

JISC CAPITAL PROGRAMME

Project Document Cover Sheet

FINAL REPORT

Project

Project Acronym	Welsh Journals Online	Project ID	
Project Title	Welsh Journals Online / Cylchgronau Cymru		
Start Date	March 2007	End Date	February 2009
Lead Institution	National Library of Wales		
Project Director	Lyn Lewis Dafis		
Project Manager & contact details	Martin Locock 01970 632885 martin.locock@llgc.org.uk		
Partner Institutions	n/a		
Project Web URL	welshjournals.llgc.org.uk cylchgronau.cymru.llgc.org.uk		
Programme Name (and number)	JISC Digitisation - Phase Two		
Programme Manager	Paola Marchionni		

Document

Document Title	<i>Final Report</i>		
Reporting Period	<i>Feb 2007-Feb 2009</i>		
Author(s) & project role	Martin Locock, Project Manger		
Date	27/3/09	Filename	final report3.0
URL			
Access	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

Document History

Version	Date	Comments
1.0	2/3/09	Draft approved for release to JISC
2.0	20.3.09	Revised following comments on draft
3.0	27.3.09	Final version after comments by JISC

Project Acronym: Welsh Journals Online
Version: 3.0
Contact: Martin Locock
Date: 27 03 2009

JISC

JISC Final Report



Cylchgronau Cymru
Welsh Journals Online

Final Report

Martin Locock

Contact: Martin Locock

March 2009

Table of Contents

Acknowledgements.....	4
Executive Summary.....	5
Background.....	6
Aims and Objectives.....	6
Methodology.....	6
User Engagement.....	11
Implementation.....	11
Outputs and Results.....	14
Outcomes.....	14
Conclusions.....	15
Implications.....	16
Appendixes.....	18

Acknowledgements

The Welsh Journals Online project was funded by JISC's Digitisation Programme, the National Library of Wales, and the Welsh Assembly Government. The project was supported by the members of WHELF (Welsh Higher Education Libraries Forum) and by the Welsh Books Council.

The National Library of Wales is grateful to the many individuals and organisations who contributed to the success of the project, including the Project Advisory Board members, Kirsti Bohata, Jasmine Donahaye, Peter Keelan, Paul O'Leary, M Wynn Thomas, and John Wright, JISC staff, particularly Paola Marchionni, Liam Earney, Alistair Dunning, and Stuart Dempster, and the publishers and authors.

It is also grateful to Jouve, CASIS, Peter Gill Associates and The ITC for their services.

It is grateful to the organisers of the following conferences for the opportunity to promote the project: the Institute for Archaeologists, CILIP Cymru, University of Aberystwyth Post-graduate History Seminar, IGLP, LDAP, NAASWCH, and WHELF.

Executive Summary

Welsh Journals Online is the most challenging digitisation project ever undertaken by the National Library of Wales. It aimed to create a website giving free searchable and browsable access to the contents of back-numbers of the major journals relating to Wales or the Welsh language. These journals form the core of the Library's collection of printed books and are its most-used resource.

The journals were chosen to represent the diversity of material available, and cover English- and Welsh-language titles including scholarly articles on topics from archaeology to zoology, poetry, fiction, reviews and obituaries. The project publishes 400,000 pages of text, from 52 titles; the 180,000 pages of Welsh content represents the single largest corpus of text in the language available on the web. Some of the titles are well-known and widely used as sources (eg *Archaeologia Cambrensis*), while others have been overlooked or are difficult to access (*Yr Arloeswr*).

The digitisation of the material required intensive work on cataloguing, scanning, and OCR conversion in order to create a resource that can be easily explored and used with a range of technologies, including text readers and mobile phones. Although the OCR text was intended mainly to allow word-searching with highlighting, it has also been exposed to users who may wish to use it. It is hoped that images, texts, or pdfs from the articles become widely used as exemplars and source material in VLEs, websites, presentations and reports, exploiting the generous licensing terms on which the material is made available.

The resource has been actively promoted to the HE sector's teaching and library staff, since they will be critical in directing students and researchers towards it. A series of roadshows to HE bodies throughout Wales was supplemented by mailings to universities elsewhere teaching Welsh or Welsh Studies, and presentations at conferences.

The website is fully exposed to Google and it is likely that many new users will find the resource through general searching of the web. For those who are unfamiliar with the journal literature of Wales some contextual help is provided in the form of factsheets; lesson plans based upon these have also been created to assist teachers wishing to use the Welsh Journals Online website to discuss the questions of copyright, searching, or referencing.

The majority of the material is covered by copyright, and licensing and rights management formed a significant part of the project. The need to control display at page level (so that where necessary a single article or photograph could be blanked) required detailed metadata to record permission, gathered in cooperation with the publishers. Of the titles included, the proportion of blanked pages is very low (less than 0.1%), but rights issues led to the exclusion of some titles completely. The Library did not offer any payment for permission and works by Dylan Thomas, Robert Graves, and R S Thomas are therefore not shown. Given that the cost-per-page of web publication is approximately £2, the payment of even minimal fees would transform the economics of mass-digitisation.

