ARTICLES

D-Lib Magazine
March 2006

Volume 12 Number 3

ISSN 1082-9873

Document Recognition for a Million Books

G. Sayeed Choudhury, Tim DiLauro, Robert Ferguson
Digital Knowledge Center
Johns Hopkins University
{sayeed, timmo, robert.ferguson}@jhu.edu

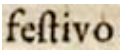
Michael Droettboom
Hillcrest Labs
mike@droettboom.com

Ichiro Fujinaga
McGill University
ich@music.mcgill.ca

Introduction

As initiatives such as Google Book Search (http://books.google.com) and the Open Content Alliance (http://www.opencontentalliance.org) advance efforts to digitize millions of books, there is great potential to make available vast amounts of information. To truly unlock this knowledge, however, it will be necessary to process the resulting digital page images to recognize important content, including both the semantic and structural aspects. Given the vast diversity of fonts, symbols, tables, languages and a host of other elements, it will be necessary to create flexible, modular, scalable document recognition systems. Document recognition involves extracting features from the images and even transcriptions of other documents in order to group diverse content.

Most commercial optical character recognition (OCR) software is designed for standard business documents. OCR software is highly successful at transcribing documents using modern printing processes. However, a great deal of content in libraries, the target content for Google Book Search and the Open Content Alliance, does not fit into this category. With many collections, the limitations of commercial optical character recognition (OCR) software become apparent. For example, with early modern English, the long "s" is recognized incorrectly as an "f": most OCR systems would misinterpret the word



as "feftivo" rather than "festivo." The Perseus Project[1] at Tufts University reports that even advanced OCR software such as Prime Recognition[2] has problems with this particular issue. Additionally, some texts feature multiple languages on the same page. There exist even more complex document recognition matters such as symbol analysis (i.e., distinguishing math formulas or chemical notations from natural language), page layout analysis (i.e., structural elements of a page such as headers, footers, paragraphs), and document matching (i.e., the ability to identify editions of the same work). For our purposes,

document recognition includes not only transcription of printed text but also extraction of the content implied by spatial layout: footnotes should not only be transcribed but associated with their markers within the text; tables should be identified and structured in the proper columns; rows, labels, block quotes, running headers, page numbers, and other conventional reading aids should be recognized and tagged. Document recognition continues until all information implicit in the page image itself has been captured.

While this paper will focus specifically on the use of Gamera[3] and its potential use in the massive digital libraries that are currently being built, the need for more advanced document recognition techniques for historical digital libraries has received a great deal of attention recently. Baird, et al. provide an excellent survey of the major challenges faced in document image analysis for digital libraries (Baird et al. 2004). Researchers have focused on a variety of areas including improving OCR for archival manuscripts (Leydier et al. 2005, Spitz 2004); improving segmentation and structural recognition of different historical document classes (Mello 2004); automatic keyword extraction from historical document images (Terasawa et al. 2005); and the development of comprehensive document recognition systems and models adapted to the unique needs of historical manuscripts (Perantonis 2004, Feng 2005).

Gamera

Since standard OCR packages cannot accommodate this diversity of content and the range of document recognition needs, Johns Hopkins University and McGill University, with partners at Tufts University and the University of Edinburgh, and with contributions from an active developer community, are developing Gamera[3] an open-source programming framework for building systems that extract information from digitized two-dimensional documents (Droettboom et al. 2003; Droettboom et al. 2002). A particular focus for Gamera has been the difficult document recognition challenges posed by cultural heritage materials. By providing a foundation for others to build upon, we hope to empower the document experts themselves to develop systems with reduced effort. As a framework for the creation of structured document analysis by domain experts, Gamera combines a programming library with GUI tools for the training and interactive development of recognition systems. One of the most important aspects of Gamera is its ability to be trained from the ground up. Most trainable commercial systems in fact use new training data to connect new phenomena to pre-existing classes of Roman print: it may be impossible to train such systems to recognize a new character set or even a single variant such as the long "s".

Gamera has been used to build preliminary systems for many different types of documents including sheet music, medieval manuscripts, eighteenth-century census data, dissertations in mixed scripts, Navajo language documents (Canfield 2005) and lute tablature (Dalitz and Karsten 2005). There are several applications of Gamera that demonstrate document recognition tasks beyond transcription. Examples include lyrics extraction from music and automatic identification of illuminations from medieval manuscripts. The University of Edinburgh has used Gamera to recognize the handwriting of particular scribes within a Middle French manuscript of Christine de Pisan.

