

JISC DEVELOPMENT PROGRAMMES

Project Document Cover Sheet

Final Report

Project

Project Acronym	BOPCRIS	Project ID	
Project Title	BOPCRIS 18 TH Century Parliamentary Papers Digitisation Project		
Start Date	1st April 2005	End Date	30th March 2007
Lead Institution	University of Southampton		
Project Director	Mark Brown		
Project Manager & contact details	Julian Ball Digitisation Manager Hartley Library University of Southampton Southampton SO17 1BJ Tel: 023 80598730 Email: j.h.ball@soton.ac.uk		
Partner Institutions	n/a		
Project Web URL			
Programme Name (and number)	CSR2 Digitisation Programme		
Programme Manager	Stuart Dempster		

Document

Document Title	Final Report		
Reporting Period	2005-2007		
Author(s) & project role	Mark Brown Project Director		
Date	28 February 2007	Filename	FinalReport.doc
URL			
Access	<input checked="" type="checkbox"/> Appendices A, H and I are not for general distribution.	<input type="checkbox"/> General dissemination	

Document History

Version	Date	Comments
1a	28 February 2007	

JISC Final Report

BOPCRIS 18TH Century Parliamentary Papers Digitisation Project

Mark Brown
Project Director

28 February 2007

Table of Contents

Acknowledgements.....	2
Executive Summary.....	3
Background.....	3
Aims and Objectives.....	3
Methodology.....	4
Implementation.....	6
Outputs and results.....	6
Outcomes.....	7
Conclusions.....	8
Recommendations.....	9
Appendices.....	9

Acknowledgements

The BOPCRIS 18TH Century Parliamentary Papers Digitisation Project has built upon the pioneering work in digitisation undertaken by the late Simon Brackenbury. The project team would like to pay tribute to Simon's commitment, energy and enthusiasm for digitisation at the University of Southampton, which was cut short by his tragic death in April 2005. We would also like to acknowledge the contribution made by the late Jeanette Cochrane, whose long association with British Official Publications, and whose foresight in the early years of programmes launched under the first Follett Report, created the momentum to build upon the unique work of Professor Peter and Mrs Grace Ford at Southampton whose commitment to making British Official Publications more accessible continues to-day through our digitisation programme..

The Project Team at Southampton, Julian Ball, Christine Fowler, Richard Wake, Wendy White and myself, have worked together as a team to bring this project to fruition. Our colleagues in the digitisation lab, including Maureen Langham, Verna Acres, Graham Caisley, Rob Boston, Patrick Dittrich, Gareth Williams, Maria Suzanna Avila and many staff from the University Temp Bank have worked through many issues and shown great commitment to the project. We would also like to thank our colleagues in Information Systems Services, particularly Oz Parchment and Jim Fuller, for providing so much system design and know-how in the setting up of the local server architecture. We have all engaged with many technical and environmental challenges to deliver the work, and have done so with both enthusiasm and determination. In this we have been strongly supported by our Steering Group, Bill Noblett (University of Cambridge), Jennie Grimshaw (British Library), Chris Pond (House of Commons), Rob Shoemaker (University of Sheffield), Paul Seward, Peter Gray, Julie Gammon (University of Southampton), Chris Owens (The National Archives), Alastair Allan (University of Sheffield) Robin Gadd (Brockenhurst College) and Stuart Dempster (JISC). Although it would be invidious to select particular individuals, I must acknowledge the support given by Bill Noblett from the University of Cambridge, and particular dedication and support of Jennie Grimshaw, Head of Official Publications at the British Library. Both have been generous and pragmatic in ensuring that content available in their libraries found its way into the corpus of the 18th Century collection.

There are a number of important technical inputs to this project which have greatly contributed to its success. The SRZ team in Berlin have provided continual support in shaping and implementing the Agora CMS software. External consultants, Michael Pidd contributed to the work on the XML framework, and web development by Mark Palmer (AnyWare).

We would like to acknowledge the work done by the JISC Collections Team, especially Lorraine Estelle, in helping us to engage with the complexities of licensing and sustainability models, and last but certainly not least, we would like to thank our Programme Manager, Stuart Dempster, for his continual support, careful guidance and sense of humour.

Executive Summary

The BOPCRIS 18TH Century Parliamentary Papers Digitisation Project has delivered 1,260,062 scanned printed and hand written pages of text and images representing the main official publications of the 'long' 18th Century. The output has include a number of technical developments, including Optical character recognition of printed texts, and word and place indexing enabled within web interface, as well as a web interface to bring together the existing BOPCRIS projects. This Final Project Report describes the approach, the activities undertaken, and the extent to which the objectives were met. It also celebrates a series of partnerships, not only with our content providers, the British Library and the University of Cambridge, but the technical input from a range of organisations and individual who have contributed their expertise and support.

