



Project Document Cover Sheet

Project Information			
Project Acronym			
Project Title	Exposing Marandet		
Start Date	1 October 2008	End Date	30 September 2009
Lead Institution	University of Warwick		
Project Director	Robin Green		
Project Manager & contact details	Stuart Hunt, Data Services and Digital Production Manager, University of Warwick Library, Gibbet Hill Road, Coventry CV4 7AL. Tel. 024 7657 5789; Email stuart.hunt@warwick.ac.uk		
Partner Institutions			
Project Web URL	http://go.warwick.ac.uk/riu/marandet		
Programme Name (and number)	<i>JISC Digitisation Programme: Enriching Digital Resources</i>		
Programme Manager	Paola Marchionni		

Document Name			
Document Title	Final Report		
Reporting Period			
Author(s) & project role	Robin Green (Project Director) and Stuart Hunt (Project Manager)		
Date	31 August 2009	Filename	MarandetFinalRptDraft.doc
URL			
Access	<input checked="" type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
1.0	31 August 09	



JISC Final Report

**Exposing Marandet: a project under the JISC Digitisation
Programme – ‘Enriching Digital Resources’ strand**

Final Report (Draft)

Robin Green and Stuart Hunt

31 August 2009

For further information contact:

*Stuart Hunt, Exposing Marandet Project Manager, University of Warwick Library, Gibbet Hill Road,
Coventry CV4 7AL. Tel. 024 7657 5789; Email stuart.hunt@warwick.ac.uk*

Exposing Marandet: Final Report (Draft)

Table of Contents

Section	Page number
1. Executive Summary	2
2. Background	3
3. Aims and Objectives	3
4. Methodology	4
5. Implementation	5
6. Outputs and Results	7
7. Outcomes	7
8. Conclusions	8
9. Implications	9

Acknowledgments

External funding for the Exposing Marandet¹ digitisation project was provided by the JISC under the 'pilot and small-scale digitisation' theme of the *Enriching Digital Resources*² strand of its *Digitisation Programme*.

¹ <http://go.warwick.ac.uk/riu/marandet>

² <http://www.jisc.ac.uk/whatwedo/programmes/digitisation/enrichingdigi.aspx>

Executive Summary

The aims of the Exposing Marandet project were to digitise and make freely available 1500 works, comprising 75,000 pages, from the University of Warwick Library's Marandet Collection of 18th and early 19th century French plays³, and to investigate opportunities to improve connections with CESAR (*calendrier électronique des spectacles sous l'ancien régime et sous la révolution*)⁴, a service created with AHRC funding that provides "a comprehensive on-line repository of French theatre resources in the seventeenth and eighteenth centuries"⁵.

The Library had previously digitised nearly 500 of these plays; when the final works from Exposing Marandet are fully processed c.50% of the Marandet Collection's 4000 plays, completing the period 1700 to 1830, will be openly accessible on the Web, hosted on a dedicated server using OCLC's⁶ CONTENTdm⁷ digital collection management software.

The project adopted a combination of in-house and outsourced approaches. Production of digital images and their associated metadata was carried out by BOPCRIS⁸ (University of Southampton). Creation of full-text page transcript files was carried out in-house, with the transcript files produced in Unicode to enable full representation of French language diacritical marks. Image files and transcript files were loaded together into CONTENTdm and made freely available. A combination of metadata and transcript content was provided within CONTENTdm to enable full-text searching against the digitised material. Descriptive metadata was contributed to the RLUK database⁹, COPAC¹⁰ and OCLC's WorldCat¹¹, and collection availability recorded in the OCLC/DLF Registry of Digital Masters¹². Project staff also identified reviews and reports in the CESAR database relating to plays digitised during the project and created additional metadata to connect the two resources.

The project encountered a number of unexpected delays and other factors which together caused a shortfall in the number of complete plays made available by project end (though the remainder of the content will be added over time). These included technical issues, slower than expected transcription throughput owing to low optical character recognition (OCR) accuracy for older material, and digitisation by BOPCRIS of 85,000 pages in total to deliver the 1500 works - 10,000 pages more than expected from our previous digitisation of the plays.

