JISC DEVELOPMENT PROGRAMMES

Project Document Cover Sheet

Final Report

Project Information			
Project Acronym	-		
Project Title	19th Century Pamphlets Online		
Start Date	1 March 2007	**End Date**	28 February 2009
Lead Institution	University of Southampton		
Project Director	Dr Mark Brown, University Librarian		
Project Manager & contact details	Grant Young, gy219@cam.ac.uk *Please address correspondence beyond project to:* Julian Ball Digitisation Manager University Library University of Southampton Highfield Southampton SO17 1BJ Phone 02380 598730 jhb@soton.ac.uk		
Partner Institutions	BOPCRIS/University of Southampton; RLUK; JSTOR University of Bristol; Durham University; University of Liverpool, LSE; University of Manchester; Mimas; University of Newcastle; UCL		
Project Web URL	http://www.rluk.ac.uk/node/71/ http://www.britishpamphlets.org.uk/		
Programme Name (and number)	Digitisation Programme		
Programme Manager	Alastair Dunning		

Document Name			
Document Title	19th Century Pamphlets Online: Final Report		
Reporting Period	-		
Author(s) & project role	Grant Young, Project Manager		
Date	31/03/09	**Filename**	PamphletsFinalReport.doc
URL	-		
Access	☐ Project and JISC internal		☑ General dissemination

Document History		
Version	**Date**	**Comments**
1.0	31/03/09	Final version

19th Century Pamphlets Online

Final Report

March 2009

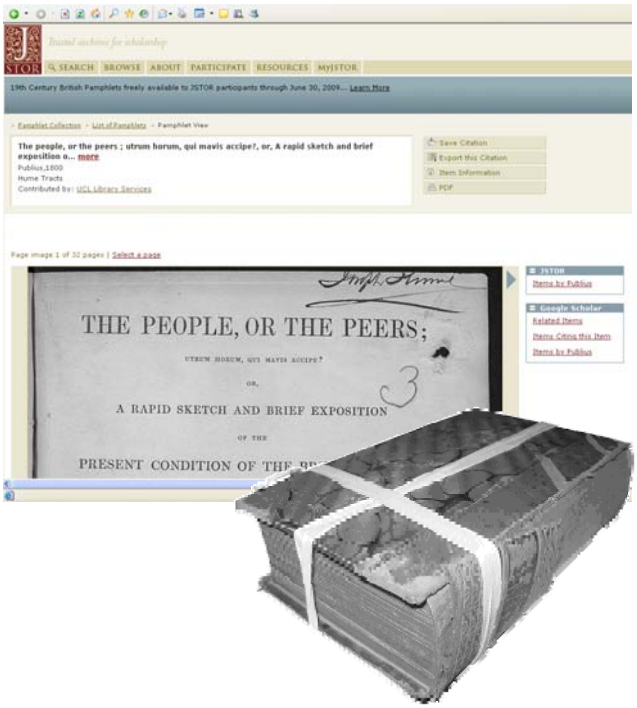


Table of Contents

Acknowledgements

19th Century Pamphlets Online was a project within phase two of the JISC (Joint Information Systems Committee) Digitisation Programme[1]. Major funding was provided by the JISC, with additional funding from RLUK (Research Libraries UK)[2]. Contributions were also made by each partner.

This has been a highly collaborative project drawing in expertise from RLUK libraries, JISC, MIMAS and JSTOR. In citing individuals we also wish to acknowledge the positive and effective partnership shown by all participants throughout across the whole project.

Many have been instrumental in the success of this project. At great risk of omission, we would like to acknowledge the following contributions to this project:

- The project management team at the University of Southampton Library: Mark Brown, Julian Ball, Richard Wake, Chris Fowler, Wendy White, Jayne Tweedle and Grant Young (on secondment from the University of Bristol)
- The scanning team at BOPCRIS and staff of Special Collections at the University of Southampton Library
- Our developer, Ed Fay, and education officer, Sarah Price
- The team at RLUK: Robin Green, Anne Poulson, Michael Mertens
- The team at JSTOR, particularly: Michael Spinella, John Kiplinger, Kimberley Lutz, Nancy Kopans, Nancy Murray, Brian LemMon, Barbara Chin and Marita LaMonica
- The team at JISC, particularly: Alastair Dunning, Stuart Dempster, Paola Marchionni, Lorraine Estelle, Emanuella Giavarra and members of the Digitisation Working Group
- The team at Mimas, particularly: Sean Dunne, Joy Palmer, and Shirley Cousins
- The representatives of the partner libraries: David Wilkins, Sheila Hingley, Katy Hooper, Barbara Humphries, Jenny Curtis, Melanie Wood, Lesley Pitman, and their colleagues
- The project's steering group: Peter King, John Belchem, Tim Leunig, Ronald Milne, Christine Paterson, James Thompson and Matthew Woollard

[1] See http://www.jisc.ac.uk/digitisation_home.html
[2] See http://www.rluk.ac.uk/

Executive Summary

The 19th Century Pamphlets Online project was sponsored by Research Libraries UK (RLUK), funded by JISC and led by the University of Southampton. Other partners included JSTOR, Mimas, and the Universities of Bristol, Durham, Liverpool, LSE, Manchester, Newcastle and UCL.

The overall aim of the project was to provide researchers, teachers and learners with online access to significant collections of 19th century pamphlets held within UK research libraries. In order to achieve this aim, the project drew on the pamphlet holdings of seven research libraries (Bristol, Durham, Liverpool, LSE, Manchester, Newcastle and UCL), choosing collections that focused on the political, social and economic issues of the day. The project scanned these collections within the University of Southampton Library's specialist BOPCRIS Digitisation Centre and then sent the datasets to JSTOR for archiving and delivery via their online publishing platform. Mimas enabled links to the digitised pamphlets to be added to the national Copac catalogue and to local library catalogues. A supporting website was developed to hold information about the collections and educational resources to support researchers, teachers and students.

The 19th Century Pamphlets Online project sought to build on previous work and expertise. It followed on from a large retrospective cataloguing project, which included many of the same partners and was also sponsored by RLUK. Metadata created within this previous project was extended and linked to the digitised pages and text. The project drew on the considerable digitisation experience of BOPCRIS, the delivery platform of JSTOR, and existing resource discovery channels available via JSTOR and Mimas (such as Google Scholar and Copac).

In addition to building on the past, the project was concerned to leave a good legacy for the future. A problem facing large consortia digitisation projects is how to preserve and sustain the resources they create. Which of the many partners will take on this responsibility? How will it be paid for? To address this problem, the UK partners chose to enter into a long (25 year) agreement with JSTOR over the care and delivery of the collection. JSTOR would preserve the data and make it available free of charge to UK users, and it would pay for this by making the content available on commercial terms to others.

Over the course of two years the project succeeded in scanning 26,041 unique pamphlets (1,000,732 pages) and ensuring their effective online delivery and discovery. Despite undertaking much research and planning prior to its commencement, the project inevitably faced challenges and changes. It was able to respond to these in a flexible and adaptive way, drawing on the strengths within the consortium and the trust that had been established between partners.

Although the main aim of the project was the production of content, it also had research and development components, and there was much learned and created through the project, which will benefit partners and the wider community of resource providers and users.

1. Background

The 19th Century Pamphlets Online project sought to provide enhanced access to a valuable but underutilised resource, building on a previous project and drawing on the considerable experience and infrastructure available within its consortia. In creating its digital collection, the project sought to adopt an efficient digitisation methodology, ensure effective resource discovery, and provide for the long-term preservation and sustainability of the content it created.

The project was intended to:

- open up a valuable but underused resource for those engaged in research or in teaching and learning activities;
- build on previous work, experience and relationships;
- explore innovative models for undertaking digitisation and for hosting and preserving digital content and develop an effective model for the long-term delivery and sustainability of the content free for the UK community.

These points are considered further below.

1.1 Opening up a valuable but underused resource

Pamphlets played an important role within 19th century political discourse, representing diverse contemporary perspectives, often polemical in nature. They are a valuable primary source that can complement other sources, such as newspapers, periodicals or parliamentary papers. However because of their ephemeral nature, they are often scarce and difficult to access and so are underused within research or teaching and learning activities.

From 1999-2002 a large retrospective cataloguing project, sponsored by the Research Support Libraries Programme (RSLP) and CURL (now called RLUK), catalogued nearly 180,000 19th century pamphlets from 21 research libraries[3]. That project greatly assisted researchers in finding pamphlets via local library catalogues and the combined academic and national library catalogue, Copac[4].

However, having discovered the existence of a pamphlet, a researcher is then often faced with the barrier of having to travel to a distant library to view it, since 19th century pamphlets are usually held within special collections and seldom loaned out. A scoping study undertaken in preparation for the project checked a random sample of approximately 100 pamphlets from each of the seven contributing libraries against the Copac catalogue. It found that between 23% and 44% of the pamphlets were only recorded as being held by that library[5]. A researcher from the University of Reading, has described his experience:

[3] See http://www2.is.bham.ac.uk/rslp/pamphlets/pamphlets.htm
[4] See http://copac.ac.uk/
[5] Young, G., *19th Century Pamphlets Online: digitisation scoping study*, (JISC, 2006), p.13. Available online at: http://www.jisc.ac.uk/publications/publications/pub_digi_scopingstudy.aspx [Accessed 20 February 2009]

> Pamphlets are an important but under-utilised historical resource. I frequently use Copac to track them down but am then faced with time-consuming and expensive journeys to look at rather short documents.[6]

It is likely that only the most determined would go to such trouble – or those who already have access to significant pamphlet collections. Where researchers, teachers or learners have access to such collections, pamphlets are often highly valued. This was confirmed by the academic speakers at the project's launch event, who emphasised the potential for developing new areas of study and enquiry.

In linking digitised pamphlets to the existing catalogue records created under the RSLP/CURL project, the 19th Century Pamphlets Online project took a further and vital step: ensuring that researchers, teachers and learners do not just discover the existence of 19th century pamphlets, but are able to access many of them directly..

1.2 Building on experience and expertise

This project is also noteworthy in the way it capitalised on previous work, experience and relationships. As outlined in the previous section, it built directly on work done within a previous large retrospective cataloguing project involving many of the same partners. The creation of catalogue information (often called metadata) is no trivial task and has frequently been found to rival digital capture in terms of its cost and complexity. Without this earlier project and the rich metadata it created, the 19th Century Pamphlets Online project would not have been feasible.

The digital capture, packaging of metadata, and generation of electronic transcriptions (via OCR – optical character recognition) greatly benefited from the expertise of the University of Southampton Library's specialist BOPCRIS digitisation unit[7], which provided centralised scanning for the consortium. BOPCRIS has significant experience in capturing historic textual material, most recently in the digitisation of a million pages of 18th century parliamentary papers for the first phase of the JISC's Digitisation Programme.[8]

In delivering and preserving the digital collection, we were able to take advantage of the infrastructure and experience of JSTOR[9], a US non-profit organisation specialising in making scholarly resources available online. The project benefitted particularly from JSTOR's search and retrieval interface, marketing activities, linking arrangements, and preservation services. These are described in more detail in later sections of this report.

19th Century Pamphlets Online project was able to make very good use of JISC-funded services and centres. For example, it seconded a project manager from the JISC Technical Advisory Service for Images (TASI, now called JISC Digital Media[10]) and engaged Mimas to extend the Copac service[11] to enable direct linking to the digitised pamphlets held within JSTOR.

[6] Personal communication. Quoted in the project's 2nd bid.
[7] See http://www.southampton.ac.uk/library/bopcris/
[8] See http://www.jisc.ac.uk/whatwedo/programmes/digitisation/britishofficialpublications.aspx
[9] See http://www.jstor.org/
[10] See http://www.jiscdigitalmedia.ac.uk/
[11] See http://copac.ac.uk/

The project was created by RLUK, and all of the UK partners within the consortium were from among its membership. It was hoped that this project would draw strength from this network and in turn provide a fruitful exchange between members. This was borne out within the project and evidenced, for example, by the smooth transfer of project lead from Bristol to Southampton in between the two stages of the bid process – or the contracting out of various workpackages to partners, such as the development of a partner database (by Bristol), educational resources (Durham) and a supporting website (Mimas). The RLUK executive itself was able to play a significant role in the management of the project's complex licensing arrangements.

