t +64 4 499 9246
f +64 4 499 9245
www.knoware.co.nz

knoware

The Knowledge Warehouse Limited
Level 22 Plimmer Towers
2-6 Gilmer Terrace
PO Box 10 541
The Terrace
Wellington 6143
New Zealand

Data, databases and data warehouses: applying recordkeeping concepts

Trish O'Kane and Clare Somerville

This article is based on a presentation to the Future Perfect conference in Wellington in March 2012.

Introduction

Recordkeepers, used to a document-centric view of records, may struggle to think of how to apply recordkeeping concepts to data. Talk of "preserving databases" and even "transferring databases" just doesn't cut it in the data world. The picture is often much bigger than individual databases. Databases don't contain records – we will explain why. However, we can use databases to create well-formed records that we can preserve and we can preserve datasets. We can create requirements: for the creation and preservation of records generated from data; for the decommissioning of databases and for the migration of datasets. This paper addresses the key concepts for recordkeepers relating to data, databases and data warehouses, discusses good practice in data management and how we can connect it up to recordkeeping metadata.

The key messages are as follows:

- Long term preservation of data requires understanding how data is created and managed
- We have to work out both what **data** and what records the business needs to create and keep
- We have to identify what data must be unchanged
- We have to define "usable" and "retrievable" for data
- To manage information assets that are data, we should adopt and adapt principles from data warehousing and data life cycle management.

The problem

Databases have replaced many semi-structured records, for example: earthquake claims data, the Register of Births, Deaths and Marriages (and Divorces!). However, we want some of that information available, ***long term*** in a ***usable format.*** Many records managers are unfamiliar with the world of structured data. Here's a disposal outcome from a draft disposal authority we saw: "When database is decommissioned, transfer to Archives NZ". A database is not a box that can be transferred from one agency to another. The disposal outcome as phrased in the example is impossible. We have to look inside the "black box" to understand why.

We have data in databases, but we want ***both*** data and records. But what is a record in the context of data? Is it the individual data item or a whole dataset? Depending on the circumstances it may be either or both. We have customers for both data and for records and therefore we need to be able to access both, now and for the long term. Information assets (useful information) come in the form of records, transactional data in databases, datasets, data marts and data warehouses.

Key Concepts

Essentially there is a different relationship between data and its metadata than there is between documents and their metadata. The following table compares key definitions in recordkeeping with those in data management:

	Recordkeeping	**Data management**
Record definition	*Public Records Act 2005* A record or class of records in any form in whole or part, created or received by a public office in the conduct of its affairs	A record is a line of data in a table in a database
Attributes of a record	• *Documents the carrying out* of the organisation's business objectives, core business functions, services and deliverables, *and/or*	• Field types ◦ Numeric ◦ Character ◦ Date/time • Composite, derived

	• Provides *evidence of compliance* with any current jurisdictional standards, *and/or* • Documents the *value of the resources* of the organisation and *how risks* to the business *are managed*, *and/or* • *Supports* the *long-term viability* of the organisation	• Contains values
Purpose of metadata	The purpose of metadata for a document is to help with discovery, context and retrieval of text-based, visual or audio content. Document metadata describes the content of the document, e.g. its function, subject areas, and provides context of Author, Date created, and Date edited etc. It provides information that helps with retrieval, such as format. Documents can range from small to huge.	The purpose of metadata for data is granular. It seeks to define field type, e.g. numeric identifier, date created, currency amount, free text, so that the values in a field of data can be properly interpreted and used, usually by calculations within a single system or when passed between systems. Data fields are typically small, and limited in the number of characters they can hold.

Table 1: Key Definitions

System administrators used to believe that there was 1 byte of metadata for every 10 bytes of data. But those numbers are changing with metadata now exceeding data. But differentiating between data and metadata can become very confusing. Whether, for example, "Date created" is data or metadata depends on the level at which it is used or applied. For a document, "Date created" could refer to the date the document came

into existence. In a customer management system, “Date created” could refer to the date that a new customer was recorded into the database.

