Metadata Creation for Digital Humanities Projects

Martyn Jessop; King's College London; London; United Kingdom.

Abstract

The Centre for Computing in the Humanities at King's College London is a research centre that conducts highly collaborative research projects with partners from the academic community and cultural heritage organizations primarily in Europe but also further a field in North America, Asia, Africa and Australia. These projects encompass disciplines such as art history, social history, linguistics, literature studies and music. These projects frequently result in the creation of digital resources utilizing a variety of technologies and methods. They provide an opportunity for CCH to research issues concerning the digitization, design, implementation and delivery of such resources. This paper focuses on issues surrounding the compilation of metadata and the effects that these have by focusing of three case study projects.

Introduction

It is acknowledged that metadata performs many roles within a digital resource but it is easy to overlook the fact that it has to fulfil these roles for different audiences of the same resource. Carefully compiled metadata greatly enhances the value of a resource to specialist users but it can also open up the same material to a much wider audience thus adding substantial extra value and helping to fulfil the wider social role of humanities computing projects. However the compilation of such metadata, even with well defined schemas, is not a simple task.

The Case Studies

This paper discusses the challenges of metadata design and compilation and the effects they have on the audience and usage of the resources produced. The discussion will be conducted through three case studies projects at the Centre for Computing in the Humanities (CCH), King's College London.

The Corpus Vitrearum Medii Aevi

The Corpus Vitrearum Medii Aevi (CVMA) is an international research project dedicated to the publication of all medieval stained glass. The publication resulting from this exploratory pilot project of glass in England and Wales is published on a website [1]. The project digitized 12,500 images of stained glass windows and has made them available via an on-line database.

The main aim of this project was to satisfy the needs of the specialized academic community but also to seek ways in which the resource could be opened up to other academics within the art history community and elsewhere. The objects within the resource are also of interest to members of the public such as local historians and to many casual viewers to whom the high quality images of the windows will appeal. As the metadata for this project was designed primarily for the academic art historian it includes information concerning provenance and detailed metadata describing the type of building and the position of the glass within the building defined by a specialized nomenclature [2]. It also includes information describing the geographical location of the building and the age of the windows.

Many of the images were taken when the glass was removed from the building for restoration purposes. When in-situ these windows are often positioned high up in the walls of the building with restricted views. The resultant digital resource therefore contains many thousands of images that show views of the windows that are far superior to those that can be seen by visiting the original site. This is clearly of great value to art history scholars but it also makes them of interest to members of the public who are interested in local history or casual tourism visits to churches to view the glass.

One of the briefs of the funding body (Arts and Humanities Research Council) was that value should be added to the project by designing the interface in such a way that it could be used by a wider audience within the general public. We had to enable the more casual viewer to discover the resources without having a knowledge of the highly specialized classifications and terminology of medieval stained glass experts. The use of location metadata and a carefully designed resource discovery tool built around it played a key role in this. This interface allows users to locate and view stained glass by county, by place name, or by pointing and clicking on a map.

A further development has been suggested as a result of comments made by members of the wider audience who have seen the pilot project. The windows contain imagery that is of great interest to those who study religious iconography [3], whether from an artistic viewpoint, as a theologian [4], or in anthropological studies. This would involve adding a further set of metadata to describe the imagery used within the windows and increase the value, and audience, of the existing images within the resource. This work is now at an early stage of development.

On-line Slide Library for Classics teaching

The second case study is a much smaller project that was conducted for the Classics department at King's College London. The study of classical civilisations relies heavily on the examination the artefacts left by those civilisations. Researchers will travel to the museum, gallery or archive where these objects reside but this is impractical when teaching involves the presentation of multiple objects from scattered locations to a class of students. The traditional method of showing these objects to a group of students has been via the use of photographic slides and a projector, frequently using two projectors so that images can be projected side by side for comparison. In 2004 a major manufacturer of the projection equipment used for these purposes announced that they would no longer be making projectors and it became clear that the technology for presenting 35mm photographic slides would soon become obsolete. The Classics department at King's approached CCH to see whether it was feasible to create an on-line slide library that could be used in lectures via the local network and PCs attached to data projectors. It would also make the images available to students for private study on the College network. This project, Humslides, was designed with a much more limited audience in mind than CVMA. It illustrates how the design and creation of even the simple metadata intended for use by a relatively small closed community of Classics department staff and students produced problems that could have limited the use of the resource even among this limited audience.

