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
tmb@iupr.dfki.de
DFKI GmbH
Trippstadter Strasse 122
D-67663 Kaiserslautern
+49 (0)631 / 205-75 400

Jutta Treviranus, UTor/ATRC/FLUID
jutta.treviranus@utoronto.ca
140 St. George St., Rm. 514
Toronto, Ontario, Canada M5S 1A1
416.978.5240

In Partnership With:
John Burns, JSTOR
john.burns@jstor.org

And other partners.

Table of Contents

Index of Tables

Index of Figures

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.

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.

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.

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).

Project Motivations

Problems to be Addressed

Document digitization is not easy. The whole process, from initial image capture to a useful output, is arcane and messy with no guarantee of usable results. Alternatives are discussed in Appendix 2, but note that none of them delivers a suitable solution.

Much of the scholarly material that would benefit from this project is complex in layout. Journals, with their multi-column layout, illustrations and complex lists and tables, auction catalogs, inventories and records, newspapers and newssheets, manuscripts and so forth contain images, multiple columns or boxes. Moreover, many of these documents are old, fragile, discolored, and in archaic typefaces. If the material is bound then even flat-bed scanning will produce distorted images. Off-the-shelf packages such as the OCR packages are not particularly good in dealing with complex layouts, and the correction process is particularly tedious. This is unlikely to change as the market for OCR is not large, and the investment of the surviving commercial companies such as Abbyy, Nuance (Scansoft) is more oriented towards the commercially more important goal of extending the languages covered than addressing the more esoteric layouts. (It should be noted that they are doing an excellent job of addressing the breadth of languages, where inexpensive software packages can OCR around 200 languages).

Though there has been an immense amount of high quality research in the document engineering field over the past two decades in both academia and industry, little of it has made it into real, deployed systems. Even after capture, the technology needed to convert the material is arcane; it requires expert users to configure it, and to develop workflows to deal with the exceptions that inevitably occur.

For the targets of this project there are really no cost-effective solutions available at this time. There are several scan-to-PDF solutions available, foremost Adobe Distiller, but they are also costly and have serious limitations (e.g., they generate low quality text and do not reflow properly). To assemble a solution an institution must procure and assemble equipment, train operators, procure several pieces of software (some of which doesn't exist), and develop exception handling and QA processes and tools. All of these require specialized skills and knowledge that is not readily available and is quite arcane. It really is beyond the scope of the average institution, and it is expensive.

Our project targets just these institutions or collections, ones with modest budgets, with material that is unique or fragile and must remain on-site, either because it is being used locally or there are restrictions on it being removed. Such institutions do not have sufficient material to justify the high set up costs of the overseas solution despite the low unit costs. In essence there are no cost effective digitization options available to the curators of small to medium collections. A capture process is needed that is fast, able to deal gently with diverse materials and resilient to operator error, paper quality, lighting variations and other factors.

Our proposed pipeline will be modular and open, and all modules will be accessible so that any particular project could use whichever parts of the pipeline it needed. As a consequence of this project it will become feasible for any collection, no matter how small or how remote, to capture and preserve their materials for a very modest outlay in terms of equipment, and by utilizing local staff or volunteers. We believe that the quality of this software solution will also make it attractive to larger operations. JSTOR, is a non-funded partner in this program and is investing in Decapod because even with its well established and highly cost effective digitization work flow, it sees this solution as an attractive path to capture singleton journal issues, isolated page replacements from libraries and to provide a means to capture so called "contributed collections", ones that are not amenable to the

normal processes and workflows. In the longer term JSTOR in particular and other archives in general could move to using a platform such as this as a primary production process.

Project Vision

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.

This is now possible thanks to the existence of well understood software and algorithmic approaches to the digitization problem and the emergence of affordable high resolution cameras. The project will deliver an out-of-the-box solution that allows local staff with modest training to easily capture their material and convert it to archive quality content suitable for deposition in online archives. The solution will deal with bound material that must be treated gently (and also, of course, single sheet material), and will trim the image down to the page boundaries and remove discolorations and other visual defects so as to deliver page images comparable to those from a flat bed scanner. Our proposed solution will accomplish the following:

- *Non-Destructive Scanning:* The system will allow the non-destructive scanning of documents, journals, and bound volumes.
- *Low Cost:* Open source software, standard laptops, consumer-grade digital cameras.
- *Competitive Quality:* When used with a high-end digital camera and good lighting, the system will be capable of generating images of quality at least as good as that obtained by Google's scanning process.
- *Portability:* All system hardware components (cameras, tripods, laptop, etc.) will fit into a small suitcase.
- *Usability by Non-Experts:* The system will require minimal operator training and be usable by non-experts such as local staff and volunteers.
- *Real-Time Scan Quality Control:* Re-scans can be expensive or impossible; real-time scan quality control catches a high fraction of capture errors while the operator still has access to the document.

Overall Solution Architecture

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. These phases are now described in more detail.

Camera-Based Document Capture

Decapod will be based on consumer grade cameras and off the shelf tripods and cradles. Page turning will be manual. This will produce a rig that can be hand carried into an archive, set up on a tabletop and can operate off battery power if necessary. It will perform adequately for A4 sized material and smaller. In order to provide the most robust dewarping and page normalization a two camera, stereo vision approach will be used. Other approaches such as structured light were considered and have some technical and cost advantages, but the project schedule and timeframe does not permit implementation of that approach. Although we suggest this particular equipment configuration for the target audience, it in no way precludes the use of other, industrial grade rigs (such as Atiz's bookscan rig) for larger projects that wish to use the software. Those rigs will not, of course, be portable or inexpensive, but Decapod does not impede the choice of alternative hardware as long as the Linux digicam libraries support it. Those libraries, as of September 2008, supported 155 brands and 1030 models.

Figure 1: A prototype scanning rig, consisting of standard tripod hardware and consumer digital cameras. The rig is portable and can be operated anywhere using a laptop computer.

A modern "prosumer" grade camera with 12 megapixels can digitize images up to A4/Letter in size and achieve the targeted quality levels of 300 dpi grayscale or 600 dpi bi-level via resolution enhancement. Smaller page sizes will naturally achieve higher resolutions, and should a higher quality be needed then a more expensive camera with a higher pixel count can be used. The proposed system will accommodate the various permutations of cameras and document sizes as part of its operation, with no special action required by the user.

For exceptionally large documents it would be necessary to use multiple images and to stitch them together. This too could be achieved using standard off the shelf hardware components, but it is felt that this project should address the most common use cases of small to medium material. As such

mosaicing and stitching will not be included in this proposal, but should such material need to be processed it is possible to simply use higher resolution cameras with no change elsewhere in the system.

The configuration chosen achieves the Decapod objectives of cost and performance. None of the existing alternative rigs meets the objectives of cost, gentle material handling and performance. Initial capture is one of the costliest steps in digital library applications. Existing projects use flat bed scanning, overhead scanning with special illumination (Zeutschel, Minolta, BookEye, I2S) and camera based scanning with special scanning rigs (Internet Archive, Kirtas and probably Google). Flatbed scanning is destructive or damaging for bound material and is very slow because of the manual page turning.

Non-destructive camera-based scanning with manual page turning is the technique used in the largest book scanning efforts and represents a good tradeoff between throughput, cost, and quality. Unfortunately, the hardware required is large and complex, and the software is not publicly available. In spite of their apparent allure, page turning robots are extremely expensive and have not demonstrated adequate reliability; all the robot projects of which we are aware or in which we have participated require skilled human supervision to correct misfeed adding to the resource overhead, cost, and complexity.

Workflow Software

The application interface will lead the users through a three-step workflow: capturing, software editing, and generate PDF. The software will guide the user through each of these phases, providing contextual tools and easy-to-understand options along the way. The user interface, designed through a user-centered process, will use language and terminology that is familiar to the user, avoiding jargon or complex image processing concepts. The application will help the user prevent mistakes and accidental data loss by providing features such as auto-save, undo, and more. In order to make it more approachable, inspiration will be taken from existing consumer-level image and document manipulation tools on the desktop and Web, such as iPhoto and Picassa. The end goal of the software is to be clear and flexible, making it easily usable by anyone with little training.

In the capture step, the user will interact with dialogs announcing the cameras are attached. From that point, the user is ready to capture single page or multi-page documents.

During book capture, scanned pages will appear in the book capture interface as they arrive from the camera. In real time, while scanning, the pages are analyzed for image quality and correct page ordering, and the interface alerts the operator to missing or duplicate pages and scan problems right away. Operators can then immediately rescan pages in order to obtain improved versions of the document, they can reorder out-of-order pages, they can suppress the warning (e.g., if the defect is in the source material), or they may be able to invoke an interactive editor that allows them to make corrections to exceptional pages on screen (see the Appendix for additional screen mock ups).

