

## Understanding metadata

As the Department of Education (DoE) transitions into an integrated electronic environment, the complexity of the information held across the Department will grow. This complexity will lead to an ever increasing need for accurate and effective information retrieval and it is against this context that metadata has recently emerged as an important and topical issue.

The potential relevance of metadata to the Department of Education (DoE) lies in the ability to share in the costs or processes of metadata creation, the ability to re-use similar data in various metadata applications, and the capacity to use metadata to search for appropriate resources across different business areas. The utilisation of metadata within DoE is unlikely to produce the outcomes or benefits expected unless there is an adequate understanding of the basic concepts behind metadata use.

This paper attempts to focus on the key issues that affect metadata application and so enable the different business units within DoE to make realistic assessments of their requirements and strategies in relation to metadata.

### Metadata and modern electronic resources

Metadata has become a loose term that now basically includes any methodology and process, formal or informal, to describe and manage data. A variety of professions use the term metadata to describe aspects of data management , but the type and nature of the metadata referred to can vary significantly.

Formal metadata applications are characterised by a schema that defines the types of data that they expect to be included (eg date created, author, format). Recent national and international metadata developments in relation to electronic resources have seen the creation of a number of high profile schema (Dublin Core, EdNA, Commonwealth Recordkeeping Metadata Set, etc).

The development of such schema is not sufficient to enable DoE to simply implement appropriate metadata applications. The actual adoption, use and maintenance of metadata requires a mature or complete metadata application model containing:

- defined and effective data standards (a definition of the characteristics and controls for the data that go into the elements, eg vocabularies, thesauri, etc)
- a set of data creation business rules (rules that ensure accuracy and consistency when the schema and standards are applied to real-world

- resources)
- sustainable and reliable data creation processes (identification of those who create and maintain metadata, bundled together with the tools, training, resourcing, procedures, and quality controls to produce the desired outcomes).

An example of a complete metadata model can be seen in modern financial systems – the schema or type of data to be created and maintained has been defined (eg creditor, debtor data, etc.); the types of data to be entered have been determined (the budget codes, standard names for creditors, transaction types, etc), the data creation business rules are in place (how do we determine who is a debtor, when does a debtor go into default, etc), and the data creation processes are in place (we know who can enter data, who authorises certain stages, who produces reports, who reviews results, etc).

Financial systems have developed data standards, data business rules, and creation processes based on the input of financial experts. Any development of a mature metadata model for other purposes should also involve appropriate professional experts.

### **Metadata schema**

There are essentially three kinds of metadata schema relevant to DoE:

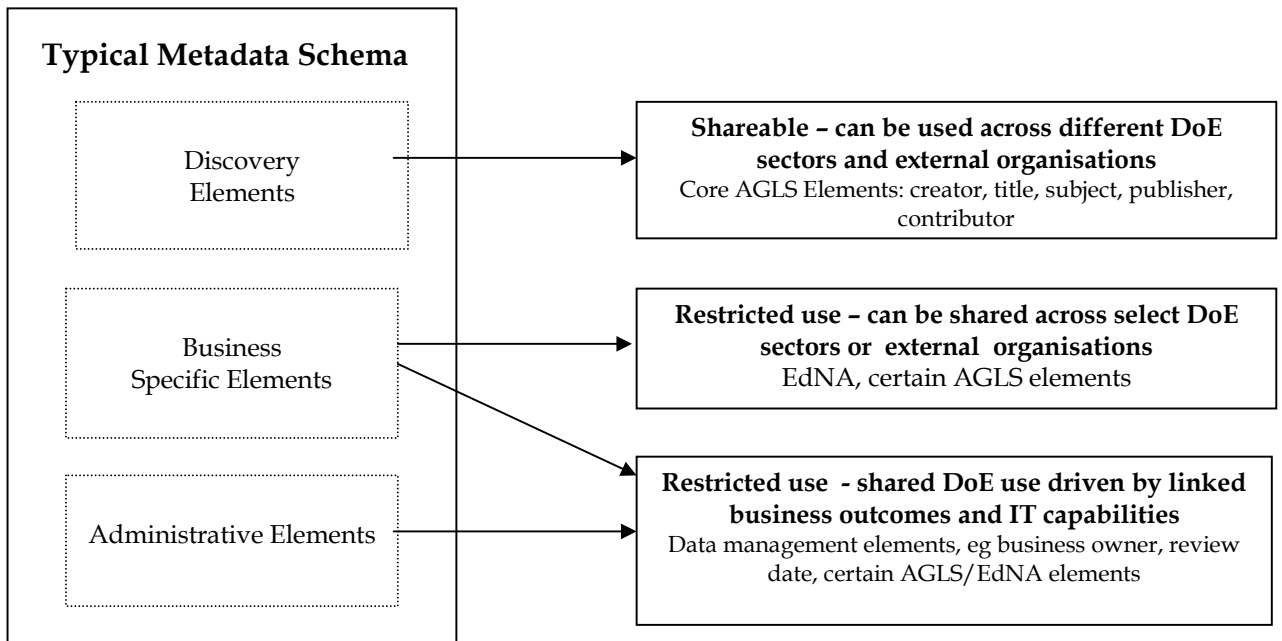
- metadata designed to help find the resource (discovery metadata)
- metadata designed to help manage the data itself (administrative metadata)
- metadata that supports outcomes specific to the business requirements of the unit or process (business-specific metadata)

Examples of discovery metadata include Dublin Core, and Australian Government Locator Service (AGLS). Administrative metadata is designed to control and manage the data resources themselves (eg information owner, date for review, etc). Examples of business specific metadata include the Record Keeping Metadata set.