Despite this, the project has largely achieved its aims in that a substantial corpus of content is now openly available to support scholarly and more general interests. The Marandet Collection largely consists of popular drama and vaudeville and so provides a unique perspective on an important period of French history. In addition, links with the CESAR resource have been augmented, and the Library has gained useful understanding of issues around digitisation, including appropriateness of methodologies, understanding of digital production and preservation techniques, and greater experience of OCR.

Conclusions from the project include:

- When carrying out a large-scale digitisation project it is important to assess whether the techniques or methodologies adopted for small-scale projects can be effectively up-scaled, as small inefficiencies can become significant issues

³ <http://www2.warwick.ac.uk/fac/arts/french/marandet>

⁴⁴ <http://www.cesar.org.uk/cesar2/index.php>

⁵ http://www.cesar.org.uk/cesar2/titles/titles.php?fct=edit&script_UOID=204511

⁶ <http://www.oclc.org/uk/en/default.htm>

⁷ <http://www.contentdm.org/>

⁸ <http://www.southampton.ac.uk/library/bopcris/index.shtml>

⁹ <http://www.rluk.ac.uk/database>

¹⁰ <http://copac.ac.uk/>

¹¹ <http://www.worldcat.org/>

¹² <http://www.oclc.org/digitalregistry/>

- Careful consideration should be given to the various advantages and disadvantages of in-house and outsourced transcription such as cost, quality control, desired throughput rates, type of content, and so on. Exploring others' experience is recommended
- Material with less apparent 'value' is worth considering for digitisation and exposure, especially where it forms a coherent body, as users can find characteristics across a collection that generate new and different interest in both that and related content.

Background

The Marandet Collection of 18th and 19th century French plays is a unique resource and one of the most significant collections of its kind in the country. There are over 4000 plays in total, c.2000 from the period 1700-1830 and a similar number for the period 1830-1900, brought together by a 19th-century French collector, author/playwright and critic, Amedée Marandet.

The significance of the collection lies in its almost ephemeral nature – it is rich in popular drama and vaudeville produced by largely unknown authors rather than major playwrights. Spanning the period before, during and after the French Revolution and Napoleonic Empire, the collection provides a distinctive socio-political perspective on an historical period that continues to capture the popular imagination and is also of scholarly interest.

The University of Warwick Library and Department of French Studies had previously acquired small amounts of funding between 2004 and 2008 to digitise a total of just over 10% of the Collection, primarily to support research and research-led teaching within the Department. The digitised plays had been made openly available over the Internet, hosted on a Library server running OCLC's CONTENTdm digital collection management software, and collaborative work had been carried out with CESAR to embed links from CESAR records to the full text of the plays. CESAR enables search and discovery of 17th and 18th century French theatre resources such as images of scenes and characters from plays, theatre plans, and playbills.

When undertaking this work the Library had become aware that very little of this type of material was currently available in digital format, and had received very positive comments from scholars in this area (e.g., Prof. Jeff Ravel, Associate Professor of History, MIT, described the digital collection as "a marvellous resource").

This interest, together with an unexpectedly high level of use of the material (e.g., some 25,000 hits between January and June 2008 alone, the majority from outside the University), indicated that further work to improve the Collection's visibility would receive support beyond Warwick, and the 'pilot and small-scale digitisation' theme of the JISC Digitisation Programme's Enriching Digital Resources strand was identified as an opportunity to create a resource that would be of significant benefit to the wider community.

Consequently, a proposal was submitted to digitise and make available a major tranche of the plays, covering the pre-Revolutionary and 1800-1830 periods. This would complete all the material from 1703 to the beginning of French Romanticism in 1830, exposing and making freely available nearly 50% of the complete Marandet collection. In addition, further work with CESAR would be undertaken to provide added value to both resources.

The proposal was approved and the JISC awarded a total of £68,996 for the 12 month (October 2008 to September 2009) 'Exposing Marandet' project.

Aims and Objectives

The aims of the Exposing Marandet project were to digitise and make freely available 1500 18th and early 19th century French plays - comprising 75,000 pages - from the Library's Marandet Collection, and to investigate opportunities to provide increased linkages between the CESAR and Marandet services in order to further enrich discovery and accessibility of material in this area. The project

would create images, metadata and full-text transcript files for the digitised content to give full access to the resource.

The target number of plays have been digitised; however, owing to a number of technical issues, the speed at which in-house transcription staff could process older material and a larger number of pages per play than expected, the target for the number of plays made fully available during the project was revised to c.750. The remainder will be loaded to the Collection over time.