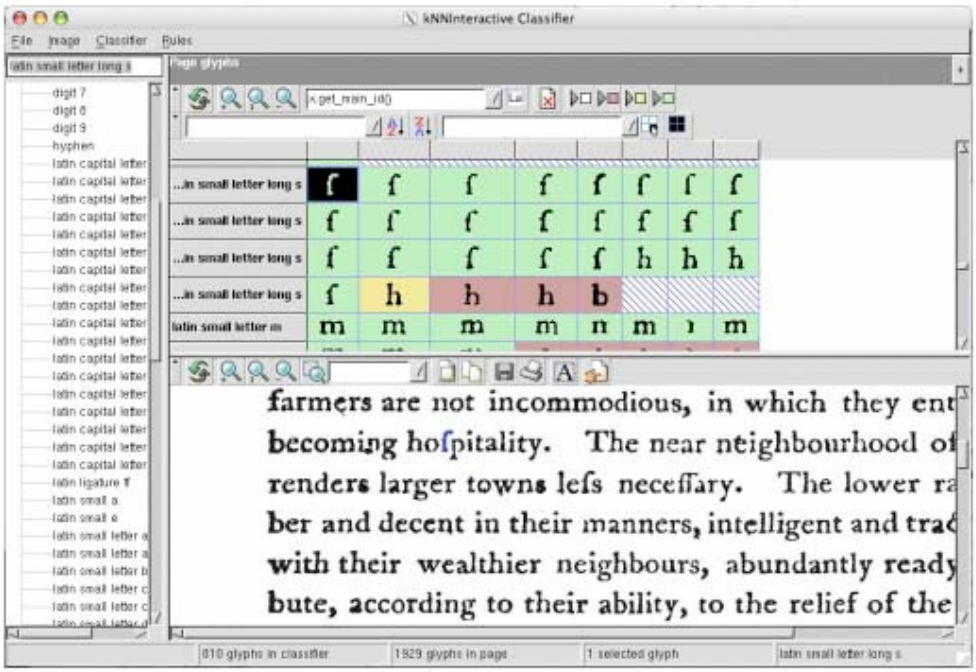


Figure 1: Gamera interface for identifying and training with early modern English characters, including the long "s".

Figure 1 illustrates the Gamera classifier interface, which includes the mechanisms for identifying characters and training the system. Without initial, specific training data, Gamera will perform poorly or require more manual corrections. As time progresses and as more text becomes transcribed, it will make fewer errors. The figure depicts how Gamera's errors can be identified and corrected by an operator (with appropriate domain knowledge), thereby building a training dataset that will improve future performance. Referring to the earlier problem of the long "s" in early modern English, Gamera did not initially recognize this character properly. However, in this representation, the error has been identified, and with sufficient training data the frequency of this particular error would diminish. This training aspect of Gamera bolsters its extensibility and flexibility, and represents an important consideration for large corpora on the order of millions of books.

Clustering similar symbols together can facilitate training of Gamera since the operator can quickly identify a class of symbols for labeling. Figures 2 and 3 demonstrate a Gamera training session before and after a pre-labeling clustering algorithm is applied.