The digitisation of the journals is the first step towards the Library's long-term vision of mounting on the web its entire print holdings relating to Wales and the Welsh people¹, and it will be maintained and added to in the future.

¹ NLW *Shaping the Future: the Library's strategy, 2008-2009 to 2010-2011*, p. 15.

Background

The National Library of Wales was established in 1907 to 'collect, preserve and give access to all forms of recorded knowledge about Wales and the Welsh-speaking peoples'. One of the primary duties of the Library is to develop its collection of periodical literature relating to Wales, which is the most comprehensive in the world. The Library is pursuing a long-term strategy of broadening access to its collection by digitisation, although it had concentrated in the past on manuscript and archival material. The Welsh Journals Online project was the Library's first mass-digitisation of printed material, intended to bring its most-used material to a new audience by exploiting the opportunity of Web technology; it would also form a corpus of reliable and wide-ranging Welsh-language and Wales-related material.

Aims and Objectives

The project identified the following aims and objectives in its first Project Plan:

Aim:

To provide remote access to the contents of the main journals relating to Wales

Specific objectives at outset:

- To digitise approximately 90 journal titles comprising 600,000 pages of text
- To secure clearance from publishers and copyright holders to allow public access to the material and to obscure articles and images for which rights are withheld
- To provide browse, word-search, article title and author search access to the content
- To allow page-scan view and, where agreed, TEI text view
- To provide subject guides and advice to users
- To promote use of the content in learning objects
- To allow users to contribute content

During the course of the project, the objectives were amended in the light of the scale and nature of the work and sustainability.

- The scale was reduced to approximately 50 titles and 400,000 pages because it was found that processing of the content was much more labour-intensive than had been estimated
- Access to TEI was provided for all showable content
- The facility to allow users to contribute content has not developed in view of the resource implications for moderation and support

Methodology

The National Library of Wales recruited and seconded a team of 12 staff to undertake the metadata creation, scanning, QA and web development required for the project, assisted by substantial technical support and advice from other Library staff. The principal out-sourced activities were design of the data creation software and workflow tools, because no staff were available at the time, and OCR processing to create TEI output, which was an area in which the Library had little previous experience. Design work was also out-sourced.

Project management

The project was managed by a specially-constituted Project Management Board with representatives of key Library sections involved in the work of the project, with two advisory committees, a Project

Advisory Board and a Technical Panel. This structure directed, supported and advised the core project team and managed the interface between project activities and other Library operations.

Rights management

The Library was aware that one of the principal challenges for the project would be the handling of copyright issues. Its policy is to treat all material published post-1900 as potentially covered by copyright.²

Therefore the project needed to undertake work in order to:

- Define who held rights over the material
- Devise ways to contact rights holders for permission
- Record the responses for future reference
- Implement the permissions in the web content

The Library decided to treat each title as a group of content, licensed by the publisher (using a version of the JISC Model Licence³ for Third-Party Material). The publisher was asked to consult with rights holders and to licence all material for re-publication except where rights holders refused. It had been assumed that publishers would hold records of rights holders names and addresses and copies of publishing agreements setting out which rights had been transferred to the publisher. It was found that this assumption was incorrect, and many publishers held little formal documentation concerning rights or rights holders. As a result it was impossible to seek copyright clearance for all material, and instead a risk management approach was adopted, where attempts to contact individual rights holders were backed up by a general licence for the publication as a whole and a takedown procedure for removing material should the rights holder come forward after its release on the website.

The journals contain much content which is potentially covered by rights but whose creator is unrecorded (including graphic design elements, editorial notices, contents pages, meeting summaries and advertisements); if this material were to be blanked in the absence of explicit consent, then the digital version of the publication would cease to reflect the nature, scope and layout of the printed edition.

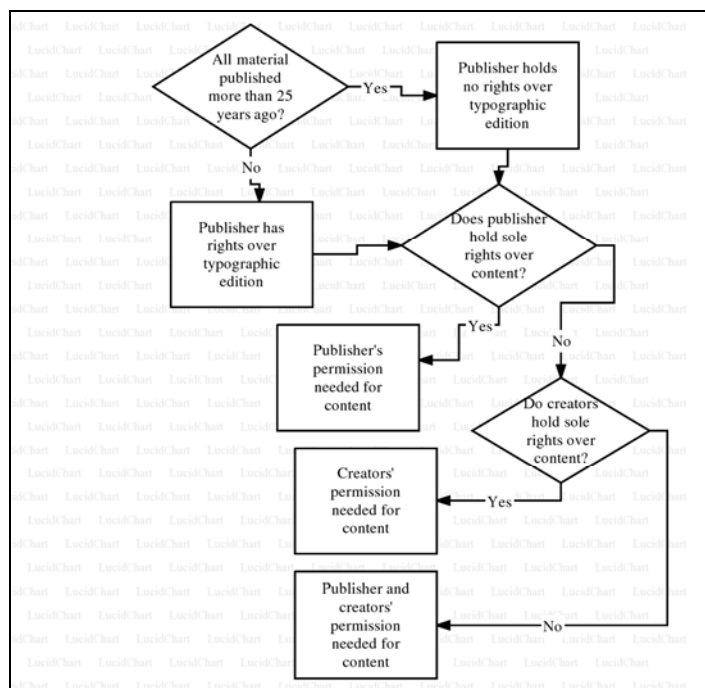
The steps taken by the publishers to contact their rights holders varied, but included:

- discussion of proposed inclusion at AGM or special meeting of members
- publication of an announcement in newsletters or in the journal
- mailing of permission forms to creators whose contact details were known

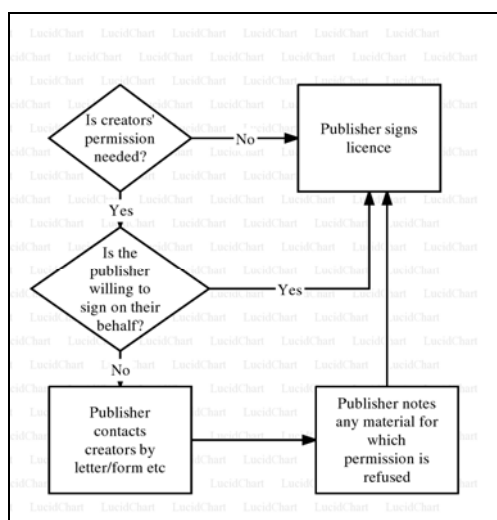
The project team assisted some publishers by undertaking the work of sending out a mailshot. The headed paper of the publisher was used and it was made clear that it was them seeking permission, rather than the Library. Where publishers provided personal data to the Library for this purpose, Data Protection forms were completed to record the nature and purpose of the transfer.

² In practice, some pre-1900 material may be in copyright and some later material may be out of copyright.

³ http://www.jisc-collections.ac.uk/model_licence.aspx



Decision tree to determine rights over typographic edition and content



Decision tree to determine whether creators are contacted before licence is signed

The rights information gathered by the project will need to be retained in the long term to provide documentation in the event of any dispute, and a Document Register was created to identify each item. The reference id of a document is cited in the rights metadata of the digital image.

Although the metadata schema used (METS/MODS) allows the definition of rights to digitise and publish, it does not directly enforce those rights. Because the proportion of material for which permission was refused was so low (between 0.1% and 0.05%), creation of blanked versions of the scans for display was handled manually.

The project found that the rights issues required a large amount of labour and administration, mainly because rights holders and publishers were unfamiliar with the documentation and legal background. Publishers were also concerned about their liability when licensing material over which they did not

hold rights, and the Library found it necessary to amend the standard licence to address these concerns. Only a small number of publishers declined to be involved with the project, the most significant being modern literary titles whose rights holders were unwilling to grant permission for their material to be republished without payment. In addition to the licensing of past content, agreements were reached with 14 publishers for the addition of further issues as they emerged from embargo (a 'moving wall').

Metadata creation

The biggest challenge of the project was the creation of bibliographic and structural metadata defining the scanned pages in terms of their intellectual content and physical arrangement. The level of granularity necessary for handling rights meant that this was a vast task. The project re-used and enhanced existing MARC21 records at title and article level held within the main Library catalogue, but much had to be created from scratch since previous cataloguing had excluded some types of content (eg front matter, reviews and fiction) and generally covered material published after 1980. The main catalogue structure could not be customised to provide the desired level of functionality, and therefore a separate database was created (Fflam) which held extracted and new MARC21 records and additional fields to cross-link the bibliographic data to individual page scans. The Library's Digitisation Workflow tool was adapted to track the project's work and also to allow management of ordering at page level.⁴ The two datasets were then brought together to create the METS record for each title.

With so many records to create (50,000 article records, 400,000 pages), the time taken on each was a critical determinant of the project's progress towards its target. As a result it proved necessary to simplify these records over time. The main Library catalogue follows AACR2 data standards, including authority-controlled author entries.⁵ It was found that creating new records in Fflam, without authority control, was significantly faster, although it was recognised that this approach would preclude some methods of searching.

The project has ended with a compromise between the creation of clean structured data to allow reliable searching (for example, all article titles are input manually and should be highly accurate) and the expectation that users will invest time in reviewing less directed search results.

The metadata was checked through a combination of manual QC and automated formatting reports.

