Report on the BooksOnline'10: Third Workshop on Research Advances in Large Digital Book Repositories and Complementary Media

Gabriella Kazai
Microsoft Research Cambridge, UK
v-gabkaz@microsoft.com

Peter Brusilovsky
Univeristy of Pittsburgh, US
peterb@pitt.edu

Abstract

The BooksOnline Workshop series aims to foster the discussion and exchange of research ideas and initiatives addressing challenges and exploring opportunities around large collections of digital books and complementary media. The third workshop in the series, BooksOnline'10 boosted a high quality program, including keynote addresses by James Crawford from Google Books and John Mark Ockerbloom from the University of Pennsylvania. From the accepted papers two main themes emerged: 1) Issues relating to building suitable infrastructure for online book collections and for the evaluation of the applications and services that make use of such collections, and 2) Opportunities relating to the social or collaborative reading and annotation of books and issues relating to the design of suitable eReaders. This paper provides a summary of the workshop and the accepted contributions.

1 Introduction

In recent years digital book content has increased dramatically through the digitization of physical books and electronic publishing. Such collections present a great value to humanity. They preserve the content that is otherwise vulnerable to deterioration or permanent loss and make it amenable to wide distribution, customization, and integration with related media. These collections are also attractive to commercial organizations for use in diverse information services and applications. As a result, millions of books have been and are still being digitized in partnership between libraries and companies such as Amazon, Google, and Microsoft. Some of the leading digitization initiatives around the world include Project Gutenberg[1], the Million Book project[2], the European Digital Library[3], the Open Content Alliance (OCA)[4], digitization at Stanford, Harvard, and Michigan University libraries, the

[1] http://www.gutenberg.org/

[2] http://www.ulib.org/

[3] http://www.euractiv.com/en/culture/parliamentarians-call-books-put-online/article-167157

[4] http://www.opencontentalliance.org/

New York public library and the Bodleian library in Oxford supported by Google, and the British Library[5] digitization of 19th Century books supported by Microsoft.

However, in contrast to the great momentum in creating on-line book repositories, there has been a lack of research initiatives that are focused on innovation opportunities and challenges created by large collections of digital books [7]. One such effort is the INEX Book Track[6], which is building a test collection and infrastructure for experiments on digitized books. This, however, represents only a first step towards enabling digitized books research. Similar initiatives are necessary in order to create a critical mass necessary to innovate in this important area. The BooksOnline Workshop series aims to provide a forum to foster such research initiatives.

Continuing on from the first two workshops, the 3rd BooksOnline workshop[7] aimed to bring together researchers and industry practitioners in information retrieval (IR), digital libraries (DL), ebooks, human computer interaction (HCI), publishing industry, and on-line book services to foster progress on addressing challenges and exploring opportunities around large collections of digital books and complementary media.

In these past three years, the BooksOnline community has been gaining strength and momentum. At BooksOnline'08[8], a research agenda was defined covering three fundamental areas: 1) Enriched digital collections, 2) Usage scenarios and user experience, and 3) Content representation and discovery services. BooksOnline'09[9] focused on more concrete actions around joint project ideas in the areas of: 1) Design and interaction models, and 2) Search and evaluation. BooksOnline10, the most successful of the three workshops so far, has focused on: 1) User aspects and design, and 2) Infrastructure issues.

Contributors and participants in the last three years included key figures - among many others - from academia and industry, including Brewster Kahle (Internet Archive), Michael Lesk (Rutgers University), Magdy Nagi (Bibliotheca Alexandrina), Gene Golovchinsky (FX Palo Alto Lab), Catherine Marshall (Microsoft Research), Bob Stein (The Institute for the Future of the Book), James Crawford (Google Books), John Mark Ockerbloom (University of Pennsylvania), George Buchanan (City University), Ivan Koychev (University of Sofia), Harald Reiterer (Univeristy of Konstanz) and Pertti Vakkari (University of Tampere).

Thanks to the contributions of the authors and participants of the workshop series, the BooksOnline community is making headway in advancing the state of the art and driving the research agenda on creating, accessing, using and exploiting online book content and complementary media.