1.3 Exploring models for undertaking digitisation, delivering digital content, and sustaining digital resources

There are many challenges involved with digitisation, especially when it is undertaken as a project by a large consortium. How can the creation of the resource be best managed when items are contributed from so many different collections? How can the digital resource be effectively embedded within the Web, so it doesn't become a "digital silo"[12] that is seldom found or used? How can the digital resource be preserved and sustained over a long period, and who will take responsibility for this? The 19th Century Pamphlets Online project sought, from its initial design, to address each of these challenges in innovative ways.

Efficient digitisation within a consortia context

The BOPCRIS unit at the University of Southampton Library had a key role in providing an expert centralised scanning service for the consortium. It carefully managed the de-duplication of collections and logistics of deliveries from and back to the contributing libraries. Wherever possible, BOPCRIS employed automated tools and techniques in order to provide efficiency and ensure quality. The methodologies used by the project are described in more detail later in this report.

Avoiding creating another digital "silo"

The UK consortium decided to partner with JSTOR, who agreed to take responsibility for delivering the collection on its behalf. JSTOR is an established provider, well known to UK scholars. Placed with JSTOR, the collection benefits from exposure alongside JSTOR's other content and from JSTOR's vigorous marketing activities. It also benefits from the linking arrangements JSTOR has with other organisations, which include Google, the History Cooperative[13], and RePEc (Research Papers in Economics)[14]. As a result, the pamphlets will appear within Google Scholar[15] and be fully indexed by Google's spider, enabling them to be found via a standard Google web search[16]. Through JSTOR's participation in CrossRef[17], references to the pamphlets in scholarly articles will also carry links through to the digitised pamphlets.

[12] A common metaphor in digital library contexts. See, for example Dempsey, Lorcan (2006) "The (digital) library environment: ten years after", Ariadne, 46, Available at: http://www.ariadne.ac.uk/issue46/dempsey/ [Accessed 20 February 2009].
[13] See at http://www.historycooperative.org/
[14] See at http://repec.org/
[15] See at http://scholar.google.com/
[16] See at http://www.google.co.uk/
[17] See at http://www.crossref.org/

In addition to the discovery and access provided by JSTOR, the project chose to commission Mimas to provide further channels into the digitised content. Mimas has developed the Copac catalogue to include direct links to the digital pamphlets within JSTOR. It has enabled searching of these linked records from an online pamphlets guide[18] and also offered linked catalogue records back to RLUK libraries for inclusion within the their own catalogues.

Ensuring long term preservation and sustainability

Preservation and sustainability are key challenges for large digitisation projects, especially those created by a consortium. The project chose to address these through its business model, in particular through an agreement made with JSTOR. The outline of the business model was agreed before the commencement of the project and has now been underpinned by a series of legal agreements.

The model and its supporting agreements are discussed in more detail later in this report but, in summary, the model is as follows. JSTOR stores a copy of the archival digital dataset created by BOPCRIS from the library collections and undertakes all the activities required to preserve this dataset, including backing-up, data checking, and migration to other formats. Contributing libraries can request copies covering their own collections, while RLUK, the JISC or HEFCE can request a copy of the entire dataset. JSTOR derives a delivery dataset, which it makes freely available to UK secondary schools, FE, HE and some other institutions. The costs of archiving and delivering the collection for UK users are borne entirely by JSTOR and funded through income it is able to generate within other markets.

As the project proceeded, the JISC asked us to find an additional UK-based store for the archival dataset. Provision has been made for this within the agreement with JSTOR and initial discussions were held with the Arts and Humanities Data Service (AHDS). However, with the closure of this service and lack of any alternative provision the project is depending upon JSTOR to provide its primary preservation store, as originally intended.

2. Aim and Objectives

In our Project Plan[19], we stated that the **overall aim** of the project was:

> **To provide researchers, teachers and learners with online access to significant collections of 19th century pamphlets held within UK research libraries.**

In order to achieve this aim, we identified **five key objectives:**

1. To digitise a wide selection of 19th century pamphlets focusing on political, social and economic issues.

[18] Soon to be placed at http://www.britishpamphlets.ac.uk/ and to incorporate content from a previous guide created by the earlier RSLP/CURL cataloguing project.
[19] Available here: http://www.jisc.ac.uk/whatwedo/programmes/digitisation/pamphlets.aspx

2. To establish an efficient consortial scanning operation.

3. To provide sustainable preservation and delivery.

4. To enable sophisticated, distributed resource discovery and access.

5. To provide models for further phases/projects.

Our aim and objectives did not change throughout the project and have been largely met. As described in detail below, the project succeeded in digitising a large and wide selection of 19th century pamphlets. These have been efficiently captured by the BOPCRIS digitisation unit and a large proportion are already available within JSTOR. The project has created multiple channels of discovery and access into the collections and will leave a legacy of lessons and resources for those who might seek to build on this project or undertake similar work.

3. Methodology

This section outlines the main tasks undertaken within the project, providing details of particular methodologies, tools and standards used. Although the JISC's final report template provides a later section for describing the project's implementation (section 5), we have included most of the discussion about our implementation within this current section. This is to enable the reader to more easily understand *how* and *why* our approaches changed from those laid out in the project's initial Scoping Study[20] and Project Plan[21]. For a full understanding of the way the project developed readers are referred to these earlier publications.

3.1 Selecting and preparing the pamphlets (Libraries)

Because it was not practical to individually select 26,000 pamphlets, the decision was made during the bid preparation to identify several collections that could be scanned in their *near entirety* (i.e. excluding non-19th century, in-copyright, or fragile/incomplete pamphlets). Apart from being pragmatic, it was felt that this approach would provide an extra dimension to the digitised collection: enabling the pamphlets to be understood within the context of particular historic collections.

Many collections were put forward for consideration by RLUK member libraries. Complete collections were chosen from Durham, Liverpool, Manchester, Newcastle and UCL, to provide a wide and balanced range of content. In addition, the bid team decided to make a *selection* of 19th century pamphlets from two of the UK's larger pamphlet collections: Bristol and LSE. Both collections have a strong political emphasis. The selections from these libraries were intended to highlight the strengths of their collections and fill in any obvious gaps left by the other collections.

The table below provides an overview of the collections we chose. Fuller details can be found in the Scoping Study and on the 19th Century Pamphlets Online website[22].

[20] Available at http://www.jisc.ac.uk/publications/publications/pub_digi_scopingstudy.aspx
[21] Available at http://www.jisc.ac.uk/whatwedo/programmes/digitisation/pamphlets.aspx
[22] See http://www.britishpamphlets.org.uk/

Table 1. Pamphlet collections

Contributing Library	Collection
Durham	**Earls Grey** *Collection belonging to family of politicians and colonial administrators*
Liverpool	**Earls of Derby (Knowsley)** *Collection belonging to family of politicians and colonial administrators*
UCL	**Joseph Hume (1777-1855)** *Personal collection of an MP;* *predominately first half of 19th century*
Newcastle	**Joseph Cowen (1829-1900)** *Personal collection of an MP;* *predominately second half of 19th century*
Manchester	**Foreign Office & Colonial Office** *Government collections focused on international relations and the Empire*
Bristol	***Selection* from pamphlet holdings** *Particularly from the National Liberal Club collection, which includes personal and party collections*
LSE	***Selection* from pamphlet holdings** *LSE is strong in party and pressure-group collections*

As the project neared its completion, we identified a need for more content, to ensure we had sufficient pamphlets to achieve our targets and make maximum use of our scanning capacity. Additional pamphlets were obtained from LSE (focusing on health, railways and canals) and Manchester (an important anti-slavery collection[23] and a selection of pamphlets related to national politics and the North West).

For the five complete collections the project devised a ***de-selection* strategy**. De-selection was made for one of four reasons:

1. the pamphlet was in copyright;
2. the pamphlet was published outside the bounds of the 19th century (we allowed a small proportion of late 18th century material but not pamphlets published after 1900);
3. the pamphlet was already digitised (or sent for digitisation) from another collection; or
4. the pamphlet was incomplete or too fragile to digitise

[23] See http://rylibweb.man.ac.uk/specialcollections/collections/guide/atoz/antislaverywilson/

The project commissioned a database from the Institute for Learning and Research Technology (ILRT) at Bristol to help libraries manage their preparations. We refer to this database elsewhere in this report as the Library Partners' Database to distinguish it from other databases. It was pre-loaded with records from Copac for each collection, enabling libraries to easily identify duplicates from other collections and to record information about the copyright status and physical condition of their own items.

The Library Partners' Database application was written using PHP[24] on top of a MySQL database[25]. SAXON[26] was used to process the MODS records supplied by Mimas and the Apache Software Foundation's HTTP Basic Authentication[27] was used to manage access to the database by library partners. The database supported a wide range of queries and results could be exported as standard comma-separated (CSV) files.

A database guide was prepared for libraries, which included a de-selection workflow and a copyright assessment workflow. The de-selection workflow was closely tied to the Library Partners' Database and so is of less value to other projects. It was developed from the workflow suggested in our Scoping Study. However, the copyright workflow is of more general application and so we have made this available as a project output. It is included within the 19th Century British Pamphlets website (http://www.britishpamphlets.org.uk) and also in Appendix B of this current report.

Whilst most libraries were de-selecting, for Bristol and the LSE a ***selection* strategy** was required. Selection was undertaken by library staff familiar with the collections, who were asked to consider for sub-collections or individual pamphlets:

1. Their relevance to themes of the great 19th century debates

2. Their usefulness in addressing gaps in the digital collection

3. Feedback and demand from collection users

4. Replacements for copies held in the smaller collections that were incomplete or too fragile to digitise.

In practice, the first two criteria were the main ones used within this project. The long lead-time required to select and prepare pamphlets meant that information about the condition or completeness of pamphlets in other collections was often not yet available. Delays in the content's online delivery also meant that it was not possible to collect and analyse information from or about online users.

Bristol and LSE used a variety of approaches to identifying potential material for inclusion, including searches via the Library Partners' Database or local catalogues and physical browsing of volumes and boxes. Bristol faced the biggest challenge, since the majority of its 19th century pamphlets are held within an offsite store and arranged in accession order rather than by subject or format.

Some of the collections have all their pamphlets bound within volumes (Liverpool, LSE, Newcastle, UCL); others have a mixture of separate pamphlets and volumes

[24] See http://en.wikipedia.org/wiki/PHP
[25] See http://www.mysql.com/
[26] See http://saxon.sourceforge.net/
[27] See http://httpd.apache.org/

(Bristol, Manchester); and one has separate pamphlets only (Durham). This meant that in some cases libraries were sending volumes that included pamphlets not intended for scanning. Slips or print-outs from the Library Partners' Database or library's own catalogue were used to indicate which items were to be scanned.

We undertook a scoping study in the Summer of 2006, in preparation for the 2nd phase of the bidding process. This included sampled surveys of the pamphlets and meetings with collection managers. Information was gathered about the collections themselves (e.g. condition and location) and also about the capacity of the library to undertake the preparations (e.g. availability of staff or timing of building work). This was used to plan a timetable for the project, which was published in our Scoping Study[28]. In the course of the project this timetable required frequent evaluation and revision in order to fit in with the workloads of partner libraries and BOPCRIS. We were able to build sufficient flexibility into our planning so that there were always pamphlets available for BOPCRIS staff to scan.

In recognition of the work that libraries were required to do for the project, they were allocated £1.50 per pamphlet scanned. This was a nominal sum and not intended to cover all costs. We expected that the real cost would be at least twice this amount and would vary from library to library. However, given the need to create a manageable budget, a per-unit cost was seen as the most practical and fair. We did not ask any of the libraries to record the value of their work, but it is certain that significant contributions were made. It is also clear that there was some variance in cost between libraries. Factors affecting the cost are likely to have included: the nature of the collection (e.g. whether bound volumes or separate items); whether libraries were supplying complete collections or making a selection (the latter is more time-consuming); and who among the staff was allocated to undertake the work. Similar future projects may wish to consider capturing real costs and exploring alternative ways of compensating contributors. In this context, setting compensation should take into account the value to the institution and its community from digitising its material and presenting it online.

3.2 Transferring and protecting pamphlets (Libraries and BOPCRIS)

As noted above, the digitisation was centralised at BOPCRIS within the University of Southampton Library. Memoranda of Understanding (MOU) between BOPCRIS and each partner library covered details of transportation, handling, reporting and the insurance of the pamphlets whilst in transit and at the University of Southampton. A consortium agreement also addressed these issues at a more general level.

