



Project Information			
<b>Project Acronym</b>	OCRopodium		
<b>Project Title</b>	OCRopodium		
<b>Start Date</b>	01 Sept 2009	<b>End Date</b>	28 Feb 2011
<b>Lead Institution</b>	King's College London		
<b>Project Director</b>	Mark Hedges, <a href="mailto:mark.hedges@kcl.ac.uk">mark.hedges@kcl.ac.uk</a>		
<b>Project Manager &amp; contact details</b>	Tobias Blanke, <a href="mailto:tobias.blanke@kcl.ac.uk">tobias.blanke@kcl.ac.uk</a>		
<b>Partner Institutions</b>	Queen's University Belfast		
<b>Project Web URL</b>	ocropodium.cerch.kcl.ac.uk		
<b>Programme Name (and number)</b>	e-Content		
<b>Programme Manager</b>	Ben Showers		

Document Name			
<b>Document Title</b>	Project Plan		
<b>Author(s) &amp; project role</b>	Mark Hedges (Project Director)		
<b>Date</b>	10 Oct. 2009	<b>Filename</b>	OCRopodium Project Plan 0.1.doc
<b>URL</b>	TBD		
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
0.1	10/Oct/2009	First draft
1.0	22/Oct/2009	Final version

Project Acronym: OCRopodium  
Version: 0.1  
Contact: Mark Hedges  
Date: 10/Oct/2009

# OCRopodium Project

---

**Project Plan**  
October 2009



## Overview of Project

### 1. Background

When digitising text-based resources such as documents, periodicals and books, a key part of the process is the conversion of paper-based analogue sources into machine-readable form using Optical Character Recognition (OCR) software. In essence, this involves obtaining a scanned image of a printed page, using software to distinguish pixel patterns within the image, and translating these into alphanumeric characters. Since the early innovations in OCR technology, both the software and methodologies have improved, and OCR has been applied successfully to a wide range of material using a variety of software and with a range of specific outcomes.

Given the potential profit to be made from OCR technologies, OCR'ing typically involves the use of proprietary software developed by commercial companies. This has a number of consequences:

- OCR'ing can be expensive, as the software has to be bought and licensed, and sometimes consultants or staff from commercial companies must be bought out in order to obtain the best results.
- The OCR software is a “black box”; institutional research staff involved in the project have limited access to what goes on inside the software, and limited control over its behaviour. They are of course able to modify various parameters under which it operates, but there is much less scope for understanding why the software performs as it does in particular cases, and thus for modifying its behaviour and improving the quality of the OCR outputs, for example by modifying the code, using specific training sets, or integrating different components. Indeed there is a reluctance of OCR software manufacturers to allow access to their code even in a collaborative environment.
- Commercial OCR engines are designed to work best with contemporary end user documents – their primary application is the scanning of company notes and documentation. The requirements for such applications are very different from the ones we have for digitising historical collections. As in other machine learning applications, much depends on how OCR systems have been tuned in their training phase. The commercial interest in historical texts, however, is not large enough for companies to train their systems for optimal scanning of such material. It is thus highly desirable to be able to customise the OCR technology; it is very difficult, however, to do this with commercial and closed software products.
- The use of “black box” proprietary software, and even more so the out-sourcing of OCR processing, leads to something of a “skills gap” or “knowledge gap” among researchers involved in digitisation, which results in a failure to appreciate the problems and opportunities OCR approaches offer the scholarly community.
- Sometimes the scanned pages (or older, analogue surrogates such as microfilms) have to be sent for OCR processing to the commercial companies themselves, as they are the only ones with sufficient knowledge of the software; however, they lack the domain-specific knowledge that the researchers have, which could feed back into the process and result in more accurate outputs. In such cases, the actual researchers involved in the digitisation project have no direct influence on the actual OCR process at all. Moreover, often researchers don't know the right questions to ask commercial OCR operators, and the operators in turn don't appreciate the unique issues in digitising historical material.

### 2. Aims and Objectives

To address these concerns, the OCRopodium project aims to trial and evaluate an alternative approach to Optical Character Recognition, using OCRopus<sup>1</sup>, a state-of-the-art, open source tool for document analysis and OCR. OCRopus is sponsored and backed by Google, among other funders, and is being developed under the leadership of Prof. Thomas Breuel from the Image Understanding and Pattern Recognition group of the German Research Centre for Artificial Intelligence, based at the

---

<sup>1</sup> <http://sites.google.com/site/ocropus/>

Project Acronym: OCRopodium  
Version: 0.1  
Contact: Mark Hedges  
Date: 10/Oct/2009

University of Kaiserslautern. OCRopus uses advanced language modelling approaches and was initially intended to be used in high-throughput, high-volume conversion projects. OCRopus is an extensible and modular toolkit rather than a monolithic system; it is possible to unplug and replace individual modules (e.g. for character identification) with improved ones, or simply with another module more suited to the particular material being processed.

