Getting a handle on exhibition catalogues
the Project CHIO DTD

Richard Light
CIMI Consultant

Abstract

This paper describes work being carried out by CIMI (the Consortium for Interchange of Museum Information) on the analysis of exhibition catalogues. This is being undertaken as part of Project CHIO (Cultural Heritage Information Online). The project plans to use the SGML (Standard Generalized Markup Language) standard to express the structure and content of source materials, including exhibition catalogues. The analysis that was undertaken led to a particular view on how exhibition catalogues (and by extension, any text-based museum information sources) could be marked up to support retrieval of extracts relevant to a wide range of queries. The process of analysis is described, and the resulting design decisions outlined. The paper concludes with an assessment of the possibilities for information retrieval offered by this approach.

Background and history

CIMI and Project CHIO

CIMI (the Consortium for Interchange of Museum Information) is a consortium of museums and museum bodies which have come together to further the cause of pooling and exchanging museum information. This is an area which has been discussed both locally and internationally for many years, but with few practical results. CIMI aims to make progress at both a theoretical and a practical level.

The theoretical (perhaps 'design' would be a better word) level involves identifying existing standards which can help museums exchange data, and in developing generalized information models that can be used for practical interchange. CIDOC (the Documentation Committee of the International Council of Museums) has produced a Data Model which aims to encapsulate museum information in a very generalized way, but this is of more value for comparing different information standards than

for actually interchanging information. CIMI's work will build on existing work such as the CIDOC Data Model, and aims to provide a set of standards and procedures which is actually implementable.

The practical level involves prototyping projects which can be used to apply the standards and models developed so far. This has two major benefits. It provides a real-life test of the design framework, offering a major learning experience for consortium members and potentially showing up any problems at an early stage. It also provides 'proof of concept' systems which demonstrate the feasibility of the design approach. These can be used to educate and inform a wider audience within the museum community.

CIMI's current major project is Project CHIO (Cultural Heritage Information Online). This project aims to deliver at least 10,000 records of objects and information about Folk Art as a searchable online resource. This will include the full text of exhibition catalogues, wall texts, etc. as well as images and the more traditional museum database records.

CIMI and SGML

Some time ago CIMI decided to adopt SGML (Standard Generalized Markup Language: ISO 887) as an approved medium for the interchange of museum information. This standard offers a framework for encoding any document: you can use any tags you like, but you must provide a description of the tags you have used at the start of the document (as a 'DTD' - Document Type Definition).

One input into CIMI's education on SGML was a pioneering DTD based on the Art Information Task Force's Categories for the Description of Works of Art. This DTD showed that SGML could be used to model even complex museum concepts, but it also showed that a 'Data Dictionary' approach to design meant that the DTD would not be appropriate for modelling free text sources such as exhibition catalogues.

During 14, CIMI set up an SGML Working Group, which brought together consortium members (those with a particular interest in encoding issues) and SGML consultants. This group has pursued the issues of general SGML dseign principles and worked towards the production of a 'CIMI DTD.'

A second attempt at a 'museum DTD' was made by the current author for CIMI during the latter part of 14. This took as joint starting points the AITF DTD (for the 'museum' concepts) and the Text Encoding Initiative (TEI) Guidelines for Electronic Text Encoding and Interchange. The TEI guidelines and DTD were developed with Humanities texts in mind: thus they have some, but not total, relevance to museum texts. This second DTD was also unsuccessful, perhaps because the attempt to combine two already complex schemes led to an unwieldy and over-generalized DTD which lacked a clear sense of purpose.

About this time (December 14), the SGML Working Group met, and decided to identify 'genres' of information that might have different information requirements. The idea was that formats such as exhibition catalogues, brochures, and sale catalogues would contain distinctive information forms that could be modelled in a DTD that would work for that 'genre'. CIMI could tackle each genre in turn, and so produce a comprehensive set of 'museum' DTDs.

An analysis meeting held last May in Washington pursued this idea, and looked in detail at the content of exhibition catalogues and wall texts. They were found to have many of the standard features of published texts, with some more museum-specific features (such as concise and formalized object descriptions - 'tombstones') embedded within them.