Scanning and OCR

The Library used two dedicated Zeutschel 5000TT Bookscanners to undertake the scanning work inhouse. These proved reliable and effective, and operators achieved rates in excess of 2,000 pages per day. The main problems encountered were:

- Binding too tight to allow volume to be fully opened
- Metadata errors (miscalculated unnumbered pages)
- Large fold-outs

The TIFF images were stored locally until QC'd and sent to archive; derivatives were then created for display.

⁴ Metadata was created in advance of scanning, so that filenames were assigned in advance. This allowed QA work on sequencing to be undertaken simply through checking the expected content of an arbitrary scan with its image (eg, scan AWJAH0020023 should show page label 55- if it does not, an error has occurred). The scanning process was managed using the physical units (a bound volume and pages in bound order), but for display purposes a logical order was created (published issues and articles). Most of the Library's journal holdings had been bound up into multi-issue volumes with the covers removed and bound at the end of the volume, but it was felt that this arrangement would be confusing to readers unfamiliar with the Library's collections.

⁵ Use of a standard form of an author's name so that all their works can be readily extracted.

OCR of the scans and the creation of TEI was undertaken by Jouve. A minimum acceptable accuracy of 99% of characters was defined (manual re-work was triggered if the automatic output fell below this level); in practice 99.3% + was achieved. The contractor was specifically required to identify diacritics unique to the Welsh language (\hat{w} , \hat{y}); this required adaptation of the standard OCR engines. In addition, the word accuracy was measured by comparing the OCR output against English and Welsh lexicons. This proved to run at 75% words recognised. The contractor also provided coordinates of each character and image on a page (Alto xml), used to control blanking and highlighting of search terms.

Presentation

The project's website presents a range of page views (optimised for standard and non-standard screen widths, Zoomify zoom for Flash-enabled computers, and OCR text), conceptualised as different views of the same page. All the pages of an article are shown (by forward arrows) before the next article begins.⁶ The pages are held hierarchically within article within issues for a publication, and are accessible from browse or searching. Links from search results highlight the occurrences of a search term on the page. A pdf version of each article is available; in addition to the scanned pages, there is a front page giving bibliographic source and a summary of permitted uses. Material for which permission has been refused is shown blanked-out in all views.

Lying behind the website is the VTL Vital DAMS (based on Fedora) which holds the METS and presentation image and text files. The Solr search engine provides the search functionality, allowing Google-type syntax and Boolean searching. The whole-text search is supplemented by specific searching within article titles, authors, and article types, providing more elegant solutions to searching a large dataset that may return thousands of matches for common search terms.

The concept design for the website was commissioned from Peter Gill Associates; the web site and other systems were developed inhouse by the Library. The hard elements of the website are bilingual and can be toggled at any point.

The website was initially intended to include Web2.0 features such as user commenting, a forum, and a personal notes area. When it came to implementation, it became clear that the legal and administrative implications of mounting user generated content on the Library's website outweighed the likely level of useful data, and this element was dropped. The principal issues were:

- Responsibility for policing nature of content
- Potential liability for libel
- Likelihood that HE users would have custom desktops within VLE
- High level of spam and unwanted content
- Low level of traffic at page level
- Absence of a definable 'user community' wishing to share comments

Archiving and sustainability

The creation of 2 Terabytes of data and 2 million files represents a significant administrative load for future management in terms of digital preservation, migration and back-up. The lifetime of the project has coincided with major developments in the Library's approach to digital preservation arising from its recognition that curation of digital assets has become one of its core responsibilities. The Digital Preservation Plan for the project has been developed using the OAIS Reference Model and will inform future operational plans and ensure that all necessary resources are in place.

The project has also put into place handover training so that the Library's commitments to maintaining and adding content can be fulfilled.

⁶ Thus, if an article has been split physically, so that after page 55 it continues on page 99, page 99 is shown as 'next' page to 55.

User Engagement

The Library's approach to user engagement was based on the identification of different target groups:

- Teachers and ICT staff in HE and FE
- Researchers and students in HE and FE
- Casual users

These groups have different needs from website design and content, and the key challenge was to meet those of the core users (providing stable urls, allowing embedding and re-use in learning objects) while still providing a usable service for casual visitors, for whom the website may be their first encounter with scholarly material.

The main channel for user engagement through the lifetime of the project was the Advisory Board (academics and HE library staff) and the members of WHELP (Wales Higher Education Libraries Forum). In the way the HE community was able to provide guidance on priorities for inclusion, relevance to areas of study, support materials and desired functionality. The importance of developing the resource as complementary to and integrated with those provided by HE libraries emerged strongly from these discussions.

The programme of public events (conference presentations, roadshows and workshops) allowed direct feedback from potential research users and led to the recruitment of a user panel to test the website as it developed. In the context of their work, the value of making pdfs readily available was clear, as was the ability to limit searches by genre.

It had been intended that there would be an extended period of usability testing and web development following launch of the website, but this proved impossible due to delays in the launch. The collection of user feedback, monitoring of user behaviour, and the further development of the website, have therefore been written into the future work of the Library.

