# The pleasures and pitfalls of creating an image collection: quality and metadata issues

**A recent project for JISC Collections – Digital Images for Education – involved acquiring 75,000 still and moving images for use by the UK education community. Creating an aggregated image collection from a wide range of vendors raised considerable logistic, quality and metadata challenges. Those challenges included the need to follow European public procurement regulations; the need to evaluate all the images provided by vendors; the consequences of adopting a common technical standard for describing the content and issues around how the various vendors adapted their disparate metadata and taxonomic standards for the purposes of the project.**

***MICHAEL UPSHALL***
Project Manager
Digital Images for Education

## Background to the project

Late in 2008, JISC Collections received funding to develop and enhance JISC's existing digital repositories, via a project called Digital Images for Education.

Under European Union public procurement regulations, vendors were invited to bid to provide still or moving images. It was immediately apparent that a wide range of image vendors would bid, and separate procurements were set up for different goals. The first invitation to tender sought collections of digital images covering world events over the past 25 years. These images would record key people and events across the world in five broad subject areas: history (British, European and world), art and creative industries, social sciences, science and geography (with emphasis on exploration).

The second invitation to tender invited bids from public sector institutions – either from museums or the increasing number of universities and colleges with niche collections of digital content. One of these, for example, was North Highland College, based in Thurso, Scotland, who were successful with a bid to provide 30,000 images of social history from the Scottish highlands.

Successful bids would be added to JISC Collection's growing portfolio of licensed image content, which is currently available in three separate collections (Newsfilm Online, Film and Sound Online and Education Image Gallery).

## Challenges of procurement

The decision on which vendors to accept was not taken by JISC Collections, but by a panel of 12 volunteers selected from UK higher and further education establishments, who together had a representative range of educational level and subject area. The volunteers were given samples from each of the vendors. After reviewing the samples, the volunteers met at an all-day session where the successful bidders were selected.

The logistics of procurement raised some difficulties; for example, European Union public procurement regulations do not allow for negotiation over price. The evaluators could only accept or reject individual bids, which made creating a balanced collection rather difficult. However, in the end, 11 bids from ten vendors were accepted.

One limitation of the procurement process was that successful vendors were selected based on sample images or films, representing only a fraction of the total images they were finally to provide. The volunteers could not review all the material, but nor could it be accepted from the suppliers without assessment and comment. It was

necessary to develop a rigorous, but cost-effective system to check the quality of the material (both images and metadata) being provided, and a team of evaluators was recruited, mainly postgraduates with experience of working with Intute, reviewing websites for educational relevance. A web-based tool, the 'holding bay', was developed by EDINA, a UK national academic data centre based in Edinburgh which enabled every image to be reviewed with the metadata alongside the image itself. The distributed panel of evaluators were able to log in remotely at hours to suit themselves, ensuring that the project was able to be completed in reasonable time.

## The evaluation process

It was calculated that if the team of evaluators had spent just two minutes on each image in the collection, it would have taken around 1.5 person-years to evaluate the content. In practice, the situation was much worse, because they were also dealing with moving images that took time to watch – the project acquired an estimated 600 hours of moving-film footage, which would have taken 28 weeks just to play through once. Risk assessment was required to balance the cost of evaluating every single image or film against a robust sampling method. As a result, it was decided that the evaluators would sample each batch of content as it was delivered. Deliveries from each vendor were split into five or six batches and each evaluator looked at a representative sample from each batch. On the basis of the sample, they decided (with the aid of the project manager) whether a more detailed examination of the batch was required. As a result, some vendors were subjected to closer scrutiny than others. In the case of two of the vendors, every single image was checked; for others, only around 10% of the images were checked. A few batches had to be resupplied for reasons of quality or missing content. One unfortunate batch was redelivered five times before successfully completing the evaluation process.

The project team accepted that using sampling would mean that a small number of errors might slip through. However, it was felt that it was not cost effective to attempt to pick up every minor spelling or coding error in the entire collection if the cost of evaluation became too significant a proportion of the cost of acquiring the content itself.

