

Documenting Business: The Australian Recordkeeping Metadata Schema

Glenda Acland
glenda.acland@uq.net.au

Barbara Reed
Barbara.Reed@recordkeeping.com.au

Ass. Prof. Sue McKemmish
Sue.McKemmish@sims.monash.edu.au

Records Continuum Research Group
School of Information Management
& Systems
Monash University
26 Sir John Monash Drive
Caulfield East, Victoria 3145

© 1999 Glenda Acland, Barbara Reed & Sue McKemmish. All Rights Reserved.

Abstract

In July 1999, the Australian Recordkeeping Metadata Schema (RKMS) was approved by its academic and industry steering group. This metadata set now joins other community specific sets in being available for use and implementation into workplace applications. The RKMS has inherited elements from and built on many other metadata standards associated with information management. It has also contributed to the development of subsequent sector specific recordkeeping metadata sets. The importance of the RKMS as a framework for 'mapping' or reading other sets and also as a standardised set of metadata available for adoption in diverse implementation environments is now emerging. This paper explores the context of the SPIRT₂ Recordkeeping Metadata Project, and the conceptual models developed by the SPIRT Research Team as a framework for standardising and defining Recordkeeping Metadata/ It then introduces the elements of the SPIRT Recordkeeping Metadata Schema and explores its functionality before discussing implementation issues with reference to document management and workflow technologies.

Introduction

Metadata has existed in record systems throughout time. Metadata, which can be generically defined as 'structured data about data', is simply a new term for the type of information that has always been captured in records and archives systems. But it is only now that the recordkeeping community has begun the process of the codification of recordkeeping metadata so it can be fully understood and employed both within and beyond our own profession.

Within the context of various metadata related initiatives in Australia and elsewhere, the SPIRT Recordkeeping Metadata project was envisaged to Proceedings of the 4th Australasian Document Computing Symposium, Coffs Harbour, Australia, December 3, 1999.

build a framework in which other sector specific metadata standards could be developed for targeted application.⁴ The major deliverable of the eighteen month Research Project, *Recordkeeping Metadata Standards for Managing and Accessing Information Resources in Networked Environments Over Time for Government, Commerce, Social and Cultural Purposes* is the Australian Recordkeeping Metadata Schema (RKMS), a high level extensible framework for specifying, standardising and mapping recordkeeping metadatas. Work is now proceeding on related research deliverables, including metamodeling of the schema in RDF* and ORM₇, the development of a User Guide to the set, and a prototype recordkeeping system that deploys the RKMS.⁸

The Context of the SPIRT Recordkeeping Metadata Initiative

In response to the policy directions announced in late 1997 as part of the Australian Government's Investing for Growth strategy⁹, a range of initiatives has been taken to support and encourage individuals and organisations to transact business electronically. They include initiatives relating to the establishment and accessibility of online government services and call centres. Information resource management initiatives are addressing challenges relating to dealing interoperably at the global level with facilitating resource description and discovery.

The thrust of government online initiatives is towards fully enabled online transactions as a significant component of service delivery. The *Electronic Transactions Bill 1999*¹⁰ is a model law which potentially provides the regulatory framework for the use of electronic communications in transactions (defined broadly to encompass all of the activities of government agencies in their roles as service providers). In the environment envisaged by the Bill, services will need to be documented and

 wondershare™

PDF Editor

instances of service delivery will need to be recorded. They will need to be clearly linked to what the agencies responsible for the services are mandated to do (what functions, activities, and transactions they are responsible for carrying out). The concerns of the national online service communities, which are closely linked to issues of client confidence, include the need for harmony, interoperability, improved access in global networked environments to services and service delivery points; and reliable, authoritative, trustworthy information about services.

To support an enterprise's business functions and activities in cyberspace, and to ensure the persistence and continuing accessibility of records of those activities that are of long term value to society, innovative, reliable and robust mechanisms are required to enable the continuing reliability and accessibility of essential evidence of business activities. Electronic recordkeeping and archival systems provide such mechanisms. They are fundamentally concerned with identifying, describing and classifying the functions, activities and transactions that records document. This can be contrasted with the fundamental concern with subject classification in library and information systems. Records document actions, not subjects - they record what an organisation does - its business transactions, the business activities of which they are a part, the business functions the activities carry out, and the broader societal purposes they fulfil. Linking records to the functions, activities and transactions they document and the agents involved is fundamental to establishing:

- who has the competence or authority to undertake the business activity or deliver the service,
- who is responsible for the business activity or service delivery,
- what instrument authorises action,
- how to access accountable information (records) about business transactions or instances of service delivery.