Metadata and data in the data warehouse

Business metadata is the link between database and users – it provides a road map for access for business users, analysts and less technical users. It establishes things such as:

- Structure of data
- Table names
- Attribute names
- Location
- Access
- Reliability
- Summarisations
- Business rules.

Technical metadata defines what data, from where, how, when etc. It is useful for developers, technical users, maintenance and growth, on-going development. It establishes things like:

- Table names
- Keys
- Indexes
- Program names
- Job dependencies
- Transformation
- Execution time
- Audit, security controls.

A comma delimited file shows each item of data is separated by commas, e.g.

0174, 000, A. “C”, “n”, “Y”, “Y”, F Dagg”, 03324

We need more information in order to know how to read and interpret this data.

The table below displays data using metadata as column headers for each field.

Values for each field are controlled by metadata rules (encoding schemes). Controlled lists restrict values in the fields for First Name (e.g. - text, up to 50 variable characters (VARCHAR), no spell check required) Gender (pick from M or F). Age is in years

and allows 1-2 (possibly 3?) numeric characters, Height is in centimetres and allows up to 4 characters with decimals rounded up to display as integers, Weight is in kilograms and allows up to 3 numeric characters, again rounded to display as integers.

First Name	Gender	Age	Age Group	Height	Weight
Alastair	M	29	mature	173	72
Nicola	F	18	youth	165	
Michael	M	20	adult	160	65
Kiri	F	19	adult	175	
Rebecca	F	17	youth	178	64
Mark	M	19	adult	170	70
Antony	M	18	youth	188	80
Sharon	F	19	adult	164	50
Kelly	F	18	youth	173	65
Peter	M	18	youth	180	70

Table 2: Data and metadata

The figure below shows five tables, with relationships between them. Categories (of products), Products related to Categories, Special offers on Products, Ordered items in a shopping card, and an Order which relates Products to a Customer. Metadata rules control the values allowed for each field and fields can be Null (empty) or Not Null (must have a value in them).

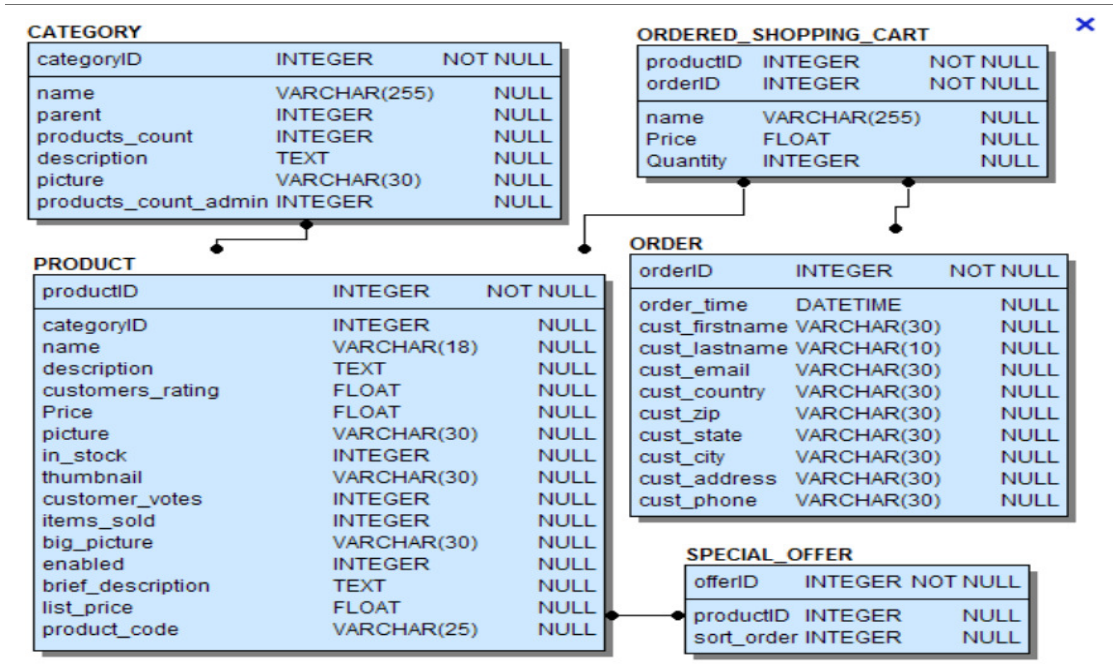

Figure 1: Tables and relationships

To make this data useable, three layers are required, namely:

