



The University of Sydney

Systematic Content Reuse in a Web Services Environment: A Model-Based Approach

Technical Report Number 554

July 2004

Edmund Balnaves and Sanjay Chawla

ISBN 1 86487 651 4

**School of Information Technologies
University of Sydney NSW 2006**

Systematic Content Reuse in a Web Services Environment: A Model-Based Approach

EDMUND BALNAVES and SANJAY CHAWLA¹

University of Sydney

A systematic approach to regeneration and reuse of digital multimedia requires a clear and comprehensive strategy on content composition, publishing runtime engine, information retrieval, and database management for loosely typed and unstructured data. In this paper a model of content reuse in Content Management System (CMS) across heterogenous multimedia resources is proposed. It is argued that the privileged focus on content reuse is what differentiates CMS from other information management systems. Furthermore the axiom of separation of content, structure and presentational form must be realised not only in a document ontology but also in the information system architecture. Content reuse on the Internet is characterised by processes of continuous transformation, and the role of the CMS may be either as a repository of content or as a facilitator in the transformation of transient content resources. A four-layer model for systematic content reuse is proposed and described in terms of a Web Services implementation. The results of a substantial field trial conducted over a three-year

Edmund Balnaves (ejb@it.usyd.edu.au) and Sanjay Chawla (chawla@it.usyd.edu.au)
University of Sydney
Madsen Building F09
New South Wales 2006
Australia

period are presented. Measures for information currency and content reuse are defined and applied in the context of the field trial with results that indicate both better content currency over time and increased content reuse.

Classification

H.3.4 Systems and Software; H.3.5 On-line Information Services; C.2.4 Distributed Systems; H.1.1 Systems and Information Theory

Keywords

Ontology; Web-based services; Content Management

1. Introduction

The Internet has joined the mass media stable as a popular form for expression and dissemination of information. More than ever people are both content creators and consumers of digital media – of both text and multimedia. In this context, the effective management of published resources becomes essential. The rapid emergence of the Content Management System (CMS) as a new class of enterprise software is a clear indication that many content publishers are placing a higher level of importance to the management of their web-based communication [Byrne, 2003; Dalton, Manning, & Gardiner, 2001].

The separation of content from style and structure has been an axiom underlying CMS development from their inception as a new class of enterprise software. CMS

products have focused on the primary generation and maintenance of website content, and their origins lie in simple content delivery mechanisms for web publishing [Fraternali, 1999; Reynolds & Kaur, 2000; Yeh, Chang, & Oyang, 2000], with a focus on effective workflow management of the production process. Content separation from style is achieved either through template-based approaches to content publication, or more recently through use of XSL (Extensible Style Sheet) templates interacting with XML document formats. Dynamic generation of personalised websites using XSL from XML content sources is also gaining popularity. While many CMS products make claims regarding content reuse, the extent of practical reuse is often limited [Stahl, 2003]. The task of Content Management does not necessarily end at the workflow organisation of the authoring process. Publishers also have a requirement to track the copyright clearance levels of content they use. Similarly, from an archivist's point of view, the management of digital content presents the dilemma of multiple copyright ownership of a given content unit [Jewitt, 2000], and the fact that much relevant audio content is not in the public domain. Aspects of the social and copyright issues of multimedia publishing of academic journals are equally applicable to the generation and regeneration of multimedia content in an Internet framework [Cox, 1998].

The management of online educational curricula is perhaps the leading area in which the challenges of content reuse are being actively engaged. There are manifest advantages to the effective reuse of curriculum objects in common teaching areas. The cost of development of curriculum content that is designed for effective visual communication in an online environment gives a clear imperative to effective reuse of curriculum materials. The framework for such content reuse is emerging with the creation of substantial

metadata ontologies, such as SCORM and IEEE LOM [Candler & Andrews, 1999; Colbert, Peltason, Fricke, & Sanderson, 1997].

Most websites and multimedia resources go through several generations of design makeover. Considerable material and human cost goes into the one-off development of multimedia resources. This investment can languish when issues of technological obsolescence intervene, or if the information itself is not current. Much of the content created for websites is discarded between generational changes.

The importance of effective content reuse is manifest in the rapid obsolescence of both technology and content assets, with longevity which is as little as eighteen months [M.E.S.O., 1995]. Such a rate of obsolescence demands the building of content reuse systems that allow not only the repurposing of content, but also the effective conversion to new runtime environments as part of a broader digital asset architecture. This difficulty is compounded by the need to publish in multiple formats. For example, an encyclopaedia may be published in print, CDROM and web-based media. Each of these has different information systems requirement to create the published result, often resulting in separate departments creating runtime, design and content elements in isolation[Norrie & Signer, 2003].

The emergence of the Content Management System has certainly addressed some of the most problematic issues of web authoring. Functionally rich Content Management Systems are now available to address primary issues of content authoring, version management, editorial workflow control and transformation of content for its target environment. The proliferation of CMS products to the market in recent years reflects the heavy demands for editorial management of a website.

1.1. Research Questions

What is lacking in current CMS models is a paradigm for content management specifically directed to systematic content reuse. Current paradigms, while they have strong foundations in XML, need to be extended to reflect the ontological and systems issues that also affect content reuse. Current industry products can claim a rich set of functions as they position themselves in the Enterprise Middleware market, but attention is needed to the content model driving future CMS development. This paper is directed to forming a richer theoretical foundation to the direction of CMS development.

An effective model for content reuse must deal with the difficulties of heterogenous content resources and the need for a flexible document object model. The model should be effective to the extent that it could significantly empower system analysts and software developers in their development of CMS. It also offers a framework for consideration of content regeneration and reuse in an information systems context that facilitates the demands for continuous short and long term translation of content to different media, formats and services.

The principle research question for this paper is formulated as follows: What approach to content management will allow effective reuse of content across heterogenous multimedia resources while tracking versional changes across language and time? The objective of this research is to:

1. Address issues of inconsistency in the presentation of the same semantic information across different media.

2. Ameliorate the economic costs of content recapture during generational changes of content.

3. Reduce the opportunity cost represented by the failure to effectively deploy content in all relevant content forms (for example through syndication).

Content management systems have achieved enterprise recognition in a space previously occupied by three classes of product:

- Knowledge Management Systems
- Document Management Systems
- Learning Management Systems

The Content Management System is discussed in the following sections in the context of these affiliated enterprise systems. The CMS is compared and contrasted with these systems.

1.3 Content Management Systems

The term “content management” is a succinct description of the demands of managing multimedia resources. The term “content” is sufficiently general to describe any digital form, media or object (“content” in the sense of a descriptive label of items that are “contained within”). It also implies the components of meta-information storage (in the sense of the “table of contents”). The term “management” can be extended to imply not only the operational management of the publishing exercise, but also the long term organisational management of the content store, the effective use of content through regeneration, repurposing and syndication, and the long term archival administration of content.

There are obvious parallels between content reuse and software reuse. Boiko [2002] draws parallels with software components and content components, choosing an object-oriented paradigm for representation of content fragments. Fraser [2002] presents a view of content reuse that draws more heavily on the object oriented paradigms, referring to content as a "component" and particular publishing instances as "specialization" of the component.. When dealing with unstructured "content" the object that results often cannot be resolved into a simple "class" and the issues involved with the organisation of content management, authoring, selection and reuse of unstructured collections of text and multimedia resources cannot satisfactorily be characterised by typifying all content reuse as a form of "specialisation". Fraternali presents high-level modelling approaches as a means of abstraction of design at the level of the website as a whole, drawing strongly on UML-style concepts [Fraternali & Paolo, 2000]. Rockley [2003] classifies "content reuse" into two methods, "opportunistic reuse" (akin to Kruegers "code scavenging" [Krueger, 1992]) where content is retrieved and reused in a "cut and paste" manner, while "systematic reuse" is achieved where the content management system automatically manages the reuse of content across target environments - what Rockley calls auto-population. Her content management view is modelled on the document management "life cycle" approach to the history of the content. Both of these reflect an information systems view of content rather than a "component" view. Nevertheless, both research into software reuse and its practical lessons will remain a constant point of reference in developing the information systems for content reuse.

There is now sufficient published discussion of the Content Management systems to discern key functional elements in common to the systems view of content management.

Abbey, Ellis, Suh, and Thiemecke [2002] characterise the CMS in terms of "workflow", "transformation" in the context of an "asset" based strategy. Fraser [2002] discuss the design of a CMS in terms of "version control", "workflow" and "personalisation".

Ceri, Fraternali and Paraboschi [1999] propose WebML as a modelling methodology for the management of complex data-intensive websites. They define an XML methodology for generalising the structures of the website and the interfaces to data sources, in the context potentially of the use of Case-like tools in a formal development environment for website design. They stress composition primitives not unlike the content units discussed above but with an integral treatment of compositional and presentational aspects of the content management. WebML is server-centric in the sense that it does not recognise multiple presentation or distribution channels for the content, and has to date been principally directed to the delivery of text-based informational content [Ceri, Fraternali, & Bongio, 2000; Ceri, Fraternali, & Paraboschi, 1999].

Fraternali & Paolo [2000] explore the visualisation and modelling of websites at the "macro" level. A content management model, to achieve the goal of both content reuse and content preservation, will necessarily provide visual presentation of the structure of information represented in a site

1.4 Knowledge Management Systems

Knowledge Management is itself a convergent discipline that has seen the integration of expert systems and Artificial Intelligence enhanced information classification and retrieval. Much has been written on the strategic importance of knowledge within an

organisation and the processes needed to describe and disseminate existing knowledge and to elicit new knowledge [Liebowitz & Wilcox, 1997]. Knowledge management systems are *organisationally* focussed and *problem* oriented [Liebowitz & Wilcox, 1997] and must often deal with the information retrieval and the classification of loosely structured text documents, particularly in the context of Internet and other Search Engines.

Knowledge management shares several common causes with Content Management. Knowledge management has a particular focus on the effective identification (automated or manually) of semantic metadata to facilitate knowledge discovery. It is also essentially directed to knowledge reuse: the economic cost of encoding the knowledge resources of an organisation is warranted by the strategic benefit yielded by the efficient application of these knowledge resources in other contexts.

1.5 Document Management Systems

Document Management Systems generally focus on the management of a document in its final form, or digitised in a form that retains reasonable legal and presentational fidelity with the original document. There is considerable current focus on establishing the legal standing of such documents [Wilson, 1997]. In some cases this will include a focus on the archival issues of document management (for legislative or other reasons), to facilitating the cost-efficient workflow management of large volumes of documents, with demonstrable efficacy [Wright, 1996].

Wilkinson [1998] provides an overview of the nature, use and control of documents. He characterises the document as a *message* [Wilkinson, 1998]. He situates Document

Management Systems in the high-context role of business enterprise document management. Wilkinson also differentiates the “document” as a finished item from the workflow process of drafting and preparation and characterises the message as comprising of *content*, *structure* and *metadata*. Wilkinson also explores technologies that support document description, production, delivery, publication, discovery and removal – a document “life cycle.”

Document Management can serve to enhance document discovery in order to maximise Document utility for reuse in the organisation [Song, Clayton, & Johnson, 2002]. In this case, the overlap with Library Systems and with Knowledge Management is clear. The deployment of technology in the construction of Document Management Systems will naturally draw on concepts of workflow management and knowledge management. However, Document Management Systems primarily focus on the commercial document assets of a particular institution [Wilkinson, 1998], as distinct from the heterogenous mix of documents managed in a Library context, and are often concerned with the management of the full “life cycle” of the document from its inception to its archiving or destruction.

1.6 Learning Management Systems

A particular class of Content Management centres on Distance Education and Online Learning. The US Defence Department has funded the “Advanced Distributed Learning Initiative” (*ADL*) for the specific purpose toward the specific goals of “content reusability, accessibility, durability and interoperability [Advanced Distributed Learning., 2001].

Their content model, Sharable Content Object Reference Model (SCORM), is one of the most recent, and comprehensive, ontologies for describing curriculum resources.

A consistent theme among the LOM ontologies is the explicit recognition of content fragmentation. The IEEE LOM refers to this granularity of content as the "aggregation level"[IEEE, 2002].

1.7 Role of the CMS

It is tempting to categorise the Content Management System with Document Management Systems. Most CMS products have at least some element of workflow management. The life cycle of web-based content is in some respects similar to that of the standard print document, and for this reason it is tempting to categorise the Content Management System with Document Management Systems. While there is commonality in the functions addressed by Document Management and Content Management, particularly in their elements of workflow and multimedia document management, there are also clear differences. In the semiotic treatment of content, however, the Document Management System focuses on the management of the content in its final expression. In this respect, the CMS is akin to the Digital Library System. The emphasis of the Document Management System is archival management and reuse of content that preserves the essential elements of its authoritative final form. The level of content treatment, that is, the "granularity" of the content, is generally that of the "whole" document. The Information Retrieval focus is on discovery of the whole document and preserving the archival fidelity of the original document. This can be contrasted against Content Management and Knowledge management, whose focus is on

the substance of the communication and the provision of efficient systems for content regeneration and reuse in variant forms and contexts. The concern of knowledge management is the effective semantic discovery of useful “pieces” of knowledge. The concern of content management is the efficient reuse and regeneration of content, sometimes in a highly dynamic environment. While there are common challenges to document management in all these contexts, the particular challenges of each are founded on this different focus. The Document Management System has a particular focus on the integrity of the management of a document through its life cycle. This contrasts with the Content Management System that may be tasked with the regeneration of content in different forms and the distribution of content in document structures through syndication.