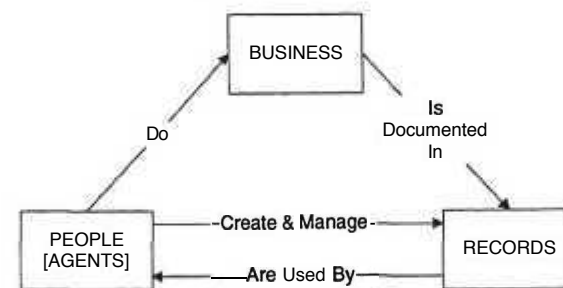
The recordkeeping community in Australia has been vitally concerned with the quality of public and corporate recordkeeping in electronic environments, recordkeeping-related issues concerning the reliability, accessibility and accountability of online activities and services, and the persistence and accessibility of records of continuing value to society. Major problems in electronic recordkeeping have been linked to the lack of controls, frameworks and standards in this rapidly evolving area. The response has been a proactive, innovative approach to the research and development role, epitomised in the involvement of the industry partners in the SPIRT 1998-99 Research Project.

The broader social context of the project relates to enabling society, government, commerce and individuals to continually access the information they need to conduct their business, protect their rights and entitlements, and securely trace the trail of responsibility and action. Maintaining authentic, reliable and useable evidence of transactions has significant social and cultural implications as records are a bastion of democratic and cultural accountability. They enable democratic rights of review and examination, and the transmission of our cultural heritage.

Framework for Standardising and Defining Recordkeeping Metadata

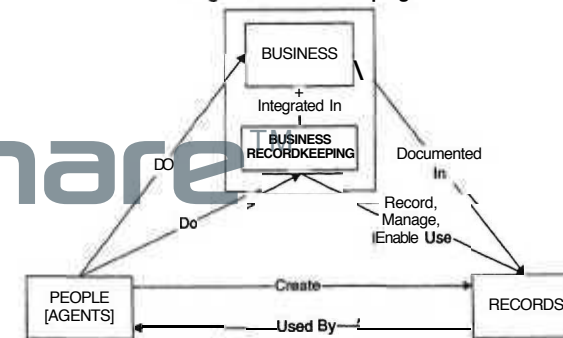
The Research Team has developed three high level models, Figures 1, 2 and 3, to provide the framework for standardising and defining recordkeeping metadata.¹¹

Figure 1: The Business



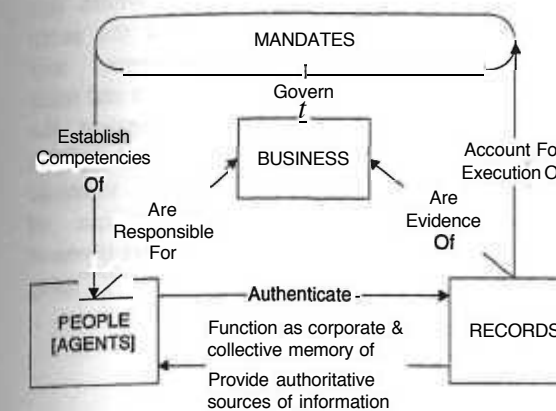
When people do business with each other, they create and manage records. The records created in the course of doing business capture the business done in documentary form. Business is here defined in the very broadest sense to encompass social and organisational activity of all kinds. A simple way of defining a record therefore is as a document that has taken part in a business process, and thereafter provides evidence of the transaction of that business. In distributed systems environments, records form a significant subset of an enterprise's digital information and knowledge resources.

Figure 2: Recordkeeping



Optimally recordkeeping forms an integral part of any business activity whatever technology is used.

Figure 3: The Business Context



People do business in social and organisational contexts that are governed by external mandates (e.g. social mores, laws, regulations, standards and best practice codes) and internal mandates (e.g. policies, administrative instructions, delegations, authorities). Mandates establish who is responsible for what, and govern social and organisational activity, including the creation of full and accurate records. Authentic records of social and organisational activity provide evidence of that activity and function as corporate and collective memory. They also provide authoritative sources of value added information. And they account for the execution of the mandate - internally and externally, currently and over time.

