Web Books:
The Fusion of Paper and Pixels

Pirker, Johanna; Wurzinger, Gerhard; Müller, Heimo

Abstract - ***Digital and online reading is now an everyday possibility and is supported by different services and devices. However, the digitalization of traditional books supports even more innovative forms of reading and experiencing the content. In this paper we introduce our solution of an interactive web book library which allows experiencing old books in an innovative and social way. First, we analyze and compare the strengths and weaknesses of existing electronic solutions, and discuss implications for adoption and digitalization of books. We will then present a digital library specialized in old books, which integrates different forms of interactivities and reading experiences. In this context we discuss the potentials of an interactive web book representation using the example of a unique Austrian encyclopedia. We conclude with a discussion of how future reading can revolutionize the way we are experiencing traditional and old books.***

Index Terms: ***E-Books, Web books, Internet books, Digital library, Interactive library.***

1. *INTRODUCTION*

ONLINE reading became an important mean of daily technology interactions. E-readers, reading devices, and digital web libraries with integrated online book readers are designed to simulate the traditional reading experience. Such technologies are being used to facilitate, enhance, and even replace traditional reading. More and more people read texts online and virtually. Especially devices such as Amazon's Kindle or Apple's IPad have revolutionized the way people experience literature and other readings [10]. New forms of reading also changed the reading behavior and the user's expectations towards the digital content. They expect realistic behavior and the look and feel of traditional books combined with interactive features such as searching, highlighting, or bookmarking.

However, still many users have a negative attitude towards technology-enhanced reading of books. As we will explain later in this article, attracting users to these new forms of reading require advanced designs of interactivities to show advantages of digital reading compared to traditional reading in a familiar setting. Also, different kinds of applications require distinct tools and possibilities to interact with the content. For the different purposes such as academic research, learning at school, or leisure, the required tools change.

In this article, we shall discuss different interactive and social technologies to enhance the traditional reading experience, shifting the focus of design away from simply simulating traditional books to enhancing reading experience with interactive and social technologies. This should not only support the new forms of reading, but should also help attracting people to read more books online and to support different application domains such as academia or research.

Our goal is to study different techniques to improve the technology-enhanced reading experience. The contributions of this work include, first, a characterization of social and activity-based interaction styles to enhance the reading experience. Second, we introduce web books as integrative solution based on these characterizations and introduce this concept in a case study. We close by discussing future advancements and ideas of web books.

2. *BACKGROUND*

Research in different areas is relevant to this work: (1) digital reading technologies and (2) online libraries. We consider each in turn.

2.1 Digital and Online Reading Technologies

One of the most popular forms of digital reading is the use of electronic books (e-books). E-books are usually formatted specifically for a reading device or software, and can be either created manually, or can be a result of a digitization process [1][9]. In contrast, digitized books are virtualized (e.g. scanned or photographed) books, which are usually stored in high-resolution formats such as TIFF and can be further processed with optical character

Manuscript received April 2, 2014.

J. Pirker is with the Institute for Information Systems and Computer Media, Graz University of Technology, Austria (e-mail: jpirker@iicm.edu).

G. Wurzinger is with the Institute for Information Systems and Computer Media, Graz University of Technology, Austria (e-mail: gwurz@iicm.edu).

H. Müller is with the Medical University of Graz, Austria (e-mail: heimo.mueller@mac.com).

recognition (OCR) algorithms [4] to be readable by machines. The OCR content can be the basis for the creation of e-books or other forms of digital representations.

Supporting machine readable content, the text could be presented in several innovative ways. However, most digital book readers are designed to present the digital content in a natural way. Studies suggest that users like to keep functionalities they used in traditional paper books also in the digital representations [4]. Hence, many digital readers assimilate the characteristics of real books, and integrate functionalities and visualizations such as page turning or inserting bookmarks, what should increase the usability and navigability of the medium [3]. In particular reading for pleasure focuses on the readers' satisfaction and enjoyment and should remind users of traditional reading experiences. For this purpose, Wilson, Landoni, and Gibb [16] suggest the integration of book design elements which meet the expectations of readers based on their experience with printed books. This includes visual elements (e.g. book cover), elements to support a sense of structure (e.g. table of content), and a sense of place (e.g. bookmarks, progress bars).