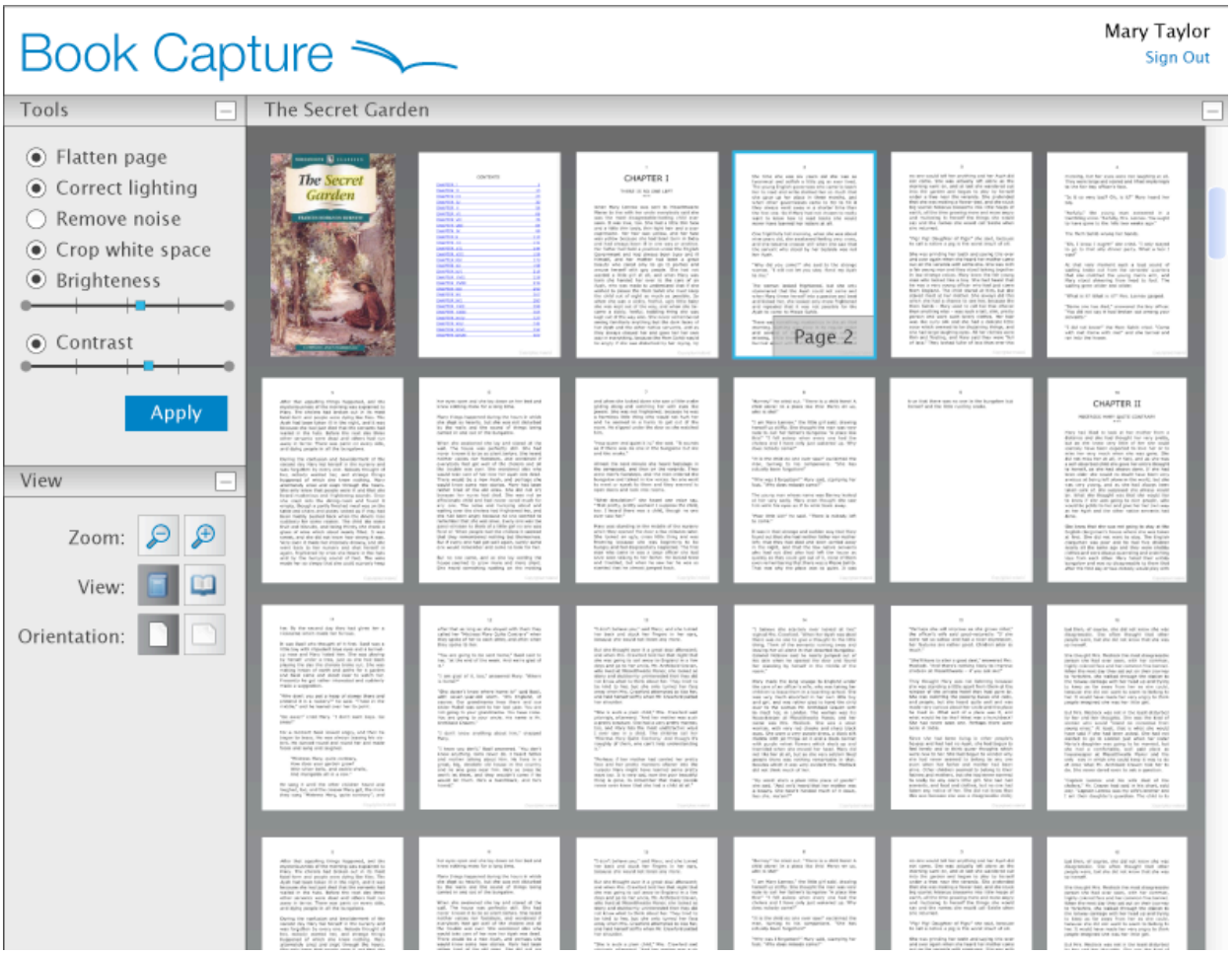


Figure 2:A mock-up of multi-pages scanned and editable in the sort and edit screen.

The editable features will include correction for the OCR, though it should be stressed that once initial capture is verified all other steps can take place offline, to avoid slowing down the onsite operations. We will leverage lessons learned about web-based OCR correction from Mechanical Turk at Amazon and reCAPTCHA from CMU that demonstrates how to leverage people to complete a mechanical task. This will tie in with the community that forms around this application -- a community that is already working with these tools that we can tap into.

Output Formats

One of the primary purposes of digitization is to make documents available for on-line reading. Page images are around 8 mega-pixels for letter/A4, and are poorly suited for web viewing because of both download size and physical viewing size. An approach is needed that allows high fidelity, compact representations of the page. The final outputs from our solution will be archive quality TIFFs and compact, searchable and reflowable PDFs that are visually faithful to the original material in terms of typography, layout and illustrations. This dual output will allow us to effectively display the material from a web site with compact representations that download and render quickly.

TIFFs
Our system will produce TIFFs with page images at up to 600 dpi resolution for letter/A4-sized bi-tonal inputs. These TIFFs can be readily injected into any standard digitization workflow, all of which have facility to have "watch directories" into which images can be deposited (In particular the Biblioteca

Alexandrina supports this facility). Those extended workflows can apply other conversions and manipulations such as high-accuracy OCR, meta-data entry and institution specific formatting.

PDFs

Decapod will generate high quality PDFs by applying token (glyph) clustering and vectorization to create document specific embedded fonts, and layout analysis to derive reflow. This will result in compact, visually faithful, searchable PDFs that can be reflowed. Our system's output will support original layout and typeface and yet allow the client to adapt the images as necessary, for example by reflowing or using text to speech. We have chosen PDF as the default output format in this release because it is the only widely used format that currently supports appropriate functionality. That is, given imperfect OCR, very imperfect font and fontsize recognition PDF is the only current format that can give a perfect "100% accurate" representation in a widely used format. It is also compatible with the sort of "digital library in a box" solutions such as Scribd, Fedora, DSpace, and eprints, and with mobile solutions. Future HTML standards will provide appropriate functionality and the architecture of our solution allows for support of such alternate formats. The PDFs (in particular the PDF/A profile with embedded fonts) will contain a visually accurate representation of the original document content with a resolution equivalent to 600 dpi binary images (for A4/letter-sized inputs). Encoding inside the PDF will be analogous to a mixed raster content (MRC) format, with token-based compression for text regions, and separate compression and representation of embedded images. The PDFs will be structured and tagged to facilitate native re-flowing and the glyph tokens will be represented as fonts to allow for searching.

Note that we have chosen 600 dpi as our project goal because it is the defacto standard for archival of bilevel (black and white) content. Although the proposed solution will readily allow higher resolutions by using more expensive cameras there is good reasons why more is not necessarily better. File size increases very rapidly as the resolution increases. According to studies by Ray Smith (the author of the Tesseract system) while he was at HP and later at Caere, OCR quality does not increase markedly above 300 dpi, and for Latin texts not at all above 400 dpi. For Asian texts there was some residual increase in quality up to 600 dpi, but not much beyond. This reflects the fact that the texts were designed for human viewing, they do not include any detail that is not usable by the human visual system, and anything above 600dpi is imperceptible to the unaugmented human eye. Moreover, the detail reaching the sensors of the cameras is limited by the optics (technically the MTF or modulation transfer function) and so increasing the pixel count of the sensors doesn't help beyond a certain level, because the physics will guarantee a certain blurriness, and to further exacerbate it, as the resolution increases the size of each pixel decreases, making it intrinsically noisier, and less light reaches it as well, again increasing the noise level. All of these considerations together conspire to deliver little or no improvement in quality above the indicated level.

It is also important to allow for an immediately useful final output without the target users having to build an extended digitization workflow (TIFFs are an intermediate and preservation format). PDF was chosen because of its widespread acceptance and because at this time it is the only format able to handle the custom fonts, searchability and reflow tagging. These properties give it the flexibility to be used in diverse ways, including mobile devices and text to speech, whilst still perfectly representing the original document. That being said, if another format is needed then a new file output stage could be used that would apply the analysis results to the generation of that output format. By virtue of the synthetic font generation process, the PDF conversion process will result in documents that are close in quality to born-digital documents.

The requirements on the system's output are:

- Visual fidelity takes precedence over any other aspect of the conversion; that is, the PDF should look nearly identical to the original scanned document.
- Tokenization should result in significant memory savings from representing the character shapes.
- If the original document layout could be analyzed with high confidence, the resulting document should contain PDF tags that permit display as a reflowable single column document.

This will be superior to current existing openly available technology for converting scanned documents into PDFs or other format. Since displays don't have the resolution and size of printed documents, documents with layouts other than simple one-column layouts require significant scrolling when displayed in page image format. Moreover, image-based formats require considerable amounts of memory and compute power to display. For example, existing JSTOR PDFs will display on the Illiad e-book reader, but they are unreadable, take literally minutes to refresh and make very poor use of screen space. We expect mobiles to continue to trade power consumption and hence battery life against CPU power and so do not expect dramatic improvements. It is also true that OCR technologies are not yet sufficiently reliable to permit fully automatic conversion of even most scanned document collections into high-quality, logically marked up text; and we do not expect early solutions to unknown symbols, mixed-script documents, separation of text and images, or the intractability of diagrams, tables, and figures.

Approaches to avoiding these issues can be found in the literature, and previous work by Breuel et al is described in an appendix to this proposal.

The particular approach that will be used by Decapod derives from the general principles used in Mixed Raster Content (MRC) to separately store text content from other content such as images, figures, tables, drawings, and math formulae (e.g., Keysers et al, 2007). Shape clustering will then merge similar or identical glyphs, much as is done in such formats as JBIG2 and DjVu. These glyphs will be converted from raster to vector using the techniques such as those in the open source potrace utility, and finally that vector representation will then be stored as an embedded, scalable font in the PDF file. The reflow analysis results will be used to embed reading order/reflow tags into the PDF.

This approach will yield far more compact and legible PDFs than the approach often used to represent scanned images where a whole page image overlays OCR.