The Library is confident that casual users will find their way to the resource from the web; its formal marketing has therefore focused on universities and library services through leaflets. In recognition of the paths that users take to reach web resources, search engine friendly metadata has been included for each page (leading to good visibility in Google), in addition to explicit linkage from Google Scholar; the resource and its contents have also been promoted by the project's enhancement of Wikipedia coverage of Welsh journals and the Library, providing another way for users to reach it.

Implementation

From the start of the project, it was recognised that there were several strands of development and implementation that would all need to be completed before the website was made live:

- Licensing agreements and permissions
- Selection of material
- Development of data structure
- Development of data standards
- Creation of data entry systems
- Metadata transfer, creation and checking
- Scanning and checking
- OCR specification and processing
- METS creation and ingest
- DAMS development
- Website design
- Usability testing
- Marketing and publicity

The project timescale proved problematic in two ways: the metadata creation work proved to be much more labour-intensive than had been envisaged, limiting the number of titles that could be included; and the dependencies within the development process led to unavoidable delays to work on ingest and presentation, with knock-on effects on usability testing and marketing.⁷

The approach towards licensing (using the publishers as the point of contact) had been decided when the project proposal was being prepared. The Library was keen to obtain permission that would allow re-purposing by users, and none of the publishers had expressed a strong preference for control of downloading. The main task in seeking licensing was to locate the relevant body or individual (which proved difficult for some of the earlier publications) and to complete the legal agreement. Only a small numbers of publishers refused permission; the most common reason for doing so was their desire to pursue digitisation of their archives themselves. The legal agreement included warranties from the publisher about the accuracy of the information they supplied, and many publishers sought clarification on the extent to which they would be liable; the Library re-drafted the agreement so that the risk was more equitably shared and assisted the publishers in contacting their rights holders. Because the Library would be handling many licences, it restricted the range of possible terms, so that publishers were given the options of allowing non-commercial re-use or unrestricted re-use, and allowing creation of derivatives or not. This simplified the presentation of the material on the web, since there were only four possible rights statements for material shown (plus 'don't show'). The Library made it clear that it preferred the rights to be as open as possible, and most publishers agreed on these terms.⁸

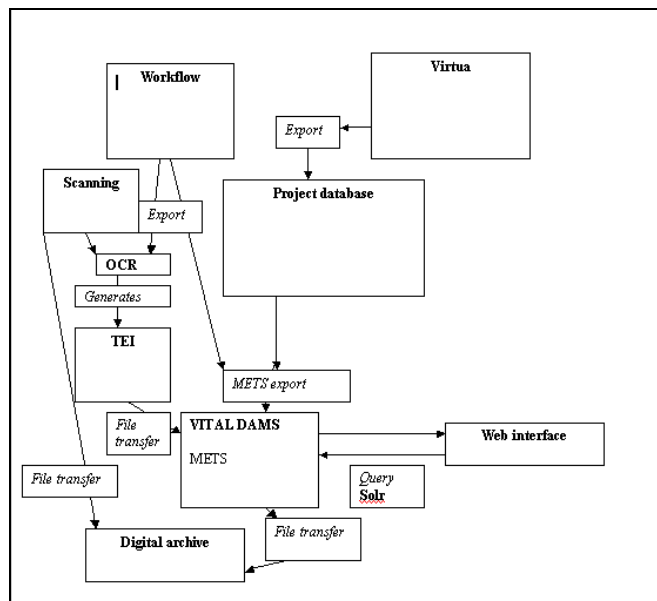
Selection of the titles was based on the priority list drawn up by the Bibliography of Wales unit at the Library, and was then reviewed by the Advisory Board. It had been thought that the list would have to be revised extensively to accommodate refusals and inability to locate contacts, but this didn't create much change; more significant was the need to reduce the number of titles because of the scope of the work. The 'long list' of 90 titles approached was reduced to 52, mainly on the grounds of progress with licensing. The coverage by language and subject remained representative of Welsh journal publishing as a whole, dominated by historical, literary and general interest titles, with some scientific and political material. One area of regret was the limited coverage of modern political magazines (eg CND Cymru's *Heddwch*) and small-press literary journals; their irregular publication history and complex formats made their inclusion more difficult than established textually conservative titles.

The data structure required to create, expose and manage the scanned content was devised to combine conventional MARC21-type bibliographic descriptive metadata with technical and structural metadata (merged into METS data for presentation). The recording of multiple variant page ordering was considered important in order to allow presentation in a logical 'single article' sequence. Underlying all the metadata was a parallel system of arbitrary record ids and structured record codes reflecting the hierarchical relationships of entities. All data elements included a Notes field for use during the data creation stage.

In addition to the data structure, data content standards were developed giving examples and guidance on how the data should be entered in order to promote consistency across the data set. In many cases these were more prescriptive than the MARC21 and AACR2 standards used in the main Library catalogue.

