A-5. UNITING LIBRARIES AND ARCHIVES: HOW AN INTEGRATED METADATA STRATEGY CAN PRODUCE A CONNECTED RESEARCH ENVIRONMENT

Richard Gartner

Centre for e-Research, Department of Digital Humanities, King's College London
26-29 Drury Lane, London WC2B 5RL United Kingdom
email: richard.gartner@kcl.ac.uk

Although libraries and archives, both key resources in academic research, are inevitably symbiotically joined in many ways, often including their administrative and physical co-location, they are usually perceived as being far apart in their approaches to metadata. For historical reasons, each domain has evolved its own standards for this, often for practical reasons dictated by their divergent functions but in many cases following traditional imperatives which have their origins in the history of their development. In the analogue era in which many of these approaches were initially conceived such disparities could operate without any significant impact on the effectiveness of their respective operations: in the digital era, however, where the boundaries between libraries and archives become much more fluid, they can present major impediments to facilitating research.

Traditionally the archive sector has concentrated on collection-level descriptions, most clearly instantiated in the archival finding aid: these documents describe collections as a whole, hierarchically dividing them into series, sub-series, folders and so on, but rarely describing individual items themselves. Conversely, libraries have concentrated on the unitary collection object, usually the book on the shelf more often employ the same approach even when describing a running of a journal (which generally receives a single entry in a catalogue as if it was a monograph).

These approaches have been forwarded into the electronic age and into the metadata standards which attempted to move their respective cataloguing traditions into formats more suitable for the imperatives of digital data. In the archival world, the Encoded Archival Description (EAD), an XML schema for encoding and exchanging information of the contents of archives, effectively translates the structures and conventions of finding aids traditionally into a machine-readable syntax but maintains the same, rigid, hierarchical approach. The library sector, on the other hand, remained firmly focussed on its item-level viewpoint when it devised the MARC (MAchine-Readable Cataloguing) standard in the 1960s; this essentially translated the conventions of the card catalogue into the machine-readable age, maintaining many of its conventions (such as the notions of main and supplementary entries) which are essentially irrelevant for digital data.

For the researcher, archives and libraries are equally important resources; in order to establish a coherent research environment which does not allow important material to become invisible it is important to devise a metadata strategy which unites both approaches. One current initiative which is attempting to do this is the European CENDARI (Collaborative EuropeaN Digital Archive Infrastructure) project, which is attempting to provide a unified enquiry environment for existing archives and resources in the areas of medieval and modern European history.

The two subject domains covered by the project have polarised emphases in their metadata requirements which correspond neatly to the archive/library divide: the medievalists are particularly concerned with complex objects at the item level (for instance, manuscripts) whereas the modern historians relate more to the discovering, to be more exact to the presently undiscovered materials in existing archives which requires sophisticated collection-level descriptions. Uniting the two into a coherent, unified metadata environment would allow the two domains to develop into a single research tool.

Some components of this environment can already be encoded in pre-existing schemas. For instance, complex item-level descriptions, for instance, are handled by the METS (Metadata Encoding and Transmission Standard) XML schema, and, at the highest level, descriptions of collection-holding institutions are handled by the Encoded Archive Guide (EAG) schema. Among these, it is necessary to design a mediating XML schema which will allow the diverse components of this environment to be linked semantically.

Such a schema is designed specifically to act as an 'intermediary' schema, that is a schema which is not intended as a final delivery mechanism for data, but as a mediator between other established schemas. Using this technique not only allows the project to continue employing schemas which have embedded themselves in their respective communities (such as EAD or METS), but to link them into a coherent whole, so reconciling to some extent their divergent metadata strategies.

Only by uniting divergent metadata methodologies in this way may the full potential of digital resources be fully realised. The approach, suggested by this project, offers a way forward but to realise its potential requires the incorporation into the XML environment of semantic features which are usually seen within the remit of the 'semantic web' and its associated encoding mechanism, RDF (Resource Description Framework). This may be done using the extensive and sophisticated linking techniques of which XML is capable: by joining schema components together using controlled vocabularies employing Universal Resource Identifiers (URIs), it is possible to encode complex semantic relationships at any level of granularity.

There are many reasons why using XML in this way, rather than encoding these linkages directly into RDF-based ontologies, may be more practical for a working, unified environment. The

atomistic approach of RDF, in which each semantic component is encoded in a single subject-predicate-object 'triple', rapidly produces information networks of great complexity involving potentially thousands of triples when objects or collections of any size are involved. Maintaining, and particularly transferring between systems, such networks is highly complex matter: because of these reasons, using the readily-packaged XML syntax is a better option in working environments.

It is because, however, the most established schemas were not designed with linkages of this type as part of their architecture which becomes necessary to employ mediating schemas of the type proposed by CENDARI. By employing these, and incorporating semantic linking features as their core design feature, it becomes possible to allow these sophisticated networks of components to be integrated into a coherent unit. In this way, a unity between the divergent strategies and methodologies of archives and libraries becomes a real possibility and the now obsolete divisions between the two can, at last, be discarded.