## Comparing apples and pears

One of the most difficult challenges everyone on the project faced was how to compare the content from widely differing collections. How could, for example, images of youth culture be compared with posters from World War I? Was there any meaningful way in which images of 1960s espresso bars (held by the Design Council archive, at the University of Brighton) could be compared to images of the war in Afghanistan? In practice, the evaluators concentrated far more on ensuring appropriate balance, which meant in broad terms that a topic received only the number of illustrations commensurate with its importance, and that specific topics were not unnecessarily duplicated.

## Achieving consistent metadata

One of the goals of the project was to ensure that images would be searchable by date and place of the event or object referred to in the image. This was easier said than done. The team realized that, perhaps surprisingly, more attention is regularly paid to the date an image was captured than to the creation date of the object being photographed. However, for the purposes of this collection, the publication date of, for example, a poster, is usually more important than the date the poster was photographed. Vendors were also required to provide GIS data for any image with a recognizable location to enable users to search by area, by country, or by specific location.

Obtaining consistent metadata from a variety of vendors proved to be extremely problematic. The project drew on images from commercial collections, such as those owned by Independent Television News (UK) (ITN), The Associated Press and Getty Images, as well as from institutional collections such as those owned by the Imperial War Museum and the Fitzwilliam Museum in Cambridge. This immediately raised the issue of consistent and appropriate metadata. For any object, metadata provides the means by which the object becomes findable. In a commercial context, images are sold (typically for illustrative or advertising purposes), hence an image of a commuter on a station

platform may be tagged 'male', 'morning', 'newspaper', without providing any information about where the picture was taken. However, in this commercial context, the metadata provided would be adequate for the intended purpose.

It may seem, in contrast, that an institutional collection would provide all the relevant educational metadata required. While it would be true to say that the institutional collections often provided excellent captions, the caption providing all the metadata was frequently contained in a single field. Trying to extract the location of a painting becomes very difficult when the caption reads, for example:

> *This is one of two panels that were part of the predella that forms the lower edge of the large altarpiece of Veneziano's 'St Lucy Altarpiece' (c.1442-48). Originally in the church of S. Lucia dei Magnoli in Florence, the altarpiece appears to have been dismantled by 1816.*

A Google-type full-text search of this caption would find 'Florence', but would not of course enable users to identify that this is a painting from the 15th century. The project team's goal was to have dates and places held in separate fields so that it would be possible to search by dates and by date ranges rather than the simple string-based search that Google provides. Both for institutional and commercial collections, therefore, quite a lot of discussion had to take place with the vendors to ensure that the metadata would be fully searchable.

## Which standards to use?

Ten vendors were evaluated as part of the project, and there were concerns amongst the project team that they would use ten separate metadata coding systems. After initial discussions with the vendors, these fears were confirmed. It is relatively unusual for commercial vendors to provide images that will be cross-searched alongside other vendors' content, so normally they do not have to worry about whether their system will be compatible with other vendors' systems.

With no single taxonomy being employed widely by the vendors, even within groups that might be expected to have a common metadata structure, and with institutions failing to follow consistent metadata standards, a choice of a single taxonomy for the project was required. This was a value judgement, but a necessary business decision for the project. The team selected the International Press Telecommunications Council (IPTC) taxonomy, for two reasons:

1. IPTC is the system used by one of the existing JISC content collections, NewsFilm Online, which we intended to merge with this project
2. IPTC is a widely used coding system for news and media content, and we hoped that commercial providers would be familiar with adding IPTC codes.

Having decided on IPTC, the team expected a few teething problems as the coding system was adopted, but it could not have been anticipated that the web itself would cause problems in this area. There are several sites on the web that provide a complete list of IPTC subject codes, and it was only realized some time into the project that one of the seemingly reputable sources, XML Coverpages (www.xml.coverpages.org), did not include many of the recent changes made to IPTC codes by its parent organization. It was a simple enough issue to correct, but this example bore out the point that the internet includes answers (to most questions) that are both right and wrong.