OCRopus works closely with another Google-backed, free OCR engine called Tesseract, originally developed by Hewlett-Packard but now released as open source. Tesseract can be used in OCRopus as a module for character recognition. OCRopus is currently the best available open source OCR system for English, when using the combination of RAST-based layout analysis with the Tesseract text line recogniser<sup>2</sup>.

The OCRopus software will be trialled and evaluated on the outputs of several completed digitisation projects carried out by CDDA. The projects selected as test cases will be deliberately chosen to include resources that adopted different OCR data capture processes (see Evaluation, Section 4.2), to ensure that the evaluation of OCRopus is as realistic as possible. Example of collections that may be used as test cases include the following: The Database of Irish Historical Statistics, Corpus of Irish Texts, Historical Hansards, Digital Library of Core Resources on Ireland, Scottish National Dictionary, Histpop. These collections are described in Section 4.2.2.

As well as a direct evaluation of the outputs of OCRopus in comparison with the outputs of the original projects, using both statistical and subjective techniques (see Evaluation, Section 4.2) we will investigate the potential that arises specifically from its status as open source; the ability to figure out why the software is behaving as it does, and the ease with which its behaviours can be modified, e.g. by specific training, incorporating different plug-ins, or modifying the code. Unlike closed, commercial OCR software, OCRopus can be adapted to the specific needs of historical collections, and can be embedded in institutional digitisation workflows that integrate other components for improving OCR performance, such as machine learning or text mining. In particular, we plan to investigate the potential for using feedback to improve the behaviour of the software – for example, when a reviewer notices that a particular printed character is misinterpreted in a consistent fashion, or the use of automated techniques such as text mining to spot gibberish and flag it.

We will not work with OCRopus as a stand-alone tool; we will convert it into a service for increased flexibility, and will implement digitisation activities, including OCR, within a collaborative, distributed (cross-institutional) and automated (or semi-automated) workflow system that is embedded within institutional practice, and which treats digitisation as a process from scanning, through OCR and mark-up, to ingest within a digital repository. The workflow system will be configurable, allowing (for example) different OCRopus plug-ins to be incorporated, different language modules to be configured, and additional processing stages (e.g. such as spelling or grammar checkers) to be included.

Such workflows would automate as much as is possible, while recognising that much of the process will require human input, such as reviewing and correcting OCR'd material. By automating those parts that can be automated, the researchers/repository staff have more time to concentrate on those things for which they are needed. The workflow(s) implemented by the project will produce outputs in a form suitable for ingest into CeRch's digital repository infrastructure – this uses Fedora, and the outputs of the workflow will conform to content models used in this infrastructure. However, this will itself be something that can be configured, as we will be developing a generic framework as well as particular implementations. We will also investigate the potential for integrating these workflows with TEXTvire, the VRE for textual scholarship being developed as part of the JISC VRE 3 programme.

Note that these two broad aims – the evaluation of OCRopus, and the implementation of semi-automated digitisation workflows – are both complementary and independent. The success of one thread does not depend on the success of the other; however, together they will provide a new and promising model – both in terms of process model and business model – for digitisation activities within academic and cultural institutions.

We see our proposal as a further step towards embedding a full portfolio of digitisation-related skills within expert digitisation centres. Reducing the dependency on commercial OCR providers and instead relying on the development of open source OCR technology, such as OCRopus and its

---

<sup>2</sup> [http://pubs.iupr.org/DATA/2007-IUPR-28Nov\\_1234.pdf](http://pubs.iupr.org/DATA/2007-IUPR-28Nov_1234.pdf)

related software, will add to the sustainability of digitised resources, as they can be more easily improved when the algorithms underlying the open source software are further improved through community development. We see the closing of such a potential skills gap in digitisation centres and the further development of institutional skills, policies and capacity to perform OCR for digitisation, as key for maintaining digital resources over time and enhancing their uptake.

### **3. Overall Approach**

#### **3.1 Approach**

The OCRopus software will be trialled and evaluated on the outputs of several completed digitisation projects carried out by CDDA. The projects selected as test cases have been chosen to include resources that adopted different OCR data capture processes (see below), to ensure that the evaluation of OCRopus is as realistic as possible.

As well as a direct evaluation of the outputs of OCRopus in comparison with the outputs of the original projects, using both statistical and subjective techniques we will investigate the potential that arises specifically from its status as open source; the ability to figure out why the software is behaving as it does, and the ease with which its behaviours can be modified, e.g. by specific training, incorporating different plug-ins, or modifying the code. Unlike closed, commercial OCR software, OCRopus can be adapted to the specific needs of historical collections, and can be embedded in institutional digitisation workflows that integrate other components for improving OCR performance, such as machine learning or text mining. In particular, we plan to investigate the potential for using feedback to improve the behaviour of the software – for example, when a reviewer notices that a particular printed character is misinterpreted in a consistent fashion, or the use of automated techniques such as text mining to spot gibberish and flag it.

