Decapod

October 9, 2008

A Proposal to the Andrew W. Mellon Foundation

By

Technical University of Kaiserslautern (DFKI)
Image Understanding and Pattern Recognition (IUPR)
and
University of Toronto (UTor)
Adaptive Technology Resource Centre (ATRC)
Fluid Project

Primary Contacts:
Thomas Breuel, DFKI/IUPR

Jutta Treviranus, UTor/ATRC/FLUID

In Partnership With:
John Burns, JSTOR

And other partners.

Executive Overview

Scholarly content needs to be online, and for much mass produced content, that migration has happened, thanks to projects such as JSTOR, Project Muse, the various publisher portals from the scholarly publishers and their service providers such as HighWire Press and Atypon. Unfortunately, the online presence of scholarly content is much more sporadic for long tail material such as small journals, original source materials in the humanities and social sciences, non-journal periodicals, and more. A large barrier to this content being available is the cost and complexity of setting up a digitization project for small and scattered collections coupled with a lack of revenue opportunities to recoup those costs. Collections with limited audiences and hence limited revenue opportunities are nonetheless often of considerable scholarly importance within their domains. The expense and difficulty of digitization presents a significant obstacle to making paper archives available online.

This project proposal arose because there is no commercial solution that is affordable, reliable and effective at digitizing content. We have assessed all of the readily available solutions and they all (irrespective of their claims) proved to be extremely expensive to buy and operate, difficult to use, unreliable, or all of the above (see Appendix 2). In several cases they were essentially unusable, are all expensive, they require extensive expertise to set up and operate, and they require a high degree of user intervention and expertise to operate. None of them provides a reasonable solution and none of them draws upon state of the art software technology now available.

This proposal outlines a solution that is primarily aimed at small to medium paper archives with material that is rare or unique and of sufficient interest that it warrants being made more widely available. It will allow in-situ capture of bound material, using non-specialized local staff or volunteers. Our solution will be useful for local historical societies, archives of local celebrities, auction catalogs, field notes, lab notes, news-sheets, and more. It will prove extremely cost effective where volunteer labor or "free" labor is available, the volumes are modest, and where setting up a full production system such as that used by JSTOR or Google is not feasible.

To meet this need we are proposing the creation of Decapod. Decapod will be an inexpensive attaché case sized hardware/software solution that can be readily procured and assembled and taken into the stacks by local staff or volunteers to quickly and unobtrusively capture the material and deliver it in usable format. It will be open-source, easy to use, and will provide an out-of-the box method of digitizing small to medium archives of scholarly material.

The project will develop new components while integrating existing components from DFKI's OCRopus system and ATRC's FLUID to create a high quality, low cost solution for digitization. The primary output from the system will be reflow tagged, PDF/A files that can present either high fidelity reproductions of the original pages or can reflow for use on any supported PDF platform, including mobile devices. Intermediate results such as archive quality TIFFs can also be generated and these can be fed into any digitization workflow via a standard 'watch folder' interface.

Decapod will deliver a complete solution for the capture of materials for which current digitization workflows are not appropriate. The deliverables will include software and suggested hardware configurations and hence allow the assembly of a complete system using off the shelf hardware components.

The software components of the proposed system are:

- camera-based document capture using advanced computer vision algorithms to create "Scanner Equivalent" page images.
- A deeply user centered and easy-to-use document capture and quality control system based on state of the art document understanding technology that removes the need for most user interaction and dramatically simplifies the interaction when it is necessary.
- high-quality scan-to-PDF conversion software that emits PDF/A with high fidelity (to the original) typefaces and embedded document layout information to permit reflow and text to speech.
- integration of all software components into an end-to-end solution

The overall flow of the system is a series of three steps. First there is the capture process, i.e. the creation of images of the pages from the physical material. The software demands at this point are primarily to ensure that the material is captured in its entirety and to sufficient quality. The next stage, which could take place later, is the generation of archive quality images and document structure information. The final stage is the generation of the usable output, which in this project is reflowable PDF/A documents.