Recordkeeping Metadata

With reference to these high level conceptual models the RKMS is presented diagrammatically (Figure 4) as essentially concerned with three classes of entities - Business entities, **People/Agent** entities and Records entities, *as well as with* the external and internal mandates which are associated with Business, People and Records entities and govern the relationships between them. Furthermore, Business-Recordkeeping entities form a sub-class of the Business entity class.

The RKMS envisions management of records, agents and business at different layers of aggregation. A taxonomy of layers has been defined. (See Figure 4)

All these entities and their complex inter-relationships require unique identifiers and standardised descriptive metadata. The RKMS enables relationships to be set up between the layers of agent, business and record in addition to relationships within the layers. Any single record may have relationships which extend through layers of

aggregation in ways which establish a rich envelope of contextual metadata.

This complexity in relationships and their fundamental importance in defining the records context has been pushing beyond the requirements of other information resource metadata sets. While the conceptual understandings of relationships is well developed, issues to do with the taxonomy of relationships, the precision of the depiction of relationships and the metadata expression of such relationships is a further fruitful area for future research.

The Recordkeeping Metadata Elements Schema - Version 3.02

A highly structured set of elements and qualifiers has been defined (note that only the elements are represented in Figure 5). The view of the Schema provided in Figure 5 presents the elements in four sub-sets. This view is derived from the conceptualisation of records in their business context as depicted in Figures 1-3 above. The RKMS inherits part of the Australian Government Locator Service set and extends it to address the sector specific needs of recordkeeping.¹²

The elements and qualifiers defined in the Recordkeeping Metadata Schema identify and describe significant features of the business contexts in which records are created, managed and used. They identify and describe the people or agents involved, and the records themselves. They also link business contexts to the people or agents doing the business and the records that document it, and they reference the mandates that authorise and control business activity. They enable description and management of recordkeeping business functions, activities and transactions that are concerned with recording, managing and enabling the use of records, e.g. transactions and activities relating to the recordkeeping functions of appraisal, control, preservation, retrieval, access and use of records. They also provide for the tracking and documenting of the recordkeeping business itself in the unique metadata elements associated here with the Records entity.

Qualifiers in the RKMS

The RKMS qualifiers allow for a more detailed recordkeeping description, providing the facility to refine the semantics of the RKMS and to add precision to the values of the metadata elements. The RKMS has adopted the DC/AGLS application of three types of qualifiers,¹³ element qualifiers, value components and value qualifiers. The metadata

community is only beginning to explore the complexity of the schemes which govern and control metadata values.¹⁴

Scalability

As mentioned above, a significant feature of this high level set of metadata is that it is scalable, i.e. when it is implemented it can apply to records at any level of aggregation, to business and recordkeeping business activities ranging from an individual transaction to the societal purpose it ultimately serves, and to agents acting at any level in organisational and social hierarchies. An Entity "switch" has therefore been included in the set. In any particular instance the Entity Switch indicates whether a **Business, Recordkeeping Business, Agent or Records** entity is being described. Within each entity, the *CATEGORY TYPE* element then functions as a handshake, introducing the specific type of entity being identified and described:

Extensibility: Inheritance of Metadata

The RKMS envisages use of metadata elements, element qualifiers and value components from other metadata sets. Within individual elements, it is possible to extend the RKMS specification by referencing other schemata, e.g. the Pittsburgh Business Acceptable Communications¹⁵ Structure layer metadata elements and qualifiers could be used to extend the **Records: PRESERVATION** and **Records: RETRIEVAL** elements. Indeed the RKMS could inherit a full range of metadata elements, qualifiers, value components and prescribed schemes from another metadata schema for one of its entities.

The RKMS also envisages inheritance of the data values from another schema. Particularly when specifying metadata associated with agents and business, it does not seek to create separate recordkeeping views of these entities. Rather it enables reference to metadata sets defined in other circumstances. The RKMS also provides for the definition and an external validation of authority for such inherited sets.

A Framework for Mapping Metadata

One significant component of the research activity undertaken during the project has been an in-depth analysis of existing records and archives metadata schemata and standards. This was accompanied by the conceptual mapping of their elements in various combinations, followed, as the project advanced, by mapping the various iterations of the Schema against these related sets. The mapping processes which informed the development of the RKMS metadata set itself, point to one of its major uses - as a framework

in which other sets, targeted for application in specific sectors, can be developed and mapped. For example, the National Archives of Australia's *Recordkeeping Metadata Standard for Commonwealth Agencies*, released in June 1999¹⁶ was developed within this framework and can be mapped against the more comprehensive RKMS. Equivalences and correspondence can thus be made between it and other metadata sets, each one being read against the standardised metadata framework provided by the SPIRT Schema. The capacity for semantic interoperability of specific implementations of metadata when mapped against a standard set is one of the resulting benefits for the recordkeeping community, nationally and internationally.