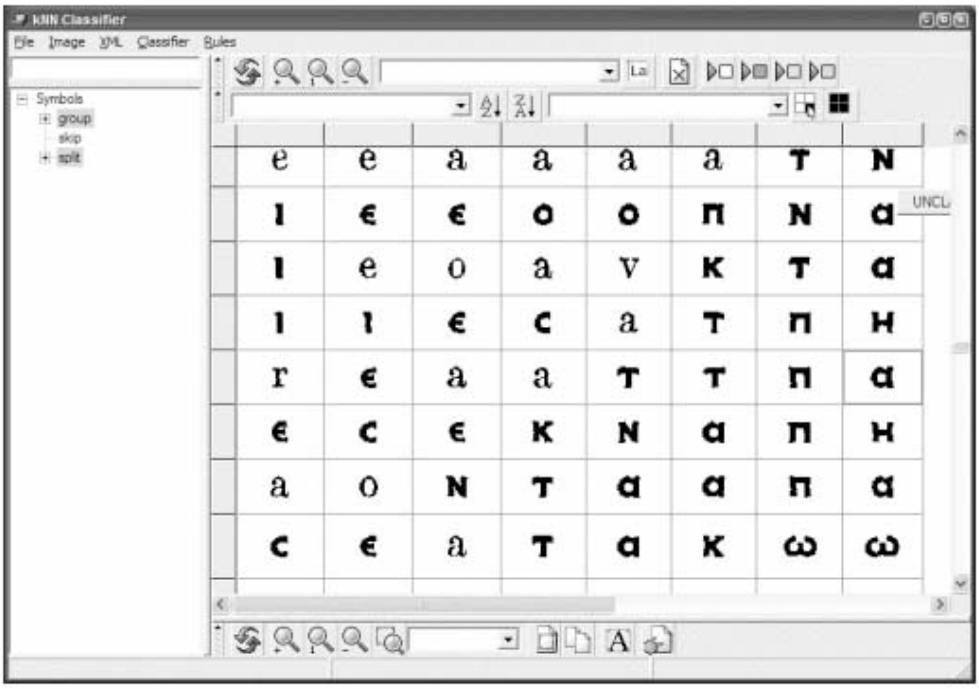


Figure 2: The beginning of a training session without clustering.

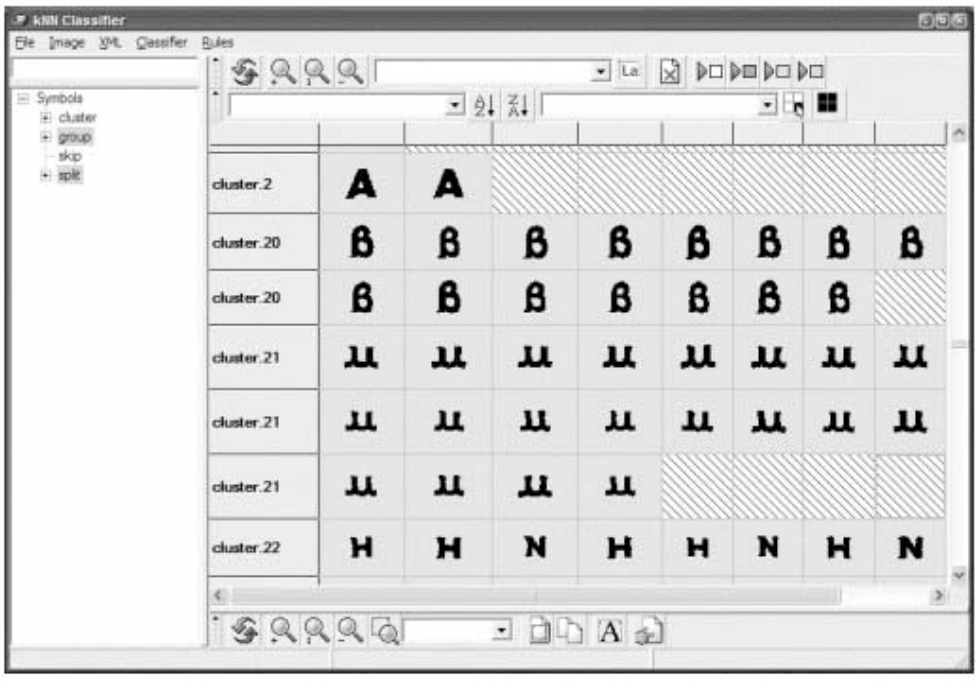


Figure 3: The beginning of a training session with clustering.

The Roman de la Rose[4] is a collaborative project led by Johns Hopkins University to create digital surrogates of a famous medieval work in French, of which there are about three hundred extant copies. An interesting characteristic of this collection is that the manuscripts are not identical, but they are derivative copies from an original work. For such collections, it would be interesting to use machine translation techniques to identify

"similar" passages from different documents across the corpus. With Gamera, an image-based concordance browsing and annotation environment is being developed that runs in most web browsers without the use of any plug-ins. When the user clicks on a "hotspot" in the image, a menu of other related hotspots is presented. Hotspots can also point to comments or other sources of information, perhaps also identified through machine translation techniques.

Medievalists who study this story have emphasized the importance of illuminated features of the manuscripts, including decorations and characters. Figures 4 and 5 depict the extraction of these illuminated elements from one of the digitized Roman de la Rose manuscripts from the Bodleian Library at Oxford.