Methodology

The project built upon the already established Marandet Plays service previously developed by the University of Warwick Library. The content produced was, and continues to be, added to this service, which runs on a dedicated Library-managed server and uses OCLC's CONTENTdm digital collection management software. CONTENTdm offers digital collection management, discovery and delivery on a platform supporting standards such as Qualified Dublin Core¹³, JPEG2000¹⁴, OAI-PMH¹⁵ and XML¹⁶.

The project was carried out with a combination of in-house and outsourced activities:

- The production of digital images, and their associated metadata, was carried out by BOPCRIS at the University of Southampton. The outsourcing of digital images production offered the most cost effective approach to this aspect of the project and used the expertise of the BOPCRIS team
- Production of full-text page transcript files was carried out by the University of Warwick Library, using temporary project staff with knowledge of French. Transcript files were produced in Unicode to enable the full representation of French language diacritical marks. In the earlier digitisation of Marandet content both transcription and production of the images were outsourced but this methodology was chosen to enable a comparison of approaches.

Once image files and transcript files were created they were loaded together into CONTENTdm and made freely available. These comprised two new period-specific collections in addition to two already populated. These can either be cross-searched or searched in isolation. The collections now available are:

- Ancien Régime Drama
- Revolutionary Drama
- Empire Period Drama
- Restoration Drama.

The Library also investigated the possibility of enhancing existing links between the Marandet plays and the CESAR service at Oxford Brookes University. This involved identifying reviews and reports in CESAR of plays digitised during the Exposing Marandet project and creating additional metadata to connect the two resources.

The project used the following standards:

Name of standard or specification	Version	Notes
TIFF ¹⁷		Preservation image format
JPEG		Web presentation image format
OAI-PMH		

¹³ <http://dublincore.org/>

¹⁴ <http://www.jpeg.org/>

¹⁵ <http://www.openarchives.org/pmh/>

¹⁶ <http://www.w3.org/XML/>

¹⁷ <http://standards.jisc.ac.uk/catalogue/TIFF.phtml>

Unicode ¹⁸		Textual transcript character set
MARC21 ¹⁹		
METS ²⁰		Accompanying digital images, including MIX, Dublin Core and PREMIS metadata as appropriate
Dublin Core		

Implementation

Prior to commencement of the digitisation the Library worked with BOPCRIS to agree the technical and metadata specifications required for the project.

Descriptive metadata for the plays to be digitised was derived from the Library's Millennium Library Management System and appropriate descriptive metadata required for digitisation was supplied by the Library to BOPCRIS together with the plays.

The metadata was also mapped to Dublin Core employing a successful cross-walk already deployed by the Library for previous digitisation projects. Both this and administrative metadata were added to CONTENTdm.

Once material was transported to BOPCRIS digital images were produced in both TIFF (for preservation) and JPEG (for Web delivery) derivative formats. Image files were produced at 300dpi greyscale except those pages that carried colour, which were produced in full colour.

Technical metadata for the digital images was created by BOPCRIS at the point of image generation. These were produced according to the MIX schema. BOPCRIS supplied the Library with METS metadata to accompany the digital images. The METS wrapper included MIX²¹, Dublin Core and PREMIS²² metadata as appropriate.

Once BOPCRIS had supplied the image and metadata files a set of transcript files, in a plain text format, were produced by Warwick project staff using the latest version of Abbyy FineReader²³ OCR software. These were created from the JPEG derivative files and not the TIFF files as the latter were apt to introduce additional transcription errors owing to their high quality and the low quality of the original documents.

The transcript files underwent quality assurance by Warwick project staff. Once image and transcript files were in a completed state they were published and made freely available on the Web with other digitised Marandet content.

Metadata was provided within CONTENTdm at both object and page level. Page level metadata included an additional transcription field to enable full-text searching against the digitised materials. Metadata for the digitised material is available for harvesting via OAI-PMH. Object level descriptive metadata, mapped from OAI-DC²⁴ to MARC21 format, is being contributed to the RLUK database and COPAC as well as OCLC's WorldCat database.

