaeological Data Service in the UK (Richards, 1997). The ADS survey of archaeological computer records of the now closed Newham Museum Archaeological Service revealed a 5% loss of data due to corruption of data stored on floppy disk. Of course, most research Institutions and online database service providers have established Business Continuity strategies directed to operational and environmental hazards. Recovery of data that is placed at hazard through lack of archival planning can also be very expensive (Chen, 2001). The trade-off between the costs of recovery against the costs of service disruption is an essential element of the risk analysis. This cost of service interruption is contingent on the nature of the organisation and its use of the digital resources, and the effect of service restrictions to ongoing operations. Typically libraries can tolerate disruption to services even over a few days through reliance on the established inter-library resource sharing networks, as demonstrated by the effective management of services during the bushfires in the Australian Capital Territory. As the period of investment in digital collection building increases, so does the exposure to financial loss caused by negotiation of alternate supply when the original supply fails.

APPROACHES TO RISK MANAGEMENT

The risk exposure associated with digital collections can be ameliorated in a number of ways. Among the emergent methods for assuring access to digital resources are the development of alternative distributed collections, market pressure by larger institutions on behalf of smaller institutions and standard approaches to escrow

Alternative suppliers and open Source Collections

The development of multiple digital database suppliers presents one form of risk mitigation. Large database vendors now offer document delivery of multiple journal collections. There are no substantial database vendors located within Australia offering such as service. "Open source" also digital libraries present an alternative delivery methodology for published journal articles. The examples of SPARC (Association of Research Libraries, 2001) and the Public Library of Science(Public Library of Science, 2001) clearly illustrate the level of concern shown by scientists and other academics for the free flow of information and the archival issues surrounding research publishing. They are an interesting challenge to the current regimen for academic journal publishing and as such are already being harnessed as alternative information systems research vehicles. They may also form an alternative digital archive resource for access to valuable research resources. While they do not represent a consistent collection replacement for purposes of digital collection building, the emergent open source collections do demonstrate the increasing ease with which substantial collections of digital resources can now be shared.

Market Pressure

Smaller to medium-sized libraries generally depend on the community pressure that can be applied in the case of a significant failure:

> If hundreds or thousands of libraries want access to a failed publisher's content, money will talk and a way will be found to provide access. The library community currently has several non-profit corporations that provide web access to old journals or that escrow and insure perpetual access to current e-journals. While it is not prudent to state that there is absolutely no risk involved with archival access to e-journals, as a practical matter the risk is minimal, and loss of access to content is unlikely as long as the present international higher education and library structure continues. (University of Texas, 2002)

That form of peer support may not be available if the service failure results from trade boycott, broader economic problems, or other trans-national externalities.

Digital Escrow

The realities of business risk were illustrated recently with the Chapter 11 filing of digital e-book supplier NetLibrary. In that case, OCLC, acting as an escrow agent, was prepared to release holdings to libraries in difficulty in CD-ROM (Letts & Walster, 2001). However, an escrow agreement is only as effective as it is periodically tested. That is, content placed in escrow must periodically be tested to ensure that it is usable – and this may well imply the availability not only of the individual resource itself but also the systems facilities to effectively retrieve the resource, and the runtime engines to effectively deliver the object to the client. The establishment of a multiplicity of escrow agreements with multiple different digital subscription suppliers has several limitations:

- Many libraries are not in a position of sufficient leverage to negotiate an agreement that would provide effective coverage
- There can be considerable costs in negotiating a specific satisfactory escrow arrangement
- Libraries may not be aware of the full spectrum of issues surrounding effective escrow coverage when signing agreements.

An escrow agreement is only as good as it is periodically tested. Given the multiplicity of potential journal suppliers, the validation of the content placed in escrow can be problematic.

Distributed digital repositories

Distributed database architectures for the dissemination and delivery of digital library objects have been evolving for some time. Paepcke *et al.* (1996) present a CORBA-based model for distributed management of queries among digital library services(Paepcke *et al.*, 1996). This was extended by Crespo and Garcia-Molina(1998) to a distributed model for digital library objects with the specific objective of long term persistence of objects (Crespo & Garcia-Molina, 1998). They examine the robust and secure transactional delivery of digital library objects, with particular attention on architectures for achieving the long-term persistence of this information. While this highlights the information system complexity of distributed library object management, it also provides an effective model for the distributed management of digital resources in the context of long-term persistence. In particular, they highlight issues relating to the:

- Distributed management of the resource to maintain persistence of the information over time
- Mechanisms for ensuring and validating the identity of digital objects.