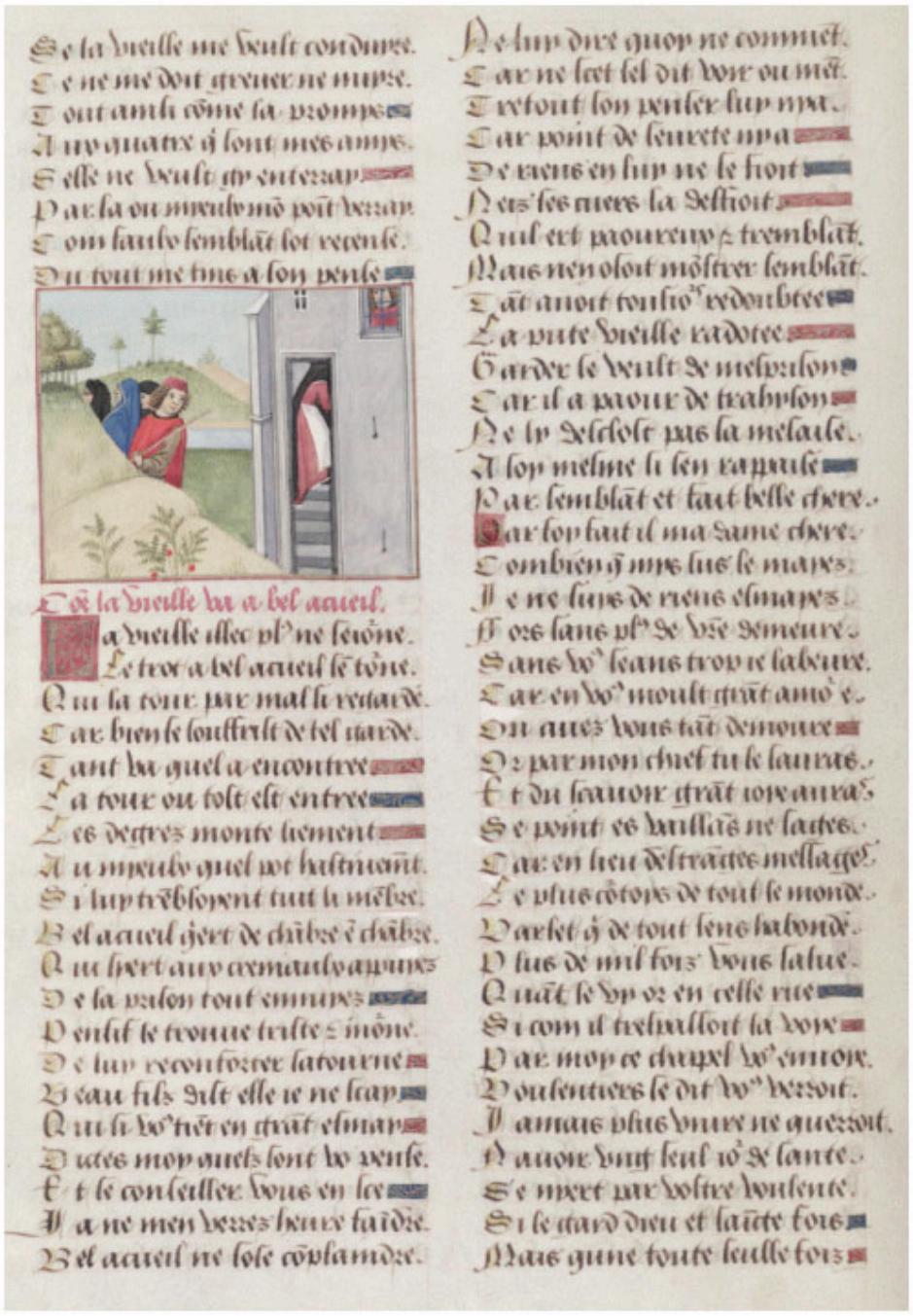


Figure 4: Folio from MS. Douce 195 Roman de la Rose manuscript.

Figure 5: Illuminations extracted from folio of MS. Douce 195 organized into three categories: decorations, characters, and illustrations.