One clear advantage of adopting IPTC was that it is a hierarchical system, so that, for example, 'homelessness' is a subset of 'social issues'; there are at least three levels of detail in the taxonomy. The eventual need to incorporate other collections into this one meant that we could at least map the top-level classification for images, and then provide a more detailed classification over time.

## Technical knowledge by suppliers

The project required XML for delivery of images. XML is widely used for the transmission of metadata, and there are plenty of freely available tools for creating and for checking XML. For those vendors who were not familiar with XML, we provided a simple tool that converted spreadsheets in Excel to XML tags. However, a surprising number of institutions had difficulty delivering valid XML. Frequently, the technical aspects of the delivery were handed to freelancers, who perhaps did not understand the requirements of the project, or who did not check the files before delivery. The process of ingesting content was slowed down

considerably since only complete batches could be accepted – the 'holding bay' would only accept content delivered in complete batches, not individual images, and the failure of one image meant that a batch had to be resupplied. One advantage of XML is that suppliers could validate material themselves before delivery, but despite the provision of customized tools that provided extensive checking, the suppliers did not always do this consistently. Often, errors were spotted by the project's evaluators rather than the vendors.

### Handling large files

One further requirement of the project was that images should be available for the education community in high resolution. Still images were provided in TIFF format at around 25MB per image, which required considerable storage capacity, but the space requirements for moving images dwarfed this. Some of the batch deliveries of moving images were over a terabyte. After some trials, the project team found that delivery by hard drive was the most reliable method for deliveries on this scale.

### Avoiding duplication

One issue that proved difficult to resolve was that of preventing two vendors delivering images on the same subject (or even the same image twice, since some of the commercial vendors source material from third parties). This was mostly an issue for vendors of news content. It did not prove possible to identify any cost-effective way by which the project evaluators could check their batch against content that had been sent to another evaluator or which had not yet been delivered by the relevant vendor. The solution agreed upon was for the moving image vendors to agree broad areas of coverage between themselves in advance of delivery, to ensure as little overlap as possible. This broad-brush approach (one vendor concentrating on UK news, another on Europe, and so on) worked within acceptable tolerances.

### Major and minor errors

Remarkably, one thing both commercial and institutional suppliers suffered from was an inability to spell. Many of the files delivered had a surprising number of elementary spelling or transcription errors. This often appeared to be for one of two reasons. Firstly, the project team encouraged moving-image vendors to provide 'rushes' (the unedited footage from which broadcasts are selected) as well as complete news bulletins. Secondly, and perhaps more fundamentally, the metadata held by vendors for their images is not typically published directly. Routinely, the publisher licensing the content, for example a television channel, will revise or rewrite metadata according to their immediate purpose. For still images which appear in a journal article or book chapter, the caption will often be rewritten to suit the requirement of the specific publication. The evaluators were asked to concentrate more on inaccuracies or incomplete metadata than on ensuring every caption was spelt correctly.

### What happens next

Currently, a new interface is being built by EDINA to host the new collection. When it launches in 2011, it will comprise over 150,000 images, incorporating the two previous JISC image collections. The project team believe this will be one of the largest general collections of images with rights cleared for educational use in existence. The content will be provided free of charge, with just a small access fee, based on the size of the institution.

Once the collection is launched, it is hoped to incorporate some crowd-sourcing features, such as a moderated feedback system, by which users can report errors. A longer-term goal is to investigate cross-searching the images alongside JISC's book and journal collections. Whatever the problems that have been (or may yet be) encountered, there are undoubted benefits to having a 'curated' collection, especially when the alternative is trying to provide access to all websites, using only Google and HTML.

### Conclusion

The collective experience of creating digital text collections means that the process of building an aggregated text platform is, to a large extent, a

familiar process. However, to build an image collection on this scale is relatively uncharted territory, where there are no straightforward answers to many of the issues raised. Nonetheless, despite the challenges encountered in building such a large and varied collection, the project has created what we believe to be a unique resource for UK education. We look forward to comments and feedback once the images are publicly available.



■ **Michael Upshall**
**Project Manager**
**Digital Images for Education**
**JISC Early Books**
**Journals and Media collections**
**E-mail: upshami@jisc.ac.uk**