The metadata schema was relatively easy to devise. The fields that were of importance to the scholars were caption, location (geographical), description, creator, keywords and date. These generated the standard problems of metadata describing historical objects. For example for location do you use the current place name or the name that would have been in use at the time, e.g. for a Roman object found in modern day Colchester should you use the contemporaneous name Camulodunum? Do you give the location where it was originally set up or the location that it is at now? For objects found in areas where a mixture of languages are spoken which do you use for the place name? How do you define dates, by century or more accurately? Historical data is rarely definitive or precise so the previous question raises an important general issue, how do you deal with ambiguity and uncertainty? How do you deal with missing values? Some of these factors can be accommodated by having a carefully prepared schema and guidance notes for the metadata editor but many revolve around scholarly issues and require not only input from experts in the field but also discussion between them to resolve ambiquities and differing opinions (or at least arrive at a context for decisions that can be stated in the metadata).

The project was intended to allow academics to contribute images that they needed for their own teaching. In this case the contributors are the primary end-users of the images, they are also the experts with whom the knowledge required to generate the associated metadata resides. The ideal workflow for this project would have been to allow academics to upload their own images and create the metadata for them. This proved impractical for many reasons. While many people were happy to provide images few were prepared to put in the long tedious hours of thinking about and typing in descriptive metadata. Five thousand slides were prepared for the project and of these we initially used 3600 on the site. Many of these had inadequate metadata, the lecturers who submitted the images knew the reference numbers of their own slides and could therefore find them and use them. This satisfied the basic level of functionality but completely failed the project's aim of opening up a shared resource as it would have been impossible to find images that were submitted by other users other than by browsing the entire collection. Even if this haphazard approach resulted in finding an image that could be of interest there was insufficient information in the metadata to confirm the identity and nature of the object.

Making the submitter responsible for the creation of the metadata also had an unexpected effect. The images were intended for a specialist audience of Classics lecturers and students so one would have expected any member of that group to produce metadata that was of value to the whole group. This was not the case as each submitter often had a very specific reason for selecting an image. The same image could be of interest to other members of the group but for reasons other than those anticipated by the contributor. For example consider an image of a courtyard containing a piece of sculpture standing on an inscribed stone plinth. This would be of interest to someone teaching about the history of the building around the courtyard, others might be interested in the sculpture as a work of art; someone else might be interested in the person depicted by the sculpture. In fact, in this case the image was contributed by a lecturer who wished to use it to study only the inscription and therefore omitted any further information from the metadata. This is understandable but seriously undermines the aim of creating a set of images that can be used by multiple users for a wide range of teaching purposes even within a specialized group. This problem was overcome by introducing an editor role between the submitter and the actual entry of the metadata into the system. This role was taken by someone who had knowledge of the subject material but was also able to see how the content of a particular image could be of value to more than just the submitter and to liaise with them to expand the descriptions within the metadata to cover the possible interests of many potential users.

Digital Image Archive of Medieval Music

The final case study is the Digital Image Archive of Medieval Music (DIAMM) [5] whose aim is to obtain and archive directly-captured digital images of European sources of medieval music. The project has created a new permanent electronic archive of over 14,000 of these images, both to facilitate detailed study of this music and its sources, and to assure their permanent preservation. The sources archived include all the fragmentary sources of polyphony up to 1550 in the UK; all the 'complete' manuscripts in the UK; a small number of important representative manuscripts from continental Europe; a significant portion of fragments 1300-1450 from Belgium, France, Italy, Germany and Spain. Such a collection of images that includes the complete British fragments has never before been possible, and represents an extraordinary resource for study of the repertory as a whole.