Content Management Systems can be considered most akin to Knowledge Management Systems, with a key differential focus in the manner of reuse. This differential focus explains the utility value of Content Management Systems in News organisations and for the management of dynamic, complex, websites and media systems [Heitmann, 1999].

The complexity of managing content in this context, and therefore the complexity in building systems for this specific focus, lies in the hybrid nature of the systems:

“Modern Web applications are conveniently described as a hybrid between a hypermedia ... and an information system.” [Fraternali, 1999, p.228].

2. Content Model for Reuse (CMR)

This Section presents a conceptual model directed to achieving a coherent approach to content regeneration and reuse. The model is organised as a layered presentation, with each layer building in complexity on the underlying framework provided by the lower layer. It aims to direct the mind of the software architect, implementer and publisher toward the long-term benefits of content reuse and regeneration in the context of the Content Management Systems that they deploy. The purpose of each layer is to present the minimally discreet set of functions that could be "substituted" by different standards encompassing a given layer. In particular, the objective of the model is to provide a systematic approach to content reuse.

A good conceptual model can assist software developers by providing a clear framework for selection of technological approaches at each layer and can encourage the use of standards-based interfaces. Such a model can serve as a basis to guide further development of systems in this area and as a benchmark for judging the functional completeness of products. As the importance of effective content reuse is elevated by the growing value of knowledge as an organisational asset (particularly for publishers, business consultants and educational institutions), organisational planners are faced with the dilemma of effective selection and development of suitable architectures for managing this asset. This process can involve the interaction between business planners and technical specialists. In many organisations content authoring is dispersed, and its capture cuts across lines of business and organisational layers. This selection process can be facilitated through the use of a good functional model for Content Management

informed by relevant standards. Privileging content reuse rather than the explicit feature set of the CMS is the most important characteristic of this model.

This paper describes the content "objects" that exist within content "ontology" as *content units*. This is substantially similar to the "fragments" of the ZyX model [Boll & Klas, 2001] and the Digital Item Declaration of MPEG [Burnett, van de Walle, Hill, Bormans, & Pereira, 2003] "components" [Boiko, 2002], "shredding XML" [Udell, 2003] and "chunks" [Byrne, 2003]. The preference for the term "content units" is to give emphasis to the economic nature of the reuse purpose, and to explicitly distinguish this from content models directed principally to representation of the final content form.

What follows is formal model for content reuse as a four-layer architecture. The lowest layers (one and two) address the content primitives – the Economic Content Unit and the ontological model for document organisation. The third and fourth layers address the public interfaces of the model to the content author and the content consumer. The first layer 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 situationally 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 situationally contingent 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, applies to all layers of the model. This model is expressed

hierarchically in Figure 1 below. This layered approach is applied in field test and evaluated in Section 3 below.

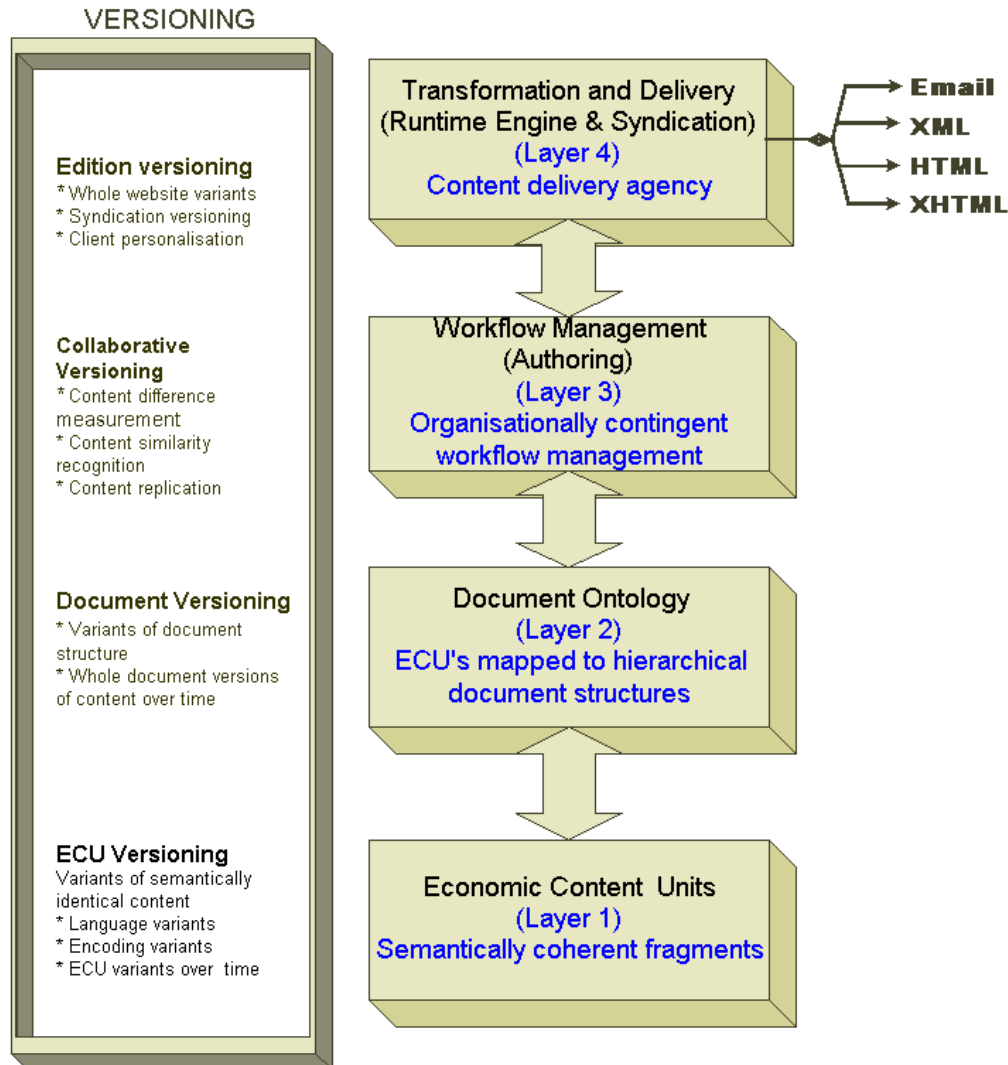


Figure 1: CMR: A model for Content Management, Regeneration and Reuse. Four layers of increasing complexity defining the management and versioning of content from its primitive form through to delivery.

The Content Model for Reuse (CMR) is a model for an information system. Content reuse requires more than a comprehensive semantic description framework (such as RDF or SCORM). The following sections provide high-level definitions of the model in terms of Web Services and a document ontology expressed in XML for each of the layers

of the model. Web Services provide an exemplary discovery and software reuse infrastructure for defining application logic accessible over Internet standards. As such they are a useful means of expressing at a high level an instance of the CMR model. These web services functions are accompanied by document definitions expressed in XML for each of the layers of the model. Other interfaces are also presented for purposes of authoring integration and workflow management, particularly at the third layer of the model, resulting in the following information systems model described in Figure 2 below.

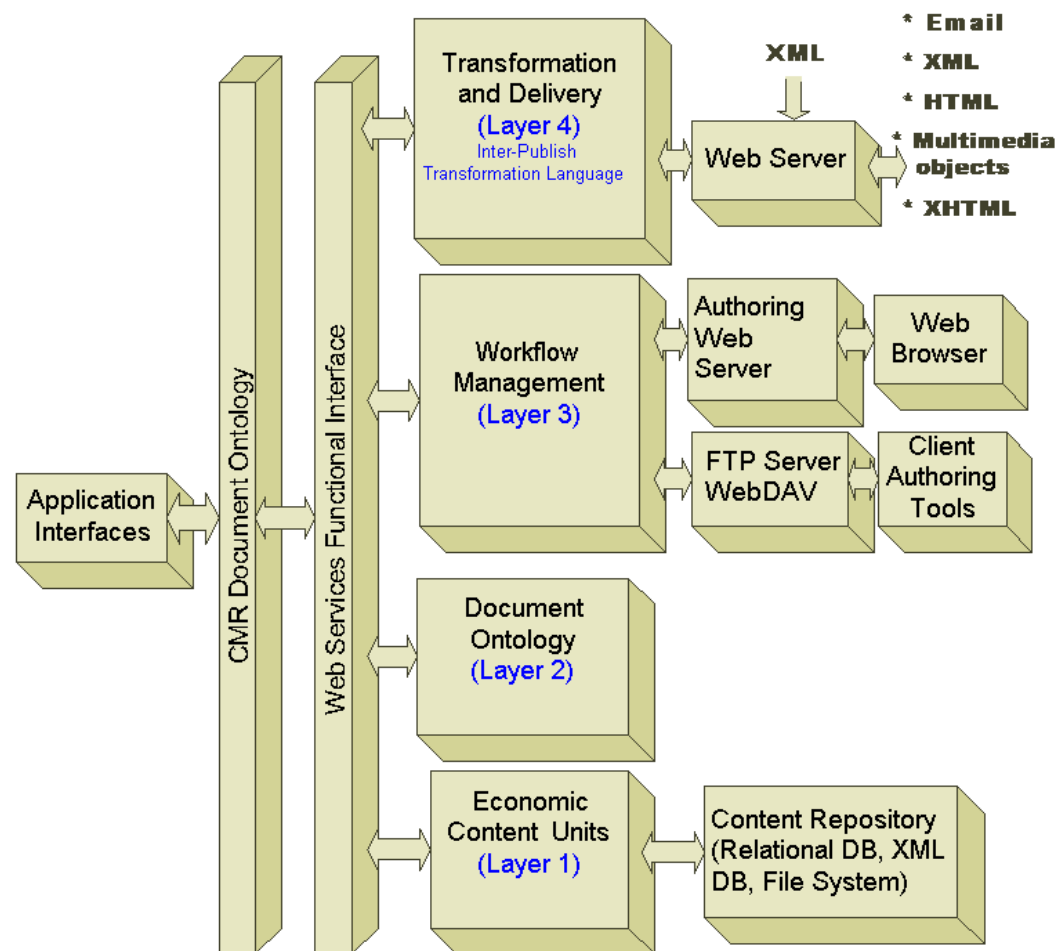


Figure 2: The CMR as an information system implemented as a CMS. The layers of the model may be implemented with a range of different technologies. This figure presents the application of the model in a publishing environment that deploys web services for content deployment at each layer of the model, a relational data store for content storage and implementation of Internet based content authoring and delivery components.

2.1 Web Services

Web Services are a key initiative of the W3C driven by very active industry-driven development of standards. While the standard is still in transition, the opportunities it offers have seen very rapid emergence of client-side and server-side development tools and a body of example implementations. The term "Web Services" has a range of meanings from the very general to the functionally specific. Lu, Dong, & Fotouhi [2002] describe "web services as "web sites that do not merely provide static information but allow one to effect some action or change in the world". However, by "Web Services", this paper is addressing that set of protocols called "Web Services" which enable the discovery and integration of business functions (for use by applications) and accessible through the Internet in a manner that Modularises and encapsulates well-defined standard interfaces, can be managed and run locally or remotely, can be ported over the Internet or intranet using standard protocols above TCP/IP (and through other delivery channels), can be discovered from central registries in a manner which fully describes their service, and exhibits "plug-and-play" characteristics at the client level [Chung, Lin, & Mathieu, 2003; Geng, Gopal, Ramesh, & Whinston, 2003].

Web services present a good example of the types of reuse and discoverability that should be typified by content reuse systems. Web services interfaces will be described below for each layer of the CMR definition. An implementation of this web services interface is an important element of the "Inter-Publish" prototype derived from the CMR model.

The web services definitions and the document ontology supporting these services are defined in full at: <http://www.usyd.edu.au/~ejb/inter-publish.wsdl> (a Web Services Definition Language description of the functions described below) and

<http://www.usyd.edu.au/~ejb/cmr.xsd> (an XML document schema describing the document exchange formats at each layer of the model).

2.2 Layer 1. Economic Content Unit (ECU)

The Economic Content Unit is the content primitive for the model. It describes a content resource (complete or a fragment) managed as a reuse object at the degree of fragmentation with is both semantically meaningful and economically effective for reuse. Integral to the ECU are the content itself and encoding and semantic information pertaining to the content. This layer reflects the following principals:

Content reuse requires the fragmentation of content in a manner that allows the separation of content from its final presentational form.

Content discovery, and semantic information that facilitates this discovery, is an essential element of content reuse.

Fragmentation of content is essential to achieve content reuse, although the granularity of this fragmentation is contingent on the reuse objectives. The results of this fragmentation are *content units*. The level of fragmentation will characterise resource considerations for the ongoing management of the content, leading to the concept of the *economic content unit*, being the optimal level of content fragmentation for the ongoing management of content in a particular context.

The motivation for such fragmentation is content reuse. This content reuse at the simplest level may be expressed in terms of the reuse of navigational and stylistic constructs in the generation of websites for purposes of consistency of presentation. Content reuse may be directed to the regeneration of content at points of generational

change of websites, or syndication of content to other agencies in an acceptable form. Perhaps the most complex exposition of the need for such content fragmentation is described by [Norrie & Signer, 2003] in their discussion of digitally augmented paper and the cross-integration of heterogeneous multimedia forms. Norrie & Singer [2003] discuss the granular dissection of content toward producing consistent content creation in multimedia forms (such as paper and CDROM delivery) and the need for a level of granularity in the content to allow cross-integration between media forms – such as from digitally augmented paper back to digital media resources such as CDROM or web. The fragmentation of content necessarily entails the management of digital encoding information and metadata concerning the content and its placement relative to other content fragments. The metadata description of content addresses both the description of the content format and the encapsulation of the content with sufficient additional information to understand the substance of the content in its context with other content units.