Gamera is not mentioned in this context as the only possibility for document recognition; there are other open-source and proprietary document recognition tools, which may be used in combination with commercial OCR packages in an overall workflow for image processing (Yacoub, et al. 2005). Rather, Gamera is mentioned in this context because it provides a useful benchmark for the types of document recognition capabilities that might be useful with a large corpus of digitized books. Additionally, Gamera's approach toward

document recognition – an open, learning-based, flexible, distributed system – may be highly relevant in this context.

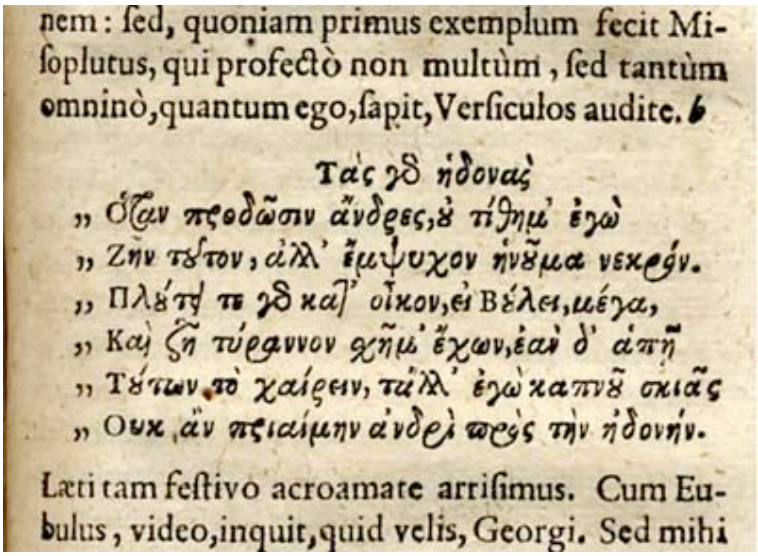
Implications of Large Corpora

Google Book Search, the Open Content Alliance, and similar efforts offer the possibility of creating digital image collections with unprecedented scale and diversity, raising major challenges for document recognition. Nevertheless, the shift from creation of small localized collections to refining small pieces of very large collections brings potential advantages as well.

Community driven efforts such as Gutenberg's Distributed Proofreaders[5] have shown how effectively groups of users can pool their efforts, many providing relatively modest inputs of labor, to create a large and efficient system for error correction (Newby 2003). Similarly, Michael Lesk has also suggested that we should be encouraging both massive digitization projects and user participation in augmenting them, pointing out that "it is hard to see, in fact, how many more specific needs could be filled other than by large volunteer efforts" (Lesk 2005). Open source systems such as Gamera provide complete access to every aspect of the system and thus allow users to train the system and tune its performance for any page layout or writing system. To date, the training occurs using a page image from a collection with (typically) only local access, and a single individual who conducts the proofreading and training. Large scale collections with users distributed across the globe provide not only new potential collaborators for error correction and cleanup but also for more efficient system training as well.

A single local collection may contain a few books from dozens of different publishers, each with their own fonts and conventions of page layout. In very large collections we may have hundreds or thousands of books from the same publisher, allowing us to create training sets more finely tuned to that set of books. The digital library system might identify dozens of different training sets for a particular set of books on a single theme but drawn from many sources. Users trying to analyze a document without a specialized training set might nevertheless locate one that is similar on which to base their new work. Multiple individuals, or groups of domain experts, could work in collaboration, sharing the results of their proofreading and training activities. Recent work by Fabio Carrera has explored this very idea, through the creation of an open source transcription architecture and tool that will allow visitors of archives to accumulate and disseminate their transcriptions across the globe (Carrera 2005). This approach would undoubtedly increase the availability of training data for any document recognition system.

Expanding upon this idea even further, it would be worthwhile to explore the potential for connections with machine translation efforts. Consider that document recognition systems could be used to assess the effectiveness of machine translation results against large-scale book image collections. In turn, instead of relying upon manually created training data, it may be possible to identify training sets from the results of machine translation processing. This combined approach could be particularly useful in automated concordance building. This iterative relationship between large-scale document recognition and machine translation would be mutually reinforcing and enriching. For further exploration of machine translation, please see David Smith's article in this same issue.


nem : ſed, quoniam primus exemplum fecit Miſoplutus, qui profectò non multùm, ſed tantùm omninò, quantum ego, ſapit, Verſiculos audite. 6