We will not work with OCRopus as a stand-alone tool; we will convert it into a service for increased flexibility, and will implement digitisation activities, including OCR, within a collaborative, distributed (cross-institutional) and automated (or semi-automated) workflow system that is embedded within institutional practice, and which treats digitisation as a process from scanning, through OCR and mark-up, to ingest within a digital repository. The workflow system will be configurable, allowing (for example) different OCRopus plug-ins to be incorporated, different language modules to be configured, and additional processing stages (e.g. such as spelling or grammar checkers) to be included.

Such workflows would automate as much as is possible, while recognising that much of the process will require human input, such as reviewing and correcting OCR'd material. By automating those parts that can be automated, the researchers/repository staff have more time to concentrate on those things for which they are needed. The workflow(s) implemented by the project will produce outputs in a form suitable for ingest into CeRch's digital repository infrastructure – this uses Fedora, and the outputs of the workflow will conform to content models used in this infrastructure. However, this will itself be something that can be configured, as we will be developing a generic framework as well as particular implementations. We will also investigate the potential for integrating these workflows with TEXTvire, the VRE for textual scholarship being developed as part of the JISC VRE 3 programme.

Note that these two broad aims – the evaluation of OCRopus, and the implementation of semi-automated digitisation workflows – are both complementary and independent. The success of one thread does not depend on the success of the other; however, together they will provide a new and promising model – both in terms of process model and business model – for digitisation activities within academic and cultural institutions.

We see our proposal as a further step towards embedding a full portfolio of digitisation-related skills within expert digitisation centres. Reducing the dependency on commercial OCR providers and instead relying on the development of open source OCR technology, such as OCRopus and its related software, will add to the sustainability of digitised resources, as they can be more easily improved when the algorithms underlying the open source software are further improved through community development. We see the closing of such a potential skills gap in digitisation centres and the further development of institutional skills, policies and capacity to perform OCR for digitisation, as key for maintaining digital resources over time and enhancing their uptake.

### 3.2 Benchmark Datasets

We will use a range of machine-readable texts produced by completed digitisation projects, to act as benchmarks for assessing the value of OCRopus for converting images containing texts into machine-readable form. We have deliberately selected as test cases resources that adopted different OCR data capture processes to make the evaluation as extensive as possible. Specifically, the test datasets will cover the following scenarios:

- Bespoke images scanned by CDDA to ensure the highest possible OCR accuracy levels. This involves using expensive specialised scanners – such as book page scanners – and post processing software to remove page curvature, reduce post-printing annotations, and remove blemishes on a page.
- Collections of image material scanned by third parties and OCR'd by CDDA. This increases the challenge of accurately capturing text, although pre-scanned images can be enhanced.
- Non-intervention OCR'ing. In these cases, 100% accuracy of the machine-readable text was not considered to be worth the considerable outlay in staff time; a suitable OCR package with an appropriate fontbase automatically creates machine-readable texts in batch mode with no user intervention.
- The development of texts where a high level of accuracy (at least 99.5%) is required; significant user intervention is needed to achieve such accuracy rates.
- The creation of texts that need to be completely accurate and contain no errors.
- The creation of an OCR'd text that exactly replicates the layout of the original scanned document.
- Materials that contain non-standard alpha-numeric data, typically texts in foreign languages or containing complex symbolisation.

The collection of test datasets may be modified during the project, but the initial list comprises:

- **The Database of Irish Historical Statistics** was the first large-scale OCR project an embryonic CDDA undertook from 1991 to 1995. The project aimed to capture most recurrent Irish census statistics from the printed census volumes from the first survey in 1821 to the last in analogue-only form in 1971. In order to reduce what was at the time regarded as a monumental task, information was only collected for larger spatial units. Thus information for Ireland's 2,000 parishes was not gathered, with the smallest enumerated unit being the barony of which there were more than 300. Initially the project intended to capture data by a double keying process but during the early stages of the work an experimental piece of software, ProLector, was released. This software was specifically designed to capture data in tabular format ignoring the potential complications of table headings, borders, and columns marked by lines. With careful scanning of the original census volumes relatively high levels of character recognition were achieved. This was particularly the case as ProLector was fully trainable. It did not come with a preset range of characters and fonts with which to match pixel patterns drawn from a scanned image. Rather the software identified a group of pixels and referred the pattern to the OCR operator to assign an alphanumeric character to. Whilst this allowed for bespoke fontbases to be developed it was time consuming.

Statistical data can quickly lose credibility if any errors are included. As a result CDDA aimed to provide a 100 per cent accurate text. This was achieved by using a variety of quantitative techniques to identify OCR errors including summing tabular data and comparing those sums with the same information provided in the original census volumes and techniques to identify unusually large or small numbers in a table. The project produced the largest historical statistical dataset of the day containing more than 32 million values.