The separation of content from its presentational style has been the explicit goal of descriptive text mark-up technologies such as SGML and XML. Published digital content may comprise a heterogeneous mixture of content encoding and mark-up strategies applied to the generation of content in a final form. Fragmentation of content across different multimedia formats offers some of the benefits of content separation from style that has so richly enhanced the management of text. In some cases this might imply the separate management of fragments of XML (for example a product description across language boundaries deployed in different XML objects). Similarly a digitally

encoded image might be used in several different resolutions across a single website (for example a company logo) but with a required level of consistency in presentation.

The objective of this model for content management is to retain importance of the internationality expressed by the publishing of content in its final form while disaggregating the content in a manner that enhances its deployment in the context of other published content. Content capture is a non-trivial process in a heterogeneous media-authoring environment. The consolidation of content into a single database framework will invariably require the interaction of a range of content formats, authoring tools and content mark-up systems. With content reuse as an objective, the first prerequisite for effective content reuse is the decomposition of content into elemental Economic Content Units (the ECUs) that can be reassembled in different styles and presentational forms. The lower the level of disaggregation, the more complex the modelling required to re-aggregate this content into another form. This essentially *economic* consideration leads to the *economic content unit*. The ECU represents the smallest fragment of content that is economically valuable and semantically meaningful to reassemble with presentation templates into a final form. The disaggregated ECU's are not orphans. They belong to one or more document structures. The information pertaining to ontological structure of binding ECUs to form semantically meaningful documents is the subject of the second layer of the model.

2.2.1 Definition

The economic content unit layer (layer 1 of the content model) describes a collection of multimedia ECU entities. Whether this collection is retained in a database or on a file system (or in a mixture of both), the ECU requires some form of unique identification.

Encoding metadata associated with the ECU (including, with text resources, the character encoding and language characteristics, contextual information such as the DTD reference for XML fragments) can then be attributed to this identifier, as can semantic metadata (to facilitate content classification, discovery and reuse) and rights metadata (for purposes of rights management and intellectual property). Usage metadata (to track usage of a particular ECU over time and over place of deployment) is also essential for purposes of empirical measurement of content reuse.

An ECU might itself represent an entire document or a fragment of HTML, a branch of an XML tree or a particular multimedia fragment (eg a GIF or JPG image, video segment, audio segment). The formal definition of the ECU appears in *Figure 3* below.

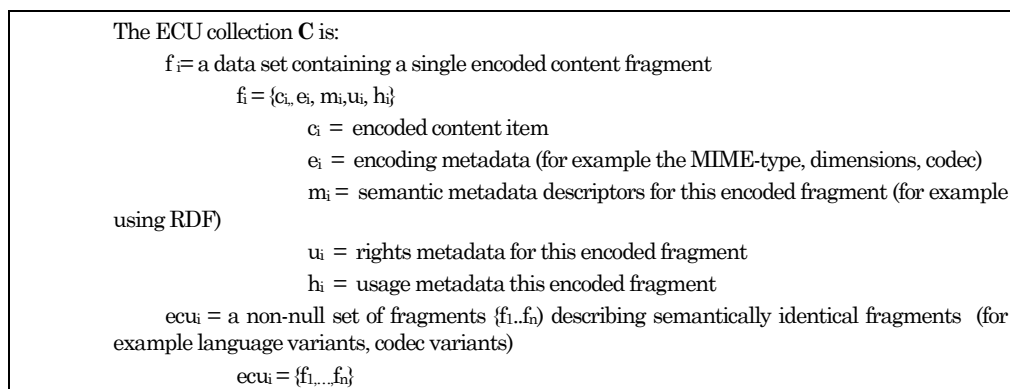


Figure 3: Definition of the ECU

Content editing and publishing functions may, at the lowest level, address a particular ECU encoded fragment f_i . This could include a single image, for a video clip at a particular encoding level, a single cell in a column of an SQL table (eg email address), a text fragment in a particular language encoded form.

2.2.2 Interfaces

This layer of the content model focuses on the integrative aspects of managing a heterogenous collection of ECUs. While public interfaces would not be expected, web services functions for the low level exchange between content services are desirable. Such content exchange requires the encapsulation of:

- (1) The digitally encoded content itself, or the referencing of content which permanently resides in external content sources
- (2) The metadata description of the content in terms of its encoding
- (3) The metadata description of the content in terms of its context in the wider document ontology (see layer 2)
- (4) Any rights, ownership, usage and descriptive metadata relevant to this content.

2.2.3 Web Services Functions

The principal application of a Web Services interface at this layer of the model is to provide a framework for information exchange at a low level with other CMS products. The following Web Services functions provide a low-level primitive for content exchange between CMS products. The *cmr_Selectionrule* (see Figure 5 below) provides an XML schema definition for static or dynamic selection of content at the ECU layer (for instance, for low-level content replication between CMS products).

Function	Request	Response	Purpose
GetECU	cmr_Selectionrule	cmr_Eculist	Fetch (optionally with lock) an ECU
UpdateECU	cmr_Eculist	cmr_Eculist	Create/update changes to an ECU

(releasing locks)			
ReleaseECU	cmr_Selectionrule	cmr_Eculist	Release a list of ECU entries

Table 1: Web Services functions at layer 1

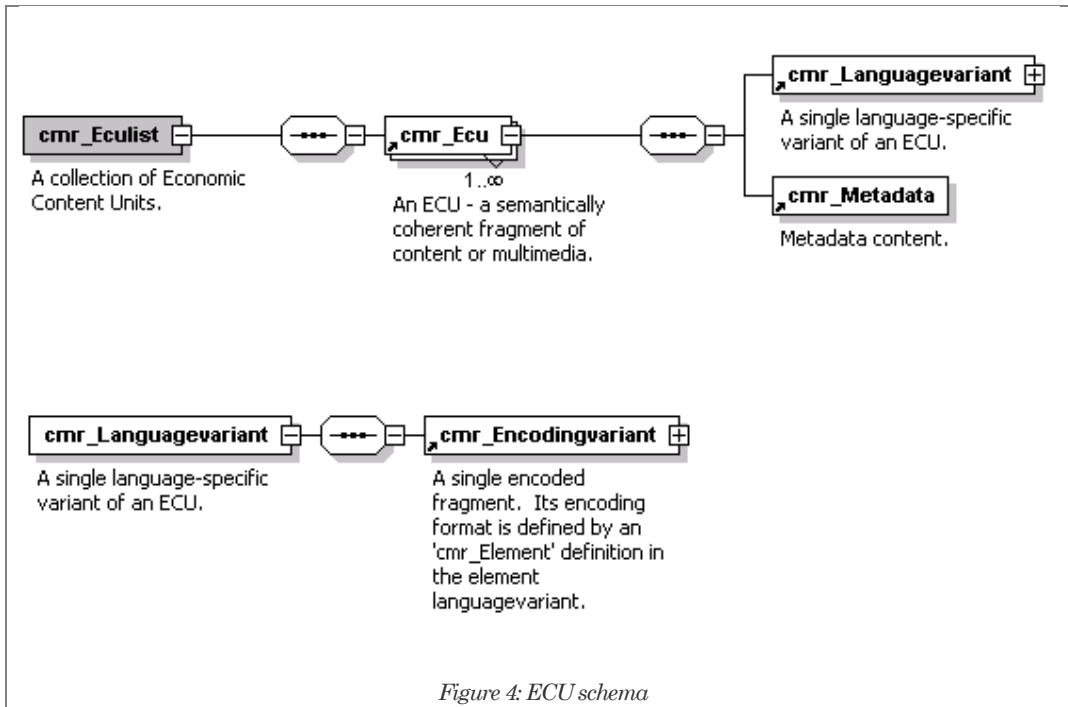


Figure 4: ECU schema

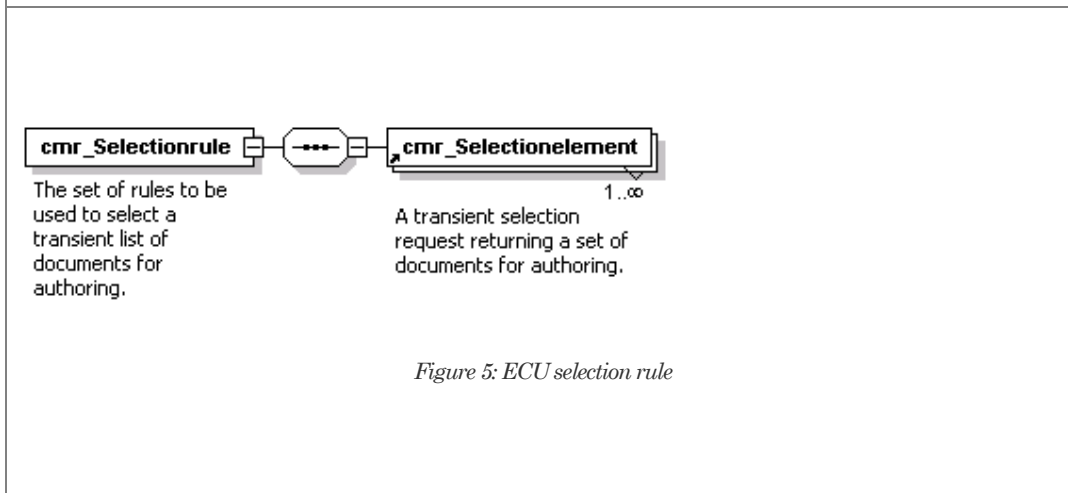


Figure 5: ECU selection rule

2.3 Layer 2. Document Ontology.

The editing, management and reuse of content is not meaningful except in a document structure. An ECU taken in isolation is an orphan until its content is understood in a document ontology that defines its semantic and presentational context. A final published multimedia document may comprise a heterogeneous set of text, images, audio and video selected for a particular level of language or media encoding. Each component of this document (the text, images, audio and video) may itself be a collection of Economic Content Units linked together in a semantically meaningful way. The final transformation that generates a multimedia document may well draw on individual ECU's for existing static document structures, which form part of the final published multimedia presentation. The acyclic network of linked ECU's comprising a particular document or partial document is the subject of the second layer of the CMR model. A content editor working on a multimedia document is selecting semantically related resources.

A Document Definition Language defines the structure of content items in the multimedia context. The flexibility implicit in SGML supporting Renear's overlapping hierarchies [Renear, Hockey, & McGann, 1999] and Boll's content fragment [Boll & Klas, 2001] provide examples for the design of complex models for representation of heterogeneous content resources. The Document Definition Language provides the framework for describing the content structures that are allowed within the system. Such definitions may be implicit (as with authoring tools like Dreamweaver) or explicit

(such as in the more generalised content model described by Boll and Klas and the one explored in this paper. This definition may be expressed in a syntactical form (such as DTD) or through a modelling interface to facilitate the definition of content items, controlled vocabularies and metadata associated with particular content components. The explicit realisation of a flexible document definition capability is essential to this layer of the model, and builds on the ECU layer to provide a methodology for content reuse.

In this model the document ontology of the CMS describes the framework for meaningful relationships between ECU's. Different multimedia types have different Document Object Models (DOM), and the purpose of a flexible document ontology at this layer of the model is to allow flexibility in brining ECU content together in different DOM structures. The encoded object represented by an *ECU* may well reside in an operating system directory rather than a database object (such as a relational table).

2.3.1 Definition

The reuse relationship layer brings together semantically related ECU's for management and transformation – for example as a "news" item or a "chapter" of a collaboratively produced book.

Figure 6 below presents a specification for a Reuse Document Type (RDT), which maps the structure of a generalised document object comprising ECU relating to metadata, design specifications and editorial content. The Reuse Document Type brings together both content and metadata, and is necessarily an acyclic tree of ECU elements and relationships - that is, while a RDT may itself participate in other RDT's, this

relationship cannot be recursive or cause a link in the document structure back to a higher point in the document tree. This specification is realised as an XML schema described in Figure 7. The purpose of the specification and schema is to provide a generalised (but acyclic) method of describing different document hierarchies organising ECU resources.

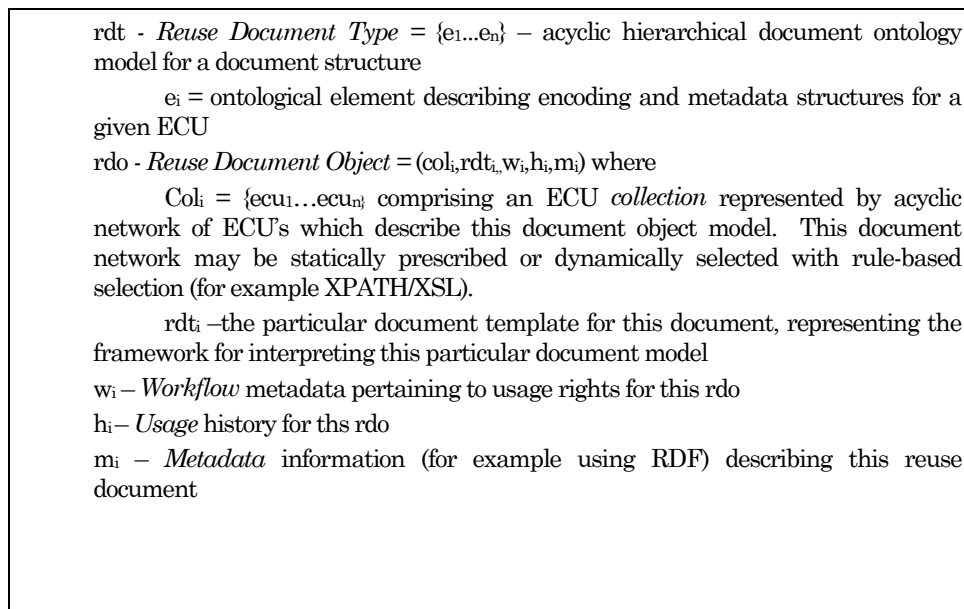


Figure 6: Reuse Document Ontology

2.3.2 Web Services Functions

Expressed as a web service, the interfaces relevant at this layer are those that support management of the document ontology. In the example below a “document type” describes a Reuse Document Object (RDO) in terms of a set of ECU “elements”. An ECU element might itself be a description of an RDO structure, giving a recursive hierarchical capability to support complex document forms. The flexibility of the management of the ontology is a key aspect of content reuse. It is the enabler for subsequent content management and reuse. The Web Services example below defines a recursive document definition structure the elements of which at any given layer of the hierarchy may be an

ECU definition, a link to another document structure or the edge of a database table connection.