Τὰς γὰρ ἡδονὰς
„ Ὅταν προδῶσιν ἄνδρες, οὐ τίθημ' ἐγὼ
„ Ζῆν τοῦτον, ἀλλ' ἔμψυχον ἡγοῦμαι νεκρόν.
„ Πλούτῳ τε γὰρ κατ' οἶκον, εἰ βούλει, μέγα,
„ Καὶ ζῇ τύραννον σχῆμ' ἔχων, ἐὰν δ' ἀπῇ
„ Τούτων τὸ χαίρειν, τἄλλ' ἐγὼ καπνοῦ σκιᾶς
„ Οὐκ ἂν πριαίμην ἀνδρὶ προς τὴν ἡδονήν.

Læti tam feſtivo acroamate arriſimus. Cum Eubulus, video, inquit, quid velis, Georgi. Sed mihi

Figure 6: A quotation of Greek poetry from Vincentius Fabricius, Orationes Civiles, Dissertationes, Epistolae, Poemata (1685).[6]

τὰς γὰρ ἡδονὰς 1165
ὅταν προδῶσιν ἄνδρες, οὐ τίθημ' ἐγὼ
ζῆν τοῦτον, ἀλλ' ἔμψυχον ἡγοῦμαι νεκρόν.
πλούτει τε γὰρ κατ' οἶκον, εἰ βούλει, μέγα
καὶ ζῇ τύραννον σχῆμ' ἔχων: ἐὰν δ' ἀπῇ
τούτων τὸ χαίρειν, τἄλλ' ἐγὼ καπνοῦ σκιᾶς 1170
οὐκ ἂν πριαίμην ἀνδρὶ πρὸς τὴν ἡδονήν.

Figure 7: A carefully transcribed version of the source for this quotation in the Perseus digital library.[7]

In Figure 6, we see a quotation of Greek poetry from Vincentius Fabricius, *Orationes Civiles, Dissertationes, Epistolae, Poemata* (1685).[6] In a large digital library, the source for many quotations will be on-line and often in a carefully transcribed form. A document analysis system in a digital library should be able to draw on this knowledge. A mature system would perform the following:

1. Identify the block quote by analysis of the layout (e.g., indentation, double quotes at each line).
2. Generate a preliminary analysis of the text and search for a probable source.
3. If a source is found, then automatically match characters in the on-line transcription to the character images. Check for variations where source and quotation may differ, keeping only those matches of transcribed character and image that seem plausible.
4. Add best examples to a training-set for this particular font and apply the reconfigured system to other texts in the same font.
5. Apply higher order markup from transcribed text to page image. In the case above, the markup of the transcribed text includes information such as author and work (this passage is from the *Antigone* of Sophocles), citation (lines 1166-1172), and the speaker is a Messenger at the end of the play.

In a large digital library, system features such as those outlined above have immense potential, for academic documents from the beginning of the printed word quote one another and especially quote core texts over and over again. Nineteenth-century

newspapers serialize novels by Dickens or other authors for which on-line transcriptions exist. Shakespeare, the Bible and other core cultural texts are widely quoted – and the more widely quoted the text, the more likely it is that a well-transcribed digital version will exist in the digital library.

Transcription represents only one component of document recognition. The presence of a large-scale book image corpus significantly raises the possibilities for these important document recognition capabilities, especially given the potential for statistical inferences or analyses. Even without transcription, it may become possible to identify "similar" documents based on layout or structure, character sets, or even shapes of words (i.e., identifying similar words without even recognizing or determining the word). By using this pattern matching approach, it might become possible to cluster similar content or documents with little or no direct transcription.

In addition to traditional metadata that can be used to discover or organize digital content, feature extraction from document recognition systems could provide augmented metadata. Traditional metadata has tended to focus on summarizing, compact, or often requested "data about data" such as authors and titles. With a broader consideration, metadata could encompass a wide range of information automatically extracted by document recognition systems. One consequence of this approach is that the actual size of the metadata could become quite large. Therefore, these new aspects or features of metadata may not be stored, but rather generated on-demand.

Conclusions

Even at this early stage, the potential and promise of bringing together large-scale book image collections and an open, distributed, flexible document recognition framework is immense. Gamera is discussed as one possible framework that meets these criteria. There are many other researchers in the field of document recognition who might offer other, even superior, alternatives. Perhaps it would be helpful to create an "Open Document Recognition Alliance" to build upon efforts like the Open Content Alliance.