Documenting Business

The RKMS as presented in this paper is modelled conceptually. As yet no implementation models have been attempted, although the **metamodelling** in RDF will enable the expression of the metadata in XML and its use for information resource description and discovery purposes as well as the description of agencies and services. Indeed the Schema is implementation neutral, defining no technological restrictions on how its elements are to be incorporated into systems, nor presuming any particular software architecture. It does not specify where, when or how metadata will be captured. The concern over time is that wherever, whenever and however metadata is captured, it will remain persistently linked to the record. Although metadata standards per se cannot guarantee such persistent associations, they can clearly demonstrate that assuring such persistence is an implementation imperative.

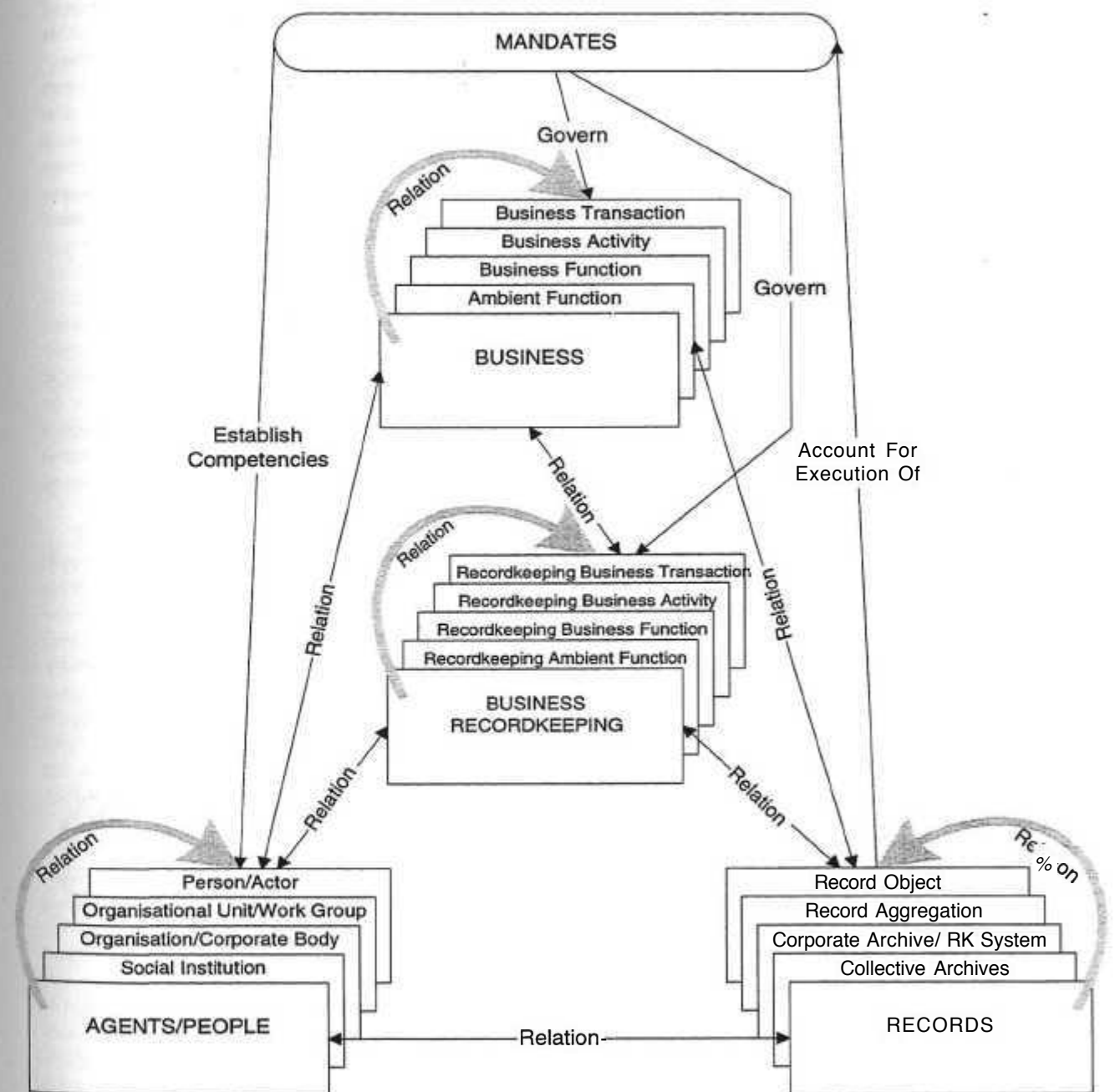
Implementors of the RKMS are enabled to identify and exploit a variety of technologies to populate the RKMS element fields. In a typical workplace, document management systems may be mandated to control the creation and dissemination of document level records, personnel systems mandated to map employees, their positions and their levels of authority, and workflow systems mapping information flows associated with business processes. Each of these aspects of seemingly disparate technologies are relevant to capturing specific metadata needed to produce reliable and authentic records over time.