- A user interface/application layer, allowing the user to browse through categories to products, creating “shopping carts” and purchase forms
- Rules and algorithms, setting out special offers, possible calculations, restricting the number of items that can be purchased in a special offer
- Data, e.g. lists of products, lists of categories, relationships between them.

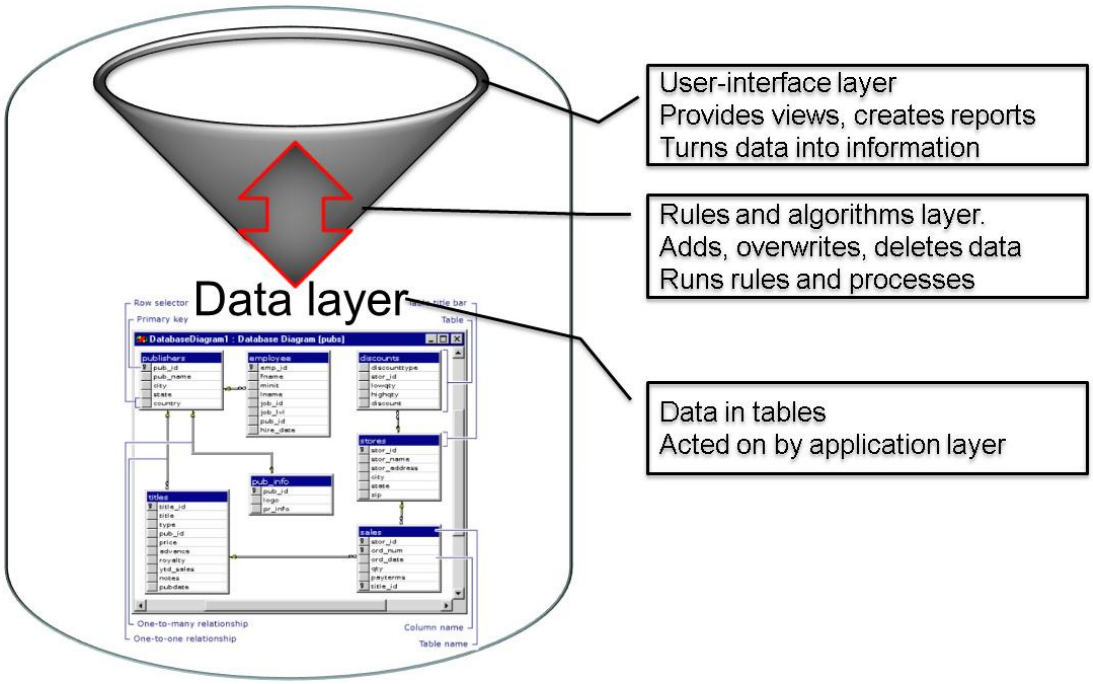

Figure 2: Layers required for useable data

Can data fit the Public Records Act definition of a record?

Given that we are “format neutral” in the management of records, then in theory some data could meet the definition of a record. To use previous examples, we assume that the Births Deaths and Marriages Register is a record, as are EQC claims data.
But even if a set of data doesn’t meet a strict definition of record, we should ask ourselves some key questions:

- If we exclude data from records management requirements, what is at risk?
- What is the impact of losing data, on the business, on individuals, on society, for the future?

The source solution is not a recordkeeping system

A source solution (a business operational system that generates and receives data) is typically not a recordkeeping system because it holds transactional data, not evidence

of transactions in context (records). In addition, the source solution isn't tamper proof. It is difficult to know exactly what the application layer is doing. Different tables and rows may be managed differently and depending on the function of the database it may not be possible to roll back to a point in time. Source solutions often overwrite 'redundant' data to run efficiently; compromises must be made between keeping a full history versus speed of operation, and business use has priority over recordkeeping.