The methodology for transferring material from and back to the libraries developed throughout the project. It was initially intended that the Library Partners Database would play a key role in logging the transfer of pamphlets, recording their stage within the workflow ("not checked", "sent", "returned" etc) and condition. During the course of the project it became clear that additional procedures should be employed.

Sometimes, for example, it became obvious that the database held incorrect information: recording that a pamphlet had been sent, when it hadn't – or vice versa. At other times the librarian and scanning staff had made quite different assessments of a pamphlet's condition, which is something inherently subjective. This necessitated the introduction of extra checking procedures and sometimes required

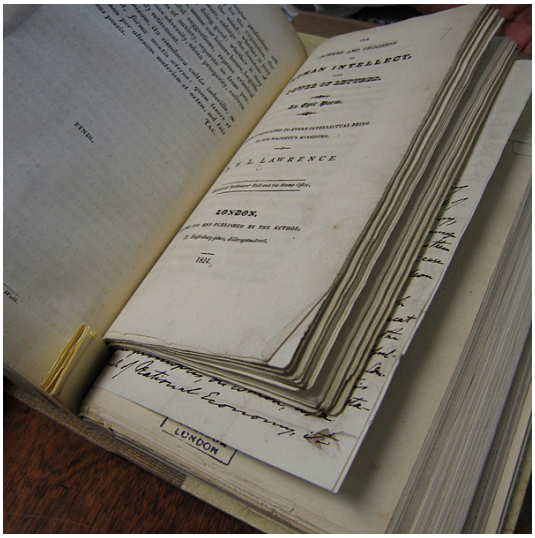
[28] Available at http://www.jisc.ac.uk/publications/publications/pub_digi_scopingstudy.aspx

discussion about the condition of individual pamphlets and what interventions might be permitted (e.g. cutting or steaming pages to allow for full digital capture). Although the project had hoped to adopt standard and automated approaches that would work for each collection and library, it soon discovered it was necessary to develop some bespoke approaches.

In order to provide extra reassurance for the contributing libraries and BOPCRIS, several additional security-related measures were implemented throughout the course of the project. A BOPCRIS staff member oversaw the packing of material at the library and the unpacking at Southampton, checking each item. Armoured vehicles were used for transporting the collections and they were housed and overseen within Special Collections (Archives and Manuscripts) within the Hartley Library. A safe was also purchased to provide further protection for the material when in the scanning laboratory during the day.

Unbound pamphlets from Bristol

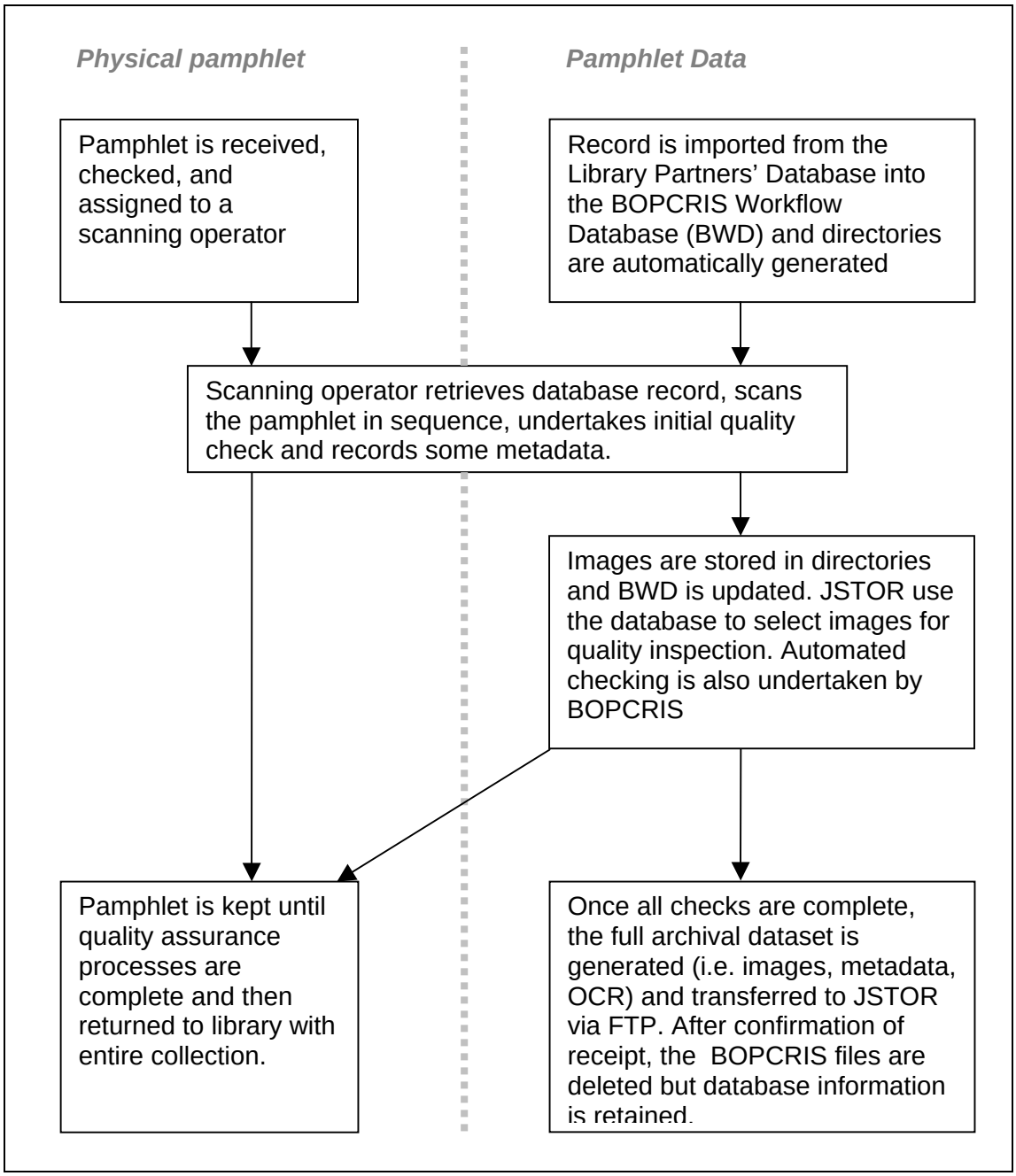
Bound pamphlets from UCL



3.3 Creating the archival dataset (BOPCRIS)

The diagram below provides a much-simplified overview of the workflow at BOPCRIS, including digital capture, metadata generation and quality assurance. The remainder of this section provides more detail on particular elements of this workflow.

Figure 1. Overview of BOPCRIS workflow



Capture specification

The project initially expected to conform very closely to the technical specifications of JSTOR, which were detailed within our Scoping Study. These required bi-tonal (i.e. black/white) scanning for plain text pages and greyscale or colour where these elements were present in the original. JSTOR's standard specifications also required rotating the page until square, cropping within the page edges, and removing any artefacts such as age spotting or library stamps.

During the initial start-up phase of the project there was much discussion between the BOPCRIS and JSTOR teams about the most appropriate standards for the pamphlet collections. The project also took advice from collection managers and academics on its management and steering groups, with several members of these groups preferring a facsimile-like capture.

As a result of these discussions we moved to an alternative specification for the *archival* images being created by BOPCRIS. Pages were captured in greyscale as a minimum, or colour whenever present, with page edges clearly in view and artefacts such as page foxing or show-through left within the image. However, in order to improve OCR and user accessibility, the pages were rotated to achieve a square text-block and some measures were undertaken to minimise the effects of show-through. JSTOR initially considered deriving a more standard *delivery* image from the archival image, but after consulting with users it decided there were benefits in retaining the facsimile style for its online delivery.

Much work has been done within BOPCRIS throughout the project to determine the best ways of handling, capturing, and recording what has often proved to be very challenging material. Detailed guidelines were prepared for BOPCRIS scanning staff to ensure a uniform approach was adopted for the scanning of the variable collections. This is a valuable document which will benefit other projects, so we are making it available on the 19th Century British Pamphlets website (http://www.britishpamphlets.org.uk/) and also in Appendix C of this report.

Metadata specification

The project had planned from the outset to adopt leading XML-based standards as described in the following table. These standards are increasingly being use to describe digitised books so were chosen to facilitate future interoperability (especially METS and MODS) and to support the long-term preservation of the content (MIX and PREMIS). However, the use of these standards posed some challenges for the project, as they are not yet completely stable - two changed during this project. Nor are they trivial to implement, typically requiring programming expertise to generate.

Table 2. Metadata Standards

	Standard	Comments
Bibliographic Metadata	**MODS** (Metadata Object Description Schema) [29]	In late 2006, Mimas began to generate MODS XML (version 3) from the library-contributed MARC catalogue records available within Copac. Given its availability, this format was an obvious choice for the project's descriptive metadata, since MODS was especially developed to hold a simplified set of MARC data for use within digital library collections.
Technical Metadata	**MIX** (NISO Metadata for Images in XML)[30]	The project adopted elements from the MIX standard to record technical information about the digital images. MIX is an encoding of the very extensive NISO data dictionary (Z39.87) [31] When the project bid was written, MIX was in draft form (version 0.2). By the end of the project it had reached version 2.0,which necessitated some changes to our metadata.
Preservation Metadata	**PREMIS** (Preservation Metadata Implementation Strategies Working Group)[32]	The project chose to use selected elements from the PREMIS data dictionary, which is intended to support the long-term preservation of digital resources. During the course of the project PREMIS moved from version 1.0 to 2.0, which required changes to be made to our metadata.
Structural Metadata	**METS** (Metadata Encoding & Transmission Standard)[33]	The project chose METS to provide its structural metadata. METS is now a well-established standard for structuring complex digital resources (e.g. publications with multiple pages) and for wrapping other sets of metadata. It is also often used with MODS, MIX and PREMIS.
BOPCRIS Metadata		In addition to the formal standards, BOPCRIS added some tags of its own to the METS metadata. These included administrative, provenance, language and

[29] See http://www.loc.gov/standards/mods/
[30] See http://www.loc.gov/standards/mix/
[31] See at http://www.niso.org/standards/resources/Z39_87_trial_use.pdf
[32] See at http://www.loc.gov/standards/premis/
[33] See at http://www.loc.gov/standards/mets/

		rights information
Delivery Metadata (JSTOR)	NLM (National Library of Medicine)[34]	JSTOR took the METS files we supplied and transformed them into their own delivery metadata standard, which is based on the NLM DTD. The METS files have been archived by JSTOR so they are available for contributing libraries or for future transformations.
Delivery Metadata (Copac and Library catalogues)	MODS and MARCXML or MARC21	Mimas have been supplied with permanent URLs from JSTOR and are embedded these within MARC records. These are passed on to contributing libraries for inclusion within their own catalogues and eventual incorporation within Copac records.

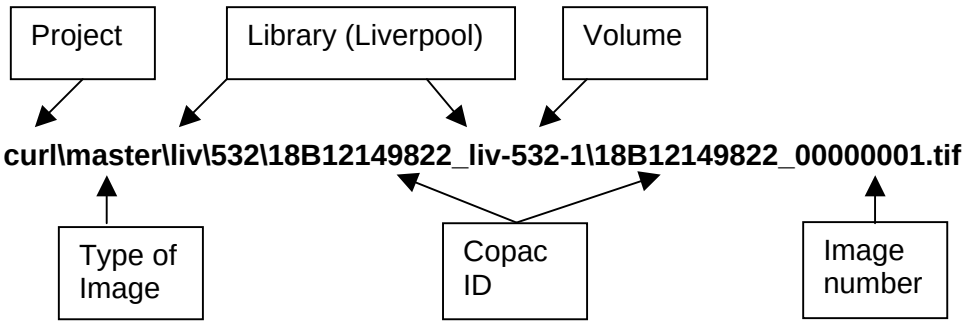
Throughout this project, BOPCRIS gained a lot of experience in dealing with XML-based metadata. This was shared with others within the Digitisation Programme. The project's metadata profile was also submitted to the METS Board for review and is now a formally registered METS profile. This profile is available at:

- http://www.loc.gov/standards/mets/profiles/00000024.html (human readable)
- http://www.loc.gov/standards/mets/profiles/00000024.xml (machine readable)

For convenience, we have also included the xml profile within Appendix D of this report.

