Digitization: surely it can't be that difficult?

With some five million collections items comprising art, film, photographs, sound, new media, writings and objects, the possibilities which digitization opens up for the Imperial War Museums (IWM) are only limited by the imagination: new ways of reaching and engaging with audiences, slicker commercial activities, online access to superbly restored film and photographic images. But without the financial clout of a Google-style business, and where systems and standards are still very much in development, heritage institutions struggle to bridge the gulf between aspiration and reality. Hiding behind the term digitization lies a mass of activities and responsibilities, all critical to its success. With several years of experience of digitizing anything from posters to motion picture film, and of capturing high-quality digital images of a range of objects from medals to missiles, IWM has now learnt, often the hard way, most of the lessons.

DAVID WALSH
Head of Digital Collections
Imperial War Museums

Digitization is the answer to everything. For a museum, it means opening up previously obscure corners of the collections, bringing in new audiences, streamlining the work of curators, enhancing academic research and, most importantly of all, creating rich streams of revenue to dig the organization out of the financial hole caused by funding cuts. That, at any rate, is the aspiration: reality, needless to say, lags some way behind this rosy picture.

It is all too possible, once reality takes over, to start believing that the rewards of digitization are nothing more than a tantalising mirage. Why should this be? Surely it can't be that difficult: everyone knows how to digitize things – we do it routinely at home when we use a scanner, and private enthusiasts happily stream films and post images on their websites. Even the smallest backroom enterprise has a functional website with an efficient online payment system, and Amazon users are well used to slick systems which tell them what they are interested in even before they realize it themselves. The issue is partly one of scale: the individual enterprise can manage a handful of digital assets without bothering with all that tedious collections-management business, while companies such as Amazon have seemingly billions of dollars to spend on developing and testing highly complex systems in order to present the simplest possible experience to the user. Somewhere in the middle sits the heritage institution, facing an audience which expects to step through a perfectly designed gateway into a virtual world where everything is available online.

Where are the standards?

"It is all too possible … to start believing that the rewards of digitization are nothing more than a tantalising mirage."

One question that is often asked at IWM is, "I suppose everything is digitized now, isn't it?" People then become crestfallen when we tell them that only a small percentage of the entire collection is online. What can we have been doing all this time? Alas, digitization is a technically imperfect process: a digital representation of a real thing can only be an approximation (albeit potentially a very accurate approximation), and there are a huge number of factors affecting its faithfulness to the original. One only needs to type 'Van Gogh Starry Night' into Google Images to see how extraordinary the diversity of renditions of the same picture can be. So digitization for a heritage institution is more than sticking something on a scanner or in front of a camera and pressing the button.

It is usually possible to convince people that, in order to achieve consistency, digitization needs to comply with set standards. However, the problem is often one of finding the right standard, or even finding any standard at all. It may seem surprising in an age when everyone is madly digitizing that, in the outside world, digitization is something of a fringe activity. The heritage institutions may have only just got started, but in commercial photography, the film and television industry, and in the business world, physical media are remnants of the past, no longer in need of conversion to a digital form. Digitization frequently has to operate in an environment where systems and equipment are not primarily designed for heritage work. Where such standards exist, they are usually conceived according to what is possible rather than what is desirable. Although the standards and techniques for the digitization of, for example, audio recordings are well established internationally, and are easily available[1], those for digitization of photographs are less well defined, and likely to be based on each institution's own working practices. Defining such aspects as formats, sizes, degree of optimization, etc., can be a fraught process in which one quickly discovers that every user's needs are different: exhibition planners need wall-sized blow-ups, researchers want rapid access to instant images, commercial departments are after beautifully finished products to sell at a premium. And they all want their chosen output in a hurry, regardless of whether or not the original item requires specialist expertise, which may be anything from the attentions of a conservator to a careful unpicking of the multiple elements which form the original masters of a film.

> "… digitization needs to comply with set standards."