As well as making a significant contribution to the rich corpus of historical documentation available free at the point of use to the UK HE and FE community, the project has helped to develop transferable knowledge and expertise in the area of digitising historical materials. Technical investigations included the exploitation of robotic scanning technique; skills, metrics and workflow, the development of an integrated hardware and software platform for management and output, the operation of evolving standards, and the implications of large-scale data storage. A subproject also investigated the application of semantic tagging to a part of the database by way of proof of concept. Issues of scalability and sustainability have also been addressed.

The report outlines the methodology employed, and shows some of the complexities involved in managing large-scale digitisation of historical materials. It also makes some recommendations for future work.

A series of technical reports are appended outlining specific work on applying standards and design principles.

Background

The project forms part of a long-term strategy begun in 1994 to provide electronic access and digitised content for the official publications collections based primarily on the Ford Collection at the University of Southampton. The aims of the strategy have been:

- exploring new approaches to digitisation printed materials, including exploiting hardware and software innovation, the implementation of standards, quality assurance and workflow management
- producing content to support research and education for the HE sector
- working collaboratively with a range of partners both in the UK and abroad to test standards and to achieve comprehensive coverage
- creating partnerships with software developers to provide linked best practice in digitisation, bringing together leading edge expertise
- exploring sustainability for publicly funded digitised output

Previous projects funded by NOF, RSLP and AHRB, focused on proof of concept in the context of resource enhancement, particularly in creating selective, descriptive records for individual documents. Feedback from previous projects indicated that practitioners wanted more comprehensive coverage and in-depth collections. The 18th Century British official publications project was a response to this need, and has explored the process for creating a large-scale, comprehensive collection of material.

In the context of the JISC Collections Strategy it can be identified as a heritage collection. It fits very well with the existing collections for the period, EEBO and ECCO, and has a link with the 19th Century parliamentary collection microfilm collection, now offered as digitised content by Proquest.

The 18th Century material is particularly important as the Palace of Westminster fire of 1834 destroyed many manuscript records, and the holdings are confined to the major research libraries, restricting access for researchers and students studying the history of political and public policy. As the project progressed, the range of content widened, and the project team have built upon the existing good relationships with potential partner libraries to build a collection which includes both a full array of published printed materials supplied by Southampton and Cambridge, and rare printed material held in the British Library.

Aims and Objectives

The objectives outlined in the project plan encompassed both proof of concept in the use of new techniques, and digitised output.

Technical Investigations

- exploitation of robotic scanning technique; skills, metrics and workflow
- integration of new hardware with software for management and output
- investigation of the operation of standards
- investigation of large-scale data storage
- integration of interface for official publications outputs from earlier projects
- investigation of scalability and sustainability
- collaboration with Stanford in sharing of technical knowledge for the 4digitalbooks scanner

Outputs

- 1,260,062 scanned printed and hand written pages of text and images representing the main official publications of the 'long' 18th Century
- Optical character recognition of printed texts
- Transcription/abstraction of titles from hand written texts
- Word and phrase retrieval with hit term high-lighting
- Subject indexing of all material to effect retrieval
- Word and Place indexing enabled within web interface
- web interface linked with existing BOPCRIS projects

The objectives did not change substantially during the project, although some of the approaches have been modified in the light of experience. One objective which was not identified clearly at the beginning of the project was the importance of developing the metadata schema, and advanced indexing. These objectives emerged from the detailed work carried out in scoping the workflow and assessing the material for retrieval. As a result of this some elements of automated indexing were introduced, and a small pilot carried out with the University of Edinburgh using semantic web techniques as proof of concept for place and person indexing.

Methodology

As well as producing a set number of pages of scanned output, the project has been investigating technical applications to scale up the models previously developed under the NOF, RSLP and AHRB projects. From the beginning it was assumed that the existing scanners and software would not be robust enough to manage the larger range of data, and a key objective was to explore technical and software solutions to manage a large-scale automated repository. There was therefore throughout a need to strike a balance between the investigation of innovation and ensuring that throughput was a level to ensure delivery.