- **Corpus of Irish Texts.** In collaboration with the University of Ulster, CDDA scanned and digitised 2,000 pages of Irish-language historical material in a poor physical state. A bespoke fontbase was developed to cope with the varied Irish scripts – both ancient and modern. This study offers a benchmark against which to measure bulk data capture of non-standard alphanumeric text and the use of poor images.
- **Historical Hansards** involved the image scanning and OCR'ing of 90,000 pages of Hansard for the Stormont Northern Ireland Parliament from 1921 to 1972. The project aimed to capture a reasonably accurate text but one that would not be of sufficient quality to be made publicly available. Users would deploy the text to search Hansard and the results would be viewed via a

scanned page image. CDDA was able to scan the images from an unused copy of Hansard however, and through the use of a book page scanner and post-processing software, it was possible to gather exceptionally high quality images. This resulted in a high quality machine-readable text. A few recurrent errors in the text – “1” recognised as “l” for example – were corrected using macros, and the text was of sufficient quality to release to users. As a result users can now view both an image of a page of debates and the text itself, which can be downloaded.

- **Digital Library of Core Resources on Ireland** In collaboration with JSTOR, CDDA created a 600,000 page archive of page images and text of Irish Studies Journals. Very high quality page images were required by JSTOR in which page sizes were standardised, post-printing annotations removed, and text bleed through ameliorated. The images were converted into machine-readable form through bulk digitisation so this study will benchmark non-intervention OCR'ing with high quality images.
- **Scottish National Dictionary** The 24 volumes of SND were scanned and OCR'd by CDDA. SND required 100 per cent accuracy in the text – and employed teams of trained proof-readers to ensure this was the case. The text also had to retain the complex layout of the Dictionary in terms of pagination, indentation and line breaks. In addition Dictionary-specific reference codes had to be recorded. This incredibly labour intensive work acts as an exemplar of the development of the highest quality text using OCR approaches and very significant manual post-processing.
- **Histpop** With the University of Essex CDDA scanned to a high standard around 1,000,000 pages of the printed census returns for the British Isles. From the scans high-quality texts were created for the census prefaces involving significant user intervention. The remainder of the material was OCR'd automatically, resulting in low accuracy levels, particularly as much of the material contained statistical tables. This study benchmarks the impact user intervention may have on a built OCR'd text.
- **Nineteenth Century Serials Edition (NCSE)** in the UK. The NCSE is a free, online scholarly edition of nineteenth-century periodicals and newspapers. It has been created as a collaboration between Birkbeck, University of London, King's College London, the British Library, and Olive Software. It was funded from January 2005 to December 2007 by the Arts and Humanities Research Council in the UK. The NCSE corpus contains circa 430,000 articles that originally appeared in roughly 3,500 issues of six 19th Century periodicals. Published over a span of 84 years, materials within the corpus exist in numbered editions, and include supplements, wrapper materials and visual elements. The OCR results for this collection have not been very good. We shall look at examples of bad OCR and see whether our approach offers any improvement.

## 4. Project Outputs

Analysis reports:

- a) An analysis of digitisation practices and processes, particularly with reference to OCR activities. This will include desired scenarios as well as ones actually carried out at CDDA/CeRch.
- b) A review of the test datasets.

Software-related outputs:

- c) A set of software components that integrate with and enhance the OCRopus software.
- d) A prototype digitisation workflow, incorporating OCRopus, in use at the partner institutions (this will not be a production system).
- e) Architectural and technical documentation, to facilitate enhancement and re-use of the software.
- f) A framework and guidelines for preservation workflows that can be adapted to other institutions.

Other reports:

- g) Case studies that evaluate our approach for each test dataset used, together with an overall evaluation report that synthesises these case studies. We will address not only functional aspects – how do the results of our approach compare with proprietary approaches in terms of the quality of the OCR outputs – but also financial aspects, e.g. business models for digitisation activities. We will develop detailed recommendations and financial case studies for other institutions so that they are able to evaluate their financial commitments in terms of their OCR strategy.
- h) Progress reports and final report.

## 5. Project Outcomes

As well as these concrete deliverables, key outcomes of the project will include:

- a) An increase in the skills of institutional staff, both technical staff and digitisation/archive staff, in the use of the open source OCR software.
- b) An embedding of digitisation workflows within institutional practice, at least as a prototype.

## 6. Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Digitisation projects at KCL and QUB/CDDA	Streamlining and improving digitisation workflows, for both greater efficiency and more effective results.	Very High
Other digitisation projects and groups interested in carrying out digitisation	Ditto – project outputs will be directly applicable to their work.	Very High
Archives and libraries at KCL and QUB	Ditto – project outputs will be directly applicable to their work.	Very High
Archives and libraries at other institutions	Ditto – project outputs will be directly applicable to their digitisation programmes.	Very High
Funding bodies	Increased visibility of work; more efficient digitisation processes; better return on investment; enhanced skills and expertise in OCR in the projects they fund.	High
Archive and library managers	Better visibility of collections.	High
ISS at King's	Provider of central IT systems and services.	Medium
Wider JISC community, including IE programme and digital repository community	Case study of integrating digitisation and OCR services within repository workflows.	Medium
CeRch	Development of data repository infrastructures for KCL	High
e-Framework	e-Framework documentation, e.g. SEs, SUMs	Medium