In addition to tagged PDF and PDF/A, Decapod will also be capable of delivering TIFF output of both the raw and dewarped page images and OCR output in the hOCR (HTML) format.

An additional format we are considering, a format that is functionally equivalent to Tagged PDF is HTML with hOCR and embedded fonts. We may optionally provide this output as well; however, HTML font embedding is not mature technology yet and there are few compatible viewers available. Therefore, tagged PDF remains the format of choice for display.

The different output formats are generated at different stages:

- TIFF output is the result of camera-based capture, dewarping, and operator scan error correction. It is the equivalent of a raw scanned image from a flatbed scanner. It is essentially an intermediate result that is not useful and is essentially the digital archive equivalent of original source material.
- hOCR output is the result of processing the TIFF image with an OCR system, by default OCRopus. The hOCR result contains text and embedded images, but any OCR errors are visible, and will be in a system font, not the original.
- Tagged PDF and PDF/A output is the result of processing the TIFF output with character shape clustering, text/image segmentation, font generation, and PDF generation.

A comparison of the different output formats is shown in the following Table 1.

	TIFF (Archive Standards)	**hOCR**	**tagged PDF or PDF/A**	**hOCR+embedded HTML fonts**
preserves fonts	yes	no	yes	yes
preserves layout	yes	partial	yes	yes
contains OCR text output	no	no	yes	yes
looks like original	yes	partial	yes	yes
text/image segmentation	no	yes	yes	yes
cluster-based resolution enhancement	no	no	yes	yes
token-based compression	no (possible with JBIG2)	n/a	yes	yes
visual fidelity despite OCR errors	yes	no	yes	yes
client support	good	excellent	excellent	limited
relative compression	1	4	3	2
viewable on small screen devices	very poor	excellent (except OCR errors)	excellent	excellent (except for lack of embedded font support)

Table 1:A comparison of the possible different output formats from the Decapod system.

Project Execution

Project Organization and Collaboration

Co-Leads

The project will be co-led by Thomas Breuel of the Image Understanding and Pattern Recognition (IUPR) research group within the Deutsches Forshungszentrum for Künstliche Intelligenz (DFKI) and Jutta Treviranus as the lead in the Fluid Project and Director of the Adaptive Technology Resource Centre of the University of Toronto. These two individuals will be responsible for leading the two distinct but closely interrelated and interdependent bodies of work required to complete the project:

1) the document services or the applications and technical solutions for camera-based document capture, OCR and PDF document generation; and

2) community building, user requirements gathering, UX design and development, workflow application development and usability testing.

Steering Committee
The co-leads will be supported by a steering committee consisting of:
o Thomas Breuel

o Jutta Treviranus
o John Burns of JSTOR
o the IUPR technical lead, and
o Jess Mitchell as Project Coordinator

The steering committee will monitor and steer project progress, coordinate collaboration activities, plan for and coordinate sustainability and coordinate communication and dissemination of results to the community.

The Steering Committee will meet monthly at minimum by teleconference or face-to-face meeting, as well as on an as-needed basis.

Work Plan - Targets

Our work plan is designed to deliver a fully working and deployable camera-based book capture and conversion system within two years. We have structured the milestones and deliverables in a way that ensures that the project partners start working and integrating their software immediately at the project start and always maintain a working system; the system is then updated incrementally with the inclusion of new functionality and improved quality and performance. User feedback and usability studies will guide the development process.

- **Year 0.5: pre-alpha release**
  - initial technology demo with affine dewarping, simple UI, simple PDF generation
  - collaboration and integration across the participants
  - establish performance baseline and testing framework
- **Year 1: alpha release**
  - basic book capture with stereo dewarping, image processing, basic UI
  - Distiller-like PDF generation
  - first useful release, basis for user testing
- **Year 1.5: beta release**
  - simplified stereo calibration
  - MRC and font clustering for improved PDF output
  - incorporation of user feedback from alpha release
- **Year 2: release 1.0**
  - deployable book scanning system using stereo, calibration, and tagged PDF generation
  - improved user interaction based on testing, feedback
  - font clustering and resolution enhancement for better PDF output
  - improved stereo calibration based on user testing

For a possible follow-on during Year 3, we would focus on improving on both capture quality and user interaction by incorporating a digital pico-projector into the system (see below Beyond Year 2).

Software Packaging and Delivery to the Community

Project deliverables will be available in the following forms:

1) a bootable ISO cd image containing an install with the Decapod software preinstalled
This makes using the software as simple as booting the ISO and following the instructions for setting up the camera; this allows end users to deploy the software with no system management or Linux experience.

2) a tar file and/or Debian package that can be installed and used directly on a standard Ubuntu system
This is a more convenient delivery format for existing Linux users.

3) versions available from the (public) version control systems and a source tar file containing all the sources for that milestone

This is aimed at users interested in contributing to the source code and development.

(Note: The software will not contain significant platform dependencies and could be packaged for Windows and OS X as well (with additional resources). However, Decapod relies on significant external software (image I/O libraries) and hardware drivers and interfaces. Based on our experience, we believe that Ubuntu installations are a simpler choice even for institutions that have no prior experience with Ubuntu.)

Community Building

Every academic library has unique holdings, and making these all available off-campus is impractical. Digital copies solve many of the distribution challenges these libraries experience. We have received positive, enthusiastic responses from informal conversations with colleagues at Cornell and NYU about this project. We believe that many higher educational institutional subscribers to JSTOR would use Decapod for ILL and digitization of selected collections. Of the more than 1800 JSTOR subscribing institutions in the US, UK and Canada, we would expect more than 900 to deploy such a system in the two years following its availability, especially when considering a proposed outreach program by JSTOR at ALA and other community forums. Assuming that these numbers scale in the same way as the JSTOR user base we would expect more than double the uptake in the non-English speaking international community. For libraries the ease and simplicity of Decapod for ILL would make it an immediately compelling solution, with digitization being an added attraction.

The following quote from Teresa Ehling, Director of the Center for Innovative Publishing at the Cornell Libraries illustrates the point.

> "I have read with interest the draft proposal submitted to the Mellon Foundation by JSTOR for a lightweight digitization system that incorporates COTS hardware components coupled with state-of-the-art software technology (Decapod). This solution promises academic libraries a cost-effective but technically sophisticated solution to a 'point and place of need' digitization program. Cornell University Library has been exploring the possibility of implementing an on-site digitization on demand initiative that would address both local as well as external requests for monographic and ms. material held in our high-density off-site storage facility, where approximately a third of Cornell's title holding now reside, and in our rare book and manuscript division. Possible applications of a cost-efficient digitization solution would be fulfillment of ILL requests for copyright unencumbered material via the BorrowDirect network. The Cornell Library would be pleased to be considered a candidate for beta-testing of the Decapod system during either the alpha or beta phases of the project."

In addition to being attractive to academic libraries, Decapod would also serve as a high quality, low cost solution for the more than 15,000 museums in the US. Decapod will fill a much-needed role for countless small collections of content worthy of digital preservation and distribution. The participating institutions on this grant will work to make Decapod available, known, and tested with a number of representatives of these communities.

Integration with Digital Libraries and Workflows

The Decapod system fulfills an analogous task in digital libraries as traditional scanner and PDF conversion software: it acquires page images from an imaging device and combines and stores them in PDF files.

These PDF files can then be input in digital library workflows for further document management and document processing applications: they can be archived, indexed and published, and used as input for more advanced content analysis, annotation, and recognition.

The difference to traditional page image capture applications lies in the nature of the input, the support for interactive quality control, and the nature of the PDF output. In terms of input, Decapod will work with digital cameras and low cost scanning rigs. For interactive quality control, Decapod will use a web and AJAX-based interface developed by user interface experts and optimized through user testing for use by non-expert users. Finally, the PDF output that Decapod does not merely encapsulate the scanned images and some approximate, searchable text, but also supports reflowing, token-based compression, resolution enhancement, and reconstructed fonts; these additional features enable and improve functionality of digital library systems, such as on-screen display and archiving.

Digital library systems will also be able to reuse Decapod components in other ways. For example, a digital library using flatbed scanner-based document capture can offer users the ability to edit and correct pages using Decapod's UI, and can invoke Decapod's PDF generation to convert its scanned images into PDF/A, including reflow information and reconstructed fonts. Integration of Decapod in this way into existing digital libraries will generally be easily accomplished using Decapod's REST interfaces.

Our system will address the problem of going from typical physical books to archival-quality PDF/A documents (optionally also outputting TIFF and/or HTML/hOCR). It will address that problem beginning-to-end. There are, however, many digital library problems we aren't addressing as part of this proposal (search, annotation, metadata capture, etc.). Institutions that already have those solutions in place can integrate Decapod into their existing workflow, simplifying a heretofore complex and costly step in capturing materials. Institutions that don't have such a solution will benefit from the existence of Decapod, a solution that will raise the possibility of solving one step of the larger digitization question confronting many organizations. So, while some groups may still have to solve the larger issue of creating digital libraries, our project aims to reduce costs overall and reduce mistakes in the digitization step. This makes it a good overall solution whether the institution wants to integrate it into an existing solution or simply start with it as a solution.