Since 2009, BooksOnline has also been fortunate to receive sponsorship from Microsoft Research to provide seed funding to selected projects. In 2009, the two projects that were awarded seed funding were: 1) George Buchanan (City University, London, UK): 'Building a Testbed for Document Annotation and Search', and 2) Monica Landoni (University of Lugano, Switzerland): 'Building a Bookshelf for Children'. In 2010, the awarded projects were: 1) W. Xavier Snelgrove and Ronald M. Baecker (University of Toronto, Canada): 'A System for the Collaborative Reading of Digital Books with the Partially Sighted', and 2) Claudia Hauff and Dolf Trieschnigg (University of Twente, The Netherlands): 'Enhancing Access To Classic Children's Literature'.

[5] http://news.bbc.co.uk/2/hi/technology/7018210.stm

[6] https://inex.mmci.uni-saarland.de/tracks/books/

[7] http://research.microsoft.com/booksonline10

[8] http://research.microsoft.com/booksonline08

[9] http://research.microsoft.com/booksonline09

In the rest of this report, we briefly summarize the BooksOnline'10 Workshop, its contributions and achievements. We conclude with plans for BooksOnline'11.

2 Workshop program

The one day workshop included two keynotes, selected paper presentations, a poster session, a break-out session to brainstorm around issues and ideas that emerged from the presentations, and a panel discussion to present and summarize the results of the break-out sessions.

2.1 Keynote by John Mark Ockerbloom from University of Pennsylvania

The day started with a deeply insightful keynote by John Mark Ockerbloom (University of Pennsylvania) entitled 'The metadata challenge: Promoting discovery, access, and usability for online books' [8]. John is a pioneer in library science and was the first person to make a substantial effort to catalog online books in a rigorous and comprehensive manner.

In his keynote, John highlighted the crucial role of metadata in organizing and facilitating access to the rich troves of information and culture that is now available to anyone in the world with an Internet connection thanks to the millions of books, serials, and other documents now digitized. He pointed out that these riches are worthless if they cannot be found, accessed, and effectively used by the readers who need them. He argued that the key to unlock these treasures is metadata. Networked computing enables techniques for making metadata more effective than ever; yet in practice, online collections all too often either do not have or do not take full advantage of the best metadata they could use.

He reviewed some of the ongoing work harnessing metadata to improve online book discovery, access, and usability. For example, online book discovery is being enhanced with concept-oriented catalogs of various kinds, including browsable maps relating millions of subjects and associated books. Copyright metadata is starting to open access to many books that had been needlessly withheld from the public, while also reducing the risk of inadvertent infringement. Structural and relational metadata and annotations, including Library of Congress (LoC) subject headings, folksonomies like PennTags, are making complex works much more usable than they were when they were represented as a mishmash of volumes.

He highlighted that using metadata effectively in multi-million-volume collections still poses special problems of scale. For example, he pointed out that subject search in catalogues for online books is not usable: "there is no way to find what would be the best book to read on a given subject". Subject hierarchies can be useful, but people typically do not want to traverse long trees. Faceted browsing methods are great for narrowing down types of books and to filter them, but again they do not work very well for subject searches and for linking related subjects and books. Solving these problems requires considered application of both library science and computer science. It also requires harnessing the collective intelligence of readers, writers, librarians, and publishers. Wise metadata management policies, including open data sharing, can promote the effective aggregation of human and machine intelligence at the scale needed. In his talk, he demonstrated and described ways in which the metadata challenges of large-scale online libraries could be met. For example, he gave a live demo of a map of online books at Penn library, which integrates various data sources, such as bibliographic data from HathiTrust, subject headings and geographic data from LoC to create an experience that allows users to search using broader or narrower terms, browse by facets, explore geographic and other semantic connections.

2.2 User aspects and design session

This session included three presentations whose common theme was on designing usable, collaborative or social experiences around digital books.