The results of this meeting were considered in more detail at a design principles meeting held in Berkeley in July. The overall conclusion was that - for exhibition catalogues and wall texts at least - the generic TEI framework provided a good starting point for CIMI. In particular, a cut-down version of TEI called 'TEI Lite' was felt to offer a reasonable set of tags without excessive overheads. Rather than try for an 'all-singing all-dancing' set of museum-specific tags, the Working Group agreed to overlay the TEI framework with only those 'access point' tags required to support the access needs of Project CHIO. The CIDOC Data Model was consulted when designing these access point tags.

Later this month [August 15] CIMI members will attend a workshop in Halifax to practise marking up real documents with the DTD that has been developed so far. This exercise will doubtless lead to further revisions of the DTD, and will help to clarify how SGML should be applied to museum texts. By Autumn 15 CIMI intends to have a Project CHIO demonstrator system that shows these texts in retrievable form alongside collections database records. This will be a further test of the design principles behind the CIMI DTD.

Analysis of exhibition catalogues

During the analysis meeting in Washington, the SGML Working Group was presented with a wide variety of document types: collections database records, press cuttings, exhibition catalogues, etc. It decided to concentrate on exhibition catalogues, since (together with collections database records) they represented the bulk of the material available for Project CHIO, and were a good example of the sort of textual material we hoped to encode. Exhibition wall texts were felt to be sufficiently similar in terms of their actual text, so they were included as well.

It was agreed that database records had different requirements, and would need a separate DTD. (The challenge there is not so much designing a DTD that reflects Project CHIO's information needs, as in mapping a wide variety of database formats - on the fly - into a single SGML structure. But that's a challenge for another day!)

Although the Group did spend time looking at the actual content found in the sample documents, much of the meeting was actually used to discuss broader needs and design issues.

DTD design issues

One 'big issue' was the approach to adopt to DTD design: should we adopt a simple DTD that represents 'chunks' of text, and allows us to mark up access points within them, or should we go for something more detailed that allows things like lists, quotations, footnotes etc. to be marked up? After looking at examples of the types of textual feature found in the sample exhibition catalogues, the group came to the firm conclusion that it was important to be able to mark up all the significant features of the source document. A simplistic 'HTML-like' DTD would not be acceptable. This means that it will be possible to display the catalogue (or selected parts of it) so that it looks reasonably like the original source. (The group did not feel that it needed to match the printed source exactly, e.g. by replicating the actual fonts used.)

Should we invent a DTD from scratch, or adopt (and adapt) an existing one? There was no enthusiasm for the idea of inventing a brand-new DTD: many of the features found in the sample documents were common to any printed book. There are already a number of DTDs around which deal with books, and so could be used as a starting point. The Text Encoding Initiative (TEI) framework for prose was found to support many of the features we noticed, and so was felt to be a strong candidate.

information resource or document archive?

Another issue was the role of the SGML-encoded catalogue. Was it to be an information resource or a document archive? If it is primarily a resource then it is acceptable to provide relevant fragments if an archive then the whole document must be encoded and made available. The group went for the 'information resource' interpretation. The project goal, in principle, is to include the whole catalogue, but if logistics prevent this we won't be too worried.

This means that a mechanism is needed to ensure that any fragments are provided with sufficient context to identify their source (i.e. the document they come from, the section/chapter within that document, and any credit line required by the document's copyright holder).

access points

We looked at the issue of 'access points', i.e. items of information that are of particular interest to people who might search the Project CHIO database. Project CHIO aims to provide information to the average paying museum visitor, so an over-academic approach has to avoided. Existing museum schemes such as the Art Information Task Force (AITF) Categories for the Description of Works of Art and the CIDOC Data Model give us an embarrassment of possible access points. The problem is deciding which to concentrate on for Project CHIO. Analysis of visitor questions by Kody Janney and Jane Sledge provided us with some candidates, compatible with the CIDOC Data Model, such as Awards, Events, Marks, Materials, Objects, People and Time-Span. The sample questions they listed showed that these simple ingredients could (all too!) easily be combined into complex queries, e.g.:

> "What inscriptions are there on coins from the reign of Hadrian?"