File naming, management, and transportation

From the point of creation, pamphlets images were held within a directory structure with individual pages numbered sequentially. The directory and naming conventions provide information about the pamphlet's provenance (contributing library), identity (based on the Copac ID of the cataloguing record), location (e.g. which volume) and sequencing. The filename itself also repeats the Copac ID in case the image becomes separated from its directory. The following example illustrates these conventions:



[34] See at http://dtd.nlm.nih.gov/tag-library/2.1/index.html

For the purposes of transportation and archiving, the files associated with each pamphlet (metadata, images, OCR text, within directories) were held together within as a single TAR file[35]. A checksum was generated using the Message-Digest algorithm (MD5)[36] to enable JSTOR to confirm the integrity of the data once received. Some of the data was transferred on LaCie 1TB drives by courier, but the bulk was sent via FTP (File Transfer Protocol).

Scanning equipment

In planning the project, we had expected to be able to use the full range of scanners available within the BOPCRIS laboratory, which included:

- **Minolta PS7000 greyscale scanners** – for bi-tonal/greyscale capture
- **Digitising Line Suprascan colour scanner** – for colour capture or bi-tonal/greyscale work requiring special support
- **Digitising Line robotic scanner** – for sturdy volumes

It quickly became clear that robotic scanning was not going to be practical for this material. The robot requires bound volumes with pages of uniform character and weight, and the volumes of pamphlets provided too much variation. However the robotic scanner proved useful for capturing some of the pamphlets when used in its manual mode.

It also became clear within a few months of scanning that for the greyscale capture specification we had chosen, the PS7000s, which were handling the bulk of the scanning, were not providing an efficient through-put. They were introducing a 30 second delay per image scan, which was significant within the context of a million pages. After careful analysis, we concluded that the best approach would be to replace these scanners with newer, faster machines. We chose the **i2S CopiBook** machine[37]. The technology available by late 2007 was of better quality and sufficiently faster to enable us to close the time gap, achieving such a good production rate that we were able to operate with one less scanning operator than planned. This and savings in other areas of our budget was able to pay for the replacement machines.

The table below shows the cumulative scanning totals, illustrating the impact of the new scanners (from early 2008).

[35] See http://www.nationalarchives.gov.uk/PRONOM/x-fmt/265
[36] See http://tools.ietf.org/html/rfc1321
[37] See http://www.iiri.com/i2s/copibook.htm

Table 3. Scanning statistics

Automating the workflow

The digital capture, OCR, and metadata generation were managed by a database (BOPCRIS Workflow Database) and an associated set of tools (MetaTool) created or adapted by the BOPCRIS metadata officer. The table below summarises these components.

Table 4. Workflow components

Component	Description
BOPCRIS Workflow Database (BWD)	This database was used to direct and monitor much of the workflow within the BOPCRIS digitisation laboratory, supporting both manual and automated tasks. It was built on the supported Oracle environment provided by Southampton's iSolutions department and employs a dependant toolset to provide some additional functionality (see further rows of table). Functionality directly provided by the BOPCRIS Workflow Database (BWD) included: BDW1. Monitoring entry/exit to the digitisation laboratory BDW2. Importing bibliographic records in MODS or Z39.50 formats

	BDW3. Assigning work and monitoring progress BDW4. Custom metadata fields to enable scanner operators to record key attributes or free-text notes at the pamphlet or page level BDW5. Status and attribute information used by automated processes BDW6. Item- and attribute- level searching BDW7. Statistical reports (daily, weekly or monthly) BDW8. Access management, including viewing/editing permissions for individuals or groups, locking of records for dual access, and Virtual Private Network (VPN) access for off-campus users (e.g. JSTOR)
BOPCRIS Workflow and Metadata Toolset (MetaTool)	This was a set of tools acquired or developed to support specific tasks within the workflow, as follows: 1. Importing of MODS XML – utilising *Retrieval Tool* (developed by BOPCRIS) 2. Preparation for image scanning – e.g. generation of image directories based on input from BWD2 (see previous row of table) 3. Support for QA of images by BOPCRIS and JSTOR using inputs from BWD4. 4. Cropping of images based on input from BWD4-5 – utilising *Page Improver* (commercial software) 5. OCR generation in .idx and .txt formats – utilising *Agora* and *ABBYY* (commercial software) 6. Extraction of technical metadata for MIX XML – utilising *Meta Extractor Tool* (developed by BOPCRIS) 7. Generation of PREMIS XML – utilising *Meta Retrieval Tool* (developed by BOPCRIS) 8. Generation of METS wrapper – *Meta Wrap Tool* (developed by BOPCRIS)

It was hoped that the project might be able to produce more generic forms of some of these tools for use by others. This did not prove possible within the constraints of the project. However, we have made the code available within SourceForge.net under an open licence, so other groups are able to take and adapt it to their own use. This is available at: http://sourceforge.net/projects/metsbuilder/ .

The commercial *Page Improver* tool[38] was used to perform several automated optimisation actions on the scanned images. A key action – and one that proved very difficult for pamphlet material – was the cropping of each page as close to its outer edge as possible, without removing any page information. BOPCRIS worked closely with the developers of this tool to ensure that this process worked as effectively as

[38] See http://www.4digitalbooks.com/software.htm

possible. As a result, the project has been able to feed into the research and development of this tool for the benefit of other users.

Assuring quality

As indicated in the workflow diagram above (Figure 1), Quality Assurance (QA) activities were conducted by both BOPCRIS and JSTOR. Scanning operators checked their own images immediately after capture so that obvious errors could be quickly addressed. Once captured, automated routines were also run to detect errors (e.g. page count, resolution, presence of page edges). All images were made available to JSTOR for selection for their QA. JSTOR used the BOPCRIS Workflow Database to identify the pamphlets they wanted to check, which were then supplied by BOPCRIS via FTP. JSTOR undertook a visual inspection, following up any issues with BOPCRIS staff. Instead of adopting a rigid sampling technique, the process was qualitative and iterative. The recording of specific attributes within the BOPCRIS database (e.g. annotations, colour or binding condition) enabled JSTOR to focus their attention on those that were likely to prove difficult to capture or were prone to error. The valuable discussions that followed the JSTOR QA enabled BOPCRIS to refine its procedures and JSTOR to adjust its expectations. As a result, both organisations have gained a considerable understanding of the challenges involved in digitising 19th century pamphlets.

3.4 Delivering and preserving the collection (JSTOR)

JSTOR's QA work and its generation of delivery metadata have already been referred to in the preceding section. Once JSTOR received an archival dataset from BOPCRIS, this was used to generate a delivery dataset and then placed into JSTOR's dark archive for long-term preservation.

At the outset of the project we had hoped to be delivering pamphlets to users within the project's first year (i.e. by early 2008). This did not prove possible due to the time required to finalise the technical specifications, define and automate the metadata schemas, refine the *Page Improver* cropping tool, and develop the JSTOR platform to provide effective delivery of the collections. A test collection was released towards the end of 2008 and a significant public release was made in early 2009 (approximately a third of the pamphlets).

As mentioned above, JSTOR initially intended to deliver the pamphlets according to its standard profile (e.g. clean, bi-tonal images). Over the course of the project it decided to preserve the facsimile look of the images we supplied. It also decided that some of the journal-centric elements of the interface (specifically the interface labelling and browse structure) would require alterations to better suit this non-journal content. These were built into JSTOR's development schedule. By early 2009 there was a large amount of content available and development was sufficiently advanced to enable a large release of data. The first half of 2009 will see releases of the remaining pamphlet content and the rolling out of further interface and search functionality within JSTOR.

3.5 Enabling effective discovery and use (JSTOR, Mimas, and libraries)

Pamphlets are a new form of content for JSTOR, so it is still exploring the best way to present the pamphlets within the broader context of its collections – and to package the pamphlet collections for sale to non-UK subscribers. JSTOR has decided to take advantage of the staged release of content in order to undertake further user research. At the time of this report the pamphlets can be found via a search or by browsing each library's collection. In time JSTOR expects to provide more sophisticated faceted browse options and to more closely integrate the pamphlets with relevant journal collections.

Mimas has played a significant role within the project: providing bibliographic records at the beginning of the project and then enhancing these records with direct links to the JSTOR pamphlets upon its completion. In order to ensure that these links are embedded within local library catalogues and the shared Copac database, Mimas is providing libraries with duplicate/replacement records containing the JSTOR permanent URLs. When these are loaded back into the Copac database they are automatically matched with holdings from other libraries and have direct links through to the JSTOR digital surrogates. In addition, Mimas are running programs to identify which other libraries (beyond the project's partners) might have physical copies of each pamphlets. They will be offered enhanced records for inclusion within their own library catalogues.

In planning the project, we intended to build a website to hold information about this project and its RSLP/CURL predecessor. The RSLP/CURL project had created a website with descriptions of each collection it catalogued, but this was taken down in 2007, after five years. We obtained permission from RLUK to incorporate the collection descriptions within our own new website and commissioned Mimas to help implement a collection search. Initially it was thought that this website would be created by Southampton and hosted by RLUK. As the project progressed, an alternative arrangement seemed desirable and Mimas agreed to take on the development and hosting of the site.

We had also planned to employ an education officer at Southampton to develop resources for the website to help support those using the collections – particularly for school users. Our plans here also changed. We discovered an experienced education officer at Durham University and commissioned her to produce research guides and teaching and learning resources.

Delays in mounting content within JSTOR meant that the work of catalogue linking, website development and educational resource development had to be staged later than we had originally planned. These have been concentrated within the last four months of the project and will continue a little beyond the end of the project.

4. User Engagement

Anecdotal evidence gathered during the bid preparation suggested that 19th century pamphlets were not as well known or used as their value might warrant. Where they were known, they were often highly valued. The LSE reported that the pamphlets guide and selected digitised pamphlets were among the most accessed pages on its website. Durham had used pamphlets in its educational programmes with great success.

The project's management group included several collection curators who were able to provide information about usage. Several members of our steering group had used pamphlets within their research and provided valuable input into the design of the project. In an effort to understand how best to present the collection, JSTOR conducted its own focus group research with US librarians and academics.

Now that much of the collection is available online, JSTOR is vigorously marketing the pamphlets and promoting them among potential users. As noted above, it is also gathering information about usage to help inform future delivery and packaging of the collection.

The project has undertaken some engagement throughout the project, with members of the team speaking at events and contributing libraries promoting the project and collections within their own institutions. The major user engagement activity was its conference, which was originally to be held in the summer of 2008, but delayed until March 2009 to ensure there was sufficient content available on JSTOR. The purpose of the conference was to draw academic attention to the collection and motivate those attending to champion the resource among their colleagues and students. In securing three well-known speakers (Laurel Brake, Brian Maidment and Miles Taylor), we were able to give the event a high profile and ensure a good attendance. The speakers highlighted the value of making 19th century pamphlets more accessible. Pamphlets had been overlooked by 19th century specialists and their availability online is expected to lead to a re-evaluation of the format and its content.

The project's education officer, Sarah Price, has developed a range of educational resources to support the use of the collections. These will be mounted on the pamphlets website and include guides for teachers, students and researchers and some sample lessons for use within schools. In the course of preparing these resources Sarah has worked with a number of school groups.

5. Implementation

The methodology section above has already provided much detail about the implementation of the project.

The scoping study had enabled us to identify a lot of the issues we were likely to face and work out approaches for dealing with them (e.g. de-duplication and copyright workflows). We also began some of the work ahead of the official start date in order to avoid delays. However, inevitably in a project of this size and complexity, issues arose that demanded attention and necessitated change.

Several changes have already been discussed in this report: adjustments to the timing of library preparations and methods of transferring and securing pamphlets; a different technical specification; upgrading of machines to achieve a more efficient production rate; a revised schedule for online delivery; and the contracting out of work we initially thought might be dealt with in-house.

Many of these issues might have posed a threat to the project and its deliverables. However, the flexibility built into the project, its strong management structure, the

goodwill of partners and the support of the JISC Programme enabled it to adapt and succeed.

One of the largest challenges the project faced was to conclude the complex set of licences required to support its business model. Because of its importance to the project – and potential interest to other projects – we discuss it here in some detail.