In a source system, the data layer is not usable without the User Interface and Rules layers, but those layers are part of the proprietary software. It isn't possible to transfer a database, because it isn't possible to "transfer" a database. To transfer a database developed in for example Microsoft Access (relatively cheap and ubiquitous software) you would have to create an export file and transfer that file to the receiving agency. That agency would have to have licences and compatible versions of the software in order to view it or manipulate it, hardware onto which to load it, and it would have to maintain software that could retrieve those versions or would have to migrate the databases to later versions of the software. The receiving agency might be reluctant to accept such databases, even when they are offered as "gifts".

Inside a database

It's a "here today - gone tomorrow" environment inside a database. Metadata is at the transaction level. For example an activity about a customer is a record. Useful questions to ask about the structure of a database are;

- Is there a Unique ID for the transaction or for the customer?
- Where and when *are*/were components located? Are there multiple data tables in one database, or multiple data tables across multiple databases
- What are the table names and column names
- Are there standard names for elements across tables?

Source / business databases

Source / business databases have a number of common characteristics. Data is stored in tables in a normalised structure. There are high volumes of data, a large number of users, and many very quick transactions. Mostly data is overwritten, and history is retained to varying degrees. For example financial databases typically can roll back to a point in time, while bespoke databases probably are not built to do so.

Data warehouse

Figure three shows data that may be incorporated into a single repository from a number of Operational Source Systems. Data warehouses are centrally owned and hold transaction-level data including historical data. Data warehouses store and access large amounts of data, and are designed for large queries to support reporting and analysis. They are a central repository for all or significant parts of the data that an enterprise's various business systems collect. Inside the Data Warehouse, a range of business intelligence activities are undertaken including large queries. They are subject to unpredictable use which can result in unpredictable pressure on resources.