Keeping the lid on demand

Controlling the increasing demand for digitized content is partly what a digitization strategy will do, but the document needs to be carefully thought out, authoritative and easily understandable. The strategy will lay down what items will be digitized, to what standard and for what purpose, couched in terms which even the most resilient technophobe can understand. A fundamental question is to decide which parts of the collection require digitization to preservation standards, and which are strictly for access, which is not to say that digitization for preservation necessarily results in a better all-round asset: a high resolution digital scan of a photographic negative may be a near-perfect digital simulacrum of that negative, but in its raw form will be less suitable for image sales than a lower resolution scan which has been colour-matched and digitally retouched to a high finish.

Where to start when selecting collections for digitization is, of course, a key issue. With millions of collections items, many of which consist of numerous pages or images, IWM is not in a position to digitize everything in short order. Even smaller organizations have to accept that the question of how to prioritize and select, and who is going to do it, is likely to be the subject of much lively discussion. Weighing the competing demands for different areas of the collection, all with different requirements and technologies, can be problematic. Selecting the right route, which may be by photography (always the case for 3D objects) or scanning (often, but not always, the preferred option for 2D items), and deciding between doing the job in-house or externally, can lead to an impossibly complicated decision matrix. Inevitably, part of the game is finding ways to simplify the calculation without departing too far from reality.

> "… how to prioritize and select … is likely to be the subject of much lively discussion."

The DAM system

Once digitized versions start coming out of the pipeline, they can be dropped into the digital asset management system (DAMS) and everyone is happy … assuming the institution has got a DAMS, and here we discover that there is no neat off-the-shelf application with the kind of large user base and robust support that we get with standard business software. Giants such as Microsoft have no interest in such a small sector, with the result that

every organization has something different, be it home-grown or a specialist software company's product. It can be a very real drain on resources setting up and maintaining a DAMS in this situation, and the heritage sector may be partly to blame for failing to unite in producing a basic specification of all the functions required so that suppliers can design their applications around it. IWM started on the road to a fully integrated digital asset and collections management system capable of handling all digital media at a time when no such system existed – and inevitably found the journey pretty bumpy. In our sector, it is usually better to avoid being a pioneer.

"… the heritage sector may be partly to blame for failing to unite …"

Perfect metadata

So now you have the digitized assets and the system to manage them, all the user needs to do is find what they want and open it up. The important word here is 'find', of course. Documentation has to be the equal partner of digitization, and one quickly realizes that what passed for good cataloguing in the analogue world often fails to pass muster in the digital, especially if rights information, essential for managing access, is not rigorously recorded. An online user, accustomed to web search engines, expects a different experience from the traditional on-the-premises researcher, and even replicating on a computer screen the experience of flipping through albums of unindexed photographs or pages of untranscribed diaries is likely to fall well short of user expectations. On the other hand, once the items are available digitally, it may make the work of documenting them much simpler and, in theory at least, opens up the possibility of crowd-sourced input.

Digitizing film: not for the faint-hearted

When everything is in place and working flawlessly, there is one last thing which may cause a carefully-implemented digital structure to come crashing down: moving images. Film is almost guaranteed to completely disrupt the smooth path at some point between digitization and customer satisfaction. This is partly because time-based media (film or audio) create their own special difficulties with access, since users want to find specific moments within potentially lengthy files, and partly because the file sizes can be so large that even the most blameless of IT infrastructures are reduced to complete helplessness. In addition, the word 'standard' appears to be entirely foreign to the digital moving image world, rooted as it is in a culture of judging images by eye rather than according to set benchmarks. In addition, the industry seems obsessed with the next big thing – high frame-rate 3D cinema, HD television, Ultra-HD television and so on. Such international standards as exist are commonly unsupported by any applications, while popular applications rely on their own proprietary formats which may be dropped on a whim. In evolutionary terms, moving image digitization is many years behind still images, and in this still-developing discipline where today's format of choice is tomorrow's Betamax, an organization has to be both light-footed and prescient in its decisions.

"… the word 'standard' appears to be entirely foreign to the digital moving image world …"

In this context, IWM's digitization under the European-funded EFG1914 project[2] of its entire First World War film collection (350 hours of film out of a total collection running to some 25,000 hours), has undoubtedly exposed our systems' weaknesses to an uncomfortable degree. Ultimately though, it has resulted in a far more robust infrastructure and, given that we can't single-handedly impose standards on the industry, a clear understanding of what compromises can be made without jeopardizing the future of these digital films.