Having identified that your enquirers want to know about (say) Materials, there are two rather different cases to consider. The first is where a material is mentioned in the running text, and you want to identify it as an interesting item of information that could be an access point:

> "... Possibly they [hooked mats] were made here [Nova Scotia] before 1840, but the craft did not burgeon until after mid-century. Before this, teriallinenmaterial was used as a ground fabric, but in the 1850s, terialburlapmaterial made of terialjutematerial fibres from India was introduced. ..."
>
> (Spirit of Nova Scotia, p.33)

This fragment contains mentions of three materials, one of which (jute) is used to make another (burlap). All three have been marked up with an SGML terial tag. If you search for 'material = linen', you would expect to be shown the point in the text where this term occurs.

The second use of access points is in describing the context, or topic, of a 'chunk' of text. In the above example, the preceding text has established that we are talking about hooked mats (this extract is taken from a 'Mat-Hooking' section within the introduction to the Textiles part of the catalogue) and Nova Scotia (the topic of the whole catalogue, but also mentioned explicitly in the previous sentence). Thus we want, in some way, to be able to say "this section/paragraph/random chunk is 'about' hooked mats". This then allows us to answer questions of the type:

> "Tell me about hooked mats from Nova Scotia"

In this case, the enquirer would expect to be directed to the complete 'Mat-Hooking' section, not to a particular point within it.

links

Our sample documents contain the usual types of intra-document links: from a point in the text to a footnote (at the bottom of the page or the end of the section), an illustration or a bibliographic citation from one section to another, etc. TEI provides a generalized linking mechanism which deals comfortably with all these cases. We assume that if we use this mechanism, it will be possible to have these links implemented as 'hot spots' which can be clicked on while browsing the text.

In addition we noted the need to point outside the document, either to a specific point within another catalogue, or to 'further information' about a person etc. mentioned in the text. (It may be, for example, that another catalogue has a biographical entry about the artist of the work being described.) The first case is also dealt with by TEI, but the second can be rather trickier. If you are pointing to a specific place where the biography is to be found, there is no problem, but in principle we want to have our markup say "link to anything you might have on this artist - no I don't know whether there is anything, or where it is stored." We have discussed (in earlier meetings) the idea of having a unique 'handle' for each access point in order to support this approach.

what's different about exhibition catalogues?

We started the meeting with the hypothesis that different types of museum information resource, such as exhibition catalogues, might exhibit unique features which distinguished them from non-museum documents and from each other. If this is the case, they would require, or at least benefit from, separate DTDs reflecting their different organisation.

At the 'macro' level we found that, while the organisation of individual catalogues varied widely, they all fell within the bounds of what the TEI defines as a standard book structure. In other words, they all have front matter, body text and back matter, and each of these parts is organised into 'divisions' (sections, chapters, etc.). Within divisions you get standard textual features (headings, lists, footnotes, citations, illustrations, etc.). So, in general, we found no immediately obvious characteristics which would distinguish exhibition catalogues from other textual sources. It's true that they have a higher than usual number of 'floating' elements (illustrations and text blocks with no obvious point of attachment within the main text), but this is just a matter of degree: TEI has a mechanism for dealing with them. Similarly, it is very important to Project CHIO that the language of each piece of text is identified clearly, but TEI also provides support to do this.

The group agreed that TEI offered a potential basis for our DTD. There is a cut-down version of the TEI prose DTD, called 'TEI Lite', and this was recommended as a starting point for customization.

'tombstones'

One feature which we did find is that most exhibition catalogues contain pithy object descriptions (called 'tombstones' within the group) which consist of a formalized set of key terms, optionally followed by a less structured free text description, e.g.:

> 2 Yarn Sewn and Hooked Mat with House
> Merigomish, Pictou County
> Wool on linen base
> Artist unknown (probably by member of MacKay family)
> Circa 1855-1865
> Dimensions: 163.0 x 82.0 cm (64 3/16 x 32 5/16 inches)
> Collection of the Nova Scotia Museum, Halifax, Nova Scotia

This is one of the earliest mats known from Nova Scotia which is both hooked and yarn-sewn with home dyed and homespun wool. The house which is depicted was built in 1852, and may be the home of the donor in which the mat was found, but which has been greatly altered through the years.