The capacity to share metadata schema, leverage production costs and reuse the data components is determined by the purpose of the schema elements. For example, it makes sense to share discovery metadata elements as data retrieval is a common objective across sectors. In contrast, business specific and administration metadata elements are likely to have little relevance to other sectors unless those specific business structures and outcomes are also shared.

It is common for individual metadata schema to contain a number of elements that cover all three objectives, eg EdNA contains metadata elements for discovery, administration, and for specific business orientated (educational) outcomes. Metadata use within DoE should permit individual business units to utilise those schema that meet their needs whilst at the same time leveraging those elements that meet common requirements.

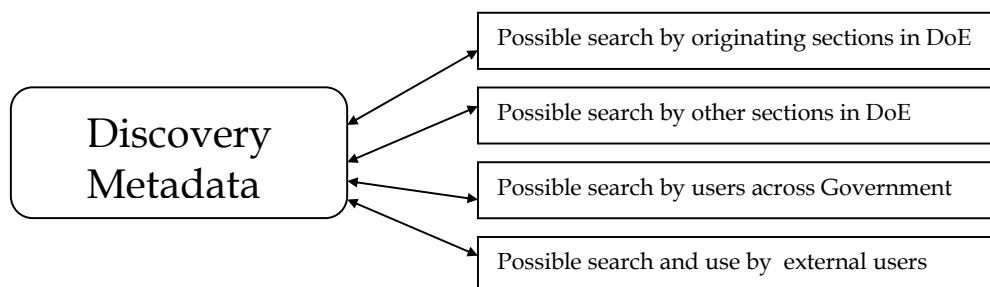
Discovery, business specific, and administrative outcomes require distinct metadata application models as they will each comprise quite different rules, standards, and data creation processes. A metadata application model that is designed to produce business or administrative outcomes will not, by itself, produce effective metadata for discovery outcomes. The following diagram illustrates the tripartite nature of metadata schema and the effect of this on internal and external metadata reuse.



*Diagram 1 - Shareable metadata elements*

### Sharing Discovery Metadata within DoE

Discovery metadata is distinct from administrative and business specific metadata in that the end-product must be useable by a variety of systems both internal and external to the Department. Metadata cannot simply be produced to meet a set of defined and specific needs because discovery needs are themselves variable (the needs of possible users and their underlying systems change over time). The discovery metadata itself must therefore be produced in such a way so as to be system-independent and neutral.