Myriam Ribière, Jérôme Picault and Sylvain Squedin (Alcatel-Lucent Bell Labs) in their talk entitled 'The sBook: towards Social and Personalized Learning Experiences' [11] presented a set of ideas on the future of eBooks. Their main motivation is to turn reading from a solitary experience to social engagement. Their envisioned eBook reader, sBook, aims to serve as a trigger of social interactions, supporting shared annotations, tagging, interlinking, conversations and learning, e.g., through collaborative learning paths. These concepts and ideas are currently being addressed in the scope of their collaborative project with the Abilene Christian University and Cambridge University Press. The goal of this project is to develop an application to facilitate cross-media and cross-community information discovery. The project has started with the implementation of an advanced eReader platform, letting students add annotations and share them, as well as including a heat map feature. The platform is used to test the acceptance of various social features in eBooks.

Jennifer Pearson (Swansea University) and George Buchanan (City University) presented their innovative system to support 'Real-Time Document Collaboration Using iPads' [9]. They exploit the light and portable properties of iPad hardware to facilitate an intimate interaction experience that gives users the ability to simultaneously mark up their own copy of a document, while instantly viewing notes made by other members of the group. They also introduced a 'look at this' referencing tool that allows group members to quickly 'point out' information within a document to other members without physically gesturing with their hands. Their work explores new interaction methods for group collaboration that incorporates the sleek design of iPads with real-time working group annotation. This allows them to incorporate the versatility and portability of paper with the instant sharing possibilities of the Internet.

In the final talk of the session, Jennifer Pearson (Swansea University), George Buchanan (City University) and Harold Thimbleby (Swansea University) in their paper entitled 'HCI Design Principles for eReaders' [10] took a step back from the immediate practical issues relating to eReader design and urged for a scientific view of current eReader technology from the systematic standpoint of basic HCI principles. Using HCI principles to think about design, they presented a detailed examination and discussion of guidelines for good eReader design and illustrated them with examples of shortcomings of some of the more popular eReader devices on the market today. For example, some of the reading functions they covered include bookmarking, annotations, page turning and magnification.

2.3 Infrastructure aspects session

Monica Landoni (University of Lugano) in her talk entitled 'Ebooks Children Would Want to Read and Engage with' [5] described an effort to build a bookshelf of electronic books for children. She argued that reading is a personal experience and, when considering children of 6-9 years old, even reading for pleasure is intertwined with learning and gaining essential lifelong skills. She believes that by involving children in the design of suitable interfaces, they can become more enthusiastic about reading. One of the first challenges, however, is to build a collection of titles that children would want to read. With this aim, she reviewed a cluster of interrelated projects, all aimed at offering children innovative and engaging eBooks, exploring their mutual implications, rationale and related expectations.

Nadia Caidi and Margaret Lam (University of Toronto) presented on 'Working with First Nations: On-Demand Book Service' (ODBS) [2], a collaboration between First Nations communities in Northern Ontario and academic researchers from the University of Toronto. Following their mandate to support the joy of reading, the aim of the ODBS is to bridge the gap between physical and digital libraries. Using the ODBS as a case study, they shared their experiences, relating them to three themes: 1) sensitivity to social context, 2) designing across cultural boundaries, and 3) the integration of content and service. The challenges highlighted include real physical demands for books, the challenge of the size of the geographic area, support for curation, ownership, control and access, and negotiation of community needs.

Ray Siemens and Julie Meloni (University of Victoria) reported on their project 'Implementing New Knowledge Environments: Building Upon Research Foundations to Understand Books and Reading in the Digital Age' [12]. This seven-year project takes a broad and holistic view of the challenges involved in developing a new digital information/knowledge environment that builds on past textual practices. These span theoretical aspects, e.g., to model electronic scholarly edition, architectural issues, interface design and test-bed construction for evaluation.

2.4 Poster session

The poster session included four contributions, including the two project proposals that were later awarded seed funding.

Fernanda Bonacho (Lisbon Polytechnic Institute) in 'Biblioteca de Livros Digitais: The privileged Space of a Transliterate Experience for Children Reading Online' [1] presented a detailed examination of the skills and abilities needed to interact with eBooks, drawing upon the unifying concept of transliteracy.