The development technologies in these areas combined with the efforts by national institutions to build interoperability in their digital collections has seen the emergence of effective distributed models of digital resource sharing (Fox & Marchionini, 1998). Information retrieval architectures such as Z39.50 provide a

framework for the distributed interrogation of information resources, and this standard has recently been extended to incorporate the "digital library object".

Chen (2001) and others have suggested models for the appraisal and archival management of digital information. They identified the importance of effective information retrieval engines as part of the archival strategy.

The economics of establishing distributed resource sharing networks are continually improving. The GratisNet Service represents an interesting innovation in distributed collection management of journal resources among health libraries in Australia. Over 300 libraries have joined in a network to provide distributed intelligent-agent inter-library loan services optimized for lowest-cost provision of health journals that are otherwise expensively or poorly supported in traditional Inter-Library Loan services. They achieve this at an annual cost per library of $110 (National Resource Sharing Working Group, 2001). This model has been taken up by three other library networks, creating an integrated resource database comprising over 400 libraries. This service is not beginning to integrate digital resource access information in the service. Other large content clearing-houses already exist and provide an effective role for dissemination of information, often at very low cost. Another example, in the humanities, is the Electronic Cultural Atlas Initiative, a project to provide an international distributed resources indexing metadata for archaeological spatial and GIS data sets (Electronic Cultural Atlas Initiative, 2003).

A key strength of digital multimedia content is the ease with which it can be copied. While this presents dramatic challenges in the management of Intellectual Property rights protection, it equally holds possibilities for long-term preservation. The more widely content is distributed, the more likely to be that a usable copy will be preserved in the long term. Most published digital journal resources are in a highly standardized digital format such as PDF (Portable Data Format) that at least ameliorates the issues of software obsolesce in the delivery systems. The implementation of effective escrow clearinghouses should not present a fundamental technological or financial challenge.

The Yale Electronic Archive (YEA) is a joint project with the publisher Elsevier, directed to establishing a substantial archive for escrow and archival management of their digital journal collection. The YEA project demonstrates that collaboration between publishers and subscribers for the establishment of a digital archive is economically feasible (Yale University Library, 2002).

Content Management and Discovery

As a member of the Knowledge Management Research Group and University of Sydney School of Information Technologies, the author has explored to ways for effectively organising digital resource databases to enhance both metadata description and document discovery. A layered approach to content description, management, and publishing provides a framework for organisation of document collections. Such an approach is an essential element to the long-term management of large digital archives in a manner that will facilitate the complex annotation of rights and access metadata associated with a digital escrow repository of resources at the national level. A systematic approach to regeneration and reuse of digital multimedia requires a clear and comprehensive strategy for content composition, publishing runtime engine, information retrieval, and database management for loosely typed and unstructured data. The privileged focus on content reuse is what differentiates useful Content Management Systems (CMS) from other information management systems. Furthermore, the axiom of separation of content, structure and presentational form must be realized not only in the document ontology but also in the information system architecture. The Content Model for Reuse comprises a four-layered model aimed at facilitating content reuse across heterogeneous multimedia resources. A proof-of-concept prototype based on CMR was implemented using a web services framework and field-tested over a three-year period.

The development of a digital archive suitable as a long-term escrow resource must address requirements for continuous transformation of digital content over time in order to ensure that the resources are usable in current digital runtime servers and client readers/players. The four-layer model provides the framework for management of a content repository that can address multiple metadata description needs of a complex clearinghouse of documents while also providing the long-term framework for continuous transformation of encoding formats

The first layer of the Content Model for Reuse addresses the issue of content mark-up and capture of a content fragment. It applies encoding and metadata standards to separation of content fragmented in a semantically meaningful and economically useful way. The second layer addresses the ontological organisation of these fragments into meaningful document structures. The third layer is the point of mediation with the content author, in the organisationally contingent workflow management of the authoring process. The fourth and topmost layer addresses the delivery of content in generated form to various runtime engines in the delivery

of content to the targeted content consumers. Content versioning is an issue that spans all layers of the model. Versioning, an intrinsic element of reuse, appears at all layers of the model. Figure 2 presents this four-layer model in the context of a digital clearinghouse of documents.