## 7. Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
<b>Staffing</b>				
Difficulties recruiting and retaining staff.	1	4	4	Key staff are already in post. CeRch/CDDA have a broad pool of knowledgeable staff on which to draw. In addition, good software development staff are proving much less difficult to find in the current economic climate.
<b>Organisational</b>				
Failure to meet project milestones.	2	3	6	Produce project plan with clear objectives, and detailed/realistic

				models of research processes. Monitor progress and re-plan when necessary.
Communications failure between partners.	1	2	2	The project partners and the creators of the OCRopus software have expressed strong support for the project.
<b>Technical</b>				
A complete solution cannot be implemented within the project timescale.	2	3	6	The absence of a complete solution is not an indication of failure, as one aspect of the project is to investigate potential problems. Project reports will address any issues that could not be resolved.
<b>External suppliers</b>				
Issues with the OCR technology (OCRopus)	2	3	6	The creators of the software are very supportive of the project (see attached Letter of Support). The software is open-source and extensible through plug-ins; thus it will be possible to adapt it to better suit the requirements of the project and of the digital material being processed (in contrast to proprietary OCR software).
<b>Legal</b>				
Status of test collections outside public domain	1	3	3	We have obtained rights to access to all collections on the basis that these are available for our particular research. As we do not foresee publishing more than evaluation results and test pages, we do not expect problems using these, though the ownership might change through the project. We will regularly monitor the situation and potentially use comparable test collections.

## 8. Standards

Name of standard or specification	Version	Notes
hOCR	Latest	An (X)HTML-compatible OCR output format used by OCRopus (see <a href="https://docs.google.com/View?id=dfxcv4vc_67g844kf">https://docs.google.com/View?id=dfxcv4vc_67g844kf</a> )
PREMIS	Latest	Used for recording audit trail metadata.
ReST	N/A	May be used for developing services for integration with digitisation workflows.

## 9. Technical Development

The technical development centres on the open source OCRopus software. OCRopus provides a rich set of OCR functionality, but it is likely to require significant enhancement – writing new modules and modifying existing ones – for it to be used successfully with the historical material that we are dealing with. As we intend that these enhancements are incorporated into the core OCRopus software, we will follow the standards that the OCRopus project uses, as far as is practical.

Project Acronym: OCRopodium  
Version: 0.1  
Contact: Mark Hedges  
Date: 10/Oct/2009

As well as evaluating and developing OCRopus, we will investigate ways of integrating the software into broader digitisation workflows. For this we will use existing software components – e.g. natural language processing software, workflow tools – rather than developing new ones. Wherever possible we will take a service-orientated approach, using ReST-based APIs, in order to maximise flexibility and re-use.

Close attention will be paid to issue tracking and version control when developing or modifying software – this is especially important as the OCRopus code base will be changing in parallel with our own work. For this, we will use appropriate tools such as Subversion and TRAC.

## 10. Intellectual Property Rights

IPR in all documents produced by the project will be retained by the authors and host institutions but made freely available on a non-exclusive licence. Modifications and enhancements to the OCRopus software will be made available to the OCRopus open source project. Any other software created during the project will be made available to the community on an open-source basis. We will respect the licence of all third party software used, most of which is open source.

IPR of digitised material used for evaluation: we plan to make available both the outputs of the original projects (where they are not available already) as well as the OCR outputs produced by the OCRopodium project, so that people can compare our results with the original results. We expect that for most if not all cases the material in question is fully in the public domain, so there will be no issue with this; in case that any resource used is not fully public, we will be able to make available samples of texts for comparison. This will be determined as the project progresses. Throughout the project we will carefully monitor changes to ownership in the collections.

## *Project Resources*

### 11. Project Partners

The Centre for Data Digitisation and Analysis (CDDA) at Queens' University Belfast: CDDA will contribute to WPs 2 and 5 (see Appendix for details), providing input on digitisation issues in general and OCR in particular. The main point of contact will be Paul Ell. The consortium agreement is in preparation but has not yet been signed.

The Centre for e-Research (CeRch) at King's College London will be responsible for all other work.

### 12. Project Management

#### *12.1. Project Management Framework*

Mark Hedges will act as Project Director, overseeing the project as a whole in the context of wider institutional, technical and JISC-related issues.

Tobias Blanke will act as Project Manager, overseeing the day-to-day activity of the project across all workpackages and liaising with project partner staff.

Paul Ell will be responsible for the overall direction of the work at QUB.

Elaine Yeates will be Site Manager at QUB, overseeing the day-to-day activities of the CDDA project team.

Notwithstanding these reporting hierarchies, the nature of the project will necessitate close liaison between analysis/development staff at CeRch and CDDA.

There will be monthly meetings, at least by Skype, to track progress.

#### *12.2. Project Staff*

Staff at Centre for e-Research (CeRch), King's College London:

**Dr Mark Hedges** (Project Director, [mark.hedges@kcl.ac.uk](mailto:mark.hedges@kcl.ac.uk)) is Deputy Director of CeRch, and will be responsible for the overall direction of the project.