Who would like to contribute to such an alliance?

Acknowledgements

We thank the National Science Foundation and the Institute for Museum and Library Services for their funding toward the development of Gamera. We also thank Greg Crane for approaching us regarding this article.

Notes

1. Perseus Digital Library, <http://www.perseus.tufts.edu>.

2. PrimeOCR, <http://www.primerecognition.com>.

3. Gamera project homepage, <http://ldp.library.jhu.edu/projects/gamera>.

4. Roman de la Rose, <http://rose.mse.jhu.edu>.

5. Distributed Proofreaders, <http://www.pgdp.net>.

6. Source <http://www.uni-mannheim.de/mateo/camenapoem/fabvi1/jpg/s097.html>. For more on this collection, see the article by Wolfgang Schibel in this issue of *D-Lib*.

7. Sophocles, *Antigone*, line 1155, <http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:1999.01.0185:card=1155>.

Works Cited

Baird, Henry S., et al. Document Analysis Systems for Digital Libraries: Challenges and Opportunities. *Proceedings of Document Analysis Systems VI: 6th International Workshop, DAS 2004*, Florence Italy, September 8-10, 2004.

Canfield, K. 2005. Athabascan Textbases: A Navajo Language Study. *Proceedings of Association for Computers and the Humanities and Association for Literary and Linguistic Computing*.

Carrera, F. "Making History: an Emergent System for the Systematic Accrual of Transcriptions of Historic Manuscripts," icdar, pp. 543-449, *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, 2005.

Dalitz, C, and T. Karsten. 2005. Using the Gamera framework for building a lute tablature recognition system. *International Conference on Music Information Retrieval*.

Droettboom, M., K. MacMillan, and I. Fujinaga. 2003. The Gamera Framework for building custom recognition systems. *Proceedings of the Symposium on Document Image Understanding Technologies, (SDIUT)*. 275-86.

Droettboom, M., K. MacMillan, I. Fujinaga, G. S. Choudhury, T. DiLauro, M. Patton, and T. Anderson. 2002. Using Gamera for the recognition of cultural heritage materials. *Proceedings of the Joint Conference on Digital Libraries, (JCDL 2002)*. 11-17.

Feng, S.L. and R. Manmatha. Classification Models for Historical Manuscript Recognition. *Proceedings of Eight International Conference on Document Analysis and Recognition*, pp. 528-32.

Lesk, Michael. The qualitative advantages of quantities of information: bigger is better. J. Zhejiang Univ. *Science* 6A, 11, pp. 1179-1187 (Nov. 2005).

Leydier, Yann, F. LeBourgeois, and H. Emptoz. Textual Indentation of Ancient Documents. *Proceedings of DocEng05*, November 2005.

Mello, C.A.B. Image Segmentation of Historical Documents: Using a Quality Index. *Proceedings of Image Analysis and Recognition: International Conference ICIAR 2004, Part II*, Porto, Portugal, September 29 - October 1, 2004.

Newby, G.B. and C. Franks. Distributed Proofreading. *Proceedings of the Joint Conference on Digital Libraries, 2003*.

Perantonis, S.J. et al. A System for Processing and Recognition of Old Greek Manuscripts (The D-SCRIBE Project), *WSEAS Transactions on Computers*, Issue 6, Volume 3, pp. 2049-2057, 2004.

Spitz, A. Lawrence. Tilting at Windmills: Adventures in Attempting to Reconstruct *Don Quixote*. *Proceedings of Document Analysis Systems VI: 6th International Workshop, DAS 2004*, Florence, Italy, September 8 - 10, 2004.

Terasawa, Kengo, T. Nagasaki, and T. Kawashima. Automatic Keyword Extraction from Historical Document Images. *Proceedings of Document Analysis Systems VII, 7th International Workshop, DAS 2006*, Nelson, New Zealand, February 13-15, 2006, *Lecture Notes in Computer Science* 3872 Springer 2006.

Yacoub, Sherif., V. Saxena and S. Sami. PerfectDoc: A Ground Truthing Environment for Complex Documents. *Proceedings of the 2005 Eighth International Conference on Document Analysis and Recognition*.

Top | Contents
Search | Author Index | Title Index | Back Issues
Previous Article | Next article
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

doi:10.1045/march2006-choudhury