Software Platform
(OCRopus - Apache 2 + Fluid - ECL 2) fully compatible, open source solution

We will be building the imaging components of the project on top of the OCRopus platform. OCRopus is an Apache 2 licensed open source project. OCRopus is not an OCR system in the traditional sense of a monolithic desktop application, rather it is a scriptable toolbox of document analysis and image processing functions: a full pipeline. OCRopus does have the four major components of a standard OCR process: (1) image cleanup and binarization, (2) page segmentation and layout analysis, (3) text line recognition, and (4) statistical language modeling. The focus of this process is to obtain high quality textual output. Decapod will reuse components already present in OCRopus to achieve a high level of quality. These components will be enhanced as part of other projects as well as being refined within this project, benefiting the overall application.

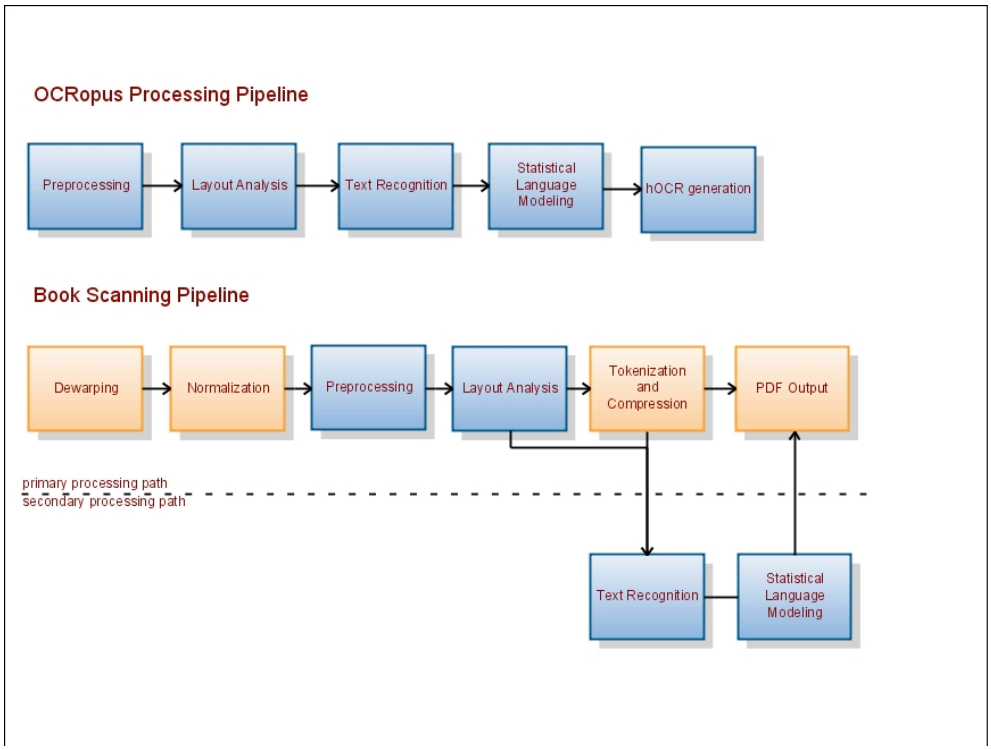


Figure 3:The relationship between the OCRopus system and the additional modules to be developed as part of this project. The first pipeline is the standard OCRopus processing pipeline. The second pipeline is the Decapod book scanning pipeline. Existing OCRopus components are in blue, components to be developed as part of this project are in yellow. Please see the text for a detailed explanation.

As part of Decapod we will develop four new modules: camera-based image dewarping, image normalization, document image tokenization and compression, and PDF output generation. The relationship between the new modules and the existing modules is shown in Figure 3. These new modules will allow us to accomplish two book scanning pipelines. One pipeline is language independent and contains the key components for generating archival quality PDF output. A second pipeline uses the OCRopus text recognition and statistical language modeling modules to generate searchable text. Both pipelines will transform the original camera-captured document images into high quality PDF output.

No major enhancements or modifications to the existing OCRopus modules are planned under this proposal, although bug fixes and smaller modifications will be carried out as needed.

Collaboration and Integration between Project Partners

Since the project will be carried out at multiple sites and by multiple organizations, collaboration and integration are important issues. All participants have experience with distributed software development, and we will be relying on standard collaborative software development tools (Subversion project hosting, web-based issue tracker, IRC channels, wiki, website, etc.).

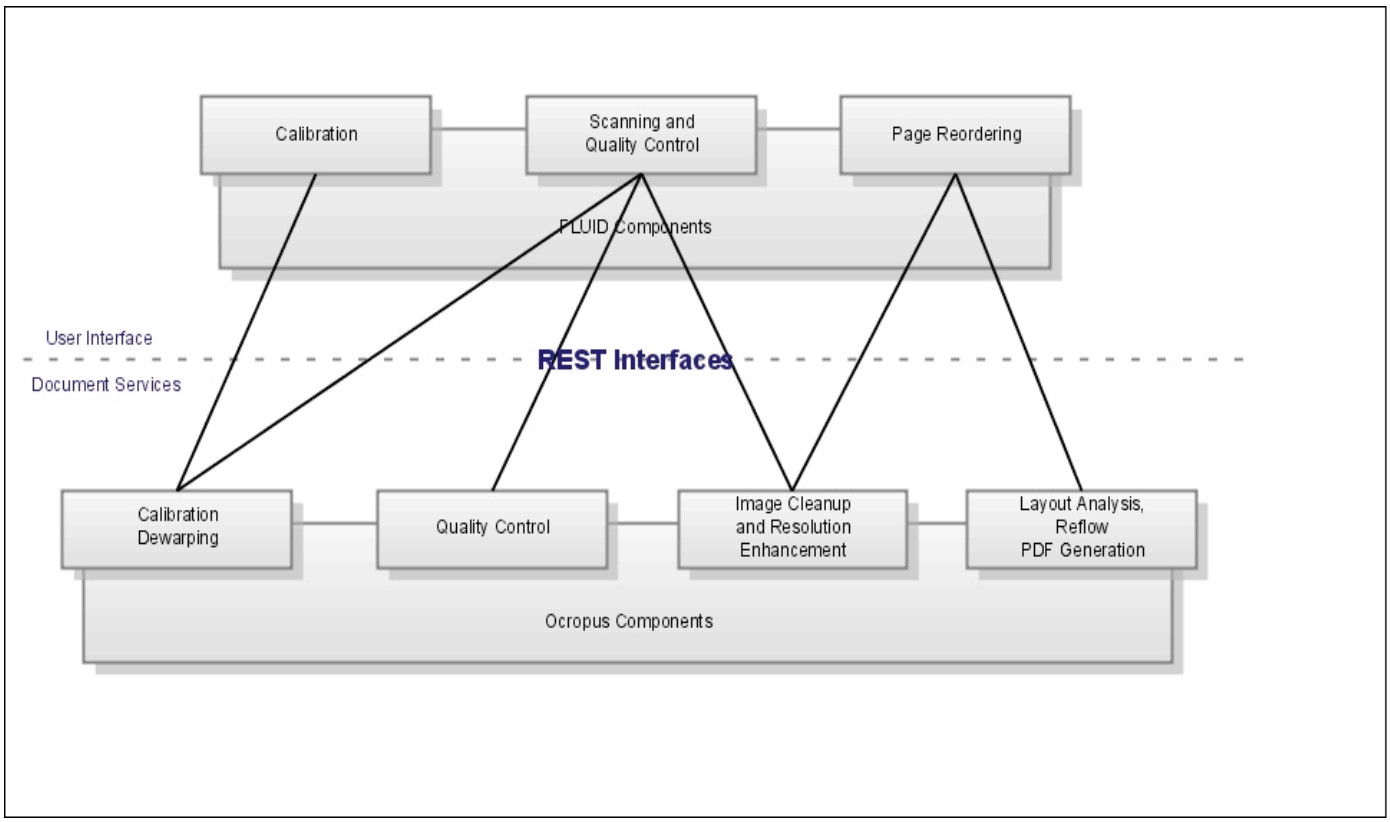


Figure 4:Architecture of the system. The user interface will be browser-based and rely on AJAX and FLUID components (Fluid carries an ECL-2 license). The back-end services will be based on OCRopus components. The user interface and back-end services communicate using a REST-based interface (with the option of streaming video for real-time adjustments during calibration).

The entire system will consist of two main parts: a browser-based user interface based on AJAX, FLUID components, and a standard light-weight web framework (TurboGears, Django), and a back-end written in C++ and Lua, based on OCRopus components. The two parts of the system will communicate via REST interfaces. All groups have extensive experience with REST-based interfaces; they are already used as part of both OCRopus and FLUID.

At the beginning of the project, IUPR will create mock implementations of the services and REST interfaces (e.g., an image dewarping service that simply stamps the string "dewarped" on top of a given image). Mock implementations are easy to create and allow UI development, system integration, and user testing to start immediately. Toronto will begin development of the UI and while development is going on we will collectively find appropriate users to test the system on. During the course of the project, the mock implementations will then be replaced by fully functional versions as the development progresses.

This architecture largely separates usability and front-end issues from back-end functionality. Furthermore, it allows back-end and front-end developer to communicate usability-related requirements in terms of changes to the REST protocols (e.g., "in order to help users avoid a problem that we observed, we require an additional REST call that gives us information about ...").

System integration, release management, server management, and integration testing will be carried out by engineers at IUPR.

Our institutions will focus on iterating often with each other, ensuring that we are in step and that the design and development reflect each others work and come together for the final solution. We will work together remotely and will have on-site working meeting as well to ensure our work is integrated.