**Dr Tobias Blanke** (0.2 FTE, Project Manager, [tobias.blanke@kcl.ac.uk](mailto:tobias.blanke@kcl.ac.uk)) is Research Fellow at CeRch, and will be responsible for project management.

**TBD** (1.0 FTE, Software Specialist) will be responsible for WPs 3 and 4, and will participate in WPs 5, 6 and 7. A new staff member is being recruited for the role. The post is being advertised currently (closing date for applications is 16<sup>th</sup> October)..

Lydia Horstman (0.1 FTE, [lydia.horstman@kcl.ac.uk](mailto:lydia.horstman@kcl.ac.uk)) will be project administrator.

Staff at the Centre for Data Digitisation and Analysis, King's College London:

Dr Paul Ell (0.1 FTE, Site Manager, [paul.ell@qub.ac.uk](mailto:paul.ell@qub.ac.uk)) is Director of the CDDA, and will be responsible for the overall direction of the QUB team.

Elaine Yeates (Project Coordinator at CDDA) will oversee the CDDA project team on a day-to-day basis.

Anthony Anderson (1.0 FTE for 9 months. IT Officer) will assess digitization workflows from an informed technical perspective (WP2).

Emma McGurk (1.0 FTE for 9 months, IT Officer) has expertise in OCR software and will test how our software developments are adopted by data input staff (WP5).

### 12.3. Training Requirements

Training (or at least expert assistance) on the OCRopus software will be needed. The OCRopus project leader has offered his support and proposed holding a workshop. We have also been in contact with SUB Göttingen, who are also investigating OCRopus. We currently plan to hold a joint workshop with OCRopus and SUB once the CeRch developer has been recruited.

## 13. Programme Support

The project team would be grateful if the JISC would:

- Provide adequate advanced notice of programme meetings and non-standard reporting requirements.
- Identify potential areas of collaboration or communication with projects in other programmes

## 14. Budget

See Appendix A

## Detailed Project Planning

## 15. Workpackages

See Appendix B.

## 16. Evaluation Plan

Timing	Factor to Evaluate	Questions to Address	Method(s)	Measure of Success
At project management meetings	Cross-partner working effectiveness	Is work progressing as expected across partners?	Agenda item for meetings	Noted that all is OK or action to address if not
End 02/10	Use case/data documentation	Do the use cases capture the practices and requirements of the researchers? Do the documents adequately describe the test datasets?	Review of docs with those involved.	Agreement of digitisation staff.
Monthly during development	Quality of code outputs	Is code stable?	Code review. JUnit testing	Successful testing (for individual WPs) and review

Monthly during integration	Progress of testing	Is testing providing what we need?	Review of testing with those involved	Acknowledged progress in integration
Monthly during user testing	Progress of testing Utility of system to digitisation staff.	Are requirements in use cases being met?	Feedback and baseline analysis undertaken by CDDA staff	Acknowledged progress in user testing
End of project	How well does OCRopus support use cases?	To what extent have use cases been supported successfully?  How does the open source OCR software compare with proprietary approaches (in terms of functionality and in economic terms)?	Feedback and baseline analysis undertaken by CDDA staff	Positive opinions of digitisation staff
End of project	Case studies and evaluation synthesis	Are these documents useful for other institutions and communities involved (or thinking of being involved) in digitization projects?	Feedback from community.	Feedback on the whole positive

## 17. Quality Plan

Output	Quality criteria	QA method(s)	Evidence of compliance	Quality responsibilities	Quality tools (if applicable)
<b>Output</b>	<b>Use Cases/dataset review</b>				
28/02/10	FFP	Peer review* and comment from Programme Manager	Acceptance by reviewers	TB, PE	
<b>Output</b>	<b>Software (quality of software)</b>				
31/12/10	FFP	Code and testing review (internal to project)	Acceptance by reviewers	TB	
<b>Output</b>	<b>Exemplar system</b>				
End of project	FFP	Review against user requirements and feedback	Acceptance by reviewers	TB, PE	
<b>Output</b>	<b>Case Studies and Evaluation Report</b>				
End of project	FFP	Peer review* and comment from Programme Manager	Acceptance by reviewers	TB, PE	
<b>Output</b>	<b>Other documentation and reports</b>				
Various	FFP	Internal project review and comment from	Acceptance by reviewers	TB, PE	

		Programme Manager.			
--	--	--------------------	--	--	--

FFP= fit for purpose

\* These documents will be shared with others in the field with appropriate domain knowledge and feedback sought.