Function	Request	Response	Purpose
UpdateDocumentDefinition	<i>cmr_Documenttypelist</i>	<i>cmr_Documenttypelist</i>	Create/Update a Document Definition
GetDocumentDefinition	<i>cmr_Selectionrule</i>	<i>cmr_Documenttypelist</i>	Fetch document definition (s)
GetElementDefinition	<i>cmr_Elementlist</i>	<i>cmr_Elementlist</i>	Create/update an Element definition
UpdateElementDefinition	<i>cmr_Selectionrule</i>	<i>cmr_Elementlist</i>	Save Element definition (s)

Table 2: Web Services functions at layer 2

2.3.3 Versioning

Just as ECU's can have variants across language, encoding and editorial dimension, so also the Document Ontology adds a layer of versioning complexity. The Reuse Document Type (RDT) may itself be versioned, with the obvious implication that changes to the document model may itself affect the management of the ECU collections that are members of an instance of a document model. The collections comprising a reuse document object (RDO) may also be expressed in editorial versions over time.

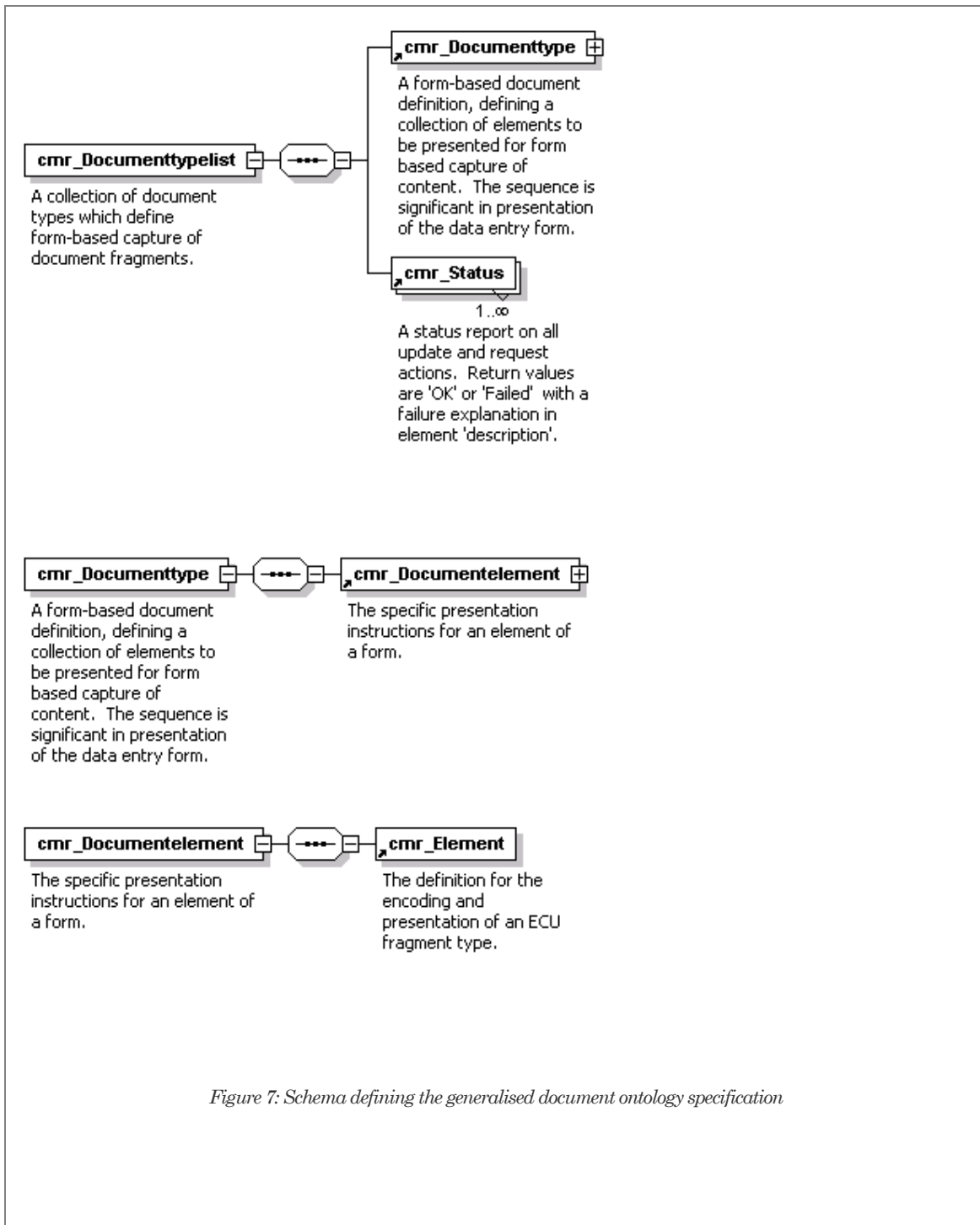


Figure 7: Schema defining the generalised document ontology specification

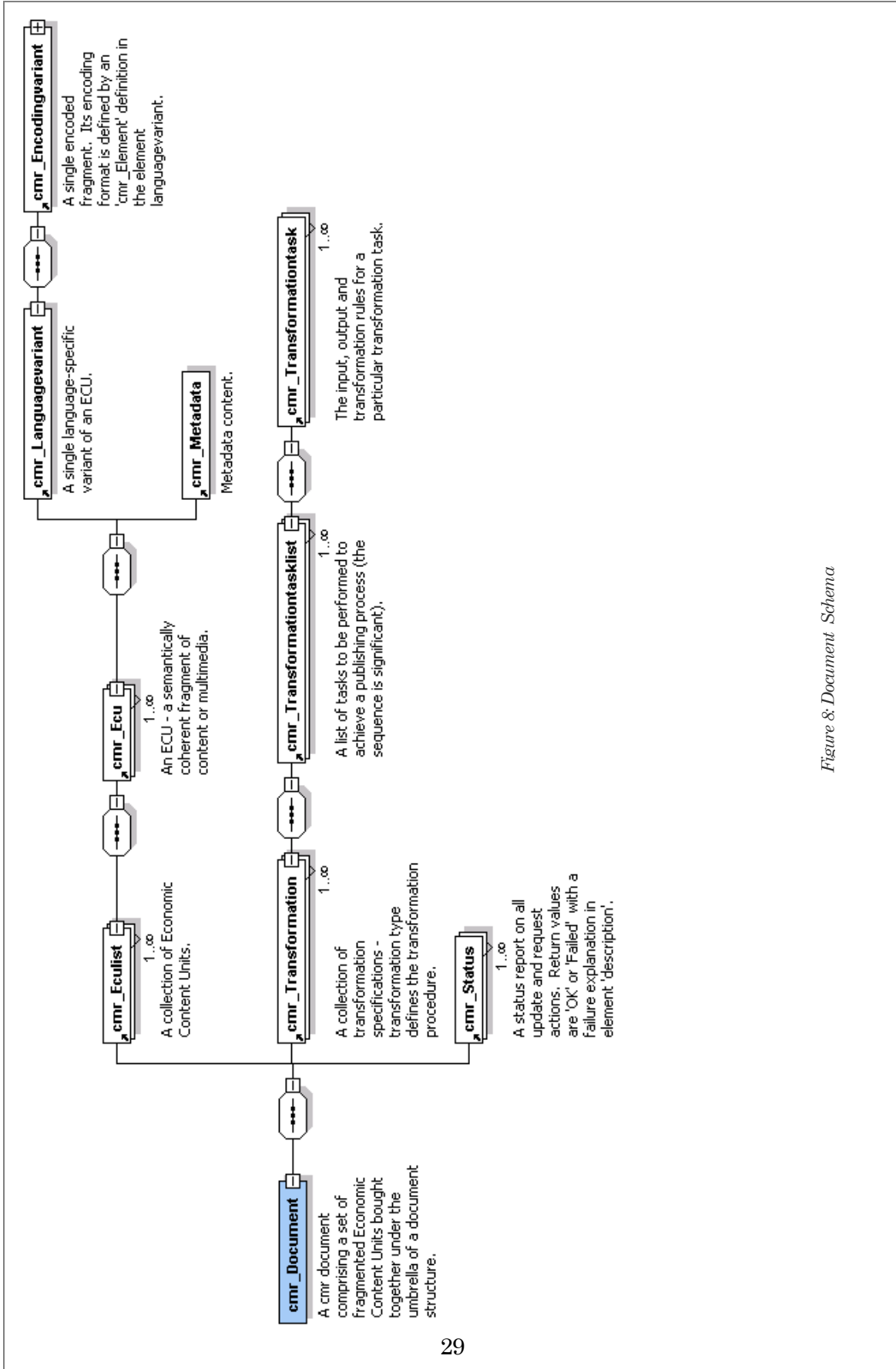


Figure 8: Document Schema

2.4 Layer 3. Workflow Management.

The Workflow layer opens the content to authoring and editing tools, placing *economic content units (ECUs)* in their document object context in a language and encoding specific manner that meets the needs of specific content authors. These public interfaces are mediated by the workflow usage and temporally defined user access rights. Characteristically, end users expect to interact with content through interfaces supplied by their existing editing tools, either browser based or through standard content delivery protocols such as SMTP (Simple Mail Transport Protocol, or Email), HTTP, FTP, WebDAV and system proprietary file referencing conventions (such as the Microsoft UNC).

2.4.1 Definition

The editor will access content in the framework of the document reuse objects defined at the second layer of the model, constrained by temporally defined access rights, based on a selection framework used to discover the content. A transformational process may be required to map encoded content to the particular editing framework used by the content editor, yielding a particular content view: This view is described in Figure 9 and described as an XML schema in Figure 10 below. The XML Schema for this layer of the model has the added complexity of workflow management (Figure 11) transient content selection (Figure 12) and static collection management (Figure 13).

view – a particular expression of an RDO document mediated by transformational rules to allow management of the content in a particular authoring language and document environment as governed by workflow rules.

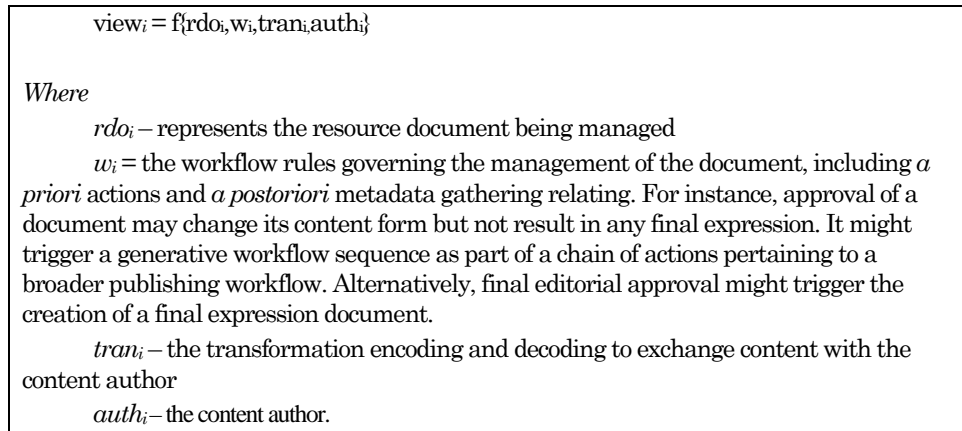


Figure 8: Content workflow framework

2.4.2 Interfaces

Rich public interfaces are required order to support the many elements of workflow and authoring. While many elements of workflow are contingent to the organizational context, there are established practices in workflow modelling which improve the flexibility of the workflow process. Included in this definition is a simple ontology for linear workflow management with data-driven over-rides.

2.4.2.1 WebDAV

WebDAV is a tool commonly used for collaborative editing of content resources. Its design illustrates some of the underlying requirements of a workflow process. WebDAV began as a protocol founded on the in-situ editing of content with the web server acting as both content repository and runtime engine. Leveraging HTTP, its first generation function was simply to provide a means of authenticated access to editing content on a website [Whitehead & Wiggins, 1998].

WebDAV could be regarded in this way as an early precursor to SOAP and Web Services (see below) It opens up the more strictly confining HTTP protocol to a more generalised exchange of information through XML that fully implements the XML

namespace. That it is implemented in an HTTP protocol environment makes the development of a server architecture interfacing through to a content store a more straightforward process.