Project Plan and Schedule

Year 1

At the end of Year 1, we will have a system that can be released to the community and that will be useful for actual document capture. We will be using early feedback from the community and user testing to drive the development of improved calibration methods and the user interface. Sample page images and documents from the community will also be used to prioritize image cleanup and other functionality. In Year 1 we will be working to build and stabilize an application framework that lays out the necessary connections for integrating the software layer with the OCRpus layer of server-based activities. The Fluid components will provide rich, client-side user interfaces for the various activities within the application workflow. The client-side components and Fluid framework will be integrated into the server-side application via a RESTful web tier. Communication between client and server will be accomplished using standard HTTP protocols and AJAX callbacks. Components will cooperate with the other application layers to provide a rich, flexible user experience for the application. The application service layer will be responsible for orchestrating the user's requests to the processing pipeline, from capture through to OCR to final PDF output.

Year 2

During Year 2, the focus is on improving capture quality, reducing the frequency of capture errors, and making the entire system easier to use.

Calibration of stereo capture systems is widely used in many computer vision applications and there are some standard toolboxes; operators can carry it out with a moderate amount of training. However, both to improve quality and reliability of overall capture, we would like to simplify calibration further, to the point where the entire stereo capture setup can be calibrated for both stereo and photometry with a single calibration target captured prior to book scanning. Part of calibration will also be to give the user feedback and instructions about the setup if the cameras are grossly misaligned or if the illumination is inadequate. This work will take into account the feedback we receive from users based on the system released at the end of Year 1.

In addition to improving calibration, we will also be improving the quality of the stereo-based capture by implementing model-based post-processing methods for textual content. Although stereo-based dewarping of pages can produce very accurate dewarped page images, the human eye is highly sensitive to even slight misalignments in linear structures in images. Model-based post-processing identifies such linear structures and corrects the misalignments.

Finally, we will begin exploring additional means of streamlining and improving the capture workflow, as well as performing extensive user testing of the capture system. One area of particular interest is the use of projectors to give users direct feedback on scanning area, scanning problems, and user interactions. We will be prototyping the use of projectors for giving users feedback on the scanning process, as well as for projecting light patterns for improved stereo matching as during Year 2 (the complete integration of projector-based illumination will be carried out in Year 3).

Following stereo processing, we will be focusing on several areas to improve the quality and compression of the generated PDF output. First, we will be incorporating text/image segmentation into the PDF generation, allowing different image processing pipelines and compression methods to be used for textual and image content (this is similar to mixed raster content approaches in pure image compression). Second, for the textual content, we will be performing shape clustering and font generation. Shape clustering groups together similar shapes and is useful for several reasons. First, shape clustering can be used for resolution enhancement and improved appearance of the scanned text. Second, shape clustering can be used for improved compression ratios by implementing token

based compression. Third, shape clustering improves OCR error rates. The output from text/image segmentation and shape clustering will be used to construct improved PDF files; in particular, shape clusters will be represented as embedded PDF fonts, and OCR will be used to assign the true Unicode character value to each of the constructed glyphs. The text/image segmentaion, font clustering, and PDF generation steps will be fully automated and require no user interaction.

In addition, we will be delivering tagged PDF output support. Tagged PDF output contains the instructions necessary for PDF viewers to "reflow" documents. Reflowing enables content originally designed for, say, letter sized output to be viewed on smaller screens or mobile devices. Reflowing is particularly important for educational uses, because students often access content from low-end devices with low screen resolution. Reflowing also enables mobile access to documents, such as reading documents on iPods. Reflowing support in the PDF output during Year 2 will be based on the existing layout analysis code contained in the OCRopus OCR system.

In terms of user interface, we will be concentrating on incorporating the results of user testing from the first year of research and development into building a user-centered interface. We will be focused on improving the existing functionality while ironing out links between our application, our components, and existing community applications that add value to our application (e.g. reCAPTCHA). Year Two will also continue the work of finalizing an application framework ensuring that future work will drop in, like modules, to add functionality easily. The up-front cost of this development will pay off in the long-term sustainability of the software.

Beyond Year 2

Year 1 and 2 deliverables of the system are based on technologies and user interactions that are well understood. In the case of camera-based dewarping, we have built prototype systems as part of prior research projects. In the case of user interactions, the interactions we are proposing are similar to those used in existing large-scale book scanning projects. In terms of PDF generation, we are using techniques from OCR and layout analysis.

For a possible follow-on during Year 3, we would focus on improving capture quality and user interaction by incorporating a digital projector into the system. In addition, we would incorporate improvements to the readability and usability of the PDF output on small screen devices.

For scanning, incorporation of a digital projector has a number of potential advantages:

- mis-scan alerts and user interface elements can be projected directly onto the work surface, reducing the need for operators to shift attention to a screen; this is expected to improve operator throughput and reduce uncorrected mis-scans
- projection of instructions and guides during calibration of the scanning setup potentially simplifies setup times and reduces the need for operator training
- structured lighting provides an additional source of 3D information, even on low-contrast page images, potentially improving de-warping accuracy and robustness further
- control over the light source permits improved dynamic range and the reduction of artifacts such as specular reflections on glossy paper
- some combinations of light source and camera permit limited forms of multispectral imaging and improved color correction

Just as with prior deliverables, we would deliver a useful system within the first six months and then improve the system and incorporate user feedback over the subsequent six months.
A second approach to improving operator efficiency that we would work on beyond year 2 is the camera-based analysis of operator actions, in particular automated triggering of scans when operators have changed pages. This avoids perhaps the most common source of mis-scans, namely operators

triggering document image capture while the page is still in motion. The interactions and gestures will be tested and optimized in collaboration with end users. The application will be tested throughout its development to ensure the end result is a solid, accessible, usable application that makes sense of the capture, edit, and production workflow.

For improving the PDF output, we would implement various forms of post-processing to the layout analysis output from the OCR system, with the goal of improving readability and usability of the resulting PDF files on mobile and low resolution devices, as well as improving accessibility. This involves various forms of post-processing that are not usually necessary for other OCR applications, such as placing floating text and image elements differently within reflowed output, marking optional hyphenation and detecting and removing extra whitespace used for justification.

Localization and Internationalization

For localization of the application and user interface, we will be relying on the internationalization features provided by established web presentation frameworks on both the client and server. All user-visible strings will be contained within property files to ensure that the translation process is easy. The server-side application will use locale resolution, provided by frameworks such as Spring MVC, to ensure that the correct language bundle is delivered to the end user. When necessary, HTML templates can also be customized for localization purposes, allowing translators to significantly adjust a page's context and structure. Client-side user interfaces will similarly render their strings in a locale-sensitive way, cooperating with the server to ensure a fully localizable experience.

The primary processing pipeline has no strong dependencies on languages or scripts and should work without modification for many languages and scripts: stereo-based page de-warping yields good results for arbitrary document images and we have already demonstrated the ability of our document cleanup, page segmentation, and layout analysis modules to operate on many different languages. The quality and performance of token-based compression and shape-clustering depends intrinsically on the scripts being used; scripts with many ligatures and few isolatable tokens (Urdu, Devanagari) necessarily yield lower compression and provide less opportunity for shape clustering.

Performing OCR is not necessary for the primary goal of the project (providing compact, high-quality, accurate visual representations of the documents; supporting on-line reading), but is important for supporting search of documents. OCRopus is designed from the ground up as an OCR engine for all known languages and scripts, and efforts are already underway to support additional scripts and languages. New scripts are supported within OCRopus by providing new text line recognizers, and new languages are supported by providing new language models. Extensibility via these mechanisms has already been demonstrated. Currently, efforts are underway to create Indic and Arabic script text line recognizers, and to create language models for all major European languages.

Comparisons and competition

The rationale for this proposal is the lack of viable digitization options is blocking the digitization and dissemination of valuable scholarly resources. The parties to project are familiar with and have been involved with many digitization projects and are keenly aware of the lack of a simple, robust solution at any price. The alternatives that are frequently mentioned fail to meet the requirements that we feel are necessary. In our various capacities and positions we have evaluated possibilities for projects both large and small, and none meet the requirements of affordable, reliable, simple and robust. One of us worked for a very large commercial company in recent times with an enormous book scanning program, during the course of which we evaluated ALL of the available commercial possibilities for capture with unskilled labor and high quality. The best solution, including robots, rigs etc was guillotining and flatbed scanners overseas. None of the page-turners avoided book damage,

There are a number of existing commercially available solutions for book scanning, including those from Atiz, BookEye, I2S, Kirtas etc. Here is a brief overview of our major findings:

- Atiz's Snapter is a single camera (monocular) document dewarping solution intended for handheld document capture. We have extensive experience with single camera dewarping ourselves and our software is used in some commercial applications. However, single camera dewarping has inherent limitations in terms of both quality and reliability because a single camera image does not contain enough information to unambiguously flatten arbitrary documents. Therefore, we are proposing the use of stereo dewarping in this proposal (this includes monocular dewarping techniques as an integral part).
- We have evaluated I2S software and Atiz's snapter software over the past year. It is our conclusion that both products require extensive training and imaging expertise to be used effectively. The goal of our proposal is to deliver software that requires little training or expertise.
- The Kirtas solution is an ultra high end, costly solution that is not at all portable and hence not comparable to the proposed solution.
- Atiz offers a hardware book scanning solution that uses off-the-shelf digital cameras and mechanical page flattening. Where available, Atiz hardware makes an excellent complement to the software we are developing. Although the Atiz hardware eliminates the need for non-linear page dewarping, the rest of our capture pipeline (affine dewarping, image and resolution enhancement, PDF generation) is fully applicable to Atiz scanning rigs and provides capabilities not available from Atiz.
- The Internet Archive offers book scanning as a service, but their hardware/software solution is not currently available separately. In our experience, libraries and digitization efforts often wish to have scanning hardware available and under their own control, and they wish to retain full control of the digital images and copyrights; outsourcing to the Internet Archive therefore does not meet their needs. The Internet Archive hardware platform is similar to the Atiz hardware platform (albeit larger and heavier) and uses mechanical page flattening; our software can therefore be used with the Internet Archive hardware platform in a way similar to the Atiz software. The Internet Archive also has a software system for scanning and OCR processing, but this software is not currently available in any form (it is not an open source project and it has not been packaged), and depends on a commercial OCR system for its final output. The primary functions of the Internet Archive software appear to be image processing of flattened document pages, an operator user interface, management of large collections of scanned documents, and distribution of OCR jobs over a network to OCR processing stations. We believe that our system and the Internet Archive are largely complementary (e.g., the PDF generation/OCR combination that we offer would allow the Internet Archive to replace their reliance on commercial OCR systems and provide a more widely alternative to their DjVu format), and in areas where there is some overlap (image processing, operator interface), we believe that the software developed as part of our project could be a useful addition to the Internet Archive software.
- Based on the appearance of occasional scan failures in Google Books, Google is using a camera-based book scanning solution with some form of dewarping; the exact method used and the hardware setup remain a trade secret. Google's hardware and software solutions are not available outside Google, and their output appears to consist only of text-backed images to support on-line search (not tagged PDF or equivalent reflowable formats).

Targets and Performance Metrics

Relevant metrics for a book scanning solution are:

- functionality and features
- cost of hardware and software
- size of hardware
- setup time for first time user and experienced user
- operator throughput
- success rate of real-time scanning error detection
- output quality and resolution
- maximum document size

We will be measuring our solution against these metrics, comparing our product with existing products as well as existing standards, practices, and user expectations (through our user testing). Decapod wants to deliver the highest quality at the lowest cost (both hardware and person-training time). We will strive to balance those requirements against existing solutions in the community by engaging with the provisioners of the existing solutions. In those collaborations we may find that Decapod solutions, because they are open and accessible, becoming desirable to a much broader community.

In addition to engaging the community around document digitization, we will engage the eventual users of this system to help us iterate on our end-user solution to ensure it is easy-to-use and functionally fits the tasks needed to capture their materials. User testing throughout development will help assure that our product is thorough, easy to use, and appropriate for the intended users.

Sustainability Plan

Sustainability is an important aspect of such a project: we want to ensure that the investment in the development of the software results in the target user group actually being able to use the software to solve their problems. We are addressing sustainability in a number of ways:

1. The system will be developed under the Apache 2 open source license, insuring that the software will remain available to anybody who wishes to use it. The Apache 2 license also permits commercial use of the software, leaving open the option of commercial maintenance of the software.

2. The newly developed components will become part of the OCRopus OCR and document analysis system and will be maintained as part of that system; OCRopus is also licensed under the Apache 2 license. Furthermore, the components are of interest for a number of other tasks (e.g., document readers for the blind, spam detection, etc.), meaning that there will be a larger community sharing in the continued development and maintenance of these modules. Documentation will be available in the form of open, Google-hosted wikis associated with the project, as well as downloadable documentation and video tutorials.

3. The proposed system reuses several OCRopus components, and these components will be extended to cover other scripts, languages, and document types under other research contracts and through community involvement.

4. The system will be developed from the start with involvement of end users and experts in digital library applications and document conversion; their feedback will be used to improve the sustainability of the system, in addition to usability and functionality.

The Fluid Project, through the Adaptive Technology Resource Centre at the University of Toronto, brings the skills and experience of user experience designers and UI developers to the project. Fluid will provide interaction design resources, including participatory user research, iterative user testing, and usability expertise. The application user interface will be built using the Fluid framework and component library.

JSTOR brings many years experience in digitization workflows and document understanding. They will manage a pilot, will provide hosting for the final solution package and maintain a list of compatible hardware.

Biographies of Principals

Thomas Breuel

Thomas Breuel is professor of computer science at the Technical University of Kaiserslautern Computer Science Department, head of the Image Understanding and Pattern Recognition (IUPR) research group at the DFKI, and a consultant in Palo Alto, CA, USA. His research group works in the areas of image understanding, document imaging, computer vision, and pattern recognition. Previously, he was a researcher at Xerox PARC, the IBM Almaden Research Center, IDIAP, Switzerland, as well as a consultant to the US Bureau of the Census. He is an alumnus of the Massachusetts Institute of Technology and Harvard University.

Jutta Treviranus

Jutta Treviranus is the Principle Investigator. Jutta will apply her extensive experience and expertise in directing large multi-partner, multi-sector projects as well as her expertise in inclusive design. Jutta Treviranus established and directs the Adaptive Technology Resource Centre (ATRC) at the University of Toronto, an internationally recognized centre of expertise on barrier-free access to information technology. She has more than 25 years of experience in the field of access technology and inclusive design. Jutta has led a large number of national and international multi-partner research networks (including The Inclusive Learning Exchange (TILE), the Canadian Network for Inclusive Cultural Exchange, the Network for Inclusive Distance Education, CulturAll, Stretch and the Barrierfree project), that have led to a range of broadly implemented technical innovations that support inclusion. She has published in many areas related to inclusive design. She is chair of the Web Access Initiative, W3C, Authoring Tool Working Group, chair of the IMS AccessForAll Specification Working Groups, Project Editor within ISO/IEC JTC1 SC36, as well as a member of a number of key advisory panels and task forces relevant to IT policy, strategy and design. Among the many awards received by the ATRC is the American Foundation for the Blind Access Award (1998), the Trophee de Libre for Open Source Development, and the Dr. Dayton M. Forman Memorial Award. Jutta holds faculty appointments in the Faculty of Information Studies, the Faculty of Medicine, and the Knowledge Media Design Institute, at the University of Toronto.

Appendix 1

References

Document Capture using Stereo Vision Adrian Ulges, Christoph H. Lampert, Thomas M. Breuel ACM Symposium on Document Engineering, pages 198-200

Document Image Dewarping Contest Faisal Shafait, Thomas M. Breuel 2nd Int. Workshop on Camera-Based Document Analysis and Recognition (CBDAR)

Finding lines under bounded error T.M. Breuel Pattern Recognition 29(1), pages 167-178

Oblivious Document Capture and Real-Time Retrieval Christoph H. Lampert, Tim Braun, Adrian Ulges, Daniel Keysers, Thomas M. Breuel International Workshop on Camera Based Document Analysis and Recognition (CBDAR), pages 79-86

Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral Images Faisal Shafait, Daniel Keysers, Thomas M. Breuel
Document Recognition and Retrieval XV

R.N. Ascher, G. Nagy: "Means for Achieving High Degree of Compaction on Scan-Digitized Printed Text", IEEE Transactions on Computers, 1974, pp.1174-1179

L. Bottou, P. Haffner, P. G. Howard, P. Simard, Y. Bengio, Y. LeCun: "High Quality Document Image Compression with DjVu", Journal of Electronic Imaging 1998

T. M. Breuel, W. Jannsen, K. Popat, H. S. Baird: "Paper to PDA". ICPR 2002.

T. M. Breuel, "The hOCR Microformat for OCR Workflow and Results", ICDAR 2007, pp.1063-1067

D. Keysers, F. Shafait, T. M. Breuel, "Document Image Zone Classification - A Simple High-Performance Approach", VISAPP 2007, pp. 44-51

R. de Queiroz, R. Buckley, M. Xu, "Mixed Raster Content (MRC) Model for Compound Image Compression", SPIE Visual communications and image processing 1999, vol. 3653 (2), pp. 1106-1117

Joint Bi-Level Image Experts Group (JBIG) Committee. Information technology coded representation of picture and audio information lossy/lossless coding of bi-level images. Technical Report 14492 FDC, ISO/IEC, July 1999.

Bass, Len, Bonnie E. John, Linking usability to software architecture patterns through general scenarios, The Journal of Systems and Software 66 (2003) 187-197

John, Bonnie E., Len Bass, Maria-Isabel Sanchez-Segura, Rob J. Adams, Bringing Usability Concerns to the Design of Software Architecture

Wakkary, R., Framing Complexity, Design and Experience: A Reflective Analysis (2005), Digital Creativity, Volume 16, Issue 2, 65-78

Wakkary, R., Niedenthal, S., Experience and Design Methods: Cross-Dressing and Border Crossing (2004). Digital Creativity, Volume 15, Issue 4, pp. 193-196

Links

http://djvu.sourceforge.net/doc/man/cjb2.html DjVu Bitonal Encoder

http://www.ocropus.org/ The OCRopus OCR System

http://pubs.iupr.org/ Publications related to this work.

50+ open source/free alternatives to Adobe Acrobat Free or open source PDF tools.

http://www.adobe.com/devnet/pdf/pdf_reference.html PDF specification.