The project uses two distinct sets of metadata; one set records information about the capture of the images the other is drawn from existing public catalogues of the source materials. The image capture metadata includes standard photographic information and details of the digital image files and their preparation. It is rather more extensive than one might expect because the creators were trying to build in a degree of 'future proofing' by including information which although limited in value now may be of greater value in the future.

The catalogues on which the content metadata were based began to be compiled in the 1950's and the process took over twenty years. During that time the collections were often re-catalogued so the information contains multiple shelf marks for the same items. Each catalogue was compiled by different people using different criteria and in many different languages. These factors alone posed many problems for the project. However the most vexing problem was that the catalogue entries are written in free prose with no standard layout design and even within a single catalogue there are substantial variations in content that reflect the specific interests of the many individuals who compiled the entries; for example some went into great detail about the bindings and watermarks while others might dwell of the history of ownership of the manuscript at the expense of anything else.

Despite these limitations the staff soon became acutely aware of the richness of the free prose entries as they worked to split them up to form the basis for the metadata extraction. There were inevitable ambiguities, pieces of information that were missing from the original catalogues, and the occasional mistake. In an attempt to overcome these and some of the problems described previously it was decided to include a wiki-style feature of the website based resource that allows scholars to add their own annotations to the images, the intention being that they could supply missing data and correct anything that was wrong or contentious. It was anticipated that although precautions would have to be taken to prevent malicious or unauthorized annotations, and that there may also have been the occasional academic dispute between scholars, this mechanism would provide a useful tool for tackling many weaknesses of the metadata. In practice it was found that there was very little use of this feature; this is probably another facet of the problems surrounding persuading users to supply metadata.

Another possible strategy for coping with the complexity and variability of the metadata was to include a 'fuzzy' search mechanism. This has proved very difficult to implement for a variety of reasons and has not yet been added.

The project would be enhanced considerably by the inclusion of full text transcriptions of the material but this is far beyond the capabilities of the most sophisticated optical character recognition software. The only way of linking text to images in this project is by physical references, for example '4th line down, three inches from the left'.

The creation of the metadata was made harder because it had to be extracted from existing, non-standard, catalogues that were established as important sources themselves. In many respects it would have been easier to have created the metadata from scratch; however it had to be compatible with the standard pre-existing reference works. The project was intended to be primarily a collection of images of music manuscripts and the aim of the metadata was purely to support the image collection. The metadata creation was a difficult and time consuming task but it has proved to be the most popular aspect of the project among the users. The original catalogues were expensive books and therefore available in only a very few institutions. The availability of the standardized metadata derived from them through the website has greatly improved access to the textual content of the original catalogues.

Conclusions

These projects are very different in their missions, content, approach and principal target audiences but by studying them it is possible to draw out a number of common themes.

Each project has four types of audience

- The principal intended audience of scholars with a high level of knowledge about the content
- Students of the subject with a more limited level of specialist knowledge
- Scholars in disciplines other than those the resource was originally intended for who find the material useful in their own fields
- Members of the public with little or no specialized knowledge of the content

The metadata must be designed in such a way that allows each audience to find and identify the object that they are interested in. This can be facilitated by providing browse and search mechanisms that work at different levels of complexity, by using constrained searches that utilize drop down menus, and by graphical navigation aids such as interactive maps of varying scales.