The tendency in present records systems is to identify by user supplied tagging a variety of data elements which are then incorporated as contextual metadata around a document located in, or linked to, the records system. Such a response is appropriate where the risks of using parallel technologies to persistently associate metadata with the record are

judged to be too great. It is a common records-centric solution - if we cannot trust other systems to be sustainable over time, metadata cannot be merely

specific levels of records aggregation, as well as specifying agent and business metadata to be associated with the record itself. Elements of the

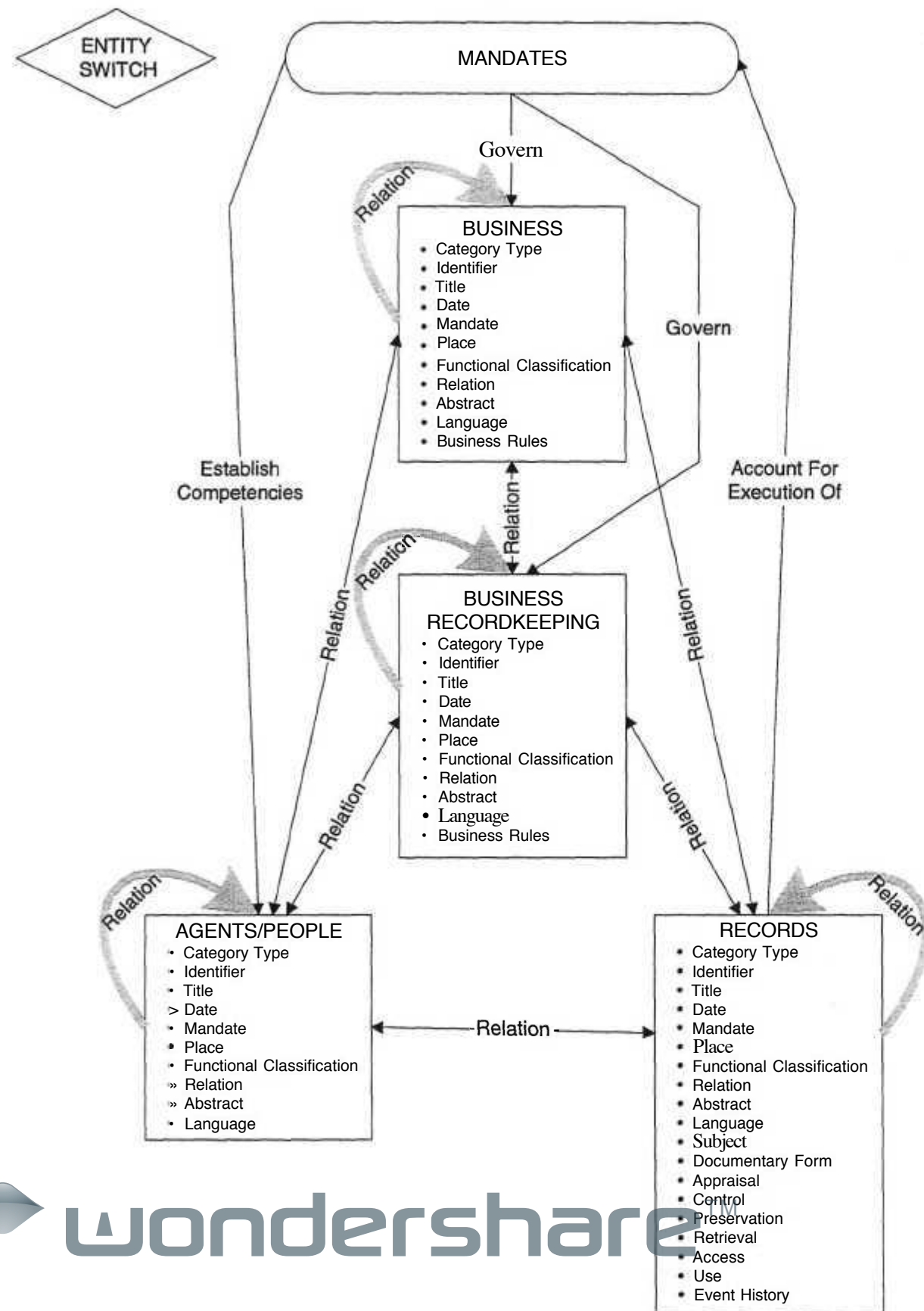
Figure 4: Coverage of Recordkeeping Metadata