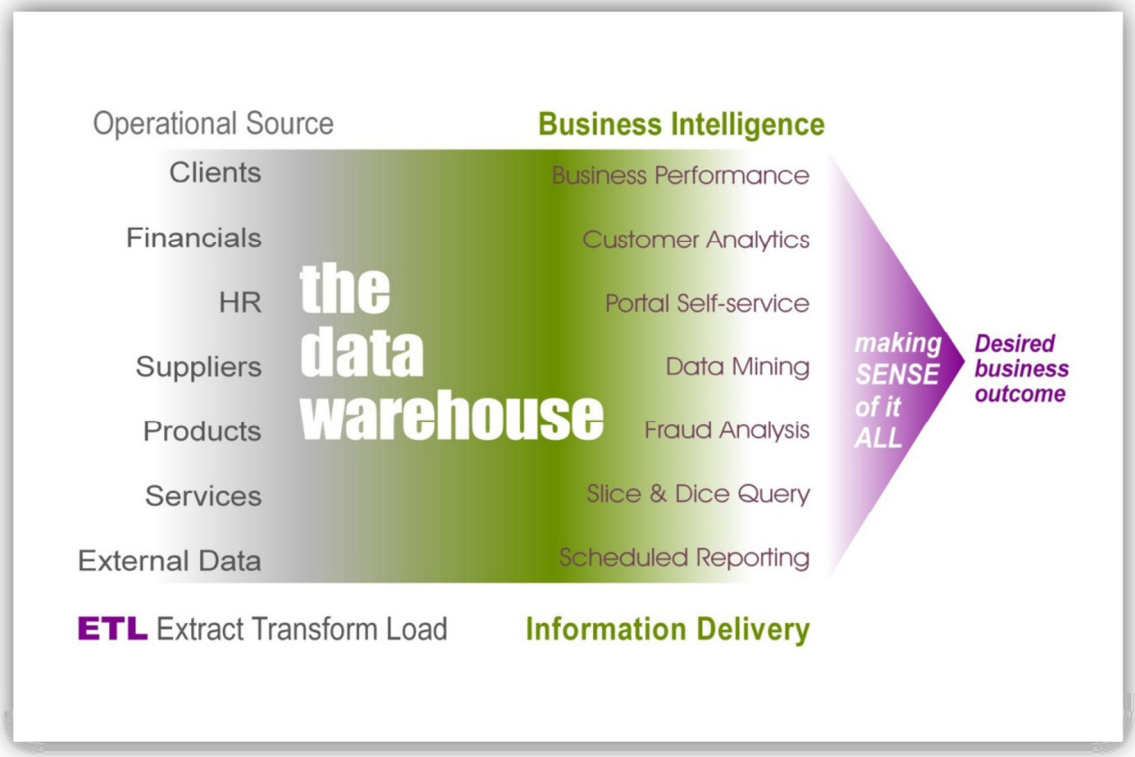

Figure 3: Source systems that may contribute data to the data warehouse

What is the simplest/most robust approach to deliver data and records from databases?

We recommend the following three approaches:

1. **Establish policy on creation of records from data**

It is important to document what authoritative records must be retained, what metadata about those records must be retained, what formats are acceptable and which (if any) records and metadata are considered transient artefacts e.g. format shifting duplicates, quality checking etc. Since most databases must get rid of transient

artefacts in order to function, it will be useful to update your disposal authorities to approve the destruction of transient artefacts as part of the normal functioning of the systems that dispose of them.

2. Create and export records from source systems

Given that databases are rapidly creating and deleting data, it is very important to identify what data tables and records are needed as important records and what can be produced via reports etc. Those records should be mapped to disposal authorities.
You are then in a position to identify which important data must be kept as records beyond normal retention in the system and which must be retained beyond system decommission. This enables planning well prior to decommission.
It is then possible to export those records in a suitable format and store in a recordkeeping system e.g. in data warehouse or EDRMS.

3. Persistently associate metadata

Appropriate metadata must be associated and retained with authoritative records. Identify data linkages between systems and retain those linkages, or consolidate metadata and associated record objects into one system, and ensure they are persistently associated. It is important to ensure migrated data/metadata/objects retain their context. and are not overwritten, for example date of migration should not overwrite date of creation, and the administrator doing the migration should not become the author.

The Future

Figure four below shows Business as Usual (BAU) transfers from source systems to both structured data warehouses and to recordkeeping systems, since both might be required for different purposes. If transfers happened during the normal course of business, then system decommissioning can be simple and quick.