## 18. Dissemination Plan

Timing	Dissemination Activity	Audience	Purpose	Key Message
ongoing	Participation at JISC programme activities.	JISC programme participants	To engage the relevant user communities & demonstrate the validity and benefits of the work	
ongoing	Liaison with other e-Content projects. We would like to increase participation in our OCR experiments	JISC programme participants	ditto	
ongoing	Presentations at conferences and workshops in various fields (e.g. humanities/scholarship, digital repositories/archives, digitisation).	Wider communities	ditto	
ongoing	Papers submitted to conferences and journals in various fields (as previous). We will target in particular Digital Humanities conferences such as DRHA and archives/digital libraries conferences such as ECDL	Wider communities	ditto	
ongoing	Demos and tutorials. We arranged a joined workshop with SUB Göttingen in Germany for 2010	Actual or potential users of digitisation tools.	ditto	
ongoing	Engagement with staff from other institutions	Actual or potential users of digitisation tools.	ditto	
ongoing	European dissemination and outreach, via SUB Göttingen, DARIAH, OCRopus.	As above, but within a European context	ditto	

## 19. Exit and Sustainability Plans

Project Outputs	Action for Take-up & Embedding	Action for Exit
Exemplar system	Continue to enhance and use system within institutional infrastructure and practice. Disseminate information about system.	Maintenance of system and its documentation.
Software components	Make available in open source code repository. In the case of OCRopus components, these will be made available to the OCRopus team for inclusion in the core source base.	
Case Studies and evaluation report	Disseminate	

Experience of project	Document in project outputs.	Make documentation available.
-----------------------	------------------------------	-------------------------------

Project Outputs	Why Sustainable	Scenarios for Taking Forward	Issues to Address
Exemplar system	Useful to College archives and other groups interested in digitisation projects	Embed in institutional infrastructure and practice at King's.  Encourage use of framework by other institutions.  Continue to develop as part of CeRch's general remit.  Continue to develop as part of other digitisation and repository projects.	Requires user community.  Requires ongoing maintenance.
Software components	Useful to developers of analogous systems.	Make available to wider community on an open-source basis.  Continue to develop as part of CeRch's general remit.  Encourage use of software components by other institutions.	Requires user community.  Requires ongoing enhancement and maintenance.
Case Studies and Evaluation Report	Repositories of practical guidance and experience that will be applicable to analogous situations in other institutions.	Disseminate documentation.	

## Appendixes

### Appendix A. Project Budget

Directly Incurred Staff	Sept 09-Mar 10	April 10-Feb 11	TOTAL £
Tobias Blanke, grade 7, 0.2 fte	£6786.68	£11239.30	£18025.98
Software Specialist, grade 6, 1.0 fte	£30849.93	£50917.20	£81767.13
Lydia Horstman, grade 5, 0.1 fte	£2331.10	£3842.93	£6174.03
Paul Ell, AC4, 0.1 fte	£7616.00	£3880	£11496
Elaine Yeates, grade 6, 0.1 fte, 9 months	£ 1835	£2339	£4174
Anthony Anderson, grade 4, 1fte, 9 months	£8999	£11473	£20472
Emma McGurk, grade 1, 1fte, 9 months	£5734	£7311	£13045
<b>Total Directly Incurred Staff (A)</b>	<b>£64,151.71</b>	<b>£91,002.43</b>	<b>£155,154.14</b>
Non-Staff	Sept 09-Mar 10	Apr 10 -Feb 11	TOTAL £
Travel and expenses	£ 5000	£ 10000	£ 15000
Hardware/software	£ 5000	£	£ 5000

Project Acronym: OCRopodium  
Version: 0.1  
Contact: Mark Hedges  
Date: 10/Oct/2009

Dissemination	£	£	£
Evaluation	£	£	£
Other (recruitment)	£ 2000	£	£ 2000
<b>Total Directly Incurred Non-Staff (B)</b>	<b>£12000</b>	<b>£10000</b>	<b>£22000</b>
<b>Directly Incurred Total (C) (A+B=C)</b>	<b>£76,151.71</b>	<b>£101,002.43</b>	<b>£177,154.14</b>
<b>Directly Allocated</b>	<b>Sept 09–Mar 10</b>	<b>Apr 10 –Feb 11</b>	<b>TOTAL £</b>
Staff	£	£	£
Estates King's	£6093.11	£9957.08	£16050.19
Estates Queen's	£735	£370	£1105
Other	£	£	£
<b>Directly Allocated Total (D)</b>	<b>£6828.11</b>	<b>£10327.08</b>	<b>£17155.19</b>
<b>Indirect Costs (E) King's</b>	<b>£23625.06</b>	<b>£38298.79</b>	<b>£61923.85</b>
<b>Queen's</b>	<b>£4081</b>	<b>£2059</b>	<b>£6140</b>
<b>Total Project Cost (C+D+E)</b>	<b>£110,685.88</b>	<b>£151,687.30</b>	<b>£262,373.18</b>
<b>Amount Requested from JISC</b>	<b>£88548.70</b>	<b>£121349.84</b>	<b>£209898.54</b>
<b>Institutional Contributions</b>	<b>£22137.18</b>	<b>£30337.46</b>	<b>£52474.64</b>
<b>Percentage Contributions over the life of the project</b>	<b>JISC 80 %</b>	<b>Partners 20 %</b>	<b>Total 100%</b>
<b>No. FTEs used to calculate indirect and estates charges, and staff included</b>	<b>No FTEs 2.3</b>	<b>Which Staff</b> Tobias Blanke, Software Specialist (KCL), Paul Ell	

## Appendix B. Workpackages

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
WP1																		
WP2																		
WP3																		
WP4																		
WP5																		
WP6																		

Dependencies between WPs:

- WP 4 depends on outputs of WPs 2 and 3.
- WPs 5 and 6 depend on outputs of WP 4.