Summary

The project was carried out via a combination of both in-house and outsourced activities. By the start of the project the Library knew exactly which material was to be digitised: this had been identified from

¹⁸ <http://unicode.org/>

¹⁹ <http://www.loc.gov/marc/>

²⁰ <http://www.loc.gov/standards/mets/>

²¹ <http://www.loc.gov/standards/mix/>

²² <http://www.loc.gov/standards/premis/>

²³ <http://www.abbyy.com/>

²⁴ http://standards-catalogue.ukoln.ac.uk/index/OAI_DC

the Library Management System where MARC records were already available. Thus the identification of materials was easily accomplished.

The scanning of materials and the creation of digital images were outsourced to BOPCRIS at the University of Southampton Library. This option was taken owing to the lack of both specialist scanning equipment and suitable trained staff at the University of Warwick. It was a far more economic use of the project funding to outsource to an already established and experienced digitisation service rather than attempt to emulate locally, at considerable cost, an already established and effective service model.

Textual transcript files were also produced as part of the project for all materials digitised. These were created in order to enable full-text searching of the digitised content. The text files were produced in-house within the University of Warwick Library. Once both image and text files were created they were loaded together, as compound objects, into CONTENTdm digital collection management software in order to enable Web-delivery of content. CONTENTdm was already in place in the Library as a tool for the delivery of digital content. It was necessary to increase the licence for the software to an unlimited level, in order to load all the digitised content created. The cost of this was derived from the institutional contribution to the project.

Issues arising during implementation

The planned schedule of delivery to Warwick of image files and their associated metadata via FTP was delayed owing to issues with the transmission of large volumes of high definition image files. This had previously been successfully tested by BOPCRIS in preparation for this phase. However, when the time came to send completed content BOPCRIS encountered transmission failures and only incomplete content reached Warwick. This necessitated investigation and testing by both parties, with the eventual adoption by BOPCRIS of CoreFTP²⁵ software, which resolved the issue.

This delay had a knock-on effect: recruitment of the Transcript Assistants was deferred from December 2008 until March 2009 as insufficient content had been received from BOPCRIS for them to work on. The Assistants did not become fully-functioning until at least mid-April. This delayed the first public availability of new content and impacted on the final number of plays the project was able to fully process.

Despite pre-project testing, creating textual transcript files with OCR proved to be a slower process than initially calculated. In particular, older French text materials with non-current characters and poor quality paper affected progress due to the increased need for human intervention. This was compounded to an extent by 85,000 pages being digitised to complete the 1500 plays, i.e., more pages for the target number of plays than the Library's previous work with other plays in the Collection had indicated would be the case.

A further technical issue arose when an upgrade to the CONTENTdm software, installed in January 2009, caused problems across its user community and significantly affected the project's workflows, e.g., for one period of 2 weeks no loading of any content was possible.

These factors together reduced the number of plays the team was able to complete during the project lifetime.

Finally, in project planning it was anticipated that the Transcript Assistant posts would be evaluated on Grade 2; however, the University's Job Evaluation Panel felt that knowledge of French - i.e., a specific skillset or qualification - would be required for the work and assigned Grade 3. This could have had implications for the budget if they had not been recruited later than expected.

²⁵ <http://www.coreftp.com/>

Outputs and Results

The end result of the project has been to make openly and freely available a coherent set of content within a single site that provides a unique perspective on a seminal period of French history. That perspective extends beyond France itself to French relationships with other countries. The type of material included in the Collection – primarily popular drama and vaudeville - and the relative obscurity of the authors means it presents the contextually important view of 'the man on the street' rather than political commentators.

The project deliverables are:

- Digitisation of 1500 currently hidden plays comprising 85,000 pages providing a scale enhancement to an already existing openly accessible and already heavily-used digital resource. Around 750 plays will be made available during the project; the remainder will be added over time. An additional benefit is that although the material itself is largely printed in greyscale, where pages carried colour BOPCRIS produced these in full colour
- Preservation quality image files, and associated preservation metadata, for the digitised plays
- High quality metadata being progressively made available in OCLC WorldCat, RLUK database and COPAC and exposed for OAI-PMH harvesting
- Collection availability recorded in the OCLC/DLF Registry of Digital Masters
- Improved linkages/interoperability with CESAR
- Project reports evaluating the chosen methodology and outcomes
- Dissemination to library and academic communities, e.g., through conference presentations.