WebDAV leverages the base protocol of HTTP for the receipt and delivery of content, adding to it the protocols for editorial access and version control. WebDAV is an illustration of one type of public interface to this layer of the model, and is amenable to presentation of a virtual view of the content resources, although still bound to a “file system” view of content structures.

2.4.2.2 Browser-based client server interface

The browser is the ubiquitous client to many information systems. Its availability on most desktops, with a (relatively) lightweight client-side application framework makes it ideal as an authoring interface to this layer of the model. The browser-based resources access paradigm is well accepted by users, as are the limitations of its graphical user interface.

2.4.2.3 Web Services interfaces

Function	Request	Response	Purpose
List Documents	<i>cmr_Selectionrule</i>	<i>cmr_Documentlist</i>	Provide a document catalogue of document items by a structured selection rule
Publish Documents	<i>cmr_Selectionrule</i>	<i>cmr_Documentlist</i>	Publish one or more documents specified in <i>cmr_Selectionrule</i> , returning a status in <i>cmr_Documentlist</i>
SaveDocuments	<i>cmr_Documentlist</i>	<i>cmr_Documentlist</i>	Create or update one or more documents specified in <i>cmr_Documentlist</i> , returning a status in <i>cmr_Documentlist</i> (and releasing any locks)
Delete Documents	<i>cmr_Selectionrule</i>	<i>cmr_Documentlist</i>	Delete one or more documents specified in <i>cmr_Documentlist</i> , returning

			a status in <i>cmr_Documentlist</i>
GetCollection Documents	<i>cmr_Selectionrule</i>	<i>cmr_Documentlist</i> <i>cmr_Collectioncriteria</i>	Get return a set of documents matching a given collection
GetCollection Criteria	<i>cmr_Selectionrule</i>	<i>cmr_Collectioncriteria</i>	Return the collection criteria for a collection
UpdateCollection	<i>cmr_Collectionlist</i>		Create / update a collection
ListCollections	<i>cmr_Selectionrule</i>	<i>cmr_Collectionlist</i>	List collections based on criteria in <i>cmr_Selectionrule</i>
LockDocuments	<i>cmr_Selectionrule</i>	<i>cmr_Documentlist</i>	Lock and edit a document
ReleaseDocuments	<i>cmr_Documentlist</i>	<i>cmr_Documentlist</i>	Release a document (no save)

Table 3: Web Services functions at layer 3

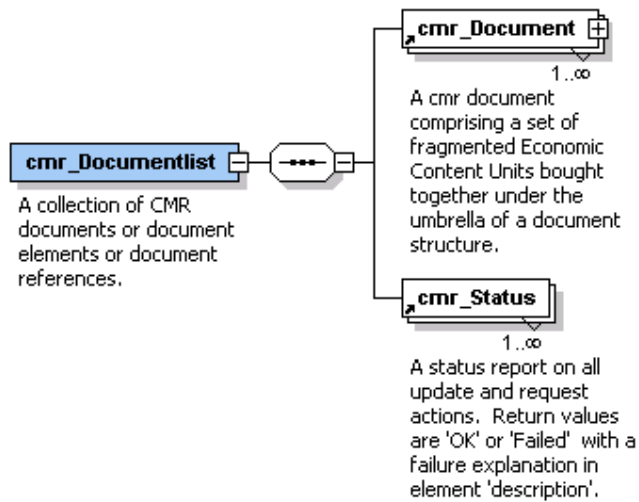
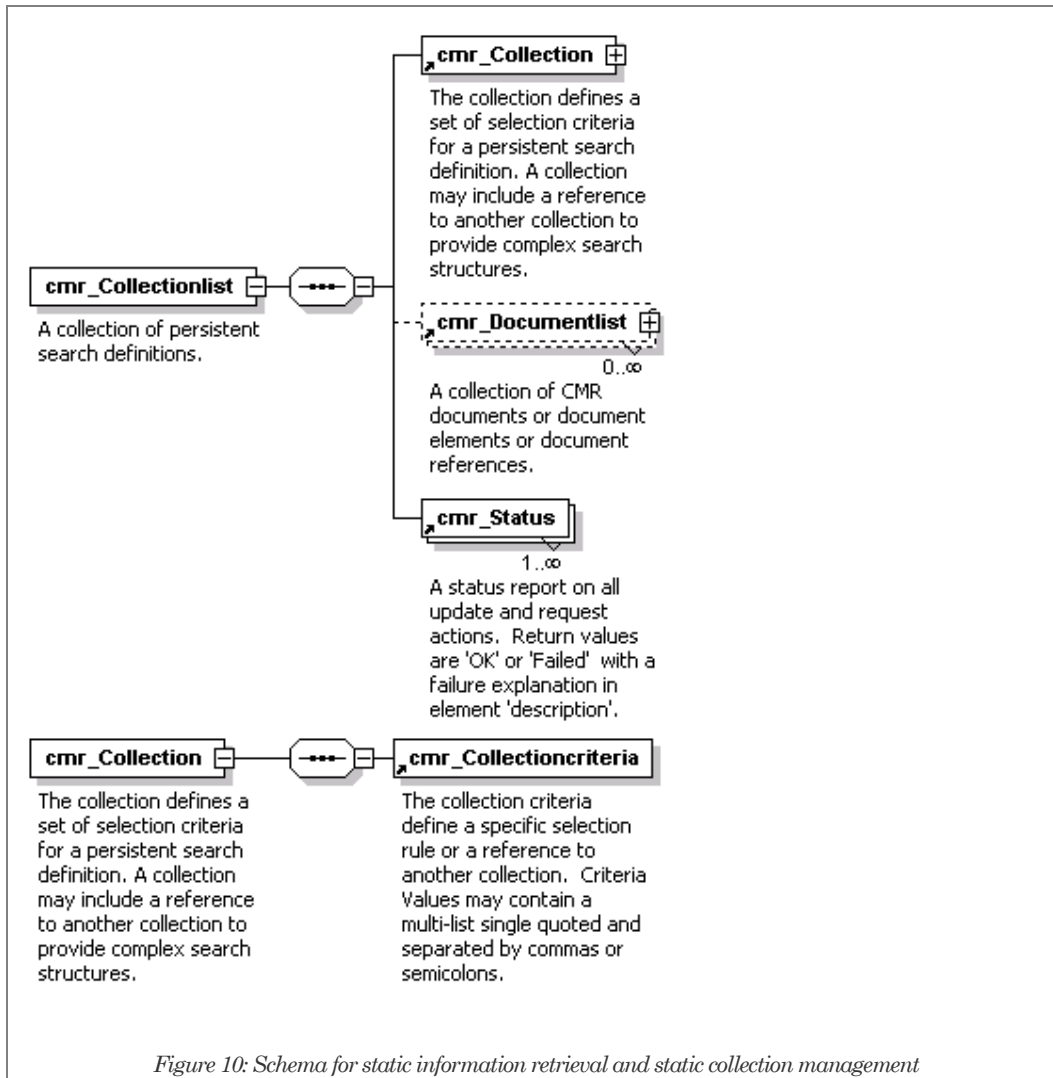


Figure 9: The document collection



2.4.3 Version control

Versioning of the content also becomes important in a collaborative work environment, as the need to view comparative changes over time becomes important. Version control of content may be required at both the content-component level and for organised collections of content. The purpose of version control at the component level of content is the support of collaborative authoring and the ability to review changes over time. The purpose of version control over broader collections of content is to support the controlled

regeneration of larger collections of content in a manner that can be consistently repeated.

2.5 Layer 4. Content Transformation and Delivery

The *content delivery* layer is the vehicle for final expression of the content. It defines the rules for interaction with the high-level presentational information system (client & server), which yields the final content version (see v_i in *Figure 11* below) to the target audience. The runtime delivery layer is quintessentially the point of mediation between the end user of the content and the runtime delivery engine(s). Content expressions may have dependencies over time in relationship to other content expressions. For instance, the content editor may expect the content to be delivered only at a particular time, or only to particular users. The content delivery system must, in this context, negotiate the issues of transient network failure, and to that degree ensure the integrity of content delivery in the form of the minimum level of coherent expression. Equally important is the role of the runtime engine in addressing the behavioural context of the content requirement of the information seeker.

Many content management implementations, particularly in a dynamic content delivery environment, make little differentiation between the information systems of CMR layer 3 and 4. A dynamic content engine may be interrogating a database dynamically and yielding a result to an end user. The point of differentiation between a dynamic, data-driven website, and a CMS is precisely the point of focus on content management in the context of content reuse. The economic worth of content management

lies in its ability to manage content in a manner that yields benefits in its effective reuse in a syndicated sense or in a regenerative sense. A dynamic data-driven website, like any other information system, need not necessarily exhibit either of those characteristics.

The merit of separating levels 3 and 4 of the information system lies in the systematic engineering for effective content reuse rather than the satisfaction of the demands of implementation of a particular information system, and for the task-oriented management of the transformation and delivery process for complex regeneration systems.

A runtime engine may be an engine for static delivery of fully composed content (such as a simple static web server) or an information system that draws dynamically off the Layer 3 systems. Some content management products (for example Vignette and Obtree) have hybrid runtime engines that integrate with existing Web Servers. Others (such as the prototype discussed in this paper) are at arms length from the actual runtime engine.

The diversity of authoring tools that feed into the multimedia production process compounds the difficulty of content reuse [Agoulmine *et al.*, 2000; Duffy, 2000]. Industry models such as the early Microsoft “White Paper” on “Content Management” explore a largely linear view of the content management process [Reynolds & Kaur, 2000], with little exploration of the issues of content reuse.

Online Learning Systems continue to be an active area of experimentation in content reuse. Just as personalisation within a commercial web delivery framework has been a focus of website design, so we are also likely to see greater efforts at the

personalisation of higher educational delivery mechanisms. Universities with a focus on building campus-wide models for effective capture of educational objects will be the best placed to begin the delivery of effective, personalised online educational delivery.

Web-based runtime engines supported by CMS software are therefore typically either:

- Serving content which is statically published through a CMS generative process
- Delivering dynamic content which is assembled for the content consumer more-or-less on demand based on the context of their information request, and which is possibly personalised for the particular content consumer.

Performance and architectural issues are the tradeoffs that keep static content serving as a popular method content delivery. A traditional website content serving static pages or with minimal dynamic generation of content still represents the most speed-efficient runtime engine [Mendes & Almeida, 1998]. The web server, acting essentially as a content file server, needs little interpretive intervention to deliver the content. Some web content delivery platforms, such as Vignette, have made a virtue of proprietary caching and generative mechanisms that deliver content as quickly as possible within the constraints of dynamic regeneration of the content in its final form. Dynamic online generation of content has considerable advantages for currency of content and personalisation, but requires itself a new layer of architectural support to meet the higher server load demand [Anderson, 2001].

Whether dealing with a static or dynamic model of content delivery, the underlying content management issues are similar – the difference between dynamic and static content delivery is essentially one of the timing for content recomposition. The final

expression of the content may be highly proprietary in form and transient in its expected longevity. The Content Management System must support both generation of content in its original form and translation to new runtime environments. An extension of this regenerative process is the long-term persistence of content in current delivery formats to address the issue of technological obsolescence.

2.5.1 Definition

The publishing process is a multidimensional transformation in one or more stages. This transformation is a process of mediation with a runtime engine, the presentation format for assembling the content and the personalisation requirements of the end user. This mediation process could be managed by a single runtime engine serving static content (and so minimal personalisation) or through a dynamic runtime engine with on-the-fly creation of pages based on the end users profile and the content resources available. The complexity of this process may be determined by the degree of personalisation required for user presentation and the point at which this personalisation occurs. A specification for this relationship is described in Figure 11 below and realised as an XML schema for content syndication (Figure 13) and content transformation (Figure 12). In some cases, the personalisation of content is managed through specific runtime engines, with increasingly complex mediation between the user and the runtime engine – such as through the medium of digital augmented paper [Norrie & Signer, 2003].

The final content version v_i expresses the delivery of fully encoded content in its published form to the end-users runtime engine based on one or more transformations that yield the final expression of content based on the contingent states of the reuse objects, the transformational template, the content usage rules, the distributive rules and finally the relevant profile of the target community.

$$v_i = f(\text{rdo}_i, \text{tran}_i, \text{temp}_i, \text{dist}_i, u_i)$$

Where

$\text{rdo}_i = \{\text{rdo}_1 \dots \text{rdo}_n\}$ that represents one or more reuse documents (rdo_i).

tran_i = the transformational template mediating the delivery of the rdo_i document collection to the user community u_i .

temp_i = the temporal rules pertaining to content usage (time/s at which the content can be considered valid)

dist_i = the distributive rules by which content reaches the relevant runtime engine, including relevant security constraints

u_i = the target user community

Figure 11: Content transformation and delivery

2.5.2 Web Services Functions

The topmost layer of the model is directed to the fruition of the previous layers of content management. There is a multiplicity of runtime engines used to deliver content, the most popular being the Web Server and email. The format of content also varies, but most popularly HTML and XML. Table 4 describes a set of simple Web Services functions that can achieve a high-level fully encoded interchange of content at full arms length – for example in a syndication process.