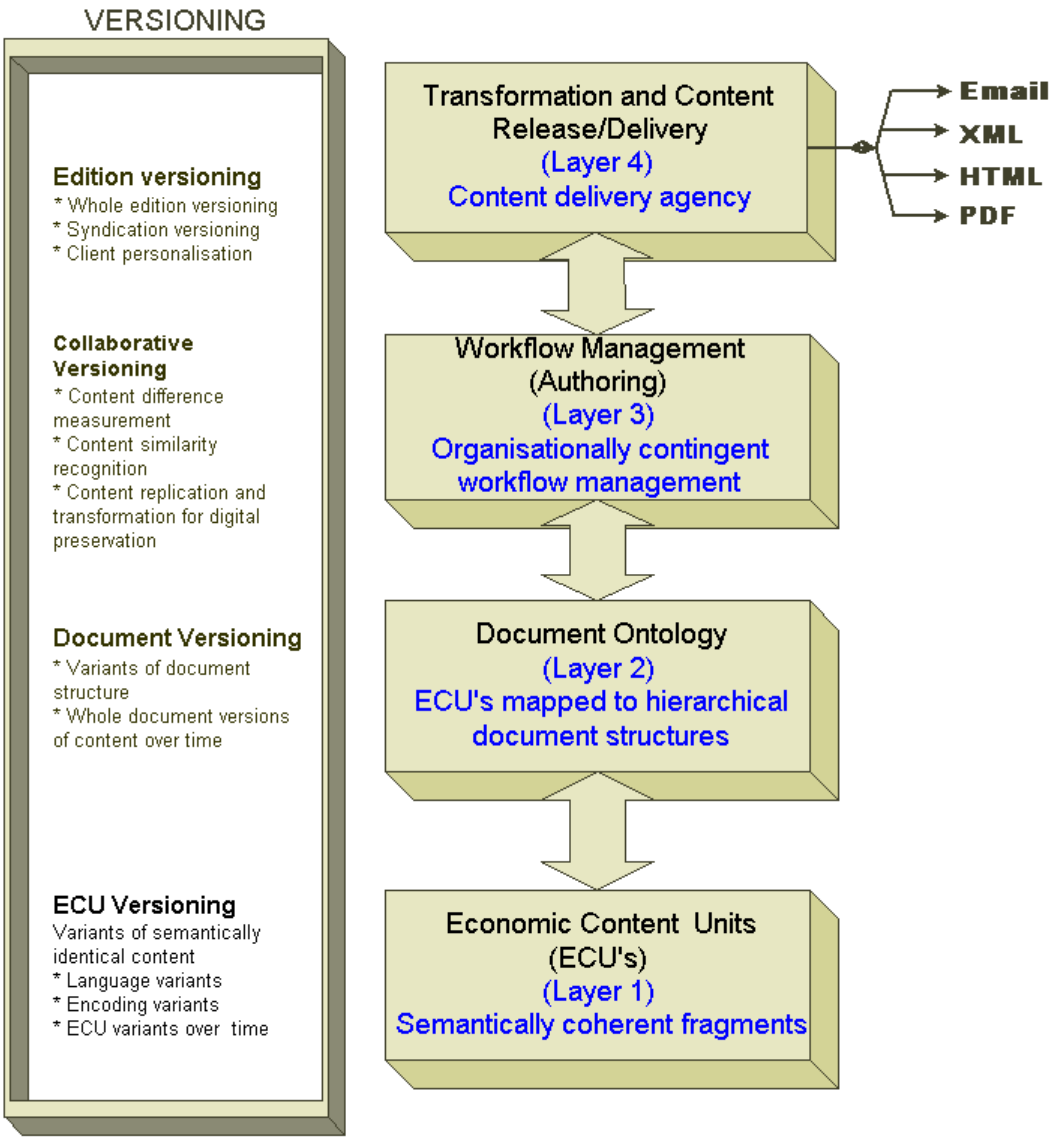

Figure 2 Clearinghouse management of Digital Resources

A PROPOSED MODEL FOR RISK MANAGEMENT OF DIGITAL CONTENT

A systematic solution is required to address the long-term risk associated with of evolving subscription-based digital collections. Figure 2 presents a proposed model for escrow arrangements combined with clearinghouse-style distributed digital archives as part of an overall BCP strategy for risk management digital resources in libraries. The model is framed around the establishment of distributed digital resource database accepting