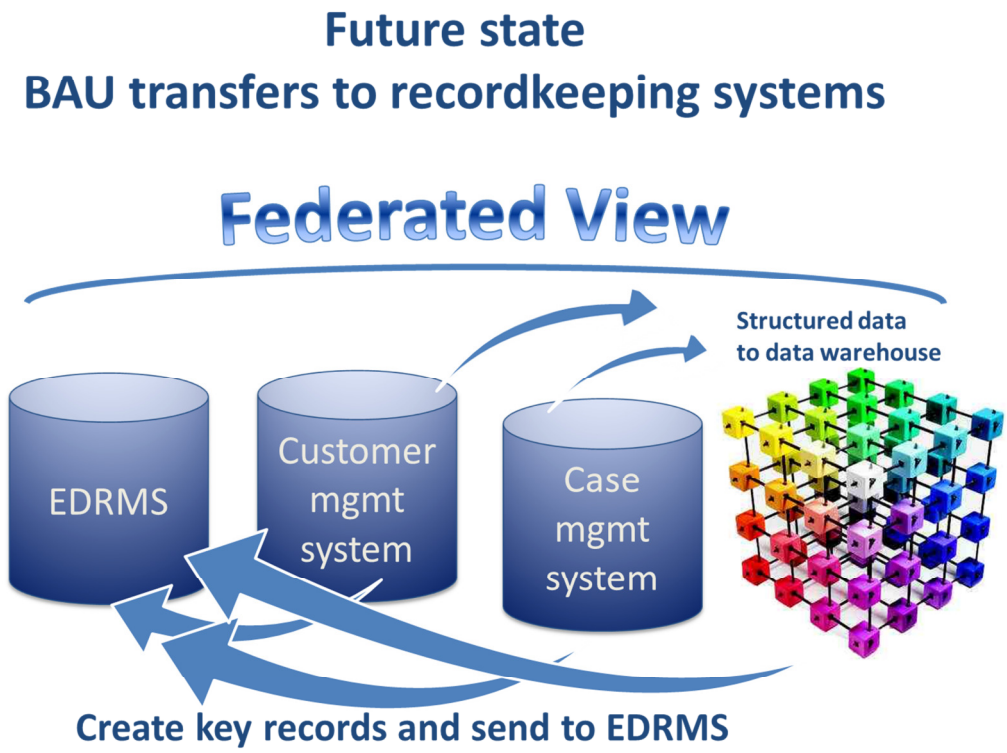

Figure 4: BAU transfers to recordkeeping systems

Data warehouses and good practice

In standard data warehouse practice, there are well-established principles for data feeds from source systems. Data must not be changed in any way, there must be no intervening processes, all changes to the data must be fully auditable and data feeds must reconcile to the source system when comparing the period of time the data covers, the number of records, the number of gigabytes etc.

Instead of transferring large amounts of data infrequently, it is better to transfer data that will be frequently used in the data warehouse, and relates to a point in time. It is easier to transfer data sets that are smaller, quicker, and easier to use. Subsets can be transferred daily, weekly or monthly.

It is typical to create a summary layer, for analysts to access. The summary layer is smaller and easier to manage. Data marts can also be created for the same purpose but for a specialized set of data.

Benefits of data warehouse

The benefits of data warehouses are that data is accessible; stored online; quick and easy to access; provides multiple sources of data; should be updated at least daily, provides a full history where it is possible to track everything. It is a tuned environment, with one version of the truth. In a data warehouse it is possible to do more – there is the freedom to explore.

However, data does not manage itself! Although much more structured than documents, data can be difficult and unruly – particularly large amounts of data. Standards and processes are essential, and roles and responsibilities must be assigned to the data warehouse team. Skills that the data warehouse team need are in the areas of data warehousing, data management, software, hardware, metadata, architecture, analysis, performance, and tuning. Ensuring the data warehouse is effective requires coordination, communication and marketing.

Data warehousing has been around for years. It has proven architectures, technologies, methodologies. It has a good infrastructure, but in the days of "big data", using data warehousing is being re-examined.

Challenges

Recent surveys have reported that data growth contributes to performance issues "most of the time", managing storage may cost 3-10 times the cost of procurement, and the average company keeps 20-40 duplicates of its data. This might or might not be comforting to those struggling with large numbers of duplicated documents. In the same way that records managers talk about needing to work with IT, business analysts have the same conversations. The following figure illustrates the risks associated with decommissioning. When systems are decommissioned, everything is at risk because everything must be either migrated or left behind.