Each project needs a metadata schema that can accommodate the inconsistency, ambiguity and contentiousness that often characterizes historical data. These show the importance, and difficulties of metadata compilation and the need to involve specialists in its creation. The metadata for CVMA project was perhaps the easiest to create because although the data was compiled from older catalogues the areas in which the metadata had to conform with these catalogues were limited. In many respects the CVMA metadata could be compiled according to a design of its own, effectively being created from scratch. The team responsible for this was very small, had full editorial control and on the whole the material was not contentious. The texts were all in English and were structured according to a standard format. The Humslides metadata was simple in design but posed a number of problems, not least of which was the fact that it had to be gleaned from a wide number of academics who were happy to hand over their slides for digitization but understandably daunted when asked to provide descriptive information for several hundred images. In many cases the slides came from two sources; small 'personal' collections and larger departmental collections. The small 'personal' collections of a few hundred slides were usually accompanied by detailed metadata that had been compiled by the submitter. The ownership of the larger departmental collections was often unclear or undefined and it proved to be far harder to obtain metadata for these more sizeable collections. The project was a pilot scheme and as such provided a test-bed for different ways of involving the contributors. Slides were put up with minimal metadata in the hope that this would encourage image submitters to contribute metadata, this worked in some cases but also resulted in complaints from users and, ironically, some of the submitters of the images themselves (who had failed to provide adequate metadata). Where metadata was provided it often reflected only the interests of the submitter and did not allow other users, to whom the images would be useful, to find them. The final project, DIAMM, experienced the greatest difficulties with metadata creation, but also produced a resource in which the metadata itself has proved to be, in the eyes of many of its users, more valuable to scholars than the content itself. The challenges here were integrate the metadata with existing public catalogues that are important reference works themselves but are of differing formats, approaches and languages. This is not an easy task and requires enormous amounts of subject knowledge, technical expertise and hard work over a protracted period of time but DIAMM shows that it can be done and does result in a very worthwhile resource.

Each project needs a workflow that maximises user contributions but ensures the creation of extensive good quality metadata that is suited to a range of potential users even within what are considered to be specialized audiences as well as the general public. It is essential that metadata is compiled by the recognized experts in their field but these are busy people who, while they may be fully committed to the aims of the project, have very little time. These projects have not found a solution to this. Allowing on-line access to metadata records through a Wiki-style service is an obvious solution but requires careful attention to system security and the issuing of passwords to authorized users. There can still be disputes between different scholars as to the content. The DIAMM project has shown that providing interactive online editing access does not in fact solve the problems of gathering contributions from hard-pressed academics. Humslides does perhaps show a possible way forward, in this case the metadata editor was a graduate student in Classics who could compile an initial entry based on his own knowledge and ask the relevant academic to comment on it. There was a marked improvement in contributions when this scheme was implemented. We were fortunate to have such a person working for us. It has been suggested that documentation could be produced to guide metadata editors. In practice this could only cover a limited range of situations and could not be used to extract metadata from the free-prose style descriptions that constitute the sources of many humanities computing projects.

Projects often go through a pilot stage and several phases of development with each phase being funded separately and not necessarily running contiguously. The aims of the project can easily change in each phase, frequently expanding or changing the focus of the target audience. An example of this is the decision to extend the metadata within the CVMA project to encompass religious iconography, this is an obvious feature of the content but is far beyond the original remit of the project. Very well designed metadata schemas may be able to accommodate this but it is more likely that the metadata will have to be extensible.

The content of metadata can be greatly influenced by the backgrounds and interests of the individuals who compile it. Their specialist knowledge contributes enormously to the success of the final digital resource. However care needs to be taken that the metadata reflects the broader purposes of the project and opens up the resource to as wider audience as possible.

References

[1] Corpus Vitrearum Medii Aevi (CVMA) at www.cvma.ac.uk, last accessed 01/03/2006.

[2] Marks, Richard, *Stained Glass in England during the Middle Ages,* London, 1993

[3] Kemp, Wolfgang, The Narratives of Gothic Stained Glass, Cambridge, 1997

[4] Rushforth, G McN, Medieval Christian Imagery, Oxford, 1936

[5] Digital Image Archive of Medieval Music (DIAMM) at www.diamm.ac.uk, last accessed 01/03/2006.

Author Biography

Martyn Jessop received his BSc from the University of London (1979). He holds the MBCS and CITP awards of the British Computer society and is a fellow of the Royal Photographic Society. He worked on university research projects in the sciences for fifteen years before joining the Centre for Computing in the Humanities at King's College London as a project manager in 2000. He has subsequently worked on a wide variety of digital humanities research projects.