Workpackage and activity	Earliest start date	Latest completion date	Outputs	Milestone	Responsibility
<b>WP1: Project management</b> Management of the project including: planning; coordination of contributors; monitoring progress; advocacy; reporting.	01/09/09	28/02/11			MH/TB (overall) PE/EY (CDDA)
1. Project Plan			Project plan	14/10/09	MH
2. Interim project report 1			Progress report	28/02/10	TB
3. Interim project report 2			Progress report	31/08/10	TB
4. Final and completion reports			Final and completion reports	28/02/11	TB Input from all

<p><b>WP2: Review of datasets and processes</b>  This WP will review and describe digitisation practices, processes and data. This work will be carried out in close consultation with CDDA staff involved in digitisation activities. Specifically, this WP will include:</p> <ul style="list-style-type: none"> <li>• Reviewing the test datasets to be used in the evaluation of OCRopus.</li> <li>• Developing use cases and process models based on the OCR and other digitisation activities exemplified in CDDA projects, paying particular attention to the projects to be used as test cases.</li> <li>• Working with researchers to determine scenarios for OCR and digitisation that have not been implemented in other projects, but which would nevertheless be desirable or useful.</li> </ul>	01/10/09	28/02/10			AA (CDDA)
1. Review of datasets	01/10/09	28/02/10	<b>Dataset review document</b>	28/02/10	
2. Use cases/processes (existing)	19/10/09	28/02/10	<b>Use case document</b>	28/02/10	
3. Use cases/processes (future/desired)	19/10/09	28/02/10	<b>Use case document</b>	28/02/10	
<p><b>WP3: Technical Investigations</b>  Investigation and review of the technologies to be used in the project, particularly of the open source OCRopus software. This will include:</p> <ul style="list-style-type: none"> <li>• Review of OCRopus functionality, with particular reference to the use cases identified in WP 2.</li> <li>• Review of the OCRopus architecture, in particular to identify extension points.</li> <li>• Review of other software that may be integrated, e.g. natural language processing modules or workflow tools.</li> </ul>	01/10/09	28/02/10	Technical review document. Small prototypes for validating individual issues (possibly)	28/02/10	TBD (CeRch)
1. Review of OCRopus	01/09/09	28/02/10		28/02/10	

2. Review of other relevant software	01/09/09	28/02/10		28/02/10	
<b>WP4: Implementation</b> Implementation of the software, including design, development, unit testing and integration of the software components. As we will follow an agile, user-driven and evolutionary approach, involving incremental cycles of implementation and evaluation, 4.1-4.4 are not distinct phases that succeed each other, but rather activities that are repeated as required. This iterative approach also implies that WP4 overlaps in time with WP5 to a significant degree.	01/01/10	31/12/10			TBD (CeRch )
1. Design	01/01/10	30/03/10	<b>Design documentation</b>	30/03/10	
2. Development	01/01/10	31/12/10	<b>Software components</b>	31/12/10	
3. Testing (individual components)	01/06/10	31/12/10	<b>Test documentation</b> <b>Tested software</b>	31/12/10	
4. Integration	01/09/10	31/12/10	<b>Integrated software components</b>	31/12/10	
5. e-Framework documentation	01/09/10	31/12/10	<b>e-Framework submissions (e.g. SUMs, SEs)</b>	31/12/10	
<b>WP5: Evaluation</b> Evaluation of project outputs by digitisation staff (see Evaluation Plan above).	01/06/10	28/02/11			EM
1. User evaluation.	01/06/10	28/02/11	Feedback on project outputs.	28/02/11	
2. Case studies (for test datasets)	01/06/10	28/02/11	Case studies	28/02/11	

3. Evaluation report. As well as synthesising the results from the case studies, this will address issues such as embedding and sustainability, as well as economic/financial considerations and models.	01/01/11	28/02/11	Report.	28/02/11	
<b>WP6: Dissemination and Outreach</b> Dissemination and outreach activities (see Section 18), with the aim of communicating the ideas and outputs of the project to a range of audiences. This will be carried out as the opportunity arises.	01/09/09	28/02/11			Various
1. Set up project website (blog-based). Note: website will be updated continually, although this is not indicated on the GANTT chart.	01/09/09	30/09/09	<b>Website</b>	30/04/09	MH (set-up) All (update)
2. Dissemination activities (to be expanded)	01/06/10	28/02/11	<b>Dissemination outputs</b>	Various	Various