Digital preservation

Finally, there is no point in digitization if the organization cannot reasonably guarantee that the resulting digital items are safely stored for the future, even if the aim is only to provide convenient access to a collection. We may be becoming accustomed to a world in which it is virtually impossible to erase one's digital footprint, (ill-advised tweets following their

authors to the grave), but in reality digital data inclines towards the ephemeral. A failed drive, a software glitch, an error in backup can all lead to sizeable data losses which, at the very least, are likely to result in time and resources wasted on recovery. If the organization is holding digital objects for permanent preservation, and it is hard to imagine an organization which does not have at least some material which only exists in digital form, then the result can be catastrophic. Digital preservation requires more than mere 'belt and braces' if the organization is not to find its digital trousers round its ankles. There needs to be as much redundancy (i.e. spare information beyond the minimum needed to hold the data) as practicable, and, as well as holding multiple copies of files and media, this may even mean using less efficient encoding formats.

"… in reality digital data inclines towards the ephemeral."

One key component of data preservation is the checksum: this little piece of mathematical magic is a calculation derived from the digital bits which make up each file, and a change in even a single bit will result in a different checksum. So at every stage where the data is moved, copied, stored, or retrieved, the checksum calculation is carried out to ensure that nothing has changed. Although checksums are commonplace in IT, they tend to be hidden from the average user, but for digital preservation it is important that the checksum is part of the digital file's preservation metadata, and ideally it should be calculated the moment an institution takes charge of a digital object.

Are you trustworthy?

The responsible institution should look carefully at the Open Archival Information System reference model (the OAIS)[3]. This standard (ISO 14721), which derives from the space industry's need to preserve large amounts of data, contains a set of high-level requirements and functions which describe an archival system capable of acquiring and preserving digital information, and of making it available to its users over the long term. Importantly, OAIS applies equally as well to a large national institution with a complex range of responsibilities, as to a small local collection with limited resources: OAIS compliance does not imply sophistication or high expenditure. The OAIS model can be a little confusing to read at first because, along with the very general nature of the recommendations (there is nothing about specific procedures, protocols or applications), it deliberately avoids using terms which may have different meanings in different disciplines. The good news is that there are handy checklists[4,5] of criteria based on the OAIS which can be used to determine if a digital archive can be considered trustworthy (and if you can't be trusted to manage digital data, then you shouldn't be doing it).

"OAIS compliance does not imply sophistication or high expenditure."

Conclusion

At IWM, we have now spent a few years falling into the various traps that becoming digital sets out for a large institution, and have dug ourselves out again, each time much enlightened[6]. The pessimists' view that digitizing the collections is something to be avoided at all costs has been firmly laid to rest. Digitization will indeed lead to sunny uplands, but only if managed very carefully, and only if ambition does not run ahead of ability. As long as everyone remembers that if a digitization project looks simple, then it probably means that something important has been forgotten, then that is a good place to start.

References

1. International Association of Sound and Audiovisual Archives, Guidelines on the Production and Preservation of Digital Audio Objects (web edition): http://www.iasa-web.org/tc04/audio-preservation (accessed 9 September 2013).
2. European Film Gateway project EFG1914: http://project.efg1914.eu/ (accessed 9 September 2013).
3. ISO 14721:2012, Space data and information transfer systems – Open archival information system (OAIS) – Reference model.
4. Trustworthy Repositories Audit & Certification: Criteria and Checklist: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed 9 September 2013).
5. ISO 16363:2012, Space data and information transfer systems – Audit and certification of trustworthy digital repositories.
6. IWM Collections Online: http://www.iwm.org.uk/collections/search (accessed 9 September 2013).

David Walsh, Head of Digital Collections
Imperial War Museums, Lambeth Road, London SE1 6HZ, UK
Tel: +44 20 7416 5248 | E-mail: dwalsh@iwm.org.uk

To cite this article:

Walsh, D, Digitization: surely it can't be that difficult?, *Insights*, 2013, 26(3), 277–281; DOI: http://dx.doi.org/10.1629/2048-7754.96