Function	Request	Response	Purpose
GetDocument	cmr_Url ist	cmr_ Encodeddocument	Deliver through Web Services a fully encoded document for presentation through another

service			
GetDocument Preview	<i>cmr_Url</i>	<i>cmr_ist</i>	Delivery through Web Services a preview version of an encoded document based on the security context of the user
GetSyndication	<i>cmr_Syndication requestlist</i>	<i>cmr_Syndication document</i>	Deliver through Web Services an XML encoded version of the document

Table 4: Web Services functions at layer 4

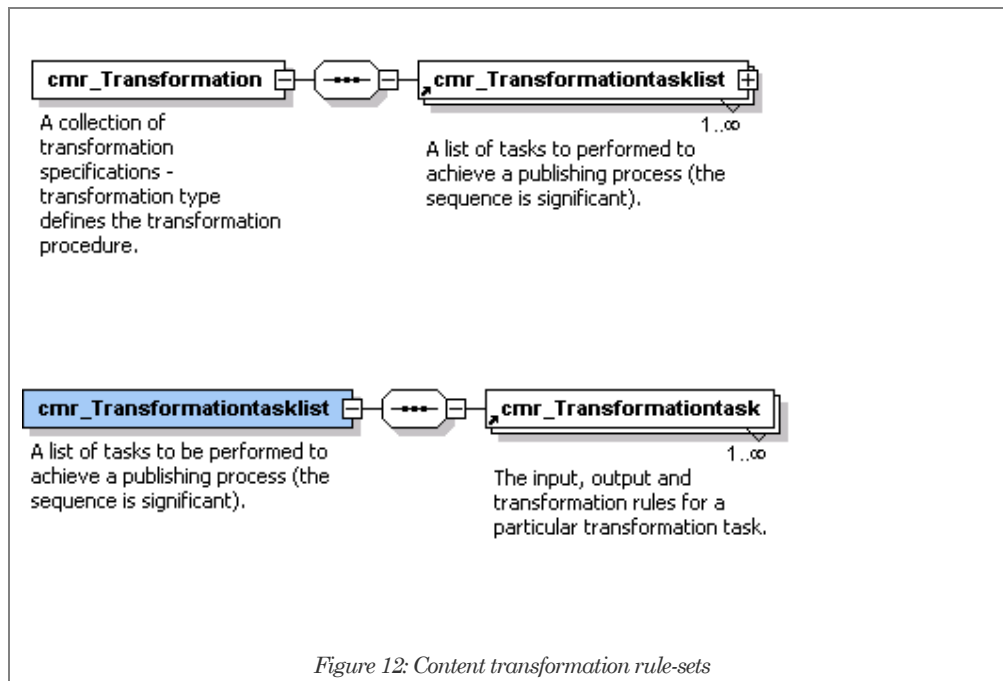
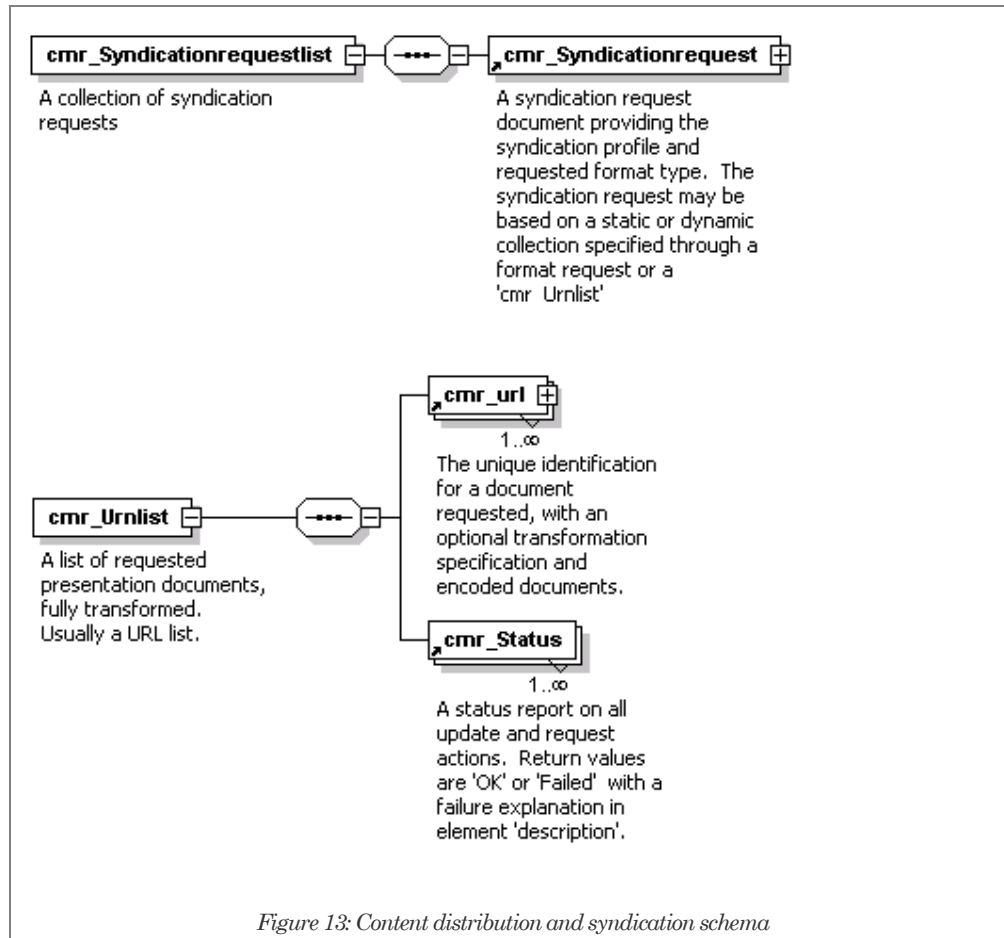


Figure 12: Content transformation rule-sets



2.5.3 Versioning

The content management layer introduces the management of versional issues at a broader level – for instance, the maintenance of consistent versional information across all ECUs for a particular version of the published content. An example of the application versional information at this level might include the archiving or rollforward/rollbackward of all content in a website to a particular moment of time. Conceptually this equates to the full edition versioning of the published product. Complexities at this level of versioning include the integration with other temporal database systems in a dynamic environment, which imply a lack of specific “version in

time” content. However, many websites are amenable to “editioning” at a particular point of time. The support of such edition versioning is an important element of humanities web publishing [Eaves, 1997].

3. Model Evaluation

This section presents a quantitative evaluation of the prototype using the content versioning and publishing metadata history for the period of the project. The purpose of this analysis is to elicit the ways in which the software facilitated the reuse of content, and the ways in which content reuse changed over the duration of the project. The results are indicative of the value of the model in content reuse. This analysis elaborates metrics for measurement of content currency also points to further avenues of research regarding content reuse.

The Inter-Publish prototype was the outcome of a sustained effort to implement the four layers of the model. The result was a content management system with a flexible ontology highly suited to reuse of heterogenous media forms. The relational design for the prototype was tested both in the higher-end (Microsoft SQL Server) and low-end (Microsoft Access) relational database environments. While the prototype went through many changes through the three-year life of the project, the underlying relational model remained relatively intact. This proved suitable for the demonstration of the fragmentation of content at the layer one and realisation of the conceptual model at the second layer (the document ontology). Many changes were directed to user visualisation of the fragmented content.

The online education concept known as “Montage” was originally piloted in 1998 as a means of linking schools in Australia and the UK in undertaking curriculum projects in areas of mutual interest. The early challenge of the project was to establish an effective framework to manage the international growth in participants, facilitated in a manner that maximised resource sharing, while sustaining a high level of quality control and content moderation. The project was also constrained by limited staff resources. The “Montage” framework provides an effective means for a small team of participants to manage this process of international collaboration. Arrangements are in train to establish similar services operating out of the UK. Each curriculum project has a project co-ordinator to manage and moderate the project. The content is agreed by the Education Department for educational appropriateness.

The project is published through the website establishing linkages between schools and their project interests – and republished using the Content Management System for reuse in various country sites and through e-news letters.

Montage has had considerable success across regional and metropolitan Australia with a program that has now been expanded to 5 countries in SE Asia as well as Argentina, England, and Finland. The program now also has a portal in the British Council UK called MontageWorld. There have been opportunities for teachers and students to travel to other countries as part of the programme. A summary table of the schools registered in 133 countries around the world (with over 5000 schools currently registered).

3.2 Resources used in the reuse analysis

The following analysis was prepared after the Inter-Publish prototype had been used in the Montage project for 3 years. Versioning, being fundamental to the ECU layer 1 model, was implemented in the first generation of the prototype. This aspect of the model had the additional benefit of allowing detailed logging of user editorial activity for the duration of the field test. This yielded valuable information on the nature of content authoring for the project, and how this content was deployed. As discussed in Section 5, the Montage project represented multiple website domains. Each of these domains was published from the central CMS server and accumulating information in a primary domain. This website resource represented educational project and information resource gateway for primary and secondary school teachers. The domain saw several name changes and design makeovers in the last two years: from www.montage.edu.au to www.montageplus.co.uk, and more recently to www.montageworld.co.uk.

The Montage project started in 1999, and from mid-2000 onward, the Inter-Publish prototype CMS was used for content management of Montage web sites. External website logs and Inter-Publish ECU versioning and publishing metadata together provide a rich picture of authoring and publishing activity for the period examined. While the versioning changes were directed toward meeting the requirements of the emergent CMR model, the downstream consequence of this change has been the collection of a history of content as it was created and deployed. This detailed version history also represents a valuable data mining resource for the purpose of analysis of the impact of use of the content management software for the three-year period analysed. The versioning table of the content management system has a snapshot of all content

changes where some element of a content item has been changed. Publishing metadata and persistent content collections provided a source over time for tracking original content use and content reuse.

The Inter-Publish usage metadata has been combined with external website usage logs to enrich the analysis of content usage. Due to a variety of downstream caching effects this data is indicative rather than comprehensive. It nevertheless represents a good profile of website trends and user behaviour on the website for the periods 2001 and 2002. The analysis also relies on Content management activity based on ECU versioning information and publishing metadata. From December 2000 the Inter-Publish prototype captured detailed versioning history for all content changes made within the CMS framework. This resource has been used to derive detailed information on rate of content change.

3.3 Website content profiles

Table 5 is a summary of publicly visible web page content categorised by type for the two principal domains managed by the British Council Australia, as at December 2000, 2001 and 2002. By 2002 the number of published web pages had nearly tripled. This change represented a large number of school registrations and the new web-based projects, as well as the reuse of content across different country sites.

	Total Docs	Web Pages	Images (gif)	Images (jpg)	Video	PDF	Other (flash, director, etc)
2001	20597	14620	4664	1248	10	2	28
2002	28365	18820	7094	2238	179	4	30
2003	33699	24980	7458	1015	179	9	58

Table 5: Published content summary for Montage

Content growth from both internal and external editors was muted in 2003 due to problems associated with relocation of the Inter-Publish and external web servers to the UK. Security restrictions on web server access prevented external submission of project and image resources and internal access by content authors, compounded by relocation of the new support team from London to Manchester and the subsequent outsourcing of server support roles, all of which effectively prevented access by both internal content authors for a period of nearly 4 months – effectively two school terms. This principally affected publishing of content to the web server – content submissions in XML were accumulated during the publishing freeze until the location and configuration server was resolved. That is, while mechanisms for content reuse and syndication were still possible, the relocation and subsequent support issues had a detectible impact on publishing for the period April through to June. Figure 13 below shows the original authoring and publishing activity for the periods 2001 through to 2003.

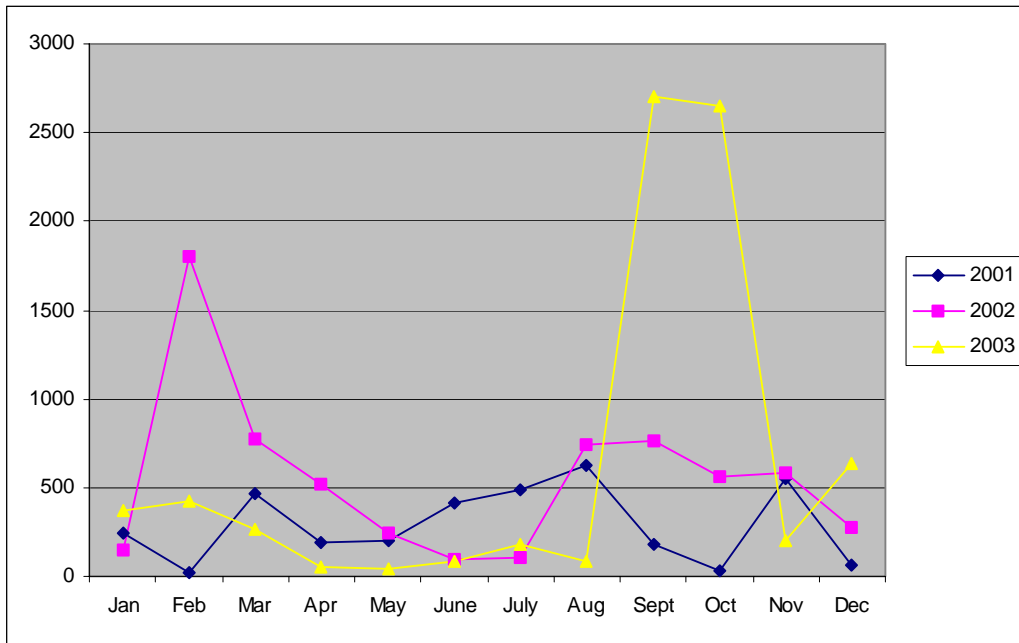


Figure 13: Content authoring profile by month for 2001 , 2002 and 2003

An important objective of content management is to maintain currency of information on the published website. In quantitative terms it is possible to measure this currency through an analysis of the versioning history of authoring changes and the publishing metadata. The following two measures are proposed as a signal of information currency on the website:

- *Information Turnover.* This ratio is intended to measure the degree to which actual content is changing excluding the effect of content reuse. This measure is calculated from the number documents for which ECU elements have been changed as a ratio of total published *primary* web pages (that is, excluding navigation indexes and pages generated as secondary content resources). The primary web page is understood to be the first web page published from the parent document in which an ECU element participates.