The shift from traditional to electronic books also results in the integration of different interactive functionalities and enhanced reading experiences. Modern e-books support different advanced interactive features. Apple's iBooks, for example, supports widgets such as image galleries, audio and video players, review questions, 3D-object viewers, and HTML objects [10]. The integration of multimedia and interactive elements can have a positive impact on the readers' engagement with the book. This may, also enhance their ability to understand and remember the information [16].

Furthermore, such interactive possibilities allow the use and application of reading far beyond traditional reading. Digital reading is an important topic in areas such as leisure, learning, and research. Different school formats already integrate digital reading models into the classroom. Some colleges already use e-book readers such as Amazon's Kindle instead of printed books and also many libraries support digital formats [19]. Fenwick et al. [10] describe the application of interactive digital books as an interactive textbook experience with Moodle integration. In learning scenarios and research it is particularly important to focus on comprehensibility, searchability, and content linking. [15] found that it is important that the digital reading device supports responsive reading techniques and the possibility of generating new texts to support reading in academia.

2.2 Digital Libraries

In online libraries and digitalization projects such as the Google Books Project, the Open Library Initiative, the Internet Archive, or the Million Book Projects a large collection of digital books can be retrieved. Their main purpose is to collect, preserve and organize books in digital form, and make them searchable and retrievable [14].

Many of these projects are converting traditional books into digital representations performing OCR on an industrial scale [5]. This mass digitalization allows users to retrieve a wide range of different books. However, the output is far from interactive content and is mainly used for preservation and discovery [6].

Another issue of traditional digital libraries is the information integration of other sources. How well digital books and digital libraries can cooperate with other information sources is an open question. There is only little work on linked libraries. In this paper we address this gap, offering a web book library solution, which is integrated into the online encyclopedia and media content collection Austria-Forum.

3. *Interactive and Linked Web Book Library*

We have seen different aspects of interactive books and libraries. In this section, we want to introduce our solution of an integrated digital library, which is a collection of interactive and integrated digitized books. Through the strong connection and integration of the web book library with the Austrian online encyclopedia and information portal Austria-Forum [2][8] a new form of reading, researching, and learning is possible.

Web books (also Interactive Internet Books [9]) are interactive and social digitized books supporting reading and searching in both, in the original high-quality scans, or in a cleaned OCR-processed content. Similar to other digital readers, our web book reader assimilates the design of real books.

Figure 1 shows the selection of a book in a bookshelf.

Figure 2 displays the reading experience in a design simulating a printed book. Adding interactive content and new features should enhance this design, but should not distract the user from main purpose, the reading experience. Hence, new forms of interactions are added as layers.

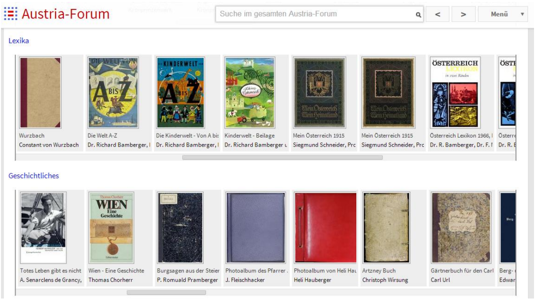


Figure 1 Book-shelve representation in the Austria-Forum web book library

Müller and Maurer [9] describe the structure of web books as a layered structure. Each book consists of four layers: (1) the facsimile layer is used to display the original version of the book. High resolution images give the users the feeling of reading the original. (2) The second layer, the OCR (optical character recognition) / text layer, