DRAL

Daresbury Laboratory
Rutherford Appleton Laboratory

Rutherford Appleton Laboratory
Informatics Department

Semantic mark-up of generalised documents

S.A. Dobson, V.A. Burrill,
J.R. Gallop
24 August, 1994
WWW/01/94

Further to Eric Thomas' report on the recent meeting about possible projects involving Informatics and one or more libraries, this note seeks to expand on the technical aspects of the project on which we seemed to converge.

"Documents" and Mark-up

An early observation is that a document – the term includes both conventional documents like books and manuscripts, but also less obvious entities like films or scientific data sets – are more than simply a surface image. A document in the abstract has a collection of attributes, of which the physical realisation is but one.

As an example, consider a manuscript. This has a physical existence and (for electronic purposes) a digitised rendition on disc. It also has a number of other attributes:

- author(s);
- publisher, location of publication, data of publication;
- keywords;
- historical data (provenance);
- relationship to other works;
- chemical composition of parchment;
- description of the contents;
- ...and so on

The "document", in the abstract, is this set of attributes arranged into a semantic description of the document. The image of the document (all that is considered by many current systems) is only one element of this description – and, indeed, may be less interesting in many cases than other elements of the description. Another way of looking this is that a document incorporates multiple views and descriptions of itself as well as its raw data.

The initial idea is to extend the definition of "document storage" to encompass documents seen in this light: as semantic description networks which include multimedia objects such as images and which may cross-reference into those objects and to other documents. This may be seen as having the form of a document "marked-up" with a description of its

content. Document storage becomes the process of holding a set of semantic networks describing documents in a data store.

Search and Retrieve

One purpose of advocating such descriptions of documents is to enhance search and retrieval capabilities in document databases. Since there is more information available about each individual document, there is more scope for searching using complex queries.

The key point is that a document is marked-up by an expert, using the expert's terminology. This helps ensure that the mark-up captures the features of the document which would be of most interest to a user. Of course, documents might have multiple interest groups – a manuscript may be of interest to linguists and chemists, for example – in which case the mark-up should include both fields. This implies a flexible, extensible and progressively refined description format.

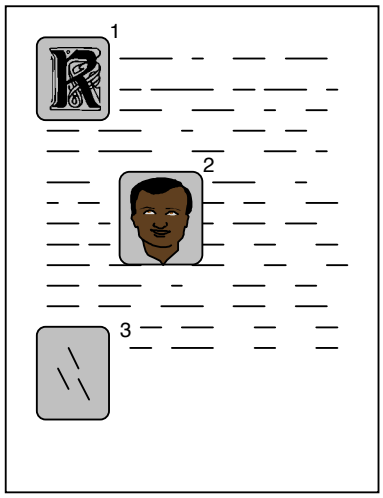
Since the mark-up of documents involves terms from specialised domains (*i.e.* "illuminated letter" or "vortex") it may be advantageous to make use of formal domain descriptions to structure the terminology being used. This may then be used as an aid to searching.

A related observation (derived from the MOVie project) is that a bare image of a document (or film) is essentially un-indexable as the image itself does not contain information in a form suitable for current query languages. The elements of the image are presented, not described. This leads us to the next point.

Images and other Multimedia Objects

The techniques of "host areas" used in many hypertext systems (including World Wide Web) can be generalised (as in MOVie) to mark-up film clips with "hot areas" which the user can select. In documents which have been extensively marked-up as described above, the types of links which can be made from an image are much improved.

Using the manuscript example again, consider the following sketch of a digitised manuscript image:



This is a simple image with three marked-up areas. Area 1 picks out an illuminated letter; area 2 an image; and area 3 a marginal note added to the manuscript after printing.

As an example of linking these areas to the document's mark-up the description may contain the fact that the manuscript contains an illuminated letter, linked in to the area of the document which surround that letter. A user examining the image might click the area to access information in the description; conversely a query to the document store of the form "find all documents with illuminated lettering" may then be satisfied, either by providing the entire document (with image, mark-up *et cetera*) providing only that part of the image and mark-up which are most immediately relevant.

Similarly the marginal note might lead to the known writer of the note, which might then lead to other facts known about that person.

In general we wish to mark-up arbitrary objects to point into (and be pointed to from) the semantic description of the document. This might include being able to play the soundtrack of a clip only when a certain person is speaking, for example. The point is that the *form* of data is related back to a description of its *content* in a form which may be manipulated.

Generalisations

So far we have concentrated on a set of techniques applied to static images such as manuscripts. However, there are several generalisations of the techniques which may be considered.

The first is the generalisation to moving images. This is essentially an extension of MOVie to include a more flexible, formal and complete description of the film clip. The existing MOVie technology be used directly, coupled to a mark-up back-end. This allows moving images to be indexed and queried.

A document, in the current context, might be regarded as any collection of information considered "as a whole". One may take the view that a collection of scientific observations (such as earth observation data or wind tunnel simulations) also constitute documents, and are so amenable to the techniques described.

In the remote sensing case, a document might be a collection of images from a satellite taken at a particular time, or a sequence of images showing the evolution of a hurricane. These documents may include both the raw, unprocessed data and pre-processed images highlighting certain phenomena. An expert in satellite image analysis could then mark-up the images to highlight interesting structures.

In the simulation case, the document might include the simulation algorithm, results, and an image of the final data set marked-up to show (for example) turbulence and shock waves. These cases obviously benefit from the use of visualisation, where that term is extended to include the mark-up of the visualised data set. The mark-up may be useful to some and, because the raw data may still be available, does not detract from the current "free-form" uses of the data.

One might also mark-up dynamic phenomena, such as (for example) an executing simulation or a functioning neural network – large data sets are not confined to physical sciences.

The third generalisation stems to an ambiguity above: what exactly *is* a document? As considered so far, a document is one or more objects collected for some related purpose, but the purpose may vary between users. Therefore, *a document is a collection of objects resulting from a query* – this may be fixed in the case of requesting all pages of a manuscript, or variable as when a scientific database is interrogated to find hurricane data. This has implications for the arrangement of data, especially for scientific data centres such as ESA.

Use in Distributed Systems

A large database is likely to be accessed remotely *via* Internet, so any set of techniques used to collect and present data must take account of network parameters and more general characteristics of highly distributed systems.

For current networks, the most obvious problem is that documents may be large entities. Images alone are large, and a document may contain many images and arbitrary amounts of mark-up, so a system must take account of the delays inherent in moving large amounts of data.

In many ways, the use of flexible document mark-up helps in this task. Since documents may be searched more precisely, there is scope for narrowing down a search at the server before moving data to the client. Moreover it may often be possible to move only portions of a document (as shown for the illuminated letter above) rather than the complete document.

Another possibility is to make use of the known characteristics of document parts to optimise their use, using techniques developed in TallShiP. For example, a film clip being played is accessed linearly: there is no reason for a client to acquire the entire clip from the server before beginning playing it, and no reason (necessarily) to retain parts already played. Thus the *type* of a (possibly multimedia) object may be exploited when determining how to move it round the network. The same comments apply to mark-up, which need not be wholly acquired before use.

Access through public systems like World Wide Web would necessitate the development of a new viewer (an external tool called through Mosaic, for example) which understood the mark-up and object types to optimise traffic.

Potential Applications

We can see many possible application areas for the sort of generalised document mark-up, storage and transmission we are advocating. In no particular order, and with a partner suggested (in a fit of optimism) for as many as possible, these include:

- manuscripts held in libraries (Bodleian)

- films and their sound tracks held by libraries (BBC archives?)
- scientific data centres holding collections of experimental or simulation data (ESA, HR Wallingford)
- collections of scientific papers, or newspapers
- annotation of teaching material for training people in analysis (MOD?)
- summarising results of data mining
- highly interactive catalogues (like the MIPS travel demonstrator)

Acknowledgements

Eric Thomas was instumental in organising the meeting from which this note emerged, and did a lot of "field work" withe the Bodleian and others. Many of the ideas described above have been crystalised by discussions with David Boyd.