> Structure: Hooked rug all home-dyed, homespun wool yarn - some parts hooked, others "yarn-sewn" with a needle. Clipped pile, very thick in the center, thinning towards the edges. Handwoven linen backing with top and bottom selvedges. Hole lower right corner - patched hole upper center edge - pile on lower edge worn away. Refer to colour plate V, page 14

(Spirit of Nova Scotia, p.35)

The information given in the formalized first part of the 'tombstone' tended to be consistent within individual exhibition catalogues (or at least within sections of catalogues), but there was no consistency between different catalogues as to what information is provided, or what order it occurs in. The topic of the free text part of the 'tombstone' tends to vary even within a single source.

We found examples of 'tombstones' for people, organizations and (museum) events as well as for objects.

The CHIO DTD: design principles

Following the analysis meeting, we looked in more detail at issues (like language representation) which matter to CIMI in general, and to Project CHIO in particular. Our overall conclusion was that

TEI Lite was a good starting point, offering generalized ways of dealing with most of the issues identified.

TEI compatibility

We agreed to start, not from the TEI Lite DTD itself (which is fixed), but from the modifications file used to generate TEI Lite. We then modified the modifications (!) to suit our own needs, removing tags that are not required and adding back in 'standard' TEI tags that we wanted. Finally, we added a set of additional tags to express the access points required for Project CHIO.

All of this has been done in the manner prescribed by the TEI Guidelines, so hopefully the result will be a TEI-conformant application. This has two positive effects: it means that we should be able to use TEI-aware software directly on these texts, and it means that our exhibition catalogues are marked up in a manner which makes them useful to the TEI (i.e. academic) community worldwide.

access points

As noted above, we have developed a set of tags corresponding to the Project CHIO access points that are not already supported by TEI. (Some, such as person, organization and place names, were already included.) These are to be used for mentions of the topic within running text, as described above:

> "... Possibly they [hooked mats] were made here [Nova Scotia] before 1840, but the craft did not burgeon until after mid-century. Before this, teriallinenmaterial was used as a ground fabric, but in the 1850s, terialburlapmaterial made of terialjutematerial fibres from India was introduced. ..."
>
> (Spirit of Nova Scotia, p.33)

Some items of information originally defined as access points could also, or instead, be considered as 'attributes' of other access points. For example, a Role is usually played by a person or organisation. To deal with these cases we have added suitable attributes, for example a ROLE attribute for all 'name'-type elements.

normalization

When these access points are found within running text, the way they are expressed will often not be useful for searching. An obvious example is a person's name. At the very least, names are hardly ever expressed in inverted form (i.e. last name first), yet this is the form which is routinely used for

searching e.g. telephone directories. Beyond that, the actual form of name given may vary (forenames given in full, as initials only, or omitted), even within a single source. One possible approach is to analyse the components of names, but this adds extra marking-up work, and is no help when the information is not there to mark up!

The approach we have adopted (already supported by TEI) is to have the option of including a normalized form of the entry as a REG ('regularized') attribute, e.g.:

This regularized form should be compatible with the TOPIC attribute described below.

broader contexts

For defining the context of larger 'chunks' of text, we adopted a different approach. Past experience has shown that problems can arise if you invent tags with a specific museum meaning, and then try to get them to cohabit with the more neutral tags used to mark up the basic structure of the document. For example, if a paragraph () is also an object description (07SC2,5,0,0,0,0,0), what do you do? If you mark it as a only, you lose the piece of information which is most important to you. If you mark it as an 07SC2,5,0,0,0,0,0 only, how do you indicate that it is also a paragraph (for example, when formatting it for a browser)? If you mark it as both, does the paragraph contain the object description, or vice versa? ...