Additional Screens from User Interface Mock-Up

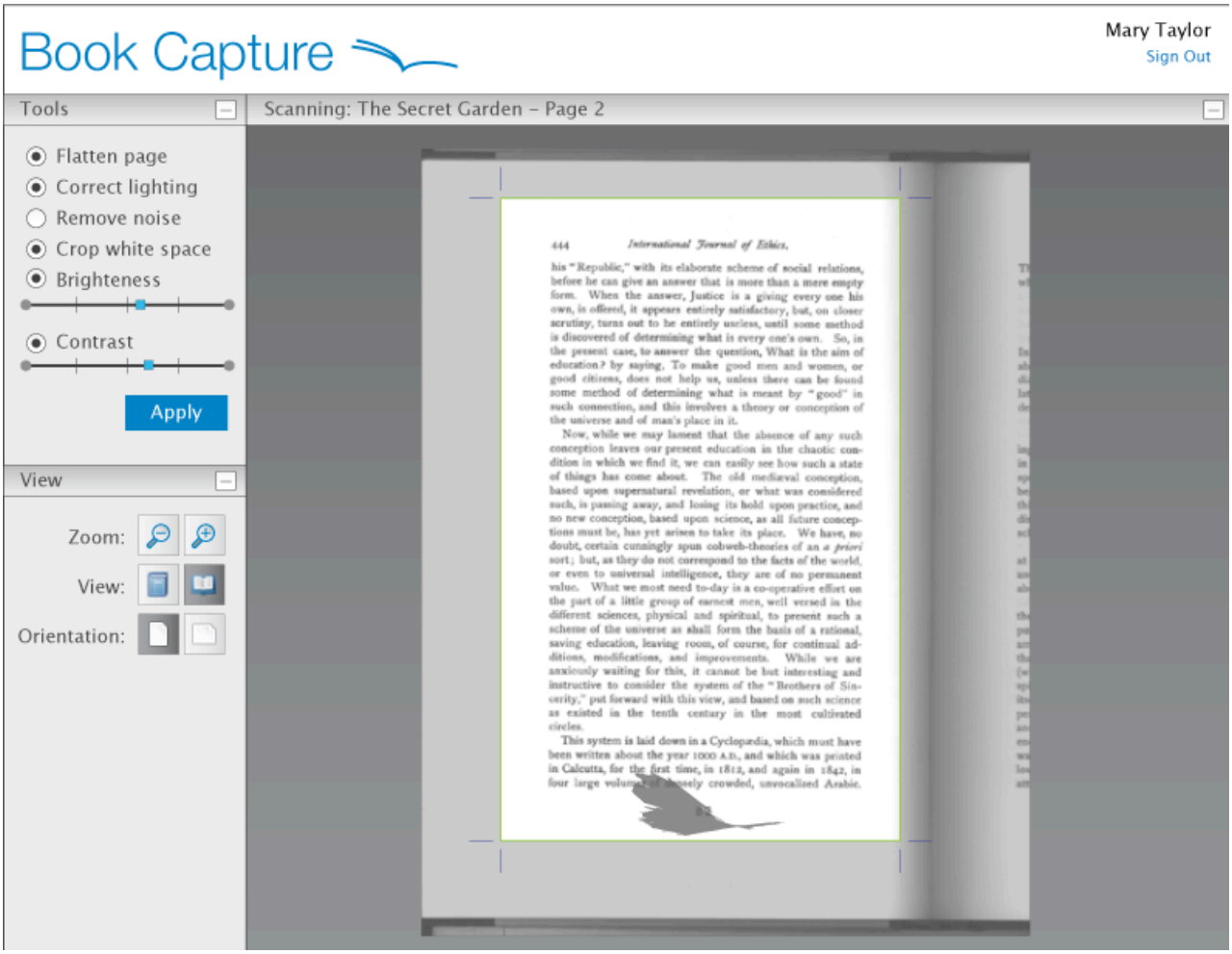


Figure 5: Interactive correction screen. Although the workflow is intended to be automated and allow capture without human interaction, occasionally, fixing a document on-screen is simpler than rescanning. The user interface will contain tools for cropping, noise removal, and "digital rescanning".

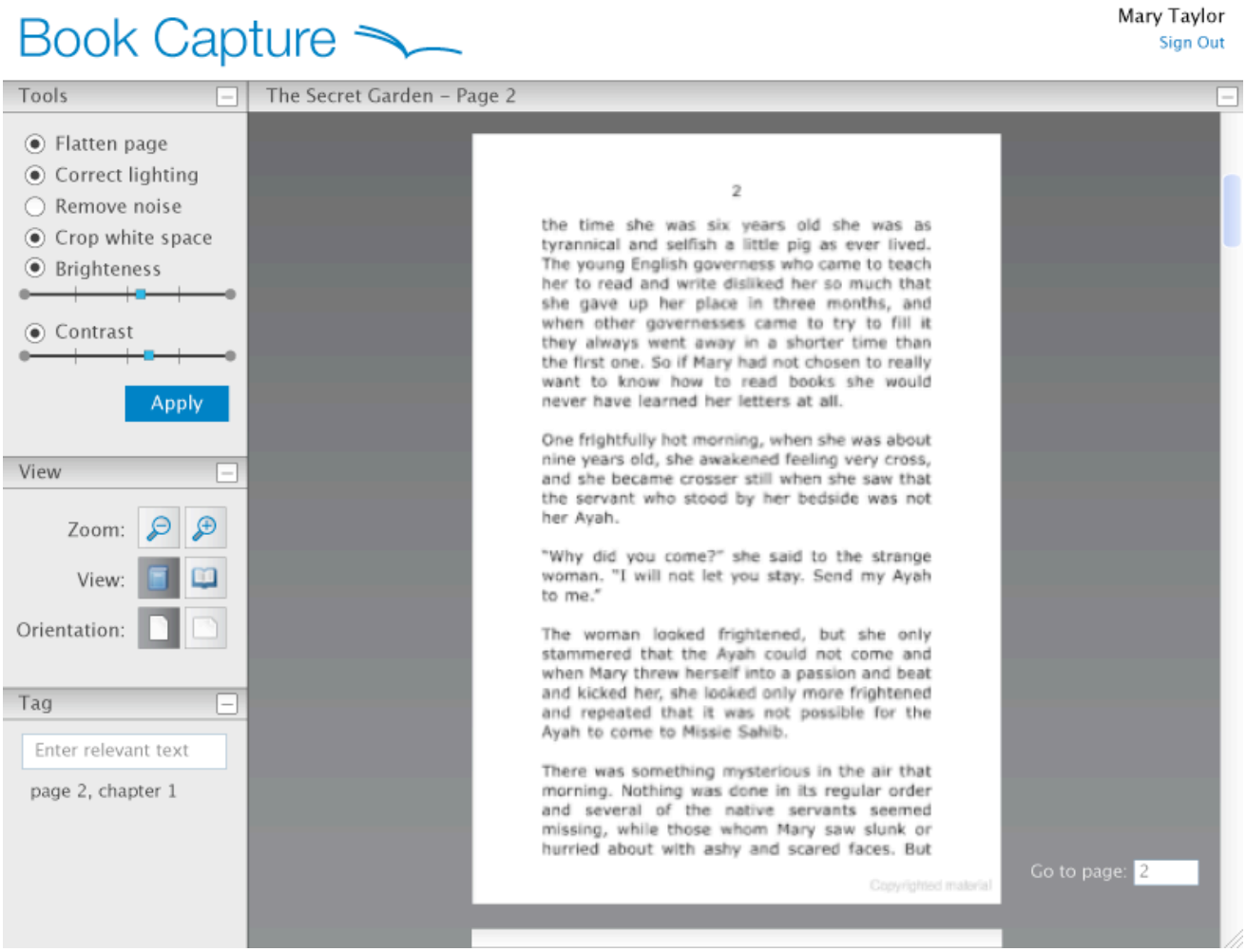


Figure 6: The user interface may also provide a high resolution scrolling preview of the entire captured book, to allow scanner operators and quality control operators to visually check the results of a book scan and make corrections.