associated with the record via pointers or links, but must be brought explicitly within the confines of the records system itself. This is the approach taken by the Victorian Electronic Records Strategy (VERS) project, amongst others. VERS defines metadata for

RKMS defined within the Records entity enable just such an approach. But whether this is the preferred solution depends on the circumstances of the implementation.

Figure 5: Recordkeeping Metadata Elements



While a variety of tags may be automatically attributed, a significant proportion cannot. An alternative approach would be to overlay the production systems with mechanisms to grab the document as it transacts business (perhaps as it is communicated beyond specified work group boundaries). Such mechanisms would associate the record with metadata from the document management system and any workflow or knowledge management systems engaged in the business process, as well as with data from the personnel system documenting the creator. This associated metadata would include system descriptions and dependencies. The associations between the record and its contextual metadata may be made by direct links into the nominated systems, by creation of metadata around the record created in robust formats lodged as discrete items into a storage location, or by embedding the metadata within the record¹⁷. As a possible means of 'future-proofing' records, these latter approaches appeal to recordkeepers, especially as the complexity of managing associations and links to disparate systems over time is, and may prove to be, an unsustainable burden for current technologies and the organisations that support them.

Such potential implementations begin to empower recordkeepers to connect with the newly emerging computer paradigms of component programming and non-proprietary, process specific program functionality. What we need to achieve are records which contain or are associated with all requisite metadata (from wherever it may be found), which are sustainable over time and over distributed network spaces. Alternative implementation strategies like this are envisioned by the RKMS, which looks ultimately to the concept of self-managing objects.

The Recordkeeping Metadata Schema encompasses more than documenting the immediate circumstances of creation. In implementing the Schema, organisations can determine the extent of the reach of their systems. If, for example, the records are only of relevance to a discrete organisational group, located within one area, the metadata may be minimal, as we can assume that contextual organisational knowledge will be implied. If a record's reach is beyond the organisation - as increasingly more documents are in distributed networks, with transactions enabled on the web via documentary carriers - then additional metadata which specifies these organisational parameters would need to be available to a wider audience to facilitate interpretation. The RKMS envisages scalable definitions of reach to be identified and configured into individual implementations through its layers of aggregation of organisation, business and record.

Defining the reach and the comprehensiveness of specific implementations will clarify for organisations the extent to which some or all of the elements are introduced and the ways in which the records created by business need to be 'bound' with metadata.

Conclusion

The RKMS uses recordkeeping understandings to make explicit connections between business, people who do business and the records which occur as a result of doing that business. It embraces traditional articulations of recordkeeping and enables future articulations. Much of the metadata work undertaken so far in electronic networked environments has been based on a passive notion of document-like information objects. The records and archives metadata community in Australia takes a different perspective in relation to records, regarding them as active participants in business processes and technologies rather than passive objects to be described retrospectively. Envisaging records as potentially self-managing information objects that act as the transactors of business has informed the SPIRT Recordkeeping Metadata Research Project. This vision links the dynamic world of business activity to the passive world of information resource in cyberspace.