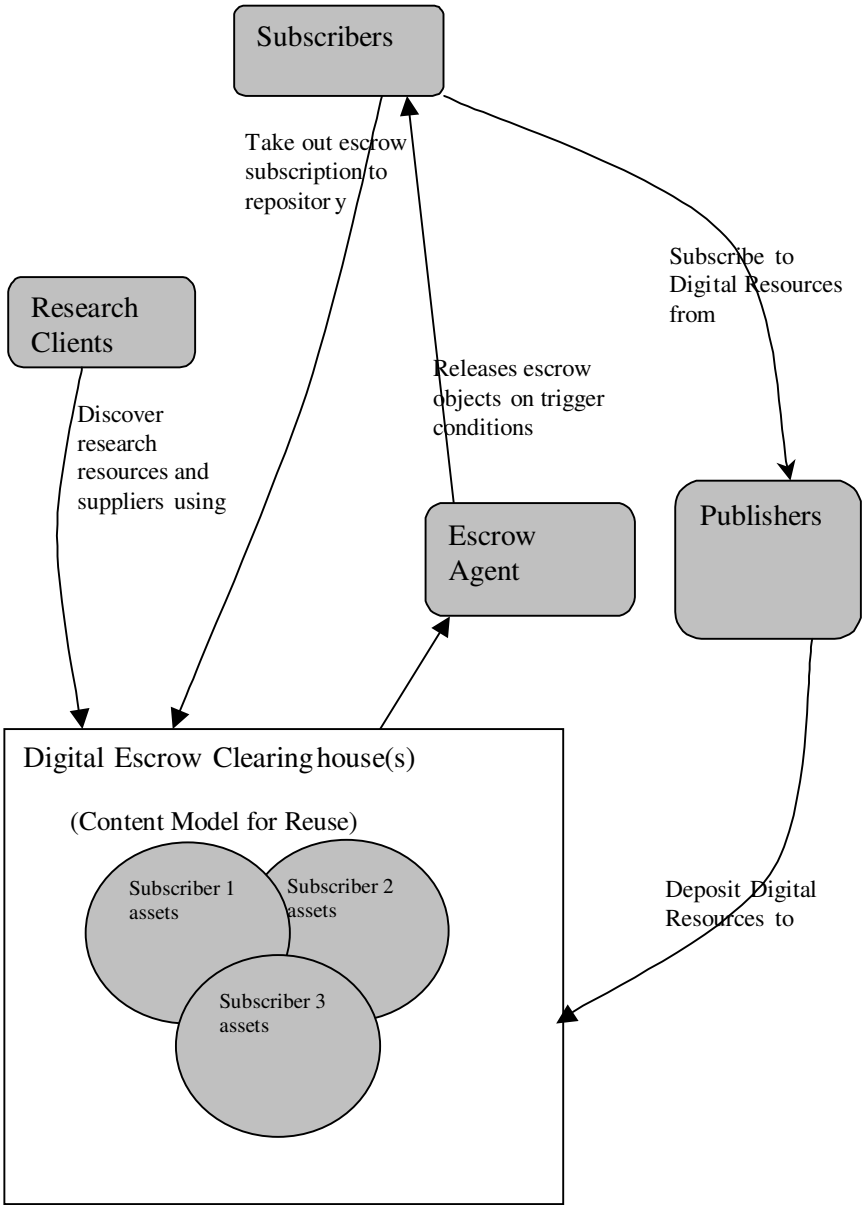

Figure 3 Clearinghouse management of Digital Resources

contributions from publishers and funded through escrow subscriptions from libraries. It proposes the integration of the escrow deposit as an ongoing element of the Document supply chain, allowing for greater long-term assurance for subscribers of collection continuity. Such an approach presents opportunities for improved document discover with referral to relevant document delivery agents. The resulting digital archive, as it matures, has the prospect of providing a rich research resource and a conduit to the originating content suppliers as the document delivery agents. Finally, the integration of the escrow deposit as part of the document delivery supply chain enhances the prospect that the escrow collection will be effective if its escrow release conditions are ever triggered.

To be effective, the clearinghouse must bring together three key distributed collection elements:

- Content identification and management. Distributed content identification servers would be required to satisfy the need for long-term archival management and discovery. An effective standard for the distributed catalogue management of digital content and associated meta-data would also be required. The framework must also support the ongoing management and continuous transformation over time of the digital assets.

- Distributed content discovery sharing with standards-based information retrieval architectures using either web-based Indexing or the Z39.50 distributed indexing.

- Content Escrow services. Digital escrow content servers should be central to the content identification and retrieval architectures. Such content servers could provide functionality in three ways:

    a) As a referral agency to the appropriate licensed or publishing server, where escrow arrangements have not been triggered.

    b) As a direct delivery agency where escrow conditions are triggered.

    c) As the deposit agency for digital content on a national or institutional grouping basis.

Without the co-operation of publishers and database vendors, an escrow approach is not achievable, and any legislation to enforce such co-operation is very likely to fail at national borders. The Yale Electronic Archives project demonstrates that escrow provision can be forthcoming, and can enhance the reputation of the supplier by providing continuity of service assurance to the client. The establishment of escrow clearinghouses may also offer new commercial opportunities for publishers and database vendors to market their collections and services, if the clearinghouse facilitated effective content discovery while still referring delivery to relevant suppliers.