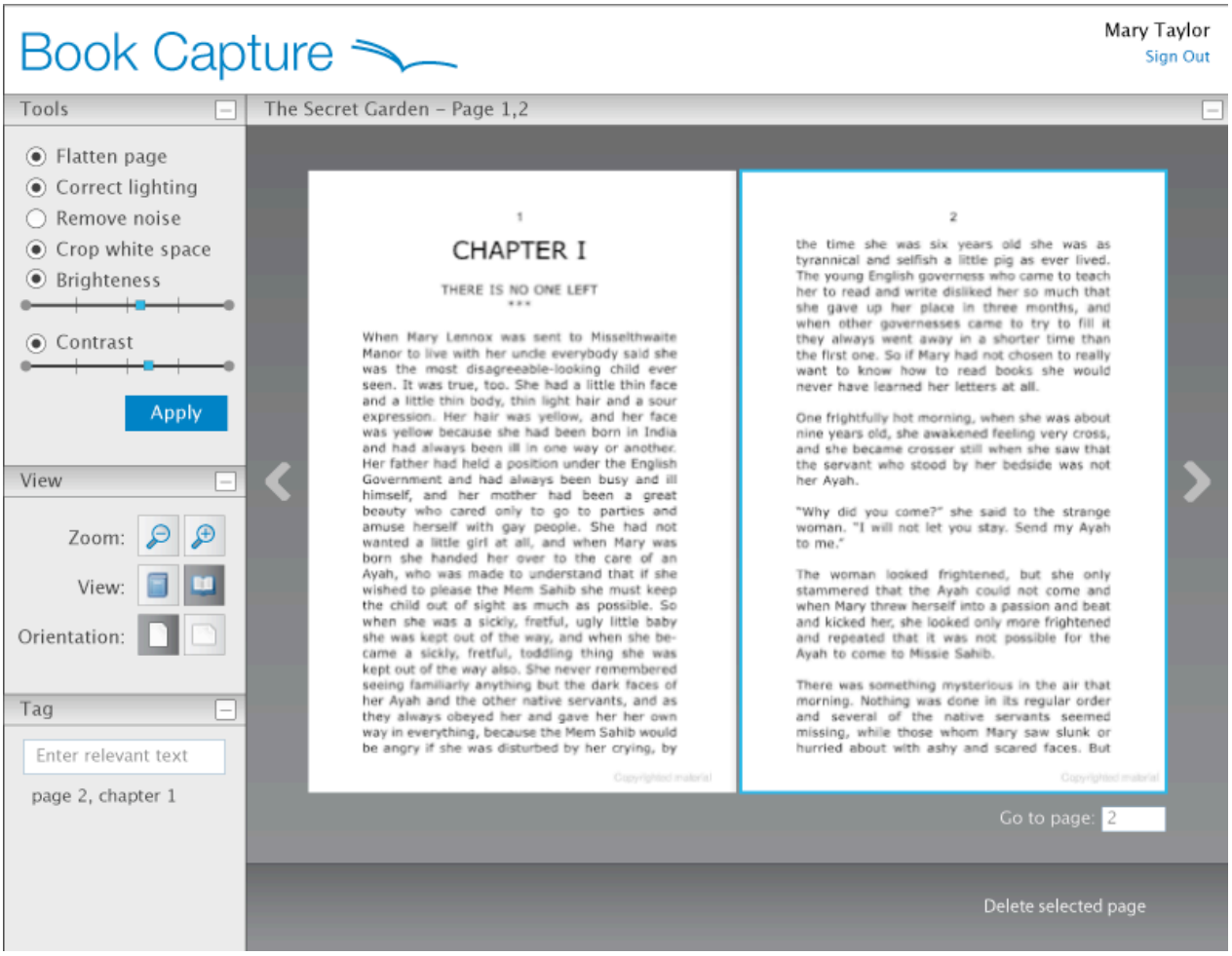
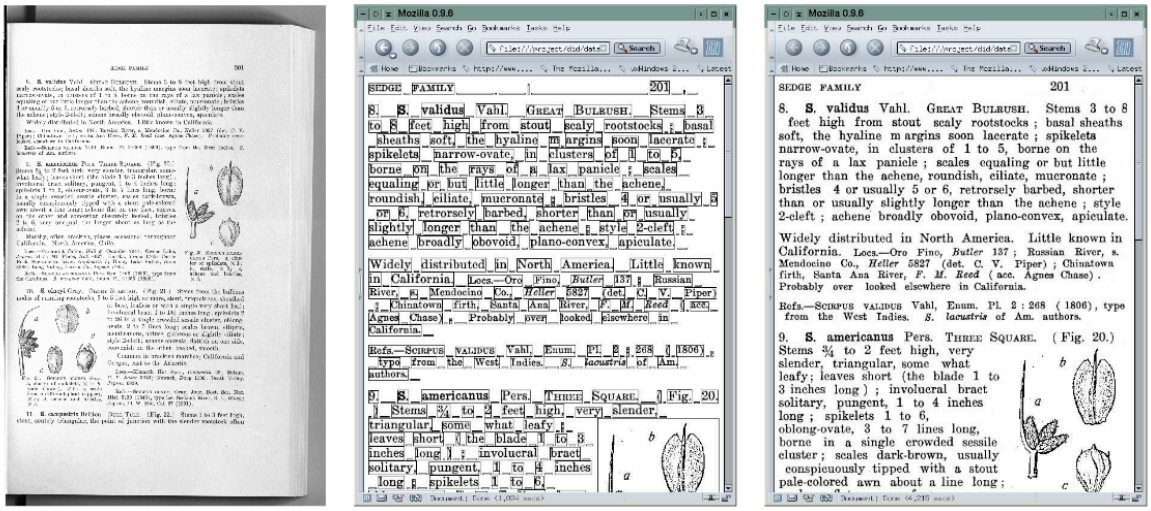


Figure 7: A non-scrolling facing page view provides another view on the scanned output and gives operators and quality control another way of checking a book for consistent layout and scanning. In addition, this view may also be useful as an end-user book reading application for users who require access to the scanned view of the image without further processing.

Examples of Object Reflowing

Several approaches to addressing this problem have been proposed in the literature. Generally, they combine layout analysis with image-based representations. One example is the Paper-to-PDA system (Breuel et al., 2002) that converts scanned documents into reflowable documents that can be read on a wide variety of devices completely without OCR.

While pioneering and already useful in its current form, the Paper-to-PDA technology has some significant limitations in terms of compression and rendering performance.



(a) (b) (c)

Figure 8: (taken from Breuel et al., 2002). (a) The original document, including figures, degraded fonts, and text requiring a complex language model. (b) The converted, reflowed document, shown with word bounding boxes (c) The rendered, reflowable document without word bounding boxes.

Appendix 2

Alternative Solutions

Although this appendix discusses alternative digitization solutions it should be noted that in the assessment of the participants, who between them have many years of experience (including running several flagship digitization programs such as Time Magazine, the MIT Press Backlist, Amazon's kindle conversion program and Hewlett Packard's scanner software program), that not one of these provides a suitable digitization solution for bound-material, on-site capture. They all fail in terms of capital cost, operational cost, reliability or functionality.

As an example, the standard procedure as used by JSTOR and others is to use overseas labor to handle the material, debind it, scan it, check and digitally edit the material up to standard and then rebind it. The other approach is to use specialized hardware to scan bound material, but technologies in these areas are costly and proprietary and none support the complete paper to reflowable, embedded font PDF solution that is proposed here. For example, Zeutschel, Google and the Internet Archive provide or use camera-based solutions, but they are heavy, bulky, costly and in the case of the internet archive and Google, not generally available. In terms of workflow software, Google, Kirtas, and the Internet Archive all have solutions, but again they are either unavailable or costly.

Off the shelf OCR packages do not do well in document structure analysis, and their remediation facilities are slow and clumsy. Camera oriented software solutions such as Atiz's snapter have proven to be remarkably unreliable in testing , and even I2S had extensive setup time and failed to perform well with a heterogeneous input stream.

In the following table we have noted DPI where possible. In the case of camera based solutions the DPI is dependent on the camera pixel count, and the document size, but should be assumed to be at least 300dpi contone, extensible to 600 dpi bilevel through resolution enhancement.

Note that these are document capture solutions, and so they are designed to be configurable and to allow integration with digitization workflows, so they may or may not include complete workflows. They will typically need other facilities to make them complete, such as compute farms, multi-engine OCR such as Prime, metadata editing facilities etc. Even where they include basic facilities it is normal to override that and build a custom solution with these as elements.

The targets for Decapod and the comparable performance figures for alternative solutions are shown in Table A2-1.

feature / metric	**Decapod targets**	**Google**	**Internet Archive**	**Atiz Booksnap**	**Atiz Hardware**	**Kirtas**	**Treventus**
camera-based	yes	yes	yes	yes	yes	yes	no
mechanical flattening	no	no	yes	no	yes	yes	yes
manual page turning	yes	yes	yes	yes	yes	no	no
cost of scanning hardware	< $1000	hardware not available at all	hardware not available at all			*~ $100k*	
cost of scanning software	open source	software not available at all	software not available at all	*$2400*	*$1595 + 2 x cameras (add $1800).*		
size of scanning hardware	carry-on luggage	*shipping crate*	*shipping crate*	*camera bag*	*large box or crate*	*large box or crate*	*large box or crate*
first time setup time	1 day						
setup by experienced users	1 hour	*1 day*	*1 day*	< 10 minutes			
scan setup training	instruction manual, interactive setup and calibration software						
scan operator training	learn as you go: immediate scan error feedback						
operator throughput (scanning)	limited by manual page turning	*limited by manual page turning*	*slower than Google*	*slower than Google*	*slower than Google*	1200/2400 pph (less misfeed op)	2500 pph. (misfeeds not included).
automatic real-time scanning error detection	> 90% of detectable errors			none			
automatic page sequence checking	yes			no			
output quality	better than Google	-	*better than Google*	*worse than Google*	*better than Google*	*better than Google*	*better than Google*
output resolution	600dpi for text	*< 600dpi*			camera based	camera based	camera based
max scanning size at 600dpi	letter size						
TIFF output	yes	yes	yes	yes	yes	yes	yes
OCR output	yes	yes	yes	yes	yes		
reflowable output (reading on mobile devices, etc.)	yes	no	no	no	no		
Manual Image correction	Minimal /one click			extensive			

Table 4:Metrics and performance targets for Decapod and several related systems. Figures that are

guessed or inferred are shown in italics. (The 600dpi target resolution derives from widely used digital library requirements and is achievable using currently available digital camera, as described above. Future improvements in available digital cameras will automatically lead to improvements in output resolution.)