There were six phases to the project, each of which overlapped as developments often took place in parallel:

Phase 1

In the first phase initial work was done to select material and create the indexing framework based on the Library of Congress Subject Headings. At the same time it was agreed with JISC to commission the DigiLine automated digital scanner as the first installation in the UK, allowing the project team to achieve a higher level of output more quickly. The use of a high capacity digital scanner was at the point of commissioning unique in the UK, and it was anticipated that it would prompt the sharing of expertise with non-UK users, particularly the Universities of Stanford and Goettingen. In the event although initial contacts were made, this did not result in any long-term links as the level and scope of the other programmes were significantly different to those at Southampton. Following the installation and testing of the DigiLine, an EU procurement was carried out for the purchase of a suitable CMS system to replace the 'daisy chaining' of the several individual software applications, which had supported the earlier digitisation projects, and which was not robust enough for the size of the new project. The choice of Agora, a large-scale repository product developed by SRZ in Berlin was the only contender able to meet the specification. The procurement was a long and complex process which took several months to bring to completion, but by December 2005 this phase was largely completed.

Phase 2

In anticipation of the technical infrastructure being in place, it was decided to go forward with the appointment of a digitisation manager, Julian Ball, to manage the repository. In this next phase, with Julian in post, intensive work took place to interface the DigiLine and the PS7000 flat bed scanners with the new Agora repository software. This proved more complex than had originally been envisaged. Although Agora had a track record in managing large-scale datasets, the structure of the parliamentary materials was more challenging than their previous projects, and much development work had to be done to try to achieve the full potential of the integrated repository material. As part of this significant work was also done to create the XML structure for managing and indexing the data which was an essential underpinning for the dataset.

This phase again lasted much longer than had been envisaged. Work was also inevitably affected by the tragic death in April 2006 of Simon Brackenbury, the project manager who had worked with the digitisation strategy at Southampton from the beginning. A Project Board within the Hartley Library was established to oversee the project and took over the detail of work which had previously lain with Simon. The Board consisted of Mark Brown, Richard Wake, Christine Fowler, Wendy White and Julian Ball.

Phase 3

Phase 3 related to the creation of content. It was found that the original page estimate for material held at Southampton and Cambridge based on sampling had been over-optimistic, and following advice from the Project Steering Group, we explored the possibility of extending the scope of the project's content to include rare material held in the British Library. This involved considerable negotiation and a rigorous assessment by the British Library of the environment and processes at Southampton. Agreement was achieved and formulated in a MoU between Southampton and the BL, and material started to be transferred in batches for scanning. The new content required specialised handling and management, and was not suitable for use on the DigiLine scanner. The workflow was adjusted to allow the new processes to be managed in parallel to completing the Southampton sources, using the PS7000 flat bed scanners. In exploring the indexing model for the new content, it was also decided to pilot sample indexing based on semantic principles. We therefore commissioned the Natural Language Group at the University of Edinburgh to undertake some research on developing personal and place name indexing.

Phase 4

This focused on web presentation and design. The approach was to build directly onto the Agora platform to create a totally integrated repository management. It was clear that this would involve considerable development work, and we therefore engaged an external IT consultant to work with the

Agora programmer to develop the potential of the integrated system. Through a careful process of analysis substantial progress was made in matching system to a potential model for delivery. Once core software work had been completed, ILRT was commissioned to conduct both a usability and an accessibility study, and the main recommendations were incorporated into the web design. It was however not possible to respond to all the requirements, largely as a result of the need for further work with the Agora element of the workflow. These limitations were significant in our decision to view the project website as a R&D platform with limited public access to the UK .ac.uk community. A more robust delivery model is under negotiation with a commercial host, which will provide a more sustainability model for ensuring access to the UK HE and FE community.

Phase 5

This phase focused on archiving, sustainability and delivery. These aspects were intrinsic to the project from the beginning, as decisions were required to ensure that the specification for the CMS had the capability of developing web access, that scanning output was archived, and that the long-term sustainability was assured. The perception of how the scanned images would be managed into the future changed over the course of the project. Options which appeared towards the latter part of the project did not exist initially. This was true both of options for 'dark' archiving, and for sustainability models. The resolution of the dark archiving was to build on existing arrangements elsewhere at the University of Southampton for large-scale data storage by placing a copy with the CCLRC. The more natural home would have been the AHDS, but at that time the AHDS felt unable to take such a large dataset into their collection. In terms of sustainability the concept of delivery directly from a platform at the University of Southampton shifted towards partnership working with the external host. There are real dilemmas in terms of sustainability of free access at point of use to the community, and experience here, and with other projects, has helped inform decisions in the second phase of digitisation projects.

Phase 6

In the summer of 2006 we applied to JISC to employ an underspend on the project to bring together earlier project material held at Manchester, whose server architecture was vulnerable. This involved enhancing the existing server capacity, and moving the database from Manchester to Southampton. The interface designed by the ILRT was replicated at Southampton so that the overall data was secure. If future funding allows, additional work could be done to integrate and enhance this data within the project website. This lies outside the scope of the current project.