Operationally the solution will require very modest training--the process can be described on a couple of pages--and will produce an output that is directly suitable for distribution and digital preservation. Advanced quality management means that the system will be highly reliable. State of the art document understanding will generate output that can be viewed on mobile devices and will be accessible using text-to-speech (TTS) systems. The combination of commodity hardware, simple process flows, ability to use local staff, and quality management will allow any collection to be digitized at modest cost. The resulting digital documents can then be shared either as a simple file archive or by deposition in existing archives such as the Internet archive and JSTOR.

The solution consists of suggested hardware configurations and three primary software components. The hardware is based on affordable, portable and easily configured digital cameras and a standard PC. We expect hardware costs for the rigs (mid-to-high DSLR cameras x 2) to cost under $1,000 with a higher end model costing roughly $6,000 to $8,000 depending on the quality of the cameras. The three-part software solution consists of a capture component and a two-part workflow component that is either local or remotely hosted. The capture component is designed to quickly identify any operator errors in the capture process and guide their correction, contributing to the ease of use. The two-part workflow component handles the organization and labeling of the page bundles into documents and then the generation of the high-quality outputs. Since access is via web technologies the software can either be local or remote.

Decapod will remove the barriers to digitization now encountered by archives of documentary material. There are many such obstacles, including cost of equipment, cost of labour, lack of digitization expertise, lack of suitable distribution formats, and lack of acceptable remediation workflows. Decapod will address them all to produce a paper-to-digital document solution that is highly effective, highly automated, and low operator interaction (apart from page turning).

The solution will address these problem areas.

1. It will allow the camera based capture of bound material by using computer vision techniques to produce flat, clean page images equivalent to those produced from a flat bed scanner.
2. It will remove the need for extensive operator intervention in the capture process by detecting scan problems and allowing the operator to rectify the scan immediately.
3. It will reduce user intervention in the conversion process by using advanced document understanding techniques to remove almost all intervention, and by reducing the remainder to very simple "1-click" operations.
4. Its PDF/A outputs will be visually faithful to the original, searchable, and widely usable.
5. It will allow the output to be viewable on mobile devices that support PDF reflow.
6. It will remove the need for deep software, hardware or digitization skills by integrating all software components into a turnkey end-to-end solution.
7. It will remove capital cost barriers by using consumer grade cameras.
8. It will reduce operational cost barriers by allowing volunteers or local staff to operate the system with minimal training or commitment.

Versions 0.5 and 1.0 will be developed and released during year 1 of our proposed

project. Version 1 will be the major milestone release of the year and will be a functioning book (bound document) digitization system that captures page images of books using digital cameras and outputs them as usable PDF documents. It will integrate components from FLUID & DFKI, implement the "book" capture functionality and the PDF generation component and provide a basic user interface for viewing page thumbnails and full pages, removing mis-scans, rearranging pages, and adding placeholder pages.

Versions 1.5 and 2.0 will take place during year 2, aimed at improving quality, improving the user experience, completing the PDF feature set and optimizing the packaging of the solution for easy install.

As each release becomes available JSTOR will use its channels into the library community to demonstrate and promote the solution, will engage community stakeholders soliciting feedback, and will provide a high degree of visibility and promotion for the solution.

The principals will be Thomas Breuel of DFKI Kaiserslautern and Jutta Treviranus of the Adaptive Technology Resource Centre (ATRC) at the University of Toronto. JSTOR will participate in community-building, pilots of the solution, and in the provision of additional resources and expertise. JSTOR is not applying for funding under this proposal. These institutions already have the staff needed with the required skills to achieve this project. In addition, much of the technology needed already exists or is well understood, so this is, technologically, a low risk project. We expect the tool to become a focus of sustained development activity from the document processing community, including commercial entities such as Google and Xerox (with whom we have existing relationships) and the academic document engineering community (whom we will further engage).

As previously noted Decapod is focused on delivering an affordable and cost effective solution to permit high quality, minimal user intervention solutions to the capture and preparation of small to medium collections. We will apply the technology advances (both hardware and software) of the recent decades to remove the usability, cost and quality barriers to such projects.