The recordkeeping metadata approach is geared to implementation in an electronic environment in which doing business electronically and delivering services online is rapidly evolving. To be able to rely on the electronic business transactions which are, according to our politicians, our future, understandings of how to ensure these transactions are reliable and robust must be built into the new enabling technologies in an integral way. We come to the electronic business table with a firm proposal for incorporation into that agenda, one which is practical and implementable in a variety of ways. The RKMS is a tool for all players concerned with authoritative and reliable documentation that provides evidence of business transactions in electronic environments.

References

- [1] Bearman, David and Sochats, Ken. 'Metadata Requirements for Evidence', draft paper dated October 1995. Paper accessed via <http://www.oclc.org/5046/conferences/metadata/requirements.txt> on 2 September 1998.
- [2] Reed, Barbara. 'Metadata: Core Record or Core Business?', *Archives and Manuscripts*, Volume 25 November 1997 No 2, pp 218-242

- [3] Weibel, Stuart. 'The State of the Dublin Core', in *D-Lib Magazine*, Volume 5 Number 4, April 1999. Paper accessed via <http://www.dlib.org/dlib/april99/04weibel.html> on 16 April 1999.
- [4] Public Record Office Victoria, Ernst and Young and CSIRO. *Victorian Electronic Records Strategy Final Report 1999*, Public Record Office Victoria, 1998

Acknowledgements

We acknowledge the contribution of other Research Team members and associates, particularly Kate Cumming, APA(I) scholarship holder, Dr Nigel Ward and Dr Linda Bird, DSTC, and Geoff Acland-Bell (for the expert rendering of the conceptual models). We also acknowledge the input of members of the Project Steering Committee and our Expert Reference Group of Australian and international reviewers.

- 1 The Project involved collaboration between Chief Investigators Sue McKemmish, Monash University, and Ann Pederson, UNSW, and their industry partners from the National Archives of Australia, State Records Authority of NSW, Queensland State Archives, the Records Management Association of Australia and the Australian Council of Archives.
- 2 The acronym SPIRT derives from the name of the Research Grant which funded the Project, Strategic Partnership with Industry - Research & Training (SPIRT) Support Grant, which provides for joint funding by the Australian Research Council and the Industry partners.
- 3 Recordkeeping metadata is defined broadly to include all standardised information that identifies, authenticates, describes, manages and makes accessible documents created in the context of social and business activity. Recordkeeping metadata so defined has traditionally been captured and managed in both recordkeeping systems and archival control systems.
- 4 Within the archival community the ACA/ASA Committee on Descriptive Standards has endorsed the SPIRT Recordkeeping Metadata Schema as a framework for the Committee's future work on the development of domain specific recordkeeping metadata and archival descriptive standards. The Chair of this Committee has recently approached Standards Australia with a proposal to develop the SPIRT Recordkeeping Metadata Schema into a Framework Australian Standard for Recordkeeping Metadata.
- 5 The term "Schema" (plural schemata) is used to mean the semantic and structural definition of the metadata used to describe recordkeeping entities. A schema describes the names of metadata elements, how they are structured, their meaning and so on. The metadata community also refers to metadata schemata as metadata sets or specifications.
- 6 The Resource Description Framework (RDF) was developed by the World-Wide Web Consortium (W3C) to provide the foundation for metadata interoperability across different resource description communities, see: <http://www.w3.org/RDF>.
- 7 Object Role Modelling (ORM) takes a conceptual modelling approach that views the world in terms of objects and the roles they play. It is very expressive, enabling a high level

of detail and rigorous analysis, and can be populated with data instances which thus allows for grounded validation.