So we have decided to abandon this approach, and find some other means of indicating the context of a 'chunk' of text. We have provided a TOPIC attribute which can qualify any element within the catalogue, e.g.:

object.NMAA.180.54" ... p

states that this paragraph is 'about' the NMAA object with accession number 180.54. The TOPIC value has two parts: a prefix ('object') which says what type of access point is involved, and a normalized term ('NMAA.180.54') which is unique to this object. (This part of the TOPIC should be the same as the REG attribute used to normalize an object's name: see above.)

The standard TEI tags allow us to mark whole sections, single paragraphs and random parts of the text we can then assign them a TOPIC in this way.

This approach is not the most elegant one possible - for example if a paragraph has two topics they must both be declared in a single TOPIC attribute, because SGML attributes are not repeatable. It also lacks the formality of a scheme where every TOPIC is taken from a centrally-controlled, formally-approved authority list. However, it is felt to represent a sensible compromise that CIMI members might actually be willing to implement!

HyTime compatibility

There is another standard lurking in the wings: HyTime is a standard (based on SGML and fully SGML-compatible) for 'Interactive Open Hypermedia.' The SGML Working Group has agreed in principle that the CHIO DTD should be HyTime-compatible. It should be possible to achieve this without changing the design principles of the CHIO DTD.

One potential benefit of adopting HyTime is that all the retrievable resources can be 'pulled together' by what is called a HyTime Hub Document. This is an SGML document which has links to all the SGML and non-SGML resources in the system, acting as the 'glue' which holds together SGML-encoded documents, database records and images.

Retrieval possibilities

Once CIMI members have marked up their exhibition catalogues with SGML tags, the next challenge will be to use these resources to answer the qusetions posed by Project CHIO's public. SGML documents are a new kind of beast in the information retrieval world: they are not relational data, and are most definitely not 'free text'. What sort of possibilities will they offer for retrieval?

'tell me about ...'

Well, the easiest sort of question to answer will be an unqualified 'tell me about' one, as mentioned above, e.g.:

"Tell me about mat-hooking"

If one or more 'chunks' of text exactly match the TOPIC requested, they can be delivered as a response. The enquirer gets exactly what they asked for: some text 'telling them about' the person, object, or whatever they expressed an interest in. The fact that it is a piece of prose rather than a database record will probably be seen as an advantage by the enquirer. The only issue we have discussed in this context is: do you give the enquirer just the 'chunk' that answers their query, or do you turn them loose on the whole catalogue of which it forms a part?

Things rapidly get more complicated. If the question above is extended slightly:

"Tell me about mat-hooking in Nova Scotia"

the same subsection from Spirit of Nova Scotia is still relevant, but it is much harder to work out that this is true. You are still looking for 'chunks' with a given TOPIC: the complication is that one of the TOPICs (Nova Scotia) will have been applied to the whole catalogue, while the other (mat-hooking) only applies to a single subsection. It is clearly the subsection, not the whole catalogue, which should be found: the question is how you define a search strategy that will deliver this result.

One approach (which is intuitively obvious to people) is to say that every chunk 'inherits' the TOPICs of all its ancestor chunks. With this model, the subsection would have (at least) three TOPICs, two of them inherited:

method: mat-hooking

object: textiles (from the section)

place: Nova Scotia (from the whole document)

Then the search becomes simple again: you are just looking for a single chunk with all the required TOPICs.

Another approach (which might be easier for a computer to implement) is to search for the TOPICs 'in place', and then see if any chunk with a relevant TOPIC is nested within another chunk with a relevant topic. Again, the smaller of the chunks is the one retrieved.

access points in running text

So far we have just talked about 'retrieval by TOPIC'. What about all of those mentions of materials, people, dates etc. in the text which have been lovingly marked up?