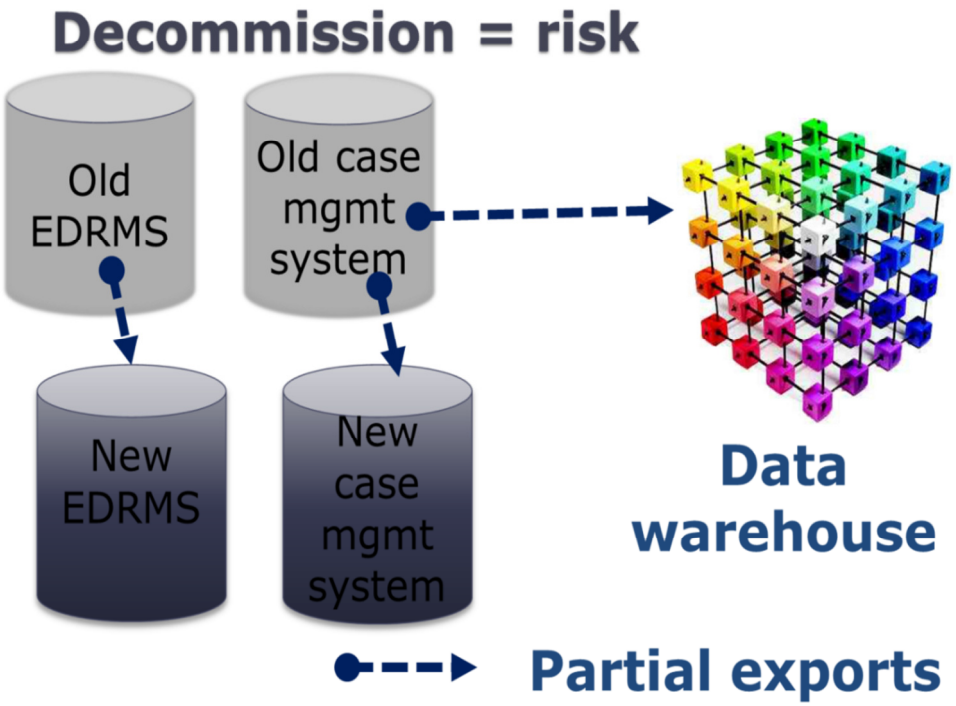

Figure 5: Risks associated with decommissioning of systems

Data lifecycle management (DLM)

DLM is about managing the flow of data, information and associated metadata through information systems and repositories, from creation and storage through to when it can be discarded. DLM recognises that the importance and business value of data does not rely on its age, or how often it is used. Data and information has value for strategic and operational business needs, for managing risk and for meeting legislative obligations. The value of information decays over time, and some information can be stored offline and some discarded. Occasionally, sometimes unexpectedly, older data may need to be accessed again, quickly, completely and accurately. DLM components show great similarity to recordkeeping models (see Figure 6).