If the escrow database is nurtured as a resource in itself, there is a greater probability that the clearinghouse will adopt new technologies at the same rate as they are taken up by authors and publishers. In particular, if the escrow agent is also a mediator for digital content discovery, there can be better confidence that it will be able to fulfil a role for long-term archival storage of its digital content in a manner that protects the rights of the content subscriber.

Libraries have demonstrated in the past very effective collaboration in building substantial distributed information resources. Similar effort should now be applied to establishing an agreed framework for escrow management of digital subscription collections on an escrow basis to provide a system for long term Digital Resource collection continuity.

CONCLUSION

Collection building of digital journals entails risks that are sometimes not recognized in typical e-journal subscription contracts. The profound added value to research represented by online subscription delivery of digital library resources is likely to see the accelerated adoption of digital-only subscription. This may well be associated with concentration of these resources through single rich online database vendors. In the long term it will only be through effective collaboration with publishers that the establishment of effective escrow management approaches to the collection integrity of Digital Libraries can be ensured. The long-term analysis of these risks should not exclude business failures due to economic misfortune, arbitrary changes at the business or governmental level in relation to ongoing access, and catastrophic events beyond the control of any single institution. This paper presents a model for managing the risk associated with digital resources based on a clearinghouse systems model supporting escrow contracts with publishers and online database vendors. Such a model may also be applicable to the management of published digital resources in other contexts.

Approaches for ameliorating these risks should be on the agenda of all organizations that are moving to digital-only delivery of their content, including digital libraries, media organizations and those responsible for archival document management in business.

REFERENCES

Association of Research Libraries. (2001). *Sparc. The Scholarly Publishing and Academic Resources Coalition.* Retrieved 20/8/2001, from http://www.arl.org/sparc/

Chen, S.-S. (2001). The Paradox of Digital Preservation. *Computer, 34*(3), 24-29.

Crespo, A., & Garcia-Molina, H. (1998). *Archival Storage for Digital Libraries*. Paper presented at the Proceedings of the Third ACM International Conference on Digital Libraries, Pittsburgh, Pa. 69 - 78.

Ekman, R. H. (2000). Can Libraries of Digital Materials Last Forever? *Change, 32*(2), 23-35.

Electronic Cultural Atlas Initiative. (2003). *Ecai Metadata Clearinghouse*. Retrieved 9/8/2003, from http://ecai.org/tech/mdch.html

Fox, E. A., & Marchionini, G. (1998). Toward a Worldwide Digital Library. *Communications of the ACM, 41*(4), 29-32.

Lawrence, S., Pennock, D. M., Rovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A., et al. (2001). Persistence of Web References in Scientific Research. *Computer, 34*(2), 26-31.

Letts, M., & Walster, M. (2001, 6/11/2001). *Oclc Offers to Bail out Netlibrary*. Retrieved 5/6/2003, from http://www.seyboldreports.com/ebooks/features/011116-oclc.html

Longstaff, T. A., Chittister, C., Pethia, R., & Haimes, Y. (2001). Are We Forgetting the Risks of Information Technology. *Computer, 33*(12), 43-51.

National Resource Sharing Working Group. (2001). *Interlibrary Load and Document Delivery Benchmark Study*. Retrieved 9/8/2003, 2003, from http://www.nla.gov.au/initiatives/nrswg/illdd_rpt.pdf

Paepcke, A., Cousins, S. B., Garcia-Molina, H., Hassan, S. W., Ketchpel, S. P., Röscheisen, M., et al. (1996). Using Distributed Objects for Digital Library Interoperability. *Computer, 29*(5), 61 -68.

Phillips, M. E. (1998). Tomorrow's Incunabula: Preservation of Internet Publications. *Lasie, 29*, 5-10.

Public Library of Science. (2001). *Public Library of Science Website*. Retrieved 4/8/2003, from http://www.publiclibraryofscience.org/

Richards, J. D. (1997). Preservation and Re-Use of Digital Data: The Role of the Archaeology Data Service. *Antiquity, 71*(274), 1057-1059.

University of Texas. (2002). *E-Journals: Frequently Asked Questions*. Retrieved 24/5/2002, from http://www.lib.utexas.edu/admin/cird/efaq2.html

Weiderhold, G. (1995). Digital Libraries, Value and Productivity. *Communications of the ACM, 38*(4), 85-97.

Yale University Library. (2002). *Yea: The Yale Electronic Archive - One Year of Progress: Report on the Digital Preservation Planning Project*. New Haven, CT: Yale.