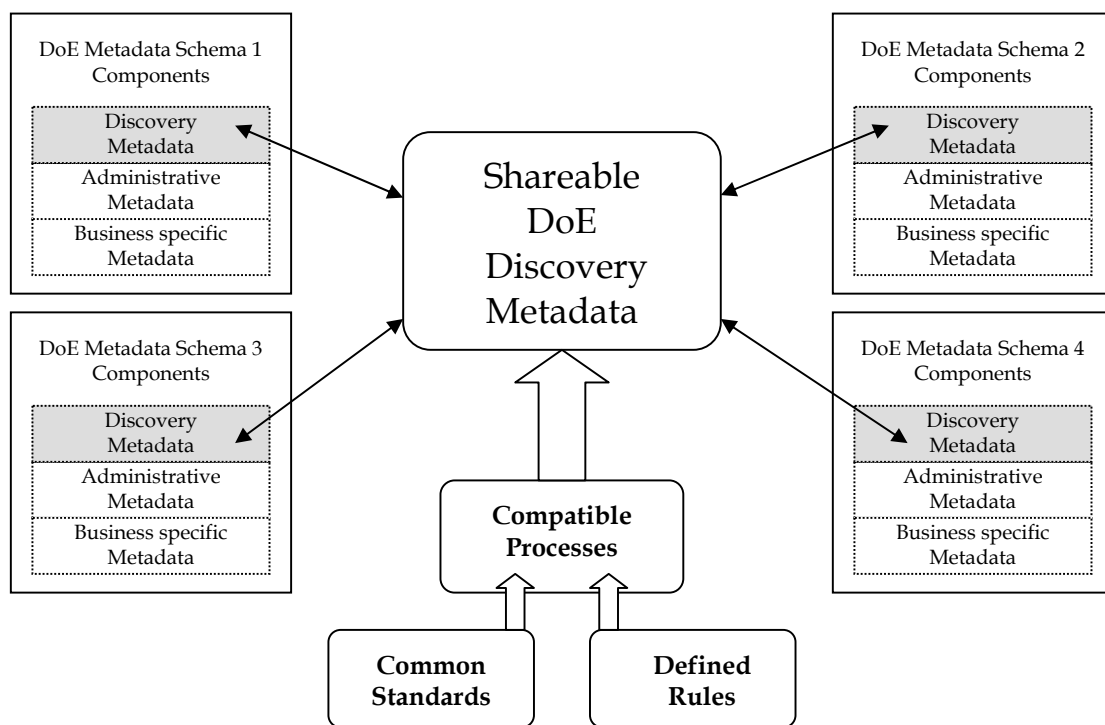


*Diagram 2 - Discovery metadata must meet a variety of current and future needs*

The production of discovery metadata needs an appropriate metadata

application model that is based on shared standards, rules and procedures. The development of a model able to produce effective discovery outcomes must leverage the knowledge and skills of information retrieval experts. To ignore this requirement is akin to building a financial system without input from financial experts.

The advantage enjoyed by the Department is that such expertise exists within the Department and that discovery metadata systems can be created that do not need to be duplicated or managed independent of this expertise. Through the use of library professionals the Department has the internal skills to produce both effective and cost-efficient discovery metadata. The nature of discovery metadata is such that its effective production requires this coordination and standardisation.



*Diagram 3 - Shared discovery metadata must also share process/standards/rules*

## Data standards

Most metadata elements in the discovery metadata schema (DC and AGLS) provide only general definitions as to the purpose of those elements, and rarely prescribe the actual data standard that should be utilised. This is particularly true of the fields for title, creator, contributor, and subject. The ability to share data in such fields requires common agreement on the type of vocabulary to be

used and it's structure. There is no point in searching for cats in a metadata schema that uses the term felines.

The creation of data standards requires an understanding of the retrieval needs of the end users. It is important to develop standards, not on the basis of how easy they are to create, but on how effective they are in retrieval. Various standards are possible, including controlled vocabulary lists (eg format types, geographic name lists, etc), thesauri (STO, Library of Congress Subject Headings, Australian Public Affairs Information Thesaurus, etc), and authority listings (eg Library of Congress Name Authorities). Thesauri are only one aspect of how data standards can and should be applied. Other aspects include data formatting (eg personal names), and keyword and phrase syntax rules.

### **Data creation business rules**

Most discovery metadata elements require a specific set of rules and guidelines in order to determine correct and consistent data values. We need application rules to determine who is the actual creator of a resource (given a variety of roles, people, and bodies often listed on a Web page), what the actual title is (given that this could be the HTML title tag, the most obvious text-based title available on the page, or the words contained in an equally obvious graphic). Likewise we need to agree on the level of specificity that should be applied in subject analysis (the subject terms and keywords used should be conceptually equal to the scope of the resource, if not the capacity to retrieve by subject is severely compromised).

Effective retrieval must utilise data that is not only correct, but that is also consistent. Consistency means that retrieval tools can add additional value to the searching process, and that users themselves can learn how to search and how to improve their own results. This type of metadata consistency requires well defined business rules and the processes to implement them.

### **Data creation processes**

The data creation processes must be based on the integration of the defined schema, standards, and application rules. For discovery metadata, this is a complex interaction that requires an understanding of the resource to be described, the data standards, and the business rules. This interaction must be based on an actual understanding of the resources, and cannot be based on the automatic creation of metadata (which is suitable only for certain administrative metadata elements). If adequate resources and processes are not provided to enable the production of effective discovery metadata, the whole point of discovery metadata will be lost and the effort wasted.

There are a number of models that can be applied to the creation of discovery metadata, but the most efficient way to deliver effective information retrieval outcomes, and to realise the infrastructure investments made, is to utilise the skills of information professionals. The production of discovery metadata is not a trivial exercise that can be managed without an adequate understanding of the full scope of the discovery resource, information retrieval principles, user needs, and the inter-relationship of the available retrieval tools and the desired outcomes.

Metadata creation can be provided through a single central service or via a distributed network of professionally-supported centres. The key here is that discovery metadata must be correct and consistent if it is to have any value. Research into the *Tasmania Online* search engine has shown that incorrect discovery metadata is not only ineffective, it is actually counterproductive in metadata-enabled searching systems, halving the relevance of items retrieved.

To meet this need for accuracy and consistency, it is vital that data standards are developed and promulgated, together with appropriate business rules and quality assurance processes. The discovery metadata processes and the standards that surround those processes should be monitored and administered through the use of those areas whose core business and expertise centre on information retrieval – the State Library and DELIC.

Lloyd Sokvitne  
Manager (Information Systems Development)  
State Library of Tasmania  
[lloyd.sokvitne@central.tased.edu.au](mailto:lloyd.sokvitne@central.tased.edu.au)  
27 November 2000