⁷ For example, appointment of the OCR contractor followed OJEU procurement procedure and an initial contract notice in August 2007 led to appointment in January 2008 and first delivery of data in August 2008; development work on ingest could not start in earnest until that time.

⁸ The Library hopes to avoid the need to re-licence content in the future for uses not envisaged at the time.



System map showing data flows

The creation of metadata required multiple stages of the workflow and the interaction of numerous systems; one implication of this process was that it was impossible to assess the full historic work rate from page to website until late in the project, which made programming and target-setting difficult.

Scanning was undertaken after the page sequence had been defined. Two scanning operators worked the machines, and all output was checked by a QA officer. It was found that the scanning could be carried out at a faster rate than metadata could be created until late in the project; balance was achieved when there were 3 metadata staff for each scanner. Because of this imbalance, the scanners were underemployed in the early part of the project.

Development of the system to hold the data and present it on the web was an iterative process in which the functional specification was agreed by discussion of desired performance and technical implications.

Outputs and Results

The key project output has been the website <http://welshjournals.llgc.org.uk> hosted as a subdomain of the main Library website but delivered separately. By the end of the project's work in June 2009, there will be 52 titles and more than 400,000 pages available (see Appendix One). This represents the single largest web corpus of scholarly material relating to Wales, fully accessible and licensed for re-use.

The website support text searching (of all text, main article text, author, and article title), advanced search (by type of article) and browsing.

The website is accompanied by contextual data including help and background information, accessibility tips, contact and takedown details, and resources for students and teachers.

Outcomes

Aim:

To provide remote access to the contents of the main journals relating to Wales

Outcome:

Achieved

<i>Objectives</i>	<i>Outcome</i>
To digitise approximately 50 journal titles comprising 400,000 pages of text	Achieved (by June 2009)
To secure clearance from publishers and copyright holders to allow public access to the material and to obscure articles and images for which rights are withheld	Achieved
To provide browse, word-search, article title and author search access to the content	Achieved
To allow page-scan view and, where agreed, TEI text view	Achieved
To provide subject guides and advice to users	Achieved
To promote use of the content in learning objects	Achieved

The project has demonstrated that the legal and technical challenges posed by the mass digitisation of 20th century Welsh material can be dealt with successfully.

The project's adoption of a risk-management approach to rights has allowed it to present the overwhelming majority of the desired material without excessive labour and cost, and the Library's commitment to ensuring takedown where required in future has allowed it to proceed largely on the basis of the publishers' licence.

The project has devised a very specific and granular system for the analysis and identification of content, driven primarily by the need to manage rights at a sub-page level and to a lesser extent by the desire to present readers with pages in a more logical sequence than the current physical arrangement. For the works covered, this may have been over-engineered; simpler and faster solutions could have been devised, had it been clear from the outset that blanking for rights would have been so rare and how much time re-ordering would consume. But a similar approach might well be appropriate for dealing with literary and illustrative material covered by commercial rights where the proportion of problematic rights holders was much higher.

The decision to combine human-created accurate catalogue content for the article titles with the machine-derived OCR proved to be a successful way of providing reliable bibliographic descriptions, vital to a research resource, with a functional search option to locate relevant material.

It is clear from user comments that the resource will become a major tool for the HE sector in Wales and beyond and will provide a valuable service to academics, students and casual users.

The ready re-use of the material in learning objects provides an opportunity for teachers to explore innovative teaching practices and to embed the principles of scholastic endeavour (resource discovery, citation and referencing) into students' behaviour.

The project has also strengthened the relationship between the Library and the publishing industry in Wales, and has raised the general level of understanding of the legal issues arising from publication.

Conclusions

The main conclusions drawn by the Library from the project are that:

- Goodwill and support of the copyright holders are essential

Unlike a conventional rights clearance process, with good documentation and payment offered, the key to obtaining permission was in explaining the purpose. Rights holders were happy to sign up once they understood that the resource would open to all without payment or registration. The principal advocates promoting the project to individual authors were the publishers themselves, who considered that the digitisation proposal formed a logical extension of their activities in physical print. Much of this support was derived from the wider role of the National Library of Wales as a respected and trusted institution. Fundamental to the relationship was clarity about what the Library intended to do and its willingness to accommodate the needs and concerns of other parties.

- Licensing requires time and labour even if the parties are willing

Although the proposed project met with almost universal support upon initial approach, the process of completing formal licensing was extremely drawn out in terms of time and arduous in terms of work. This was particularly true of the publishers, who often required four or more committee meetings to move from agreement in principle to signing a licence; the willingness of the Library to send a representative to attend such meetings and discuss the issue in an open and honest way was critical. Similarly, the preparation of a layperson's commentary to the formal agreement was found to be necessary to allay concerns about the nature of the legal rights being sought. A negative corollary to the time taken to resolve permissions was that practical digitisation work had to be commenced in advance of licensing, with the risk that some of this work might prove fruitless. In the event, this only affected a small number of titles, but the risk could not be minimised further. Because most of the publishers had no paid staff and changing responsible officers, it was vital for the project to maintain good and accurate records and to send timely reminders.