- *Website Turnover.* This ratio introduces the effect of content reuse in measuring the turnover of content on the website. The content turnover figure measures the number of web pages changed (through primary web page generation or content reuse) as a ratio of total published web pages.

The Information Turnover is a measure of authoring activity and actual content change. Taking as an example a website of 100 pages. An ECU is published to a single content page on the website but its title is referenced in two other web pages for navigational purposes. A modification to this ECU would result in an Information Turnover ratio of 1/100 while the Website Turnover would be 3/100. In principle the higher the turnover ratios, the better the indication of information currency in the website. The Information Turnover is a useful metric in measuring the authoring activity. The Website Turnover metric provides a measure of the degree to which content reuse is contributing to currency of information in the website.

These ratios are naturally contingent of the type of site that they measure. A large established website with a rich body of archived and static material may show a low percentage result, but the measurement of turnover over time will be indicative of the relative currency of information. Figure 14 shows the ratios for the Montage field trial.

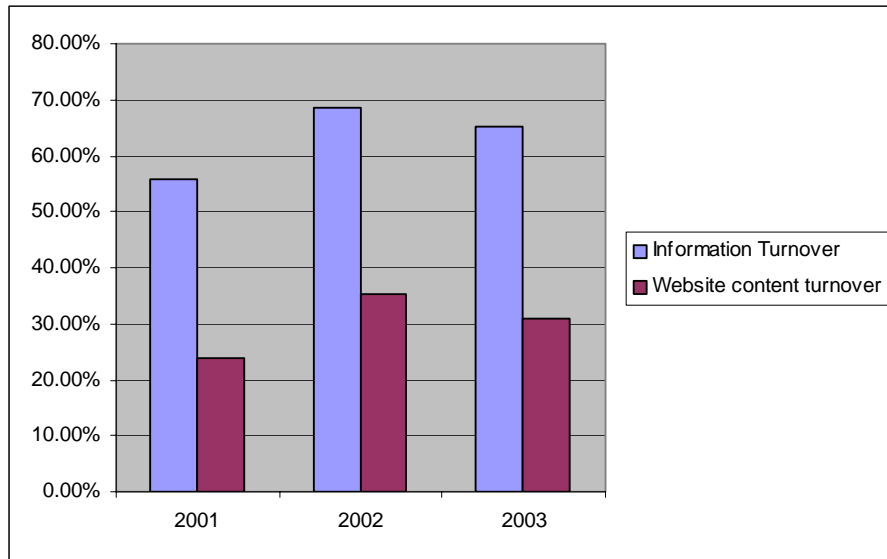


Figure 14: Website Content Turnover and Information Turnover

The improvement between 2001 and 2003 both in Website Content Turnover and Information Turnover, despite a substantially increased pool of published pages, is indicative of the benefits of operating within a content management framework and of the effects of content reuse. They also suggest a sustained improvement in the currency of published text content. It should be noted that this growth in and currency of content was achieved despite static or declining staffing resources in relation to the project at different times of the British Council funding cycle for the project. For the first year of the project the British Council had one person working full time on content management and authoring. Thereafter the individual country sites undertook day-to-day content administration on an ad-hoc staffing basis. The workflow process was important in alerting staff as to content actions required. The slight decline in 2003 is attributable to the publishing freeze during migration of the server infrastructure to the UK (see above).

In addition to the manual editing of web pages, a substantially larger number of pages are frequently updated as a result of scheduled publishing and syndicating activities of the Content Management System Inter-Publish. This aspect of content reuse is the subject of further quantitative evaluation in the next section.

3.3 Reuse analysis of Content Management

These indicative metrics (Information Turnover and Website Content Turnover) justify further analysis of the underlying contribution of original editing and reuse to the currency of information on the website. In many cases, the Inter-Publish CMS was used for static web publishing. The versioning history in Inter-Publish tracked both the editorial history of content changes and the usage of ECU content through publishing metadata and usage of ECU's through static content collections. Together, this content versioning history and publishing metadata provide a means of quite granular analysis of content publishing. Content was typified in two categories in the analysis below:

Primary content. ECU content that was attributed only to a single website location or was used once only in publishing (usually in the parent document to the ECU) was classified as "primary content."

Reuse content - the content was used two or more times in publishing actual content (as distinct from metadata concerning the content). Metadata content reuse, such as automated site maps generation, was not included in this category.

Inter-Publish associates content units through a document definition called a "document type". These "document types" were further classified into a "content type" to bring together documents of similar type. While many document types are pre-defined to

accommodate the most frequently used content capture types, the prototype allows flexible user definition of other document type definitions.

The definition of an ECU is realised in the cmr schema across potential encoding and language variants for a given ECU. *Figure 15* shows a partial xml snippet of a published multilingual fragment expressed in the cmr xml schema and generated as a separate web page for each language instance. The final expression of this content is two separate web pages. The language variants are described by the element “cmr_languagevariant”.

```
<cmr>

<auth sessionid="PSX100000000000092"><cmr_status status="OK"/></auth>

<cmr_documentlist>

  <cmr_document title="about us" sectionname="About"

    documenttype="Document" divisionname="E-Link Japan Website" contenttype="Document"

    documentstatus="Published" datelastedited="17/12/2002 1:45:00 PM">

    <cmr_status status="OK" />

    <cmr_ecu contentid="10000000019543">

    <cmr_languagevariant language="ja-jis">

      <cmr_encodingvariant

        elementtype="TextArea"    fragmentationlevel="2"

        element="PageContent"    transferencoding="Shift-Jis"

        encoding="Shift-Jis"><!-- content -->

      </cmr_encodingvariant>

    </cmr_languagevariant>

    <cmr_languagevariant language="en">
```

```

    <cmr_encodingvariant      elementtype="TextArea"

        fragmentationlevel="2"      element="PageContent"

        transferencoding="utf-8"      encoding="utf-8">!-- content -->

    </cmr_encodingvariant>

</cmr_languagevariant>

<cmr_ecumetadata

    element="Keywords" elementtype="MetaData"

    cmr_transferencoding="utf-8" encoding="utf-8">elink

</cmr_ecumetadata>

<cmr_ecumetadata

    elementtype="Date" element="PublishedDateTime"

    transferencoding="utf-8" encoding="text/plain" >

    17/12/2002 28:58:21PM

</cmr_ecumetadata>

</cmr_ecu>

</cmr_document>

</cmr_documentlist>

</cmr>

```

Figure 15: Multilingual ECU fragment

A document returned from a web services function will encompass one or more ECU content fragments and associated metadata. For instance, the ECU's for the multilingual Japanese pages described above are contained in the following document XML. The "auth" element is present for purposes of session management between web

services clients. The “cmr_ Documentlist” element contains the site context information for the document (the website division and section, a title for the ECU container document, content and the document definitional framework (the document type and content type), as well as metadata at the workflow level, wrapped in relevant workflow status information (see Figure 16 below).

```
<cmr>

<auth sessionid="PSX100000000000092"><cmr_status status="OK"/></auth>

<cmr_Documentlist>

<cmr_document title="about&#32;us" sectionname="About"

documenttype="JapaneseDocument"

divisionname="E-Link Japan Website" contenttype="Document"

documentstatus="Published" datelastedited="17/12/2002 1:45:00 PM">

<cmr_status status="OK" />

    <!-- ECU CONTENT FRAGMENTS APPEAR HERE -->

</cmr_document></cmr_Documentlist >

</cmr>
```

Figure 16: Workflow document container

There was progressively increased use of the CMS for image as well as text management through 2002. This trend continued in 2003, with inclusion of additional document types to meet specific publishing requirements. Teacher/School and project resources were largely sourced through online submission recommendations that were moderated and published.

Figure 17 presents a dissection of content on the CMS that was reused in web-based publishing. This was determined through an analysis of the ECU versioning history and publishing metadata for the years 2001 through to 2002. ECU content that was attributed only to a single website location or publishing destination, or only participated in a site map role, was treated as “Primary” content. ECU content that has been used in two or more locations on the website (other than as metadata) or for two or more content presentational types has been classified as “Reuse Content”.

An example of content reuse is the School content type. This content type represents a series of linked content units describing a school subscription to the Montage site. The content units are used in three ways:

To populate the online search engine with the XML school profile for (a) searching and categorising schools on the public website, and (b) online maintenance and updates by the schools (schools were emailed a username and password on registration).

To generate email mail out lists for manual creation of emails

To generate system publishing of email mail outs personalised through Inter-Published and with a strongly designed HTML layout.

To research and report on school subscriptions within particular countries and regions.

Content within each document type was directed either to "primary content" or "reuse content". Figure 17 below analyses the number of reuse instances for each document type in 2001 and 2003. The high reuse levels of School and Teacher entries can be attributed to email mail outs and could be considered a trivial reuse of the school entries, although

e-Zines were at times personalised to use the teach and school name in the email. Nevertheless, there is a clear trend to content reuse. This was principally achieved through the use of the Inter-Publish Transformation Language and associated dynamic and static content selection rules, which could be scheduled to distribute updated website content in the form of content indexes, categorised web pages and outbound syndication of content (such as projects) in XML form.

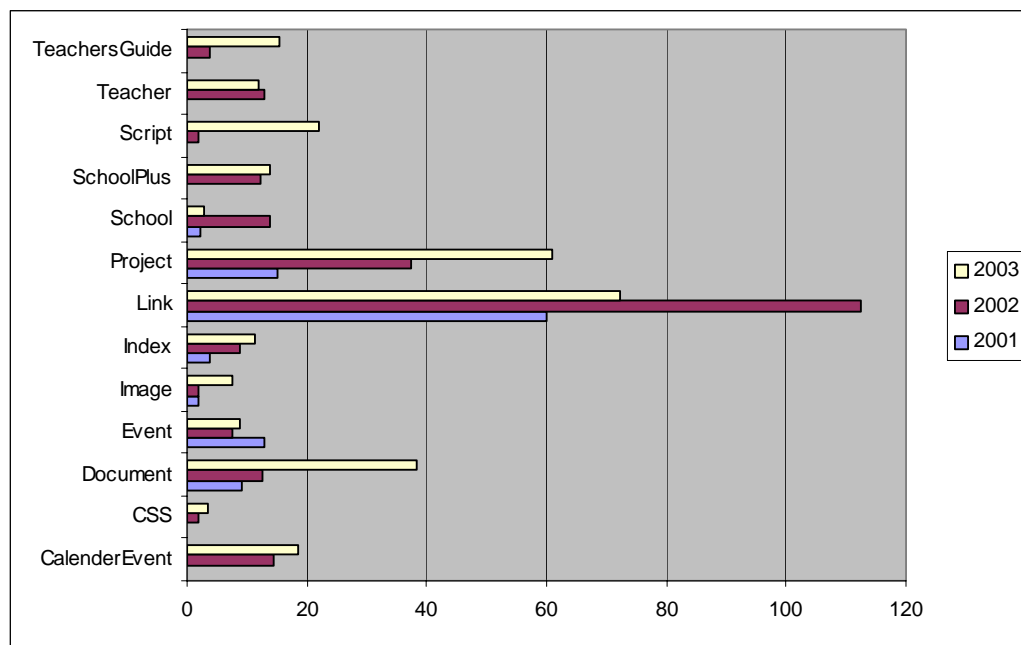


Figure 17: Content reuse by Document Type

Figure 18 summarises content reuse of ECU fragments for the three years of the field trial. The underlying content reuse shows clear progress over time. Reuse of content for generation of content other than web pages was an increasing feature in use of the website, as well as regeneration of content elements such as events at strategic points in the website. This content reuse figure includes all forms of content reuse (including syndication).

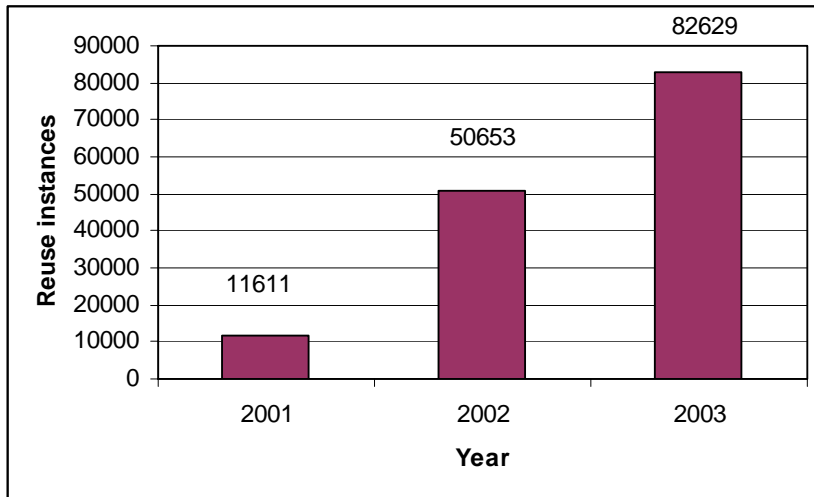


Figure 18: Content reuse by Content Type

Flexibility of the document ontology emerged as an important element of the utility of the software to the British Council, and for the Montage project alone more than 58 document types had been defined for various aspects of website content management – many of these to facilitate workflow management of original publishing.