The bid and scoping study had identified the need for appropriate licences, but it was only during the pre-project planning stage, in December 2006, that it became clear that a number of contingent agreements would be needed to underpin the creation and management of the digital collection. The process was made more complex by the requirements of the funder (JISC), the large number of project partners involved, the necessity to obtain permissions from some outside of the formal partnership, the commercial elements of the business model, and the long duration (25 years in one instance) of some of the agreements.

The first task was to determine a framework for the agreements. A partners' meeting was held, brokered by JISC Collections, in which the complexities of licensing were discussed and referenced to existing JISC licensing practice. This envisaged RLUK as the project sponsor being licensed to licence the content to JSTOR with JISC and HEFCE as interested parties. A diagram was drawn up to explain the relationships (see Appendix E).

Although JISC Collections were careful to emphasise that they were acting in an advisory capacity, the role of JISC as funder in effect drove the licensing strategy with the JISC legal advisor acting as the primary designer and drafter. Integrating the different interests of JISC, HEFCE, JSTOR, RLUK, the libraries supplying content, the library undertaking the digitisation, and in two cases the owners of the material deposited, was a complex undertaking, especially as the agreement was designed to run for 25 years. It was therefore inevitable that the process of agreeing terms and drafts would be extended. Although the framework of principles was agreed relatively quickly, and was used throughout the project as a benchmark, translating that into legal documents took far longer than had been originally envisaged. It is important to emphasise therefore that the project proceeded on the basis of trust and goodwill of the partners based on their position as members of the RLUK consortium.

Our second task was to complete a consortium agreement while the licensing discussions were continuing. For this we relied on the framework document without needing to have the detailed legal agreements in place, relying in effect on trust and goodwill among partners. We took the decision to create a standard consortium agreement, which could set out roles, responsibilities, procedures and practices, and expectations for the RLUK partners, reserving the licensing relationships with JSTOR, JISC and HEFCE to the forthcoming legal agreements. This was a pragmatic solution to the dilemma presented by the long drawn out legal processes which could have left the consortium without a statement of collective responsibility and practice.

The delay in concluding the legal agreements did not therefore restrict our ability to carry through the project, or act as a brake on our collective commitment. There was also some advantages in the sense that issues arose during the project implementation stage which might not have been covered had the agreements all been concluded in advance. These issues included agreements on how much material could be hosted locally by a library in addition to the whole collection being delivered by JSTOR, requirements for long-term preservation and access, the

balance of benefits between exclusivity and non-exclusivity, ownership of IPR, restrictions on accidental copyright infringement, and indemnities. It would clearly have been impossible, and probably undesirable, to agree these all in advance. This raises issues for JISC in terms of what it is practical to require of consortia in advance of starting work on a project.

The project team kept in constant touch with the partners alerting them to any issues which arose in the course of negotiations, and some of the issues outlined above arose as a result of these discussions. The process was steered through by the JISC legal advisor who provided the drafts of the various agreements to align their core provisions, and ensure that the interests of JISC (and by extension HEFCE) were assured. Southampton as the lead partner also took legal advice on a number of issues. As a result the supporting agreements for the partners and the owners were redrafted, and then referenced into the main agreement. The agreements were then passed to the other partners to secure legal agreement, and some further modifications were made. On reflection it might have been more effective to bring in local legal advice earlier to restrict the possible implications of conflicting interests. It must be emphasised, however, that legal agreements are complex, time consuming and potentially expensive. The level of payment to JISC Collections for legal services was capped by JISC, but there was no direct payment from the project to JSTOR or to the partners' legal advisors.

6. Outputs and Results

The project intended to create a number of tangible outputs. Some of these were related to the conduct of the project, such as management reports and project presentations, but most were intended to leave a significant lasting legacy:

In our project plan we anticipated the following outputs:

- **A substantial dataset** representing approximately 23,000 19th century pamphlets or a million pages. This would exist in two forms: (1) as high-quality, standards-compliant data and (2) as a web-quality, easily accessible data.

- **An online pamphlet collection** with good browsing and searching and appropriate contextual information, mounted on the JSTOR platform.

- **Enhanced catalogue records** providing links to the pamphlets within JSTOR from Copac and individual library catalogues.

- **A supporting website**, reviving content from a preceding project, providing links into the digital pamphlets and resources to help researchers and educators use them effectively.

- **A key event** to engage potential users and encourage them to champion the resource among others.

- **Resources for those undertaking similar projects**, including software, documentation and models.

Substantial dataset

Over the course of this project we digitised 26,041 unique pamphlets (1,000,732 pages). The table below shows how many we were able to digitise from each collection.

Table 5. Pamphlets scanned by the project

Contributing Library (and collections)	**Est. pam.**	**Actual pam.**	**Est. pages**	**Actual pages**
Durham (Grey)	1,160	945	75,478	36,452
Liverpool (Knowsley)	1,209	1,494	51,745	77,271
UCL (Hume)	3,528	4,847	148,881	214,452
Newcastle (Cowen)	1,579	1,896	45,796	71,463
Manchester (FCO and additional selections)	3,149	5,075	109,281	196,425
Bristol (selections)	6,250	5,019	284,375	187,133
LSE (selections)	6,250	6,765	285,000	217,536
	23,125	**26,041**	**1,000,556**	**1,000,732**

The estimates are from the project plan. The variance is due in large part to the sequencing and duplication of the collections. Later collections had more material excluded because a high proportion had already been scanned. Note that these statistics give an average pamphlet size of 38 pages.

Online pamphlet collection

The pamphlets are available within JSTOR and will be freely accessible to UK users for at least 25 years. At the time of writing this report, three of the seven library collections are available. The remainder will be added by the middle of 2009.

The pamphlets are currently available via JSTOR's standard browse and search interfaces, including an advanced search. We expect additional browse and search functionality to be added over the next few months, including a faceted search and subject-based browsing.

The pamphlets collection was the first substantial non-journal collection added to JSTOR, so has promoted it to consider alternative ways of presenting its content.

Enhanced catalogue records

JSTOR is providing Mimas with links through to each pamphlet. These are being incorporated into MARC records and given to partner libraries to enable them to update their catalogues. As a result, instead of requesting fragile originals, users from the partner libraries will be able to click through to digital versions at their desktops. Other libraries holding copies of the pamphlets are also being offered hyperlinked records for their catalogues.

The enhanced catalogue records are being carried through to the Copac database, enabling this to act as another entry point to the JSTOR collection. A further means of access is provided via the pamphlets website (see below).

As a result of this project, hyperlinked records are being added to Copac for the first time.

Supporting website

Mimas has also developed a website for the project, which can be found at http://www.britishpamphlets.org.uk/ . This is intended to provide a focus for 19th century pamphlets on the Internet. It revives content from the previous 19th century cataloguing project, which had been removed from the web, and adds collection-based searching, resources for researchers and teachers, and documentation for those undertaking similar digitisation projects. Mimas has agreed to host and maintain the site for at least five years.

Key event

A day conference was held at Liverpool University on 20 March 2009. This was attended by nearly 60 academics and librarians. It focused on the value of the collection for research. Speakers included Laurel Brake (Professor of Literature and Print Culture at Birkbeck), Brian Maidment (Research Professor in the History of Print at the University of Salford) and Miles Taylor (Professor of History and Director of the Institute of Historical Research), who stressed that pamphlets had been a neglected resource and the availability of a substantial number online was likely to have an important impact on 19th century studies. A session on digital collections for scholars was provided by Michael Spinella (Executive Director of JSTOR), Mark Brown (Chair, RLUK) and Alastair Dunning (JISC Digitisation Programme Manager). The conference also included a formal launch of the collection and a reception at the University's refurbished Special Collections and Archives, where the original pamphlets were on display. A colourful brochure was prepared to highlight the pamphlets – a copy can be accessed on the BOPCRIS website.[39]

Resources for those undertaking similar projects

Those involved in this project gained a lot of experience, which will benefit future projects in many institutions. Several resources created for the project have been

[39] See http://www.southampton.ac.uk/library/bopcris/uos_001pamphletsawfinallow.pdf (3.5MB download)

made available for others. These are appended to this report and available via the pamphlets website:

7. Outcomes

In addition to the specific outputs described in the previous section, the project has achieved a number of outcomes: some intended, others not.

Digitisation experience

We gained significant experience in digitising 19th century pamphlet literature. Much of this learning is reflected in the Technical Specifications included in Appendix C. Rather than using part-time or temporary labour, as is common with large-scale digitisation projects, BOPCRIS chose to recruit scanning staff for the full duration of the project. There were some staff changes during the project, but several five stayed throughout the project. It is regrettable that the considerable expertise they have built up over this period is now lost.

Metadata experience

Significant experience was also gained in working with the metadata standards we chose. The approach we took has been documented in Appendix D, where it will serve as an exemplar for others. The officer responsible for developing the metadata has now joined one of the partner libraries as their new Digitisation Manager, so this experience is being retained within the community.

Licensing experience

As noted above, the licensing proved much more complex than we had anticipated. So an unintended outcome is the considerable experience we gained in the processes of drafting and negotiating licences and working closely with lawyers. Our goal in this was to achieve the balance required to deliver the content as freely as possible, whilst providing sufficient legal protection for partners and protecting the business interests of JSTOR, which would ensure the sustainability of the collection. We believe this was achieved, but time will confirm this.

Cooperation among RLUK libraries and with JSTOR

RLUK had hoped that the project would provide a good opportunity for its members to work closely together and would provide benefits for all involved. This proved to be the case, with the strong relationships between partners providing sufficient strength to overcome the challenges the project faced. Although we have been unable to move to further phase of the 19th Century Pamphlets Online project, we hope to build further on the relationships established within the project. We note that the project has fed into the digitisation strand of RLUK's strategy and that two members of the project team are actively involved within RLUK's Digitisation Think Tank.

The project was also to develop a very good relationship with JSTOR, which it hopes will lead to further work with UK libraries.

Impact on JSTOR and Copac services

As noted previously in this report, in seeking to embed our content within the existing JSTOR service and to utilise the Copac service, we have prompted the further development of both of these services. This has been of direct benefit to this project and the user community, but it is also expected to benefit future projects – as more content is added to JSTOR and Copac acquires further enhanced catalogue records.

Testing of a new model for preserving and sustaining digital resources

The business model chosen by this project to sustain and preserve the content has attracted much interest. It was highlighted at the Digitisation Programme's Cardiff Conference in 2007[40], and is the focus of current research into sustainability models (by Ithaka[41]) and preservation (by the University of London Computing Centre and Portico[42]).

Impact on research and teaching

One of the most important outcomes envisaged by the project cannot be evaluated at this point in time, but belongs to the future:

- Increased use of 19th Century pamphlets, leading to advances in scholarship, more stimulating teaching, and a greater understanding of this form of literature

8. Conclusions

The 19th Century Pamphlets Online project has succeeded in its goal of "providing researchers, teachers and learners with online access to significant collections of 19th century pamphlets held within UK research libraries."

It has met its objectives of:

1. digitising a wide selection of pamphlets,
2. establishing an efficient consortial scanning operation,
3. providing sustainable preservation and delivery,
4. enabling sophisticated, distributed resource discovery and access, and
5. providing models for further phases/projects.

Although the project avoided the complexity and risks involved in developing its own delivery platform for the pamphlets, it faced many challenges – particularly the logistics of coordinating seven overlapping collections, maintaining a high rate of production for difficult material, and putting in place the licensing agreements necessary to safeguard the preservation and delivery of the collection over a very long period of time. The project has met all these challenges and learned much in the process.

[40] See http://digitisation.jiscinvolve.org/digitisation-conference-2007/
[41] See http://sca.jiscinvolve.org/2009/02/11/scaithaka-business-models-and-sustainability-case-studies-workshop/
[42] See http://digipressurvey.jiscinvolve.org/

When the project management team reviewed the project in preparation for this report, it felt that the biggest challenge within the project was to manage the complexity of a large partnership. However, it also noted that this large and broad partnership provided much opportunity within the project and proved to be one of the most important assets. We would encourage those undertaking similar projects to invest in partner relationships, developing clear understandings and expectations as early as possible and ensuring that there are good channels of communication in place.