For the purposes of Project CHIO I would suggest that these are viewed as properties of the smallest enclosing chunk with a TOPIC attribute. For example, the section describing materials that we used earlier occurs within the mat-hooking subsection (to which I have added a second TOPIC):

> " *TOPIC="method.mat-hooking//object.hooked mats"*
>
> .. Possibly they [hooked mats] were made here [Nova Scotia] before 1840, but the craft did not burgeon until after mid-century. Before this, teriallinenmaterial was used as a ground fabric, but in the 1850s, terialburlapmaterial made of terialjutematerialfibres from India was introduced. ..."
>
> (Spirit of Nova Scotia, p.33)

We can, in effect, add all the access point terms marked up within the subsection to the access points implied by its TOPICs:

method: mat-hooking

object: textiles (from the section)

place: Nova Scotia (from the whole document)

material: linen

material: burlap

material: jute

(Note that, if the REG attribute has been used to provide a normalized

version of an access point term, this is what should be used here.) This allows us, in principle, to use this subsection to answer questions such as:

"What types of object were made from burlap?"

"Tell me about materials used in hooked mats."

However, the nature of the association between individual access points and the overall subsection cannot always be clearly deduced. For example, if 'India' were marked up as a 6ceName, this could give a (false) answer to the question:

"Where were hooked mats made?"

Two ways round this problem are: be very selective as to what is marked up (only mark 'obviously relevant' access points) or go to the other extreme and do a complete semantic analysis of the text. I am not sure that the former approach will completely avoid 'false drops', and it might lead to missed matches. The latter approach is out of the question: marking up the access points and basic text structure is quite enough overhead for this project, and it is not clear that a search engine would be sophisticated enough to make good use of the additional markup anyway.

there's always free text

As well as these two types of structured information (TOPICs and access pointsin the text), we have the text of the document itself. While we are obviously hoping that our careful marking up and normalization work will lead to improved precision of retrieval via controlled terms, it would be silly to ignore the rest of the text as a potential resource for answering questions.

If an enquiry uses a specific word or phrase which has not been separately marked up, it should be possible to search for it throughout the text. Each word has a context, because it will fall within an SGML element of some sort. This is better than searching through a totally unstructured text. If the required word falls within an element with a defined TOPIC, then it can be thought of as an additional 'property' of that chunk of text. This allows non-marked-up words to be used as part of a complex query with more than one parameter, e.g.:

"Tell me about objects associated with sacking."

However, you have the usual problem with free text that you cannot specify that you mean the material 'sacking', as against the gerund.

authority files

The approach of naming TOPICs and normalizing access point entries will be most effective if it is applied consistently by all participants in Project CHIO. This suggests the need for some sort of authority file giving the 'approved' form of each term. We are going to investigate how far this can be achieved in practice as the project proceeds: there are clearly logistical difficulties in coordinating markup which is taking place simultaneously at several sites. Existing authorities (such as the Art and Architecture Thesaurus, and the Unified List of Artists' Names) will be used wherever feasible, but they won't cover all the ground. Certainly we have fought shy of designing a system which requires out-and-out consistency from the start (e.g. via what SGML calls Formal Public Identifiers). There will probably be a process of review and harmonization once some real exhibition catalogues have been marked up.

Another way in which authority files are potentially useful to this project is in the expansion of queries. If an enquirer asks for "material = wood", they are probably going to be interested in all mentions of different types of wood ('pine', 'oak', etc.). A materials thesaurus can take 'wood' and return all the more precise terms: these can then be added to the query.

We decided that this was the only feasible way to support query expansion: the alternative of adding broader terms within the text was rejected as inefficient and too labour-intensive.

query language

The retrieval side of Project CHIO will have to solve the intriguing question of what query language(s) to use. In principle, every question should be put, equally, to the collections database and to the SGML documents in the system. This requires as a minimum a common front-end, and facilities for mapping queries to at least two different query languages. In its later stages, Project CHIO will be using a Z3.50 application profile to support querying: it will be interesting to see what that can offer.

Providing a common front-end for queries still leaves the question of what query language to use for the SGML side of the project. At the present time (Summer 15) the SGML world has not standardized on a single query language, although I understand it is moving towards adopting one. The SQL language used by relational databases is probably not relevant: the query language adopted needs to express SGML document properties, and complex combinations of such properties. However, the fact that we are not planning to use the more arcane possibilities offered by SGML may make the task simpler.