- 8 For background information on the project, see Sue McKemmish, Adrian Cunningham and Dagmar Parer, 'Metadata Mania: Use of Metadata for Electronic Recordkeeping and Online Resource Discovery' in Place, Interface and Cyberspace: Archives at the Edge, Proceedings of the 1998 Conference of the Australian Society of Archivists, Fremantle 6-8 August 1998. Canberra. Australian Society of Archivists. 1999, pp 129-144; and Sue McKemmish and Glenda Acland. 'Accessing Essential Evidence on the Web: Towards an Australian Recordkeeping Metadata Standard.' Paper for AusWeb99 Conference. Available at: <http://ausweb99.scu.edu.au/aw99/papers/mckemmish>. For details of project outcomes, visit the project web site at <http://www.sims.monash.edu.au/rcrg/research/spirt/index.html>
- 9 Information on the Investing for Growth strategy can be found at: <http://www.dist.gov.au/growth/html/infoage.html>
- 10 For further information see: <http://law.gov.au/ecommmerce/interim3.html>
- 11 The Project Team, in developing a simple but high level framework model for recordkeeping metadata given as Figure 1, used as an example of effective visual representation the INDECS Community's "Model for Commerce" as derived from David Bearman, Eric Miller, Godfrey Rust, Jennifer Trant and Stuart Weibel, 'A Common Model to Support Interoperable Metadata: Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities.' D-Lib, Vol.5 No.1, January 1999. Available at: <http://www.dlib.org/dlib/january99/bearman/01bearman.htm>
- 12 For details of the Australian Government Locator Service see <http://www.naa.gov.au/govserv/agls/>
- 13 The Australian Government Locator Service (AGLS) Manual for Users, Version 1.1: 1999-06-09, Office of Government Online, National Archives of Australia provides details of the application of these types of qualifiers - see <http://www.naa.gov.au/govserv/agls/>
- 14 Simon Cox has written an excellent discussion paper for the DC community on issues relating to structure, authority and qualification in DC: <http://www.agrc.csiro.au/projects/3018CO/metadata/dc-guide>
- 15 The Business Acceptable Communications model is described at <http://www.sis.pitt.edu/~nhprc/meta96.html>
- 16 See <http://www.naa.gov.au/govserv/techpub/rkms/intro.htm>
- 17 The interactions of the Monash University, Archives and Records Faculty team (Sue McKemmish, Frank Upward, Barbara Reed and Chris Hurley) with David Bearman, particularly at the Managing the Records Continuum intensive workshops in Melbourne and Canberra in June 1996 informed the ideas presented here.

Effective Reuse of Textual Documents Containing Tabular Information

L.E.Hodge, W.A.Gray and N.J. Fiddian

Department of Computer Science,
Cardiff University,
Cardiff, UK. '

{scmleh|wag|njf}@cs.cf.ac.uk

Abstract.

This paper presents an overview of a toolkit that can facilitate efficient reuse of tables appearing within textual documents [1]. In order to effectively reuse information contained in these documents, it is important that we process the accompanying text as well as any tables that appear. From this text, it may be possible to extract metadata such as descriptions of table content and related formulae, mappings and constraints. This metadata can then be exploited to enhance the value of extracted tables during their subsequent reuse. In this paper we present a discussion of the techniques used to process tables and associated text, both of which rely heavily on the use of regular expressions. Our techniques for locating tables utilise similar visual clues used by other table processing techniques discussed in the literature [5,6,7,8], although our approach to exploiting them is quite different. Our tools have been designed to provide a high level of support for the numerous types of table layout encountered in plain text tables, an area that has previously been somewhat overlooked.

1. Introduction.

The growth of the Internet has provided increased access to large bases of textual information that has potential for reuse. Of these documents, some of the most useful (in terms of reuse in information systems) are those that contain information in the form of tables. Tables offer a powerful mechanism for presenting related information in a clear and concise manner. The very nature of tables makes them suitable for reuse since it should be (relatively) straightforward to convert a table from a document into an equivalent table in a spreadsheet or database to facilitate integration with other information systems. Also, the fact that an author has gone to the trouble to present information in a tabular form, would indicate that they felt there was value in presenting the information in this way. This indicates to some extent

that the content of the table may have value that can be reused.

In this paper, we focus our discussion on the processing of documents in the form of plain text files (i.e. those that contain only ASCII text) as these are the most difficult type of human readable source document to process. Previous work on processing tables in plain text documents has resulted in a number of successful techniques for locating and processing tables. Unfortunately correct understanding of plain text tables has not always been possible using these techniques due to the large variety of layouts that can occur [2]. To overcome these problems, we have developed techniques that utilise the positions of components within the table to determine correct table layout.

Although previously, work has been undertaken to process tables that appear in textual documents, the textual component itself has been ignored. We feel that the text that accompanies tables can be useful in reuse as it often contains information relating to the content of the tables. In [3] we discuss the type of information that may be present in the text along with techniques to facilitate its location and extraction. This information can be employed as a form of metadata to augment any extracted tables and has two basic forms; textual descriptions of table content and technical descriptions i.e. formulae, mappings and constraints.

In this paper, we discuss the techniques we have developed to provide effective processing of both the table and the accompanying text.

¹ In our toolkit, we also support processing of documents that contain embedded mark-up such as HTML and LaTeX. Processing here is simplified by exploiting embedded tags.