Although an important aim of the project was to create a substantial amount of content, there has always been a further dimension to the work we've undertaken: that of research and development. The project considered what technical specifications might best suit this kind of content. It tested a centralised method for scanning multiple collections with overlapping content. It tested a semi-commercial business model for ensuring the preservation and sustainability of the data it created. It has enhanced several existing services and tools and left resources to support those undertaking similar work.

9. Implications

Our initial bid made reference to a potential series of projects of which this was Phase 1. The project has been successful in creating a partnership based on innovation, shared contribution and experience. We hope that this will provide a legacy of knowledge and experience, which will be retained and disseminated. Much of this has been embedded within individual partner institutions, but it is also being shared more broadly, through further work between partners and more formally through RLUK's Digitisation Think Tank.

Given the reduced resources available to the JISC for further digitisation work, we welcome its emphasis, in the 02/09 e-Content call on exploiting capacity and infrastructure and building on existing clusters such as 19th century content. This project will have hopefully provided some input into this new approach. At the same time the project has been contributing to the Ithaka work on creating case studies for digital preservation, helping to broaden understanding of the type of collaborative partnership development achieved between RLUK libraries and JSTOR.

One of the implications of the 02/09 call is the move away by JISC from being seen as a major funder of mass digitised content for the UK community. The first two calls not only provided a mass of content, but funded innovation where research and development as well as high quality delivery across a range of complex formats could be promoted free from the constraints of commercially-led practice. It is unclear where and when the next tranche of funding will appear, but it is hoped that the practice established by this project and its innovative sustainability model will help to inform future publicly funded calls.

10. References

This report has referred in several places to the project's Scoping Study and Project Plan. These are both available on the JISC website and on http://www.britishpamphlets.org.uk/ .

- **Scoping Study**
  JISC: http://www.jisc.ac.uk/media/documents/publications/digicurlscopingstudy.pdf
  Pamphlets site: http://www.britishpamphlets.org.uk/

- **Project Plan**
  JISC: http://www.jisc.ac.uk/media/documents/programmes/digitisation/pampp.pdf
  Pamphlets site: http://www.britishpamphlets.org.uk/

Appendices

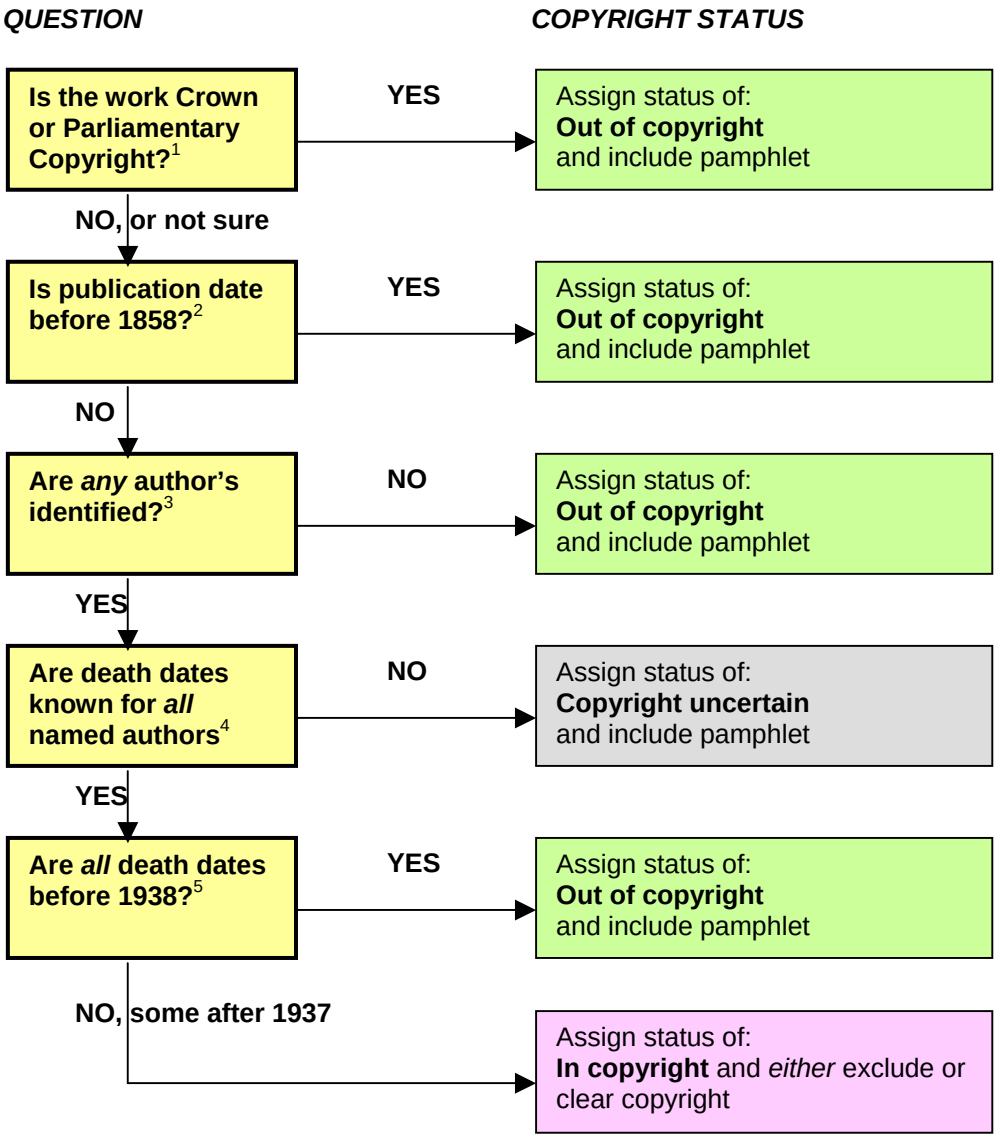
A. Glossary
B. Copyright Workflow
C. Scanning Guidelines
D. METS Metadata Profile
E. Licensing Diagram

A. Glossary

AHDS	Arts and Humanities Data Service. See http://ahds.ac.uk/
BOPCRIS	University of Southampton Library's specialist digitisation unit. See http://www.southampton.ac.uk/library/bopcris/
Copac	National merged catalogue of major university, specialist and national libraries in the UK and Ireland. See http://copac.ac.uk/
CURL	Consortium of Research Libraries in the British Isles. Now called RLUK (see below)
DPI	Dots Per Inch – a measure of image resolution
FTP	File Transfer Protocol – a means of transferring data via networks
JISC	Joint Information Systems Committee. Funding body. See http://www.jisc.ac.uk/
JSTOR	US non-profit provider of digital resources for the scholarly community. See http://www.jstor.org/
Mimas	National provider of data and bibliographic services. See http://www.jstor.org/
OCR	Optical Character Recognition. An automated means of generating electronic text from images of printed pages.
RePEc	Research Papers in Economics. See http://repec.org/
RLUK	Research Libraries UK. Major grouping of UK and Irish research libraries. See http://www.rluk.ac.uk/
RSLP	Research Support Libraries Programme. Funding programme. See http://www.rslp.ac.uk/
TASI	Technical Advisory Service for Images. JISC-funded advisory service. Now called JISC Digital Media. See http://www.jiscdigitalmedia.ac.uk/

B. Copyright Workflow

This chart was developed in 2007 to the project. It presents a series of questions that are intended to enable librarians to determine the copyright status of a pamphlet as simply and quickly as possible. The notes provide information about the underlying assumptions. Note that if were used for another project, some dates would need adjusting.



Notes:

1. Crown and parliamentary copyright = 50 years after publication.
   For definitions of these, see http://www.opsi.gov.uk/advice/index.htm
2. This is a conservative cut-off based on the possibility that an author may have published at 20 yrs and lived to 100 yrs. 1858 represents 150 years prior to 2007, when this copyright workflow was prepared.
3. Copyright in anonymous works = 70 years after publication.
4. If an author is named but their death-date is not given, then copyright is uncertain. Libraries may wish to investigate further by checking library catalogues or sources such as the Library of Congress Authorities http://authorities.loc.gov/, Oxford Dictionary of National Biography http://www.oxforddnb.com/, or Karlsruhe Virtual Catalogue http://www.ubka.uni-karlsruhe.de/hylib/en/kvk.html.
5. If all authors have died before 1938 then the work is out of copyright (e.g. 1937+70yrs=2007). If any author died in 1938 or later then the work is still within copyright and should be excluded or cleared.

C. Scanning Guidelines

See following pages.

UNIVERSITY OF Southampton

Report

Title: **19th Century Pamphlets Online Scanning Guidelines**

From: BOPCRIS, Hartley Library, University of Southampton, Southampton, SO17 1BJ

Tel 023 80598730

Email bopcris@soton.ac.uk

Date: March 2009

Contents

Introduction

The following guidelines were used by BOPCRIS within the 19th Century Pamphlets Online digitisation project. They are based on a particular form of publication (pamphlets) and particular equipment and requirements, but may be of use to others undertaking similar projects.

1. Book Scanners

1.1. Due to the fragile nature of the material, all pages were scanned by hand using the following book scanners within the Hartley Library scanning laboratory :

- PS7000 book scanners from Kodak[1]
- SupraScan book scanner from I2S[2]
- CopiBook book scanners from I2S[3]. These scanners are lit by two desk lamps with cold output
- Robotic Scanner from 4Digital Book operating in manual mode[4]

2. Image capture

2.1. Binding line margin

- Maintain at least 3-5mm between the binding line and the printed text. If this is not achievable on the CopiBook book scanners, pass the pamphlet on to the SupraScan scanner, which can scan deeper into the gutter. If the gutter is still not achievable, note this in the Scanning Database in the 'Binding' field.

2.2. Blank pages

- Scan all pages, including blanks. Scan the backs of fold-out maps to the same page size as their fronts.

2.3. Board, spine, front and back matter

- For pamphlets bound in original bindings, scan boards and spine in colour . Capture all front and back pages, which might include a handwritten table of contents. These images are saved within the following directories:
    - \master\library identifier\volume\front
      (for all front matter including the spine, front board and front matter)
    - \master\ library identifier \volume\back
      (for all back matter ending with the outer back board)

2.4. Damage

- Any physical damage to pages, board and spines should have been noted in the 'Notes' field within the Database by contributing libraries before the pamphlets were transferred. The damage should then have been verified by scanning laboratory upon arrival by the addition of the phrase 'Noted' within the 'Notes' field.
- Any damage not noted by the contributing library and any subsequent physical damage that occurs to a pamphlet, book or binding while at scanning laboratory, must be detailed in the 'Notes' field starting with the text: 'SCANNER NOTE ...'.
- Notes should be made of any tears that are >10mm and any damage that could lead to further damage. Reference should be made to the actual page number of the pamphlet and not the tiff number. The note should reference both the front and back of the page, for example 'page 4/5 tear noted'. Any damage caused during the scanning must be

[1] http://www.konicaminolta.co.uk/business-solutions/products/monochrome-systems/product-overview-discontinued-products/book-scanner-ps7000.html

[2] http://ww.i2s-bookscanner.com/en/products_SUPRASCAN.asp

[3] http://www.iiri.com/i2s/copibook.htm

[4] http://www.4digitalbooks.com/default.htm

noted, for example, 'SCANNER NOTE: Front board became detached during scanning process.'

2.5. Duplicates

- The libraries' preparations included checking for duplicates within their own collections and in other library collections. Only scan duplicates where a previous copy has been sent and marked as damaged.

2.6. File formats

- Files are saved in baseline TIFF 6.0 format.

2.7. Foldouts

- Scan foldout maps, diagrams or tables as one image. For 400dpi resolution can be maintained up to A1 (81.4" x 59.4cm). For paper sizes larger than A1, the dpi will be reduced to 300dpi.

2.8. Frame size

- Set a frame size for each pamphlet approximately 10-20mm beyond the three out edges and 10mm over the binding edge. The left and right frames should be identical in size and kept constant throughout the entire scanning of the pamphlet. For pamphlets with a large number of pages, the gutter may increase due to a relaxation of the binding. In these circumstances, the frame size should be large enough to accommodate this increase.

2.9. Glass use

- Pamphlet pages scanned on the PS7000 up to A3 size are scanned beneath 4mm float glass. This enables all page edges to be captured without any interference from fingers and to reduce rippling. With the PS7000 scanners, the glass is hand held horizontally across the pamphlet. The glass on the SupraScan book scanner is hinged and operates only in a horizontal position. Pages larger than A3 are scanned if appropriate with glass. The CopiBook scanners have a non-reflective glass plate which lowers automatically onto the pages.

2.10. Page masking

- Insert a thin black card (approximately 30cm square) beneath pages to provide a black outer edge. Insert the card up to a maximum of ten pages.

2.11. Rotation

- Scan all text in the reading orientation.

2.12. Scanning resolution

- All pamphlets are scanned at 100% (uninterpolated) using one of the following book scanners:
    - PS7000 Grey scale – 8 bit 400 dpi
    - SupraScan Grey scale – 8 bit or Colour-24bit 300 dpi
    - CopiBook Grey scale – 8 bit or Colour-24bit 300 dpi
    - DL Scanner Grey scale – 8 bit 300 dpi

- Pages containing only black and white print or writing should be scanned in greyscale. If a page contains any colour element (annotation, print, coloured paper) the whole page should be scanned in 24-bit colour. If the paper and print/writing cannot be distinguished in grey scale, the page should be scanned in 24-bit colour. Do not treat colouration due to aging as a colour element.

2.13. Scanning sequence

- Scan all pages in the order presented within the pamphlet or volume. If pages are bound out of sequence no attempt will be made to make a correction. If there appears to be missed page, a blank will not be added. If pages of a pamphlet have been mis-bound amongst or between other pamphlets, these pages will be scanned in the correct pamphlet page sequence, rather than the book order sequence.

2.14. Serial scanning

- Sometime are series of pamphlets have been treated by libraries as a serial publication and assigned a single catalogue record. In this case, all pamphlets in the series should be scanned sequentially.

2.15. Show-through

- Show-through can occur from four sources:
  - Bleed-through of ink from the back of the page
  - Over printing caused by the transfer of wet ink from the page opposite
  - Show-through of print from the back of the page
  - Show-through of print from the pages beneath due to paper transparency
- Show-through can be reduced at the scanning stage by the following processes:
  - On the PS7000 book scanners, reduce bleed-through by making a contrast adjustment.
  - Where there has been overprinting through the transfer of wet ink from the opposite page, use the SupraScan book scanner which has a 'softer' light source.
  - Where show-through is due to paper transparency, insert white paper beneath individual pages to enhance the contrast of the print.
  - Black paper beneath individual pages can be used to block show through from underlying pages.

2.16. Tables

- If a table or diagram goes across two pages, scan these as two separate pages.

2.17. Text alignment

- Pages are scanned to obtain relatively horizontal text. If pages are bound or printed on the skew, the scanner operator will shift the page to align the text horizontally.

2.18. Vertical binding alignment

- Where individual pamphlet pages are bound within a volume at differing vertical locations, the frame size should encompass the full extent of the vertical shift.

3. Metadata recorded during capture

3.1. Introduction

An in-house scanning database was utilised by the scanner operators to track the scanning of pamphlets, carry out QA and to record metadata items noted during the scanning process. Selected items are carried into the METS record.

3.2. Additional material

- If a pamphlet is accompanied by additional material, for example a letter, the tiff filenames of these pages should be noted in the Scanning Database within the 'Additional' field. The binding sequence should be maintained: for example, if the letter precedes a pamphlet, it is given the first filenames in the sequence. This metadata is included within the xml.

3.3. Annotations

- Annotations are defined as any additional word, letter, stamp, seal, underlining, embossment, intentional mark or additional print that has been added to a page subsequent to printing or binding. This does not include material added during a repair. If an annotation is present, the corresponding tiff number should be noted in the Scanning Database in the 'Annotations' field. The database makes no distinction between different kinds of annotation. This metadata is included within the xml.

3.4. Associated matter

- Associated matter is defined as material printed to accompany the pamphlet but which lies outside the main body of the text. Examples include adverts, petitions or invitations. Adverts within the body of the text should not be regarded as associated matter. The tiff filenames of all images containing associated matter should be noted in the Scanning Database within the 'Associated matter' field. This metadata is included within the xml.

3.5. Binding/Cropping

- For pages where a gutter of 3-5mm cannot be established, the individual pages should be noted in the Scanning Database in the 'Binding' field. Loose pages that are individually scanned and consequently have no gutter, should not be marked as a 'Binding/Cropping' issue. This metadata is included within the xml.

3.6. Clipping

- If any page clipping has occurred during binding that results in the loss of printed or hand written text, the corresponding tiff number should be noted in the Scanning Database in the 'Clipped' field. Loss of text due to clipping may have taken place along the three outer edges or bound into the gutter during binding. This metadata is included within the xml.

3.7. Colour

- The tiff filenames of all pages containing any colour component should be noted in the Scanning Database within the 'Colour' field. This will include coloured paper, text, images, maps, diagrams and annotations. This metadata is included within the xml.

3.8. Errata

- The tiff filenames of all pages containing errata slips or printed errata should be noted in the Scanning Database within the 'Errata' field. This metadata is included within the xml. Location of Addendum material should not be recorded.

3.9. Folded pages

- The tiff filenames of all pages containing any folded pages should be noted in the Scanning Database within the 'Folded' field. This metadata is included within the xml.

3.10. Foreign language

- For all complete pages within a pamphlet with text in a language other than English, the tiff filenames should be noted in the Scanning Database within the 'Foreign language' field. A drop-down box will enable a foreign language to be defined. English is the assumed default and no entry need be made to the database. This metadata is included within the xml.

3.11. Imagery

- The tiff filenames of all pages containing any printed imagery associated with the pamphlet text and provided by the author should be noted in the Scanning Database within the 'Imagery' field. This should not include printer's embellished letters or lines, or arrows. This metadata is included within the xml.

3.12. OCR issues

- Any bleed through, show through, wet ink transfer, marks, or foxing that is likely to limit the OCR of the printed text should be noted in the Scanning Database. List the tiff filenames within the 'OCR issue' field. This metadata is included within the xml.

3.13. Pagination

- The tiff filenames for all pages within a pamphlet that are incorrectly paginated (due to either printing or binding errors) should be noted in the Scanning Database within the 'Pagination' field. This metadata is included within the xml. Pages should be scanned in the order of the pamphlet, and no corrections made. If an incorrect pagination affects the remaining pages of a pamphlet, all the pages following the error should be noted in the database. Where a pamphlet uses a numbering sequence not starting at zero (for example beginning on page 267) and the sequence is correctly maintained beyond this, this should not be marked in the 'Pagination' field.

3.14. Rotation

- Pages printed or bound in a landscape position, for example tables and images, are rotated by the scanner software at the time of scanning to a read orientation. Images are then saved in this read orientation. The tiff filenames of all pages rotated should be noted in the Scanning Database within the 'Rotation' field. This metadata is included within the xml

3.15. Tables

- The tiff filenames of all pages containing any tabular row and column data, or list data, or bulleted lists should be noted in the Scanning Database within the 'Table' field. This metadata are included within the xml. Do not include Table of Contents or indexes or hand-drawn tabular data, but include worked equations. This metadata is included within the xml.

3.16. Watermarks

- These should be noted in the 'Notes' field using the standard text 'water marks'

4. Post-processing

- Prior to OCR the images are automatically rotated with proprietary software to give a horizontal text-block and resaved in this form. Due to the nature of the printed text, this does not mean that every line will be completely aligned or straight.
- Images are then cropped by the same proprietary software to the three outer edges of the page and within the binding line. This will result in some black edging included with the images.
- Images are then saved with LZW compression.

D. METS Metadata Profile

This is the project's METS profile. It is reproduced from
http://www.loc.gov/standards/mets/profiles/00000024.html


```
<?xml version="1.0" encoding="UTF-8" ?>
- <METS_Profile xmlns="http://www.loc.gov/METS_Profile/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.loc.gov/METS_Profile/
      http://www.loc.gov/standards/mets/profile_docs/mets.profile.v1-2.xsd">
    <URI LOCTYPE="URL">http://www.loc.gov/mets/profiles/00000024.xml</URI>
    <title>RLUK 19th Century Pamphlets METS Profile</title>
    <abstract>This profile specifies the use of METS to provide metadata for the project to digitise
        19th Century Pamphlets under Phase 2 of the JISC Digitisation Programme. Further
        information about the project can be found at:
        http://www.southampton.ac.uk/library/bopcris/
        http://www.jisc.ac.uk/whatwedo/programmes/programme_digitisation/pamphlets.aspx The
        materials to be digitised comprise single pamphlets, sometimes bound into a volume,
        sometimes not. The item level is defined to be a pamphlet. One METS document should be
        present for each digitised pamphlet. If the pamphlets were bound into a volume, the front
        and back matter, including tables of contents, of the volume have also been scanned. In
        such cases, one METS document should also be present for each volume, containing the
        images of the volume and references to the METS documents for the pamphlets contained
        within the volume. The digital files for each pamphlet should comprise: master images
        (TIFF) and OCR full-text in two formats (plain text and word co-ordinated XML). The digital
        files for each bound volume should comprise master images only. All digital files are
        packaged together with the METS and extension metadata into a single digital object
        (TAR) for delivery and storage.</abstract>
    <date>2008-08-06T00:00:00</date>
  - <contact>
      <name>Ed Fay</name>
      <institution>University of Southampton</institution>
      <address>BOPCRIS Digitisation Centre, Hartley Library, University of Southampton,
          SO17 1BJ, UK</address>
      <phone>+44 (0) 23 8059 3575</phone>
      <email>E.Fay@soton.ac.uk</email>
    </contact>
    <related_profile>None</related_profile>
  - <extension_schema ID="MODS">
      <name>Metadata Object Description Schema (MODS)</name>
      <URI>http://www.loc.gov/standards/mods/v3/mods-3-2.xsd</URI>
      <context>mets/dmdSec/mdWrap/xmlData</context>
      <note>Bibliographic metadata at the item level, supplied in MODS by COPAC
          (http://www.copac.ac.uk) There will be one instance per pamphlet, and none per
          bound volume. This metadata will be contained in a dmdSec linked to the top level div
          in the logical structMap. MODS metadata will conform to version 3.2 of the
          schema.</note>
    </extension_schema>
  - <extension_schema ID="MIX">
      <name>NISO Metadata for Images in XML (NISO MIX)</name>
      <URI>http://www.loc.gov/standards/mix/mix20/mix20.xsd</URI>
      <context>mets/amdSec/techMD/mdWrap/xmlData</context>
      <note>Technical metadata at the file level, extracted from standard file information and
          TIFF headers. There will be one instance per master image file listed in the fileSec.
          This metadata will be contained in an amdSec linked to the relevant file element. MIX
          metadata will conform to version 2.0 of the schema.</note>
    </extension_schema>
  - <extension_schema ID="PREMIS">
      <name>PREMIS Data Dictionary for Preservation Metadata</name>
      <URI>http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd</URI>
      <context>mets/amdSec/digiProvMD|techMD/mdWrap/xmlData</context>
```