Throughout the project we have scoped activity to apply information environment standards. In some areas this has proved challenging as the standards themselves, particularly those associated with XML descriptors contained in MODS and METS, are not stable, and have been under continual development in the course of the project. In general we have found that the buying in of external expertise, as was the case with the XML schema, the specification for dark archiving, the data to web design and the standards for usability and accessibility for the web interface itself, worked effectively as long as those working with the project were conversant with the standards or able to apply them. The major challenge in this sphere has been working with SRZ, whose business profile had not so far required them to meet these standards. Although they have been responsive to our needs, this may indicate potential limitations when working with technical partners outside the UK.

Implementation

The above summary outlines the process of implementation as part of the discussion of the methodology as the two are integral.

The detail of the implementation is contained in a number of specialised reports attached as Appendices to this report.

Outputs and Results

The 18th Century British official publications project has delivered its core requirement, 1,260,062 pages of scanned content as a single corpus. The actual content has been wider and more diverse than originally envisaged, and now includes semi-unique material from the British Library of particularly value to researchers and scholars in the field. Although the delivery and sustainability strategy is not yet fully agreed, we are confident that the requirement for the collection to be available free at the point of use to the HE and FE communities will be met, and that the research website developed by the Project Team will remain as a development platform for potential further work. The migration of the vulnerable earlier projects to a more robust platform at Southampton has also been achieved. The project has therefore offered a blueprint for can be achieved in the area of digitising historical material from major library collections.

The project, however, was designed as a research project in its own right, and there are outputs associated with the work which have charted the context in which the digitisation of historical materials can be viewed. The six reports outlined above are therefore also key outputs.

Outcomes

During the course of the project there have been significant changes in the model for mass digitisation, particularly the Google Print Project and the strong showing by commercial services based in India. One of the outcomes, therefore, is the mapping of the why, what and how of mass digitisation processes based in the UK. The key element here is the nature of the content itself. The model developed at Southampton is of particular relevance to material which needs curatorial care and assessment, is semi-unique, and where digitisation cannot be based on disbinding of duplicate copies. Careful assessment of the appropriate handling, scanning technique and indexing has been integral to the workflow, with some material within the same area matched to different approaches. In terms of the production of scanner output only around 60% of the material could be scanned using the DigiLine scanner. Flat bed scanners with enhanced speed were also necessary to complete the work. This confirms the understanding that the Google Project, for example, rejects material not suitable for mass scanning, and is creating critical mass rather than a critical corpus.

This in turn has an effect on workflow. One of the key objectives of the project was to exploit robotic scanning technique, developing skills, metrics and workflow, and to integrate the hardware with repository software for management and output. The project has successfully integrated both the robotic and flat bed scanned output into the workflow provided by the repository software, but the level of development needed to profile the output through the OCR and indexing process was much greater than originally conceived. In part this reflected the necessity of profiling correctly the complexities of the Agora software, but it also reflected the challenge of providing consistent indexing for a range of documentation with different internal structures. In both cases a high level of technical input was required, both from SRZ and local programmers. The automation of the workflow through the repository process was achieved, but significant effort was required to adjust the standard metadata schema to each format, and to ensure that the data collected reflected the required standards.

An aim was also to investigate indexing techniques for mass scanning. This was to be based on two levels. At the higher level a selective approach was taken with the creation of broad level Library of Congress Subject Headings and up to 10,500 detailed catalogue records for key documents as defined by the project advisory group and researched by the project officer, including item level abstracts and LCSH. The amount of material however inevitably limited the extension of this approach to the whole corpus. The aim therefore was to explore techniques for automating indexing.

Three approaches were taken to this. Firstly the structure of the each class of document was translated into XML for incorporation into the metadata managed through the Agora repository. This created certain key elements for each document, and at the highest level, date, session, regnal year, were used as generic indexing key for searching across the whole collection. In order to automate the process of manual input a local software enhancement was introduced to define areas of text relating to days within the Journals so that they could be correctly linked to the correct date, and an enhancement to the Agora software was developed to provide a table to automate the linking of dates to sessions and regnal years. Indexing of the Sessional Papers was manually carried out to define paper type and standard categories of content. The third level of indexing was a pilot use of semantic indexing techniques commissioned from the Natural Language Group at the University of Edinburgh to devise a tool for extracting names of people and places. One of the outcomes of the project

therefore was to highlight the limitation of automating standard indexing, and the potential for semantic indexing.