- Investment in digitisation is best repaid if re-use is enabled and promoted

The Library's intention is that the work of digitisation of the journals should not need to be repeated by itself or others in the foreseeable future, and it has therefore endeavoured to support the exploitation of the resource by current and new technologies. Alongside this technical openness (open web content, exposure of the OCR-derived text, multiple page-view formats, stable url handles for bookmarking to deep content) has been the desire to permit as broad use as possible through seeking generous licensing terms from rights holders. Although some publishers were concerned about the loss of control that resulted, it is clear that this approach offers the best prospect of allowing the resource to become embedded in future teaching and research practice. The Library would have

been reluctant to invest in undertaking digitisation of material under terms that precluded its imaginative and creative re-use.

- Cross-searching of large data sets requires faceted or structured searching, or there will be too many hits

The general principle of accumulating unified data sets which can be searched as a single operation is sound in theory, but in practice may have unfortunate effects. Although locating every appearance of a search term anywhere in the text may seem at first glance to be helpful to users, a much better service is provided by searching to locate significant content through prioritisation of some material because of its nature. The two challenges are to encode information scent into the data (for example, by identifying article titles as a special high-value group) and to encourage users to exploit the power of non-standard searching by limiting searches to 'article titles only' or 'within a single publication'. It is to be hoped that users will over time become more sophisticated in their approach to resource discovery, but at this point the website design needs to actively highlight alternatives.

- Scanning and OCR of modern texts can be highly accurate

Much discussion about the practice and limits of digitisation has focused on the difficulties of high-end custom scanning of unique material for preservation purposes. In that context, the digital surrogate must replicate as well as possible the physical original for all purposes. Mass-digitisation of modern printed text has a different philosophy and can concentrate on the delivery of readable and searchable text suited to the needs of 90% of users. By optimising the scanning for high-contrast greyscale images and de-skewing, it is possible to generate highly accurate OCR text without manual re-work. Non-standard diacritics can be successfully handled by adaptation of the OCR engine, although performance is variable.⁹ If (as here) the principle function of the OCR is to expose the text to searching, a level of inaccuracy can be accepted as a reasonable compromise.

- Metadata creation is the largest and most complex task

Both in terms of the digitisation process and the exposure of the content, the creation of consistent accurate and structured metadata is fundamental to success. From the initial definition of the issue numeration and pagination, through the hierarchical description of content, to the management of scanned pages for delivery to users, the data from each stage has to be valid. In addition to the work involved in creating such data, there is the additional work of human QA and automated validation to reduce the number of errors. The creation of data structures and standards and the training of the team in their application is therefore crucial. The management of metadata creation lies at the core of a successful project, since it determines the amount of material that can be included and the ease with which it can be explored by users.

Implications

The Library's experience of the project has clarified its approach to its general aim of digitising its printed holdings. The key role of the rights regime in determining the necessary data structure, and the resources that need to be committed to managing permissions during digitisation and into the future will be taken into consideration when setting priorities for material to be covered. The availability and accuracy of existing descriptive metadata for the material is another key factor: where electronic records do not exist, they will need to be created before digitisation can commence.

⁹ The circumflexed w and y have often been printed using custom type and other solutions that make the character less uniform and recognisable.

Digitisation has impacts on a wide range of Library activities and the management structure must take this into account; the dependencies and workflow may lead to bottlenecks at some stages of the process which may require the diversion of other staff.

The need to consider project legacies at an early stage has become clear from the explicit programming of post-project activities, with resource and technical implications; the Library will seek in future to extend established procedures rather than custom project-specific methods.

The Library has created a workflow that could be readily exploited in order to present further material on the web, if permission was granted and resources for cataloguing and scanning were available. It will seek external funding in order to develop its journal coverage, alongside other digitisation proposals.

Because the pages and publications have stable urls, cross -searching and deep-linking from other resources can be readily supported, allowing the creation of mash-ups. The Library will seek to enable and promote innovative re-use and integration with other of its resources, and will monitor user behaviour to identify scope for further development.