Monica Landoni (University of Lugano) in 'Evaluating E-books' [6] detailed the Active Reading task of the INEX Book Track, which aims at exploring, through user studies, how and why readers use eBooks in specific scenarios.

W. Xavier Snelgrove and Ronald M. Baecker (University of Toronto) in 'A System for the Collaborative Reading of Digital Books with the Partially Sighted' [13] (seed fund award) proposed a collaborative reading environment for digital books to record and re-play the audio of a person reading aloud a book, specifically designed with partially sighted users in mind.

Claudia Hauff and Dolf Trieschnigg (University of Twente) in 'Enhancing Access To Classic Children's Literature'[10] (seed fund award) presented a proposal to enrich text-only children's books with illustrations to make them more appealing and with links to background information to make the books more accessible.

2.5 Keynote by James Crawford from Google Books

The highly anticipated second keynote of the day was given by James Crawford (Google Books) entitled 'The Present and Future of Google Books' [3]. Despite the obvious limitations

[10] http://research.microsoft.com/en-us/events/booksonline10/hauff-trieschnigg.pdf

on information sharing, the talk did not disappoint and generated a very lively discussion. It even attracted additional audience from other workshops.

James's talk covered various aspects of the Google Books project, which "has the modest goal of scanning all of the world's books, converting them to digital form, and making them searchable and accessible". To date over fifteen million books, containing over five billion pages, have been scanned and digitized. While this is an impressive number, he pointed out that scanning is only the beginning of the challenge.

One part of the challenge in making books searchable and accessible is that a scan produces an image of a page, and often a blurred or partially obscured one at that, but searching requires a digital representation of the text on the page. Converting the image to text is also critical to creating a good reading experience since the text can then be reformatted to match the display size and the user can control the font size and layout. This is especially important for tablet devices and smart phones.

Another part of the challenge is that a search query will often match thousands or even tens or hundreds of thousands of books. Consider for example the query 'to be or not to be'. This line has been quoted in an untold number of books (as well as in this workshop report). How should the top ten matches be chosen to return on the first page of search results? Although much work has been done on search rankings for web pages, these techniques do not apply well to books. The Google books teams has had to invent largely new techniques to rank book results.

A third challenge is that copyright law was not written with the digital world in mind. In the days of print books the cost of a print run was sufficiently high that when books went out of print they relatively rarely came back into print. However, in the digital world, millions of out of print books can be made available as eBooks. For books that are out of print, but in copyright, the largest expense in creating eBooks is the cost of accurately identifying the owner of the digital rights. This cost has emerged as an important non-technical challenge.

In addition to the above challenges, James also covered a variety of other topics, including how OCR errors are being fixed through captcha's and how metadata correction is a good candidate for a crowdsourcing application. Interesting challenges highlighted include the diversity of dates, both when books are published and dates that appear in the content, the diversity of the languages and even issues on page numberings. The issue of how to rank books also got a mention due to the lack of incoming links (that PageRank builds on), and challenges in how to judge the *importance* of a book. Finally, James also looked at some of the new opportunities that arise from the emerging digital books corpus, ranging from social collaboration, to linguistic analysis and the evolution of languages, to other new areas that are only beginning to be discovered, including books as corpus of all human knowledge or using structure in ranking books.

3 Conclusions and Looking Ahead

The success of the BooksOnline'10 workshop was reflected not only in the quality of the contributions and the turnout of its participants, but in the liveliness of the discussions throughout the day. We left the workshop both with a sense of achievement and great excitement in looking towards the future.

We are now planning the next workshop, for which we plan to shift the focus more explicitly and deliberately to exploring the role of social media and crowdsourcing in the context of online books. Both social media and crowdsourcing have been key in defining new user experiences on the Web and thus we aim to jump-start the BooksOnline community into embracing and exploiting these phenomena in order to make digital and online reading more widely accepted and popular.