It became clear during the course of the project, and is evident in the reuse figures presented above, that extensive content reuse (through replication, syndication and regeneration) was confined to only a few of the many document types defined. In particular, link information (that is links to other educational resources and websites), projects, schools and particular categories of webpage content. While there was increasing use of the prototype for multimedia content management – this was principally directed to use of the software as a multimedia resource repository. While there was emergent reuse of image content, content reuse was primarily directed to text content in this project.

The Web Services framework became increasingly important as the principle point of mediation between the CMS and other services at the third layer and fourth layers of the model. Outside RSS syndication, some projects were transferred using web services to the other CMS product used by the British Council – ObTree. Similarly, XML interfaces were used to transfer content from ObTree to Inter-Publish. Moreover, the exchange of information from local client databases (such as outlook) to Inter-Publish, achieved through the CMR web services layer proved an effective means of content gathering – centralising disparate resources, including database resources, in the central CMS clearinghouse. Once again, much of the focus of such content clearinghouse activity centred on specific categories of content, and the continuous transformation of content.

6. Conclusion and Future Work

The CMR approach to content reuse was tested through a rich prototype development process that explored the implementability of concepts in the model. This prototyping research process was directed to a non-trivial web-publishing environment to explore the complexities of the core concepts of the model. The outcomes of the Software Development Research were a prototype that provides a rich presentation of the functions at each layer of the model, exposed through Web Services, a document ontology at each layer of the model. A variety of interfaces were implemented at each of the layers of the model, and particularly at the workflow layer, to demonstrate the flexibility of the model to different interfacing standards.

By the use of versioning history and detailed authoring and publishing activity to track the outcomes of the project in facilitating content reuse as against primary publishing and the aspects of the model which facilitate this process. This review established that over the period of the field experiment, there was evidence of increasing content reuse, as distinct from simple site editing. There was a focus on the reuse of particular content types, an issue that was highlighted by the layered management of content and the ECU chunking process. In addition the prototype analysis indicated that the framework allowed a high level of information currency, sustained despite significant growth in published web content during the duration of the project.

Content reuse was realised in many different ways. Beyond syndication and reuse of informational content across multiple sites and media, content reuse became an increasing element in site layout design, increasing even further the abstraction of the logical site structure from particular implementation instances and semantically situated informational content.

The limits of content reuse tested in this project are obviously constrained by the needs of the particular case study tested. Further research would be valuable in the context of a client with a different domain of interest and for a client making more extensive use of syndication.

Web Services were explored as alternative expression of each layer of the model and to serve as a means of cross-integration with other applications (such as the ObTree CMS also used by the British Council). Through extension of existing XML and URL interfaces the architecture readily supported the SOAP-based approaches represented by Web Services for information interchange.

VB6, being familiar to the author, was ideal for the RAD-style prototyping used in this project. The inefficiencies of VB6 in string processing, however, do not make it an ideal language for scalable performance of the application layer, and the Inter-Publish prototype would benefit from redevelopment in a performance-oriented software architecture. Obviously with greater development resources, transfer and implementation to other server architectures would be valuable, with the possible implementation in CORBA objects supporting the defined interfaces rather than a VB6 application.

The implementation of content in the ECU framework also offered possibilities in the use of text mining techniques to facilitate the generation of descriptive metadata to aid the authoring process. An interesting avenue for further research in the extension of text mining techniques at the authoring stage to enhance the metadata categorisation and automatic classification of content structure provided by the CMR framework.

This Software Development Research culminated in a sophisticated application which implements both the functional elements of the model and a detailed Web Services interface for further extension of the application. The evaluation demonstrated a strong reuse in particular content items, indicating also what may be a strong distinction between content types. Extension of this analysis in other publishing environments could yield further insights into this aspect of content reuse, especially if extended to other domains of publishing. Nevertheless, the evaluation of content reuse in the Montage project clearly demonstrates "systematic" rather than "opportunistic" content reuse [Rockley, 2003].

In summary, the authors are confident from this evaluation that the CMR model for content reuse that can be implemented, and can meet the needs of a challenging environment for content publishing and reuse as well as showing interesting avenues for further research in content reuse.

Given the small size and budget supporting the Montage project, it is arguable whether the Montage project could have been achieved without the extensive support delivered by the project for both site replication and ECU-level content distribution and reuse.

4 Improvements and Future Work

While the Inter-Publish prototype clearly transformed the capability for content reuse in the field trial environment, the prototype could be significantly improved in a number of ways. A further Post Implementation Review survey indicated a gap between utility and ease of use, and further enhancement of the end user design for content authoring would help the commercial transition of this software. The field trial and empirical analysis of content reuse activity indicated that there were clear boundaries between types of content that were commonly reused and types of content directed solely to primary generation. Further research would be valuable, perhaps through field trials in other domains, to expand further on this differential between content types and its implications in design of content management and reuse systems.

As a prototype, the performance of Inter-Publish could be considerably improved. The development in a Microsoft VB6 framework facilitated rapid prototyping but is not an ideal platform for scaling to very large document sets. The planned enhancements

include the redevelopment of the system architecture, and the extension of the web services model, enhancement of the user interface, integration of further text mining in the authoring process, and possible integration of automatic translation interfaces to enhance the multilingual versioning elements of the system. Further experiments comparing the relative merits of XML database engines against relational databases for fragmentation of content for reuse would also be valuable.

Empirical work conducted as part of this project focused on a particular content reuse environment intending to verify the efficacy of the proposed model and evaluate the effectiveness of the Inter-Publish reuse application. The field trial was invaluable as a complex and challenging website in which to evaluating the prototype and give impetus to its development.

Further studies are needed to determine other aspects of Inter-Publish effectiveness. Such studies may include determining Inter-Publish methods for automatic fragmentation of content from existing websites and content resources, as well as automatic mapping systems to facilitate fragmentation of content from well structured existing document ontologies.

REFERENCES

ADDEY, D., ELLIS, J., SUH, P., & THIEMECKE, D. (2002). Content Management Systems. Birmingham (UK): Glasshaus.

ADVANCED DISTRIBUTED LEARNING. (2001). Sharable Content Object Reference Model Version 1.1. Retrieved 6/8/2001, from <http://www.adlnet.org/>

AGOULMINE, N., DRAGAN, D., GRINGEL, T., HALL, J., ROSA, E., & TSCHICHHOLZ, M. Trouble management for multimedia services in multi-provider environments. *Journal of Network and Systems Management*, 8,1 (2000), 99-123.

ANDERSON, K. M. (2001). Data scalability in open hypermedia systems. Paper presented at the Proceedings of the tenth ACM Conference on Hypertext and hypermedia: returning to our diverse roots, February 21 - 25, 1999, Darmstadt Germany. 27-36.

BOIKO, B. (2002). *Content Management Bible*. NY: Wiley.

BOLL, S., & KLAS, W. ZyX - A multimedia Document Model for Reuse and adaption of Multimedia Content. *IEEE Transactions on Knowledge and Data Engineering*, 13,3 (2001), 361-381.

BURNETT, I., VAN DE WALLE, R., HILL, K., BORMANS, J., & PEREIRA, F. MPEG-21 : Goals and Achievements. *IEEE Multimedia*, Oct/Dec 2003 (2003), 60-70.

BYRNE, T. (2003). *The CMS Report: Web Content Management Products and Practices* (Electronic No. 4th Edition): CMS Watch.

CANDLER, C. S., & ANDREWS, M. D. Avoiding the great train wreck: Standardizing the architecture for online curricula. *Academic Medicine*, 74,10 (1999), 1091-1095.

CERI, S., FRATERNALI, P., & BONGIO, A. (2000). *Web Modeling Language (WebML): a modeling language for designing Web sites* [<http://www9.org/w9cdrom/177/177.html>]. Paper presented at the Ninth International World Wide Web Conference, Amsterdam, May 15 - 19, 2000.

CERI, S., FRATERNALI, P., & PARABOSCHI, S. Data-driven one-to-one Web site generation for data-intensive applications. *Very Large Data Bases. Proceedings of the Twenty-Fifth International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers. 1999 (1999), 615-626.

CHUNG, J.-Y., LIN, K.-J., & MATHIEU, R. G. Web Services Computing: Advancing Software Interoperability. *Computer*, 36,10 (2003), 36-37.

COLBERT, M., PELTASON, C., FRICKE, R., & SANDERSON, M. (1997, August 18-20, 1997). The application of process models of information seeking during conceptual design: the case of an intranet resource for the re-use of multimedia training material in the motor industry. Paper presented at the Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques, Amsterdam, the Netherlands. 73-81.

COX, J. E. The Changing Economic Model of Scholarly Publishing: Uncertainty, Complexity, and Multimedia Serials. *Library Acquisitions: Practice & Theory*, 22,2 (1998), 161-166.

DALTON, J. P., MANNING, H., & GARDINER, K. (2001). *Managing Content Hypergrowth*. Cambridge, MA: Forrester Research.

DUFFY, K. Content management comes of age: streamlining the information creation and delivery process. *Information Management & Technology*, 33,4 (2000), 183-191.

EAVES, M. Behind The Scenes At The William Blake Archive: Collaboration Takes More Than E-mail. *The Journal of Electronic Publishing*, 3,2 (1997), <http://www.press.umich.edu/jep/03-02/blake.html>.

FRASER, S. R. G. (2002). Real-World ASP.Net:Building a Content Management System. Berkeley, CA: Apress.

FRATERNALI, P. Tools and approaches for developing data-intensive Web applications: a survey. ACM Computer Survey, 31,3 (1999), 227 - 263.

FRATERNALI, P., & PAOLO, P. Model-Driven Development of Web Applications: The Autoweb System. ACM Transactions on Information Systems, 18,4 (2000), 323-341.

GENG, X., GOPAL, R. D., RAMESH, R., & WHINSTON, A. B. Scaling Web Services with Capacity Provision Networks. Computer, 36,11 (2003), 64-72.

GRUBER, T. R. A Translation Approach to Portal Ontology Specifications. Knowledge Acquisition, 5,2 (1993), 199-220.

HEITMANN, J. Content management systems for television production. EBU Technical Review,280 (1999), 24-34.

IEEE. (2002, 15/7/2002). Draft Standard for Learning Object Metadata. Retrieved 23/11/2003, from http://grouper.ieee.org/p1484/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

JEWITT, C. Digital chaos and professional standards. Multimedia Information and Technology, 26,2 (2000), 148 - 150.

KRUEGER, C. W. Software Reuse. ACM Computing Surveys, 24,2 (1992), 131-183.

LIEBOWITZ, J., & WILCOX, L. C. (1997). Knowledge management and its integrative elements. Boca Raton, Fla.: CRC Press.

LU, S., DONG, M., & FOTOUHI, F. The semantic web: Opportunities and challenges for next-generation web applications. *International Journal of Information Research*, 7,4 (2002).

M.E.S.O. (1995). Multimedia Educational Software Observatory. Retrieved 14/12/2002, from <http://europa.eu.int/comm/education/meso/>

MENDES, M. A. S., & ALMEIDA, V. A. F. (1998). Analyzing the impact of dynamic pages on the performance of Web servers. Paper presented at the CMG Proceedings, Proceedings of the 1998 24th International Conference for the Resource Management & Performance Evaluation of Enterprise Computing Systems, CMG. Part 1 (of 2), Dec 6-11 1998, Turnersville, NJ, USA. 539-547.

NORRIE, M. C., & SIGNER, B. (2003). Issues of Information Semantics and Granularity in Cross-Media Publishing. Paper presented at the CAiSE'2003, 15th International Conference on Advanced Information Systems Engineering, June 2003, Klagenfurt/Velden, Austria.

RENEAR, A., HOCKEY, S., & MCGANN, J. (1999). Panel: What is text? A debate on the philosophical and epistemological nature of text in the light of humanities computing research. Paper presented at the ACH-ALLC '99: Association for Computers and the Humanities and the Association for Literary and Linguistic Computing Joint annual Conference, June 9-13, 1999 in Charlottesville, Virginia. (<http://www.ach.org/abstracts/1999/hockey-renear1992.html>).

REYNOLDS, J., & KAUR, A. (2000). Content Management. Retrieved 25/7/2001, from <http://www.Microsoft.com.technet/ecommerce/contmgt.asp>

ROCKLEY, A. (2003). *Managing Enterprise Content: A unified content strategy*. Boston: New Riders.

SONG, Y., CLAYTON, M. J., & JOHNSON, R. E. Anticipating reuse: documenting buildings for operations using web technology. *Automation in Construction*, 11,2 (2002), 185-197.

STAHL, F. (2003). *Empirical Aspects on Content Management*. Paper presented at the Third Open Source Content Management Conference, Cambridge, Massachusetts (Harvard Law School).

UDELL, J. Databases get an XML infusion. *Information Age* (2003), 61-64.

WHITEHEAD, J., & WIGGINS, M. WEBDAV: IETF Standard for Collaborative Authoring on the Web. *IEEE Internet Computing*, 2,5 (1998), 34-40.

WILKINSON, R. (1998). *Document computing : technologies for managing electronic document collections*. Boston: Kluwer Academic Publishers.

WILSON, J. Electronic document management; Electronic documents. *Computer Law & Security Report*, 13,2 (1997), 124-125.

WRIGHT, M. A. Automating the business office. *Patient Accounts*, 19,10 (1996), 2-4.

YEH, J. H., CHANG, J. Y., & OYANG, Y. J. Content and knowledge management in a digital library and museum,. *Journal of the American Society for Information Science*, 51,4 (2000), 371-379.