Appendixes

Appendix 1: Publications included

Title¹⁰	Dates included	Subject group	Main language of contents
Archaeologia Cambrensis	1846-1999	History and geography	English
Arloeswr, Yr	1957-1960	Literature	Welsh
Barddas: Cylchgrawn y Gymdeithas Gerdd Dafod	1976-2007	Literature	Welsh
Bathafarn: Cylchgrawn Hanes yr Eglwys Fethodistaidd yng Nghymru	1946-2003	Religion	Welsh
Brycheiniog	1955-2003	History and geography	English
Bulletin of the Board of Celtic Studies	1921-1993	History and geography	English
Bwletin Cymdeithas Emynau Cymru	1968-2003	Religion	Welsh
Cambria: a Welsh Geographical Review	1974-1991	History and geography	English
Cambrian Law Review	1970-2006	Law, politics and education	English
Cardiff Naturalists' Society, Reports and Transactions	1867-1986	Science and engineering	English
Cardiganshire Antiquarian Society Transactions	1909-1938	History and geography	English
Cennad: Cylchgrawn y Gymdeithas Feddygol	1980-2001	Science and engineering	Welsh
Ceredigion	1950-2004	History and geography	English
Cofiadur, Y	1923-2002	Religion	Welsh
Collections Historical and Archaeological Relating to Montgomeryshire	1876-2002	History and geography	English
Contemporary Wales	1987-2001	Law, politics and education	English
Cristion	1983-2006	Religion	Welsh
Cymmrodor, Y	1877-1951	History and geography	English
Cymru	1891-1927	General	Welsh
Efrydiau Athronyddol	1938-2000	Law, politics and	Welsh

¹⁰ This list ignores title changes for a publication; on the website each title form is held as a separate entity, and the pre-1900 issues are listed separately.

		education	
Flintshire Historical Society Journal	1906-2003	History and geography	English
Ford Gron, Y	1930-1935	General	Welsh
Fflam, Y	1946-1952	Literature	Welsh
Gower: Journal of the Gower Society	1948-2005	General	English
Gwent Local History	1977-2006	History and geography	English
Gwyddonydd, Y	1963-1996	Science and engineering	Welsh
Hanes cerddoriaeth Cymru	1996-2004	Literature	Welsh
Heddiw	1936-1942	Literature	Welsh
Journal of Pembrokeshire Historical Society	1985-2004	History and geography	English
Journal of the Welsh Bibliographical Society	1910-1983/4	Literature	English
Journal of Welsh ecclesiastical history	1984-1992	Religion	English
Journal of Welsh religious history	1993-1984	Religion	English
Llafur: the journal for the Society for the Study of Welsh Labour History	1972-2004	History and geography	English
Llenor	1922-1955	Literature	Welsh
Lleufer: Cylchgrawn Cymdeithas Addysg Gweithwyr yng Nghymru	1950-1980	General	Welsh
Minerva	1993-2003	History and geography	English
Morgannwg	1957-2006	History and geography	English
National Library of Wales Journal	1939-2005	History and geography	English
Nature in Wales: a Natural Science Journal for Wales and the Borderland	1955-1987	Science and engineering	English
Pembrokeshire Historian: Journal of the Pembrokeshire Historical Society	1959-1981	History and geography	English
Presenting Monmouthshire	1956-1976	History and geography	English
Proceedings of the South Wales Institute of Engineers	1859-1998	Science and engineering	English
Radnorshire Society Transactions	1939-2004	History and geography	English
South Wales Record Society Publications	1981-1994	History and geography	English
Studia Celtica	1960-2000	History and geography	English
Tir Newydd	1935-1939	Literature	Welsh
Traethodydd, Y	1845-2006	General	Welsh
Transactions of the Honourable Society of Cymmrodorion	1892-2005	History and geography	English

Wales	1937-1959	Literature	English
Welsh Book Studies	1998-2007	Literature	English
Welsh History Review	1959-2001	History and geography	English
Welsh Outlook	1914-1933	Law, politics and education	English

Appendix 2 Glossary

AACR2	Anglo-American Cataloguing Rules (2nd Edition) data content standard for bibliographic data http://www.aacr2.org/
DAMS	Digital Asset Management System software to store, retrieve and disseminate (display) image, text and other files
MARC21	Machine Readable Catalogue data. MARC21 version data content and structure standard for bibliographic data http://www.loc.gov/marc/bibliographic/
METS	Metadata Encoding & Transmission Standard XML data standard for metadata http://www.loc.gov/standards/mets/
OAIS	Open Information Archival System a structure for documenting the actions and responsibilities required for the maintenance and development of a digital resource
OCR	Optical Character Recognition: automatic extraction of text content from images
OJEU	Official Journal of the European Union publication for public procurement notices
QA	Quality Assurance activities to ensure that a process meets the required quality standard
QC	Quality Control evaluation of the output of a single process to determine whether it meets the required standard
TEI	Text Encoding Initiative: an XML schema for the presentation of text-based content http://www.tei-c.org/index.xml
VLE	Virtual Learning Environment networked integrated software application for teaching materials, coursework and supporting documentation eg Blackboard, Moodle
Web 2.0	Model of web services that moves beyond a unidirectional publisher-to-user relationship and promotes user-to-user and user-to-publisher interaction (User Generated Content)
XML	Extensible Markup Language: a language allowing the definition and use of structural and descriptive tags http://www.w3.org/XML/