In parallel with the work on indexing there has been continual attention on the quality of OCR output. One of the observable changes in terms of user perception is the expectation for high quality OCR. This poses significant challenges for the digitisation of material from the 18th Century, where consistent spelling and orthography were yet to be established. The project investigated a range of OCR options in order to increase the accuracy, but the total accuracy remained in the region of 98%. This aspect is carried in more detail in the accompanying report.

Official publications are an essential element in the study of our economic, social and political history, in particular the evolution of public policy and government. They have wide application for both research and teaching, both in the UK and abroad. The use of material from the project will assist researchers working in the 18th Century, and selectively students at HE and FE level developing their understanding of the period through access to original documents. Material rarely seen has been digitised together with the main series of parliamentary papers creating a rich collection of material for researchers to mine.

Conclusions

The project has shown some of the complexities involved in managing large-scale digitisation of historical materials. In two technical areas, the use of the robotic scanner and the implementation of automated repository management software, significant progress was achieved in creating an automated repository able to handle large-scale scanning and data creation. The limitations of this approach relate to the nature of the content itself. In order to ensure an appropriate level of metadata covering both the identity and structure of the document, there has to be an interplay between knowledge of the material, its structure, relationship with other documents in the collection and likely profile for access and retrieval, and the capability of the automated system. This places a high reliance on the quality and appropriateness of the XML.

In terms of indexing and OCR output the perception intrinsic to projects such as the [ECCO Text Creation Partnership] and other projects handling pre-19th Century materials that non-standard orthography and spelling taxes assumptions about what indexing and OCR can provide. As more and more digitised content becomes available for use in a research context, user expectations are growing. The Steering Group has had an interesting debate as to whether the OCR should be displayed to allow researchers to understand what they are retrieving, and why there may be variable output. One important conclusion is that 100% accuracy is very difficult to achieve without a process of quality assured text recreation, an approach which underpins the work done by the Oxford ECCO project or that of the History of Parliament Trust. Nevertheless the work done to create some automated indexing for a mass dataset, and the use of the Gothic OCR tool showed that specialist software applications could have value. The pilot work done on semantic based indexing at Edinburgh also shows that this approach may have value in creating specialist tools; these have however to be linked to priorities for researchers for searching. In terms of indexing therefore one conclusion is that there is no single process which can provide a full retrieval process, and that if a semantic approach is to be made viable, considerable additional research will be necessary.

The issue of sustaining such a collection on open access to the JISC community in the future has absorbed considerable effort as the funding model precludes continuation funding for sustainability. The initial view that the collection could be hosted directly at Southampton had to be re-examined, not least because during the course of the project user expectations of delivery shifted upwards. The model which is now being examined relies upon a licensed approach to ensuring content is hosted on a robust commercial platform, with rights to exploit the content outside the JISC community. The work with licenses has taken us into uncharted territory, and one which was not foreseen at the beginning of the project. We expect the model to deliver the requirements for sustainability, but there are also wider issues to consider. These include long-term responsibility for dark archiving and digital preservation, the potential limitations to re-use implied by licensing content commercially, and the restriction on delivery outside the JISC community. It is important that these issues are addressed when further publicly funded digitisation projects are being planned.

As well as the sustainability of the content there is also the question of the sustainability of knowledge and expertise. At the beginning of the project there was an indication that as part of the digitisation

investment, JISC was creating centres of knowledge and expertise, and in the case of Southampton, the potential for community led mass digitisation. In the event this concept was not reconcilable with the financial and tender processes inherent in JISC funded projects. This raises the issue of the value of community investment in infrastructure which is then not optimised, and in the acquisition of knowledge and skills which could potentially be lost.

It is as yet too early to know whether the collection will achieve the level of interest and use which we hope for it. Certainly it contributes to a growing corpus of material available on-line for researchers to use, for academics to repurpose for their teaching needs, and for students to use directly. We have learned a lot in the course of the project, but ultimately its value can only be proved in terms of the value to the community.

Recommendations (optional)

We have three recommendations for the community.

Further work is needed to determine historical collections which could be suitable for digitisation without being moved outside the UK. This is best done through strategic partnerships to help obtain content from within the sector.

There should be investment in semantic based indexing for mass digitisation in order to develop the potential for more sophisticated retrieval.

JISC Collections should be given a specific role of manage the licensing and business models for ensuring digitised content is made available free at the point of use to the UK community.

Appendices (optional)

- A. Technical report
- B. Completion report 18th Century
- C. Completion report legacy data migration
- D. Financial breakdown
- E. Project final budget
- F. Usability evaluation
- G. Accessibility evaluation
- H. Web interface documentation
- I. Natural Learning Group report

MLBmlb.draft1.28 February 2007