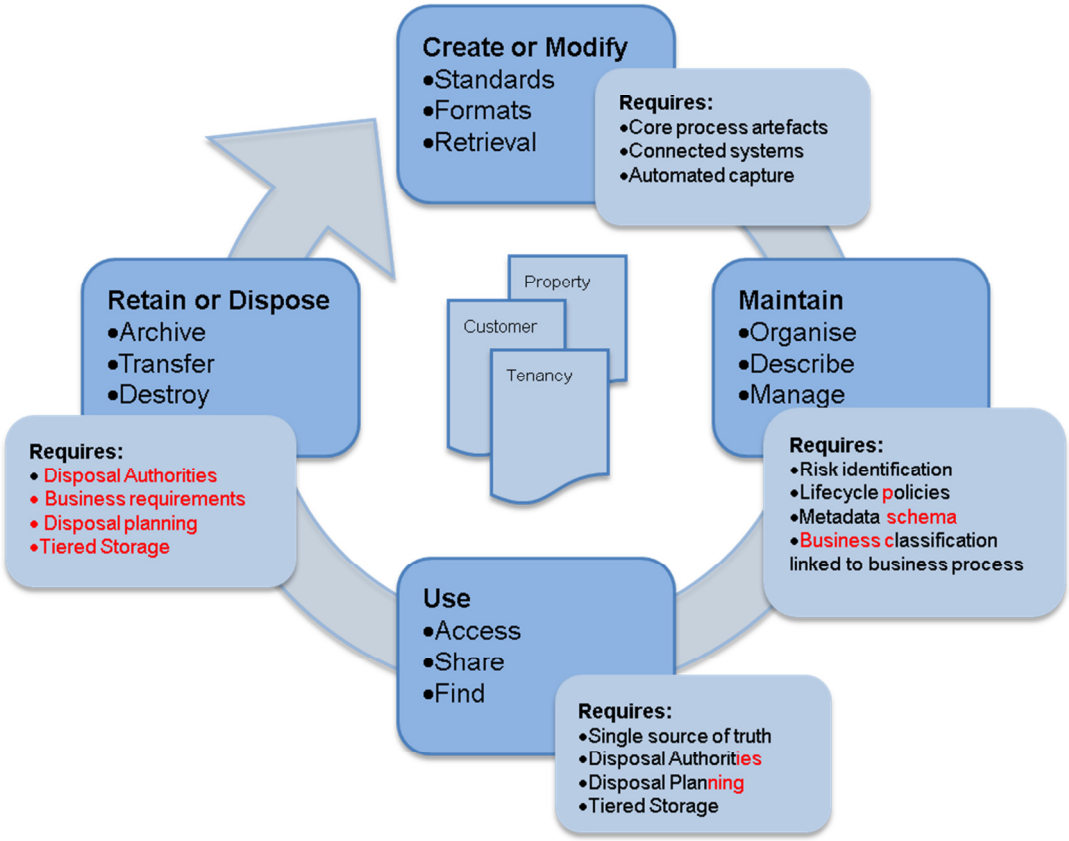

Figure 6: DLM components

Create and maintain

The Archives New Zealand Create and Maintain recordkeeping standard is applicable to data as well as to documents. The first principle requiring planning and implementation is clearly appropriate:

Principle 1: Recordkeeping Must be Planned and Implemented

1. Responsibility must be assigned from the Chief Executive down
2. Policies and procedures must be in place
3. Responsibilities must be defined and resourced
4. A recordkeeping programme must be in place including monitoring

The following tables show Principles 2-4, comparing compliance to the requirements between source solution systems (databases) and data warehouses.

Requirement	Data base	Data Warehouse
1. Functions and business activities identified and documented	☑	?
2. Records of business decisions and transactions must be created	?	☒☑
3. All records of business activity captured routinely into an organisation-wide recordkeeping framework	☒	?
4. Training provided	☑	☑

Table 3: Principle 2 Full & accurate records of business activity must be made

Requirement	Data base	Data Warehouse
10. Authentic: accurately documented creation, receipt, & transmission	☑	☑
11. Reliability & integrity, maintained unaltered	☒	☑
12. Useable, retrievable, accessible	☒	☑
13. Complete, with content & contextual information	?	☑
14. Comprehensive, provide authoritative evidence of all business activities	?	☑

Table 4: Principle 3 Records must provide authoritative and reliable evidence of business activity

Requirement	Data base	Data Warehouse
15. Identified & captured in recordkeeping framework	☒	☑
16. Organised according to a business classification scheme	☒	☑
17. Reliably maintained over time in recordkeeping framework	☒	☑
18. Useable, accessible & retrievable for the entire period of their retention	☒	☑
19. Contextual and structural integrity maintained over time	☒	☑
20. Retention & disposal actions systematic	☒	?

Table 5: Principle 4 Records must be managed systematically

Conclusions

A system that holds authoritative records, must be capable of recordkeeping, or must be made capable, or must transfer records to a recordkeeping system. The business owner should make the decision (with advice from IT) on whether to improve the capability of a system or transfer records. Data warehouses show us what can be done and give us some ideas on how to do it.

Looking ahead to the future, the following points are critical:

- The future state of data should be accurate, relevant and timely delivery of data and trustworthy information, where it is needed, in formats most appropriate to business need and future. This implies that data will be managed through its lifecycle to ensure there is a single source of truth, meeting the needs of users, and can be accessed in a timely manner. "Everyone who needs it" will include mobile workers, which is another reason why format is an important consideration.
- Information should be found quickly, whether it's old or new.
  Finding old or new information quickly implies: information is described

by metadata so that it has context and meaning, and systems use that metadata to locate relevant content and deliver information to users.

- There must be clear guidelines for systems and processes. We should keep what's needed for only as long as it's needed, and keep it in the right format so that it is useable. Clear guidelines imply that principles will be agreed on, which will decompose into business rules for systems. For example: that X information should be kept for Y years, then disposed of by following Z procedures
- Data must have recognisable value and appropriate levels of management, to address:
    - Business need: so we know what's important, and when it's important
    - Risk: so we're clear about what to manage, and how
    - Regulatory framework: so we meet legislative obligations
- Processes are in place to manage how business need is determined, and how risk and legislative obligations are managed.

Contacts: clare.somerville@knoware.co.nz
Trish.okane@knoware.co.nz