The outputs of the work will extend beyond project end as owing to technical issues, the difficulty of the content for accurate OCR, and the digitisation of 85,000 pages against 75,000 scoped – more than 13% above initial estimates - only c.50% of the proposed 1500 plays have been fully processed and made available. However, the remainder will be processed over time by the Library as part of its ongoing stewardship and development of the Collection

Outcomes

The project has been broadly successful against its aims and objectives. Although there has been a shortfall in the number of plays made available by project end against the target, a substantial body of new digital material representing a type of content that has previously had very little digital representation is now easily and openly accessible. This is accompanied by exposure of metadata in a range of services providing access to the material.

The most significant outcome of the project is that what was effectively a small niche resource has become a substantial service spanning teaching, research and general interest, one that may encourage other institutions around the world with related collections to convert and expose their content should opportunities arise for collaborative work.

Within Warwick, the newly available material will enable the Department of French Studies to support a plan to submit a proposal in the 2009/10 session for research funding to study theatre from 1800-1815, and to continue and extend its innovative approach to research-led teaching. Other institutions, such as Nottingham, Oxford Brookes (home of CESAR), Goldsmiths and MIT offer courses that could also draw on the new content.

The added linkages with CESAR will bring increased use of that resource, together with increased referrals from CESAR to the Marandet collection for material up to 1800 (CESAR's end-date). This is of particular benefit as CESAR is not funded to enable development and this work will attract new interest in the resource.

The Exposing Marandet project proposal also identified the following key stakeholders and desired outcomes arising from the project:

Stakeholder	Outcome
JISC	Project accountability for commitments and funding
University of Warwick	Project completed successfully maximising benefit from institutional financial and other contributions
University Library	Maintain trusted project leader/partner reputation
Warwick lecturers and researchers	High quality difficult-to-access content easily available to inspire teaching and research
CESAR	Improving its service
Wider academic and general community	Access to new high value content
Other digitisation projects	Project learning relevant to own activities

It is felt that these outcomes have been achieved.

In addition a number of less tangible outcomes have been achieved:

- Experience of managing the Library's largest digitisation project to date, specifically in up-scaling methodologies adopted for earlier small-scale projects
- Increased knowledge of digital production techniques and developing specifications for contractors
- Increased awareness of digital preservation issues
- Greater experience in the use of OCR both technically and in the levels of human intervention required for textual materials.

Conclusions

The external interest generated from initial internally-funded digitisation of around only 10% of the Marandet material demonstrates that making available a critical mass of a less obviously 'valuable' collection can attract new audiences. For example, there may be characteristics across a collection that stimulate different interests and connections from those relating to individual works.

It may therefore be worthwhile to digitise and expose appropriately-sized samples of different types of content and gather information about resulting usage, one benefit of which would be to support approaches to potential funders.

When carrying out a large-scale digitisation project it is important to ascertain that the techniques or methodologies adopted for small-scale projects can be effectively up-scaled. Any manual processes or interventions will be magnified by a greater volume of material. This is seen in the case of producing transcript files for 85,000 images by OCR which is a time-consuming process. Such techniques or methodologies need to be integrated into a standardised workflow to ensure efficient and timely production.

It is possible to undertake transcription work in-house. There can be significant cost savings, together with more control over quality assurance as processes can be managed more immediately and smoothly to suit needs. There are disadvantages, particularly where temporary staff are used, with overheads in management time from recruitment to oversight, and dealing with variations in throughput and quality. Any such issues are likely to increase with the age of the material, for which extra time should be factored in.

This has to be assessed against the initial efforts of building understanding with an external service provider, and the requirement, especially for larger projects, to work to their schedules and time flows

– this could be an issue for shorter-term projects. If multiple projects are planned, or older material is being digitised, outsourcing with a satisfactory agent is likely to be a more optimal approach unless a permanent/skilled in-house team is in place.

Higher Education institutions planning digitisation activity should actively explore the existing high quality service available within the sector, which offers expertise as well as interest in collaboration and partnership, as an alternative to the high overheads of acquiring in-house equipment and technical expertise.

Implications

A number of institutions in the UK, France and beyond hold content related to the material in the Marandet and CESAR services. Should funding become available there is scope to build on the interest in the Marandet Collection and work with partners to create a major collaborative digital resource with the Marandet material at the centre and related resources (including CESAR) around it.

There is also considerable potential for work to be done to further develop linkages between the Marandet collection and CESAR.