Examples of where social media is making headway in promoting online book usage is LibraryThing.com and Amazon's book service that integrates with Kindle. Crowdsourcing has thus far been primarily used in aid of building benchmarks for the evaluation of book search engines at INEX. However, these are merely the tip of the potential opportunities that such social engagement models and platforms can offer to online book services. Since social engagement and supporting infrastructure are clearly vital to the future of digital library platforms, the workshop will aim to unearth these potentials and lead the way to defining a research roadmap towards advancing both theory and development.

Acknowledgments We would like to thank ACM and CIKM for hosting the workshop, in particular Mounia Lalmas, Aijun An and Jimmy Huang for their outstanding support in the organization.

We would also like to thank members of the program committee for all their dedicated work in shaping the program: Birte Christensen-Dalsgaard (State and University Library, Denmark), Bruce Croft (University of Massachusetts Amherst, US), Gilles Falquet (University of Geneva, Switzerland), Edward Fox (Virginia Tech, US), Norbert Fuhr (University of Duisburg, Germany), Gene Golovchinsky (FX Palo Alto Laboratory, Inc., US), Ananda Gunawardena (Carnegie Mellon University, US), Ivan Koychev (Sofia University, Bulgaria), Monica Landoni (University of Lugano, Switzerland), Birger Larsen (Royal School of Library and Information Science, Denmark), Ray Larson (University of California, Berkeley, US), Michael E. Lesk (Rutgers University, US), Paolo Manghi (CNR, Italy), Catherine C. Marshall (Microsoft Corporation, US), John Ockerbloom (University of Pennsylvania, US), Prakash Reddy (Hewlett-Packard), and Ian Witten (University of Waikato, New Zealand).

Special thanks are due to the paper authors, the invited speakers: James Crawford and John Mark Ockerbloom, and all the participants for a great and lively workshop.

We are also grateful for the sponsorship received from Microsoft Research, including a £3,000 award to provide seed funding for selected projects based on ideas and collaborations proposed at the workshop.

Details about the workshop are at `http://research.microsoft.com/booksonline10/`. Contributed papers are at `http://portal.acm.org/citation.cfm?id=1871854`.

References

[1] Fernanda Bonacho. 'Biblioteca de livros digitais': the privileged space of a transliterate experience for children reading online. In Kazai and Brusilovsky [4], pages 39–42.

[2] Nadia Caidi and Margaret Lam. Working with first nations: on-demand book service. In Kazai and Brusilovsky [4], pages 29–32.

[3] James Crawford. The present and future of Google Books. In Kazai and Brusilovsky [4], pages 37–38.

[4] Gabriella Kazai and Peter Brusilovsky, editors. *BooksOnline'10: Proceedings of the Third Workshop on Research Advances in Large Digital Book Repositories and Complementary Media*, New York, NY, USA, 2010. ACM.

[5] Monica Landoni. Ebooks children would want to read and engage with. In Kazai and Brusilovsky [4], pages 25–28.

[6] Monica Landoni. Evaluating e-books. In Kazai and Brusilovsky [4], pages 43–46.

[7] M. Lesk. *Understanding digital libraries.* Morgan Kaufmann, San Francisco, CA, 2005.

[8] John Mark Ockerbloom. The metadata challenge: promoting discovery, access, and usability for online books. In Kazai and Brusilovsky [4], pages 1–2.

[9] Jennifer Pearson and George Buchanan. Real-time document collaboration using iPads. In Kazai and Brusilovsky [4], pages 9–14.

[10] Jennifer Pearson, George Buchanan, and Harold Thimbleby. HCI design principles for ereaders. In Kazai and Brusilovsky [4], pages 15–24.

[11] Myriam Ribière, Jérôme Picault, and Sylvain Squedin. The sBook: towards social and personalized learning experiences. In Kazai and Brusilovsky [4], pages 3–8.

[12] Ray Siemens and Julie Meloni. Implementing new knowledge environments: building upon research foundations to understand books and reading in the digital age. In Kazai and Brusilovsky [4], pages 33–36.

[13] W. Xavier Snelgrove and Ronald M. Baecker. A system for the collaborative reading of digital books with the partially sighted: project proposal. In Kazai and Brusilovsky [4], pages 47–50.