<note>Technical metadata at the file level, extracted from standard file information. There will be one instance per file of any kind listed in the fileSec. This metadata will be contained in an amdSec linked to the relevant file element. The PREMIS components will be contained within seperate METS elements, as suggested by the "Guidelines for using PREMIS within METS" at http://www.loc.gov/standards/premis/premis-mets.html techMD: premis:object for each file digiProvMD: premis:event containing information about file derivations digiProvMD: premis:agent containing information about software packages PREMIS metadata will conform to version 2.0 of the schema.</note>
</extension_schema>
- <extension_schema ID="descriptors">
<name>BOPCRIS Descriptors Metadata</name>
<URI>http://www.southampton.ac.uk/library/bopcris/xsd/descriptors/bopcris-descriptors-0-2.xsd</URI>
<context>mets/dmdSec/mdWrap/xmlData</context>
<note>Descriptive metadata at the page level, recorded by scanner operators at the point of scanning. This schema indicates attributes of a page, for example that it contains text in a certain language, imagery or tabular data. There are 12 such attributes enumerated within the schema. There will be one instance per item page, indicating language as a minimum. Language information uses the controlled vocabulary ISO 639.2. This metadata will be contained in a dmdSec linked to the page level div in the logical structMap. BOPCRIS Descriptors metadata will confrom to version 0.2 of the schema.</note>
</extension_schema>
- <extension_schema ID="provenance">
<name>BOPCRIS Provenance Metadata</name>
<URI>http://www.southampton.ac.uk/library/bopcris/xsd/provenance/bopcris-provenance-0-1.xsd</URI>
<context>mets/amdSec/digiProvMD/mdWrap/xmlData</context>
<note>Administrative metadata at the object level, indicating the provenance of the item. This schema contains information about the source library, collection and shelfmark of the original item. There will be one instance per pamphlet, none per bound volume. This metadata will be contained in an amdSec linked to the top level div in the logical structMap. BOPCRIS Provenance metadata will conform to version 0.1 of the schema.</note>
</extension_schema>
- <extension_schema ID="rights">
<name>BOPCRIS Rights Metadata</name>
<URI>http://www.southampton.ac.uk/library/bopcris/xsd/rights/bopcris-rights-0-1.xsd</URI>
<context>mets/amdSec/rightsMD/mdWrap/xmlData</context>
<note>Administrative metadata at the item level, indicating the copyright status of the item. For a full explanation of the copyright status, see the project documentation. There will be one instance per pamphlet, none per bound volume. This metadata will be contained in an amdSec linked to the top level div in the logical structMap. BOPCRIS Rights metadata will conform to version 0.1 of the schema.</note>
</extension_schema>
- <description_rules>
Bibliographic records will conform to the descriptive specifications of COPAC at the time of export (2007).
For further information consult COPAC: http://www.copac.ac.uk/
</description_rules>
- <controlled_vocabularies>
- <vocabulary ID="ISO_639_2">
<name>ISO 639.2 Codes for the representation of names of languages-- Part 2: alpha-3 code</name>
<maintenance_agency>Library of Congress</maintenance_agency>
<URI>http://www.loc.gov/standards/iso639-2/</URI>
- <context>
BOPCRIS Descriptors metadata uses ISO 639.2 language codes to indicate the presence of text in a certain language on a page.
</context>
</vocabulary>

```
</controlled_vocabularies>
- <structural_requirements>
  - <metsRootElement>
    - <requirement ID="OBJID">
        <p>OBJID must contain a project specific ID.</p>
        <p>BOPCRIS generated a unique ID for each pamphlet based on: the local holding ID from the bibliographic record, library identifier, volume identifier (if applicable, otherwise the default value 000 was used), and another number to indicate whether the pamphlet is a duplicate item.</p>
        <p>Bound volume identifiers are generated from library identifer and volume identifier only.</p>
        <p>Library identifiers are identical to the institution codes used to identify libraries within the COPAC database. Volume identifiers are derived from the information about the volume available within the bibliographic record, in most instances, or from another source, such as a barcode, where necessary.</p>
        <p>Pamphlet ID form: holdingID_library-volume-duplicate Examples: 1234567890_tst-000-1 27000123456_bri-1800-2 19B39982X_liv-531-1</p>
        <p>Bound volume ID form: library-volume Examples: liv-531 ucl-A70</p>
      </requirement>
  </metsRootElement>
  - <metsHdr>
    - <requirement ID="metsHdr">
        <head>Applies to: pamphlet and bound volume METS documents.</head>
        <p>metsHdr must be present, indicating creation and modification dates. At the point of delivery these will be identical, as METS documents are generated as an export at the end of the creation process for a digital object, and are not subject to ongoing revision.</p>
        <p>Sub-element agent (role="CREATOR") will indicate the source of the object at BOPCRIS.</p>
      </requirement>
  </metsHdr>
  - <dmdSec>
    - <requirement ID="dmdSec_Biblio">
        <head>Applies to: pamphlet METS documents only.</head>
        <p>A dmdSec with attribute ID="dmdSec_Biblio" will be present containing the bibliographic record of the item.</p>
        <p>The bibliographic record will be in XML (MIMETYPE="text/xml") and MODS format (MDTYPE="MODS").</p>
        <p>The record will be contained within the METS document using mdWrap.</p>
      </requirement>
    - <requirement ID="dmdSec_page">
        <head>Applies to: pamphlet METS documents only.</head>
        <p>dmdSecs, one per physical page, will be present, containing descriptive metadata at the page level. attribute ID will be "dmdSec" + "_" + physical page identifier (see requirement ID="file_naming"). Examples: dmdSec_00000001 dmdSec_00000007 dmdSec_00000017</p>
        <p>Metadata at this level will be in XML (MIMETYPE="text/xml") and BOPCRIS Descriptors Metadata format (MDTYPE="OTHER")</p>
        <p>This metadata will be contained within the METS document using mdWrap.</p>
      </requirement>
  </dmdSec>
  - <amdSec>
    - <requirement ID="amdSec_Object">
        <head>Applies to: pamphlet METS documents only.</head>
        <p>An amdSec with attribute ID="amdSec_Object" will be present containing administrative metadata at the item level.</p>
        <p>This metadata will include: rightsMD ID="rightsMD_Object" containing rights metadata for the item in XML (MIMETYPE="text/xml") and BOPCRIS Rights Metadata format (MDTYPE="OTHER"). digiProvMD
```

ID="digiprovMD_Object" containing provenance metadata for the digital object in XML (MIMETYPE="text/xml") and BOPCRIS Provenance Metadata format (MDTYPE="OTHER").
These metadata will be contained within the METS document using mdWrap.
</requirement>
- <requirement ID="amdSec_PREMIS-AGENTS">
<head>Applies to: pamphlet METS documents only.</head>
An amdSec with attribute ID="amdSec_PREMIS-AGENTS" will be present. In cases where post-processing actions have been performed on images, such as cropping and/or OCR, this section will contain premis:agent elements indicating the agents that performed such actions.
</requirement>
- <requirement ID="amdSec_file">
<head>Applies to: pamphlet and bound volume METS documents.</head>
amdSecs, one per file, will be present, containing technical information encoded in multiple extension schemata.
The attribute ID will be constructed from "amdSec" + "_" + file ID (see requirement ID="file"). Examples (pamphlet item): amdSec_MASTER_00000001 amdSec_TXT_00000007 amdSec_IDX_00000017 Examples (bound volume): amdSec_MASTER-volume-front-00000001 amdSec_MASTER-volume-back-00000001
amdSecs for every file will contain PREMIS metadata: All files will have a premis:object element. In cases where files are the parent or child of another (due to OCR processing), this will be indicated in the premis:object. The event of derivation will be recorded in a premis:event element, linked to a premis:agent.
The premis:object for each file will indicate the format, and should indicate a registry providing format information when available.
amdSecs for master image files will additionally contain MIX metadata.
</requirement>
</amdSec>
- <fileSec>
- <requirement ID="fileSec">
<head>Applies to: pamphlet and bound volume METS documents.</head>
There will be one file group per file type. fileGrp attribute ID will indicate the relevance. Master Images MASTER Plain text OCR TXT Word co-ordinated XML OCR IDX
Pamphlet items will contain all three file groups, bound volumes will contain only master images.
</requirement>
- <requirement ID="file">
<head>Applies to: pamphlet and bound volume METS documents.</head>
There will be one file element per file, referencing the file location using FLocat . FLocat elements will be LOCTYPE="URL" and use xlink:href to point to files. File locations will be given relative to the path of the METS document.
As the digital object is packaged into a single TAR file for delivery, relative paths allow the TAR package to be unpacked anywhere and, providing the directory structure is maintained on unpacking, all paths should remain accurate.
file ID will be constructed from fileGrp ID + "_" + physical page identifier (see requirement ID="file_naming"). Examples (pamphlet item): MASTER_00000001 TXT_00000007 IDX_00000017 Examples (bound volume): MASTER_volume-front-00000001 MASTER_volume-back-00000001
file GROUPID will be identical to the physical page identifier (see requirement ID="file_naming"). Example (pamphlet item): 00000001 Example (bound volume): volume-front-00000001 volume-back-00000001
file AMDID will link to the amdSec containing the technical metadata for the file, as indicated in requirement ID="amdSec_file".

file CHECKSUM and CHECKSUMTYPE will be present. Checksums will be calculated using MD5.
file MIMETYPE will be present.
</requirement>
</fileSec>
- <structMap>
- <requirement ID="structMaps">
<head>Applies to: pamphlet and bound volume METS documents.</head>
There will be a logical and physical structure map.
In the case of pamphlet items: The logical and physical structure maps will be identical, except the logical structure map will also contain ID linkages to relevant metadata sections.
In the case of bound volumes: The physical structure map will contain only the image files comprising the covers. The logical structure map will also contain pointers to the METS documents of pamphlets contained within that volume.
</requirement>
- <requirement ID="structMap_logical">
<head>Applies to: pamphlet and bound volume METS documents.</head>
In all cases, the following attributes will be present: ID="structMap_logical" TYPE="logical"
In the case of pamphlet items: Top level will contain one div: ID="logical_root" This div will be linked to dmdSec_BIBLIO and amdSec_OBJECT. This div will contain page level divs (see requirement ID="div_page").
In the case of bound volumes: The top level will contain one div: ID="logical_root" This div will contain further divs: TYPE="section" These divs will contain, either: Page level divs (see requirement ID="div_page") Or: mptr elements, pointing to the pamphlets contained within the volume.
</requirement>
- <requirement ID="structMap_physical">
<head>Applies to: pamphlet and bound volume METS documents.</head>
In all cases, the following attributes will be present: ID="structMap_physical" structMap TYPE="physical"
The top level will contain one div. This div will contain page level divs (see requirement ID="div_page").
</requirement>
- <requirement ID="div_page">
<head>Applies to: pamphlet and bound volume METS documents.</head>
Page level divs represent a physical item page.
Page level divs contain a single fptr element for each file that constitutes a representation of that page, in all formats. There will be one fptr present for the file of each type that represents that page. A page level div will contain a mets:fptr for every equivalent file of each type that is a representation of that page.
div TYPE="page"
ORDER and ORDERLABEL will be present, equal to the physical order of the page. In the case of pamphlet METS documents, this will be equal to the sequential page number. In the case of bound volume METS documents, this will be equal to a sequential number beginning at 00000001.
</requirement>
</structMap>
</structural_requirements>
- **<technical_requirements>**
- <content_files>
- <requirement ID="file_naming">
METS documents will be named by the OBJID + ".xml"
There will be one file per file group per physical page.
Files will be referenced from within METS documents using their path relative to the location of the METS document.
Filenames will be 8 characters in length, plus extension.

Files will be sequentially numbered, starting at 00000001.
In the case of pamphlet items: Content files will be arranged by sub-directory according to file group. Examples: ./master/00000001.tif ./txt/00000007.txt ./idx/00000017.idx
In the case of bound volumes: Content files will be arranged by sub-directory according to their relevance to the volume. Examples: ./volume/front/00000001.tif ./volume/back/00000001.tif
The physical page identifier is constructed from components of the relative path and filename (minus extension). Example (pamphlet item): 00000001 00000007 00000017 Examples (bound volume): volume-front-00000001 volume-back-00000001 The physical page identifier will be used in construction of METS element IDs for those elements relating to files or their metadata. See requirement ID="file" for example.
</requirement>
- <requirement ID="master_image_files">
Master image files will be in TIFF 6.0 format.
TIFF files will be compressed using LZW.
</requirement>
- <requirement ID="OCR_files">
OCR output will be present in plain text and word co-ordinated XML format.
Word co-ordinated XML is in IDX format. This is a derivative of Abbyy FineReader SDK XML output, generated by the OCR workflow software: Agora (SRZ Berlin).
IDX XML files contain: A <milestone unit> indicating the dimensions (width and height, in pixels) of the source image. Individual word locations <w> given in pixels relative to the dimensions of the source image. Word locations contain co-ordinates of: left (l), top (t), width (w) and height (h). Individual words are contained within sentences <s>. Sentences are contained within paragraphs . Sentences and paragraphs themselves do not contain co-ordinates.
</requirement>
</content_files>
</technical_requirements>
- **<tool>**
<agency>BOPCRIS</agency>
- <note>
The digitization workflow at BOPCRIS is co-ordinated by a relational database. This database is also used by scanner operators to capture metadata for mapping to BOPCRIS Descriptors format. METS documents are generated as the product of a combination of: An export from this database Extraction of technical metadata from digital files (standard file information, and TIFF headers) The bibliographic record for the item.
These tools are currently for internal use only.
</note>
</tool>
</METS_Profile>

E. Licensing

This diagram provides an overview of the project's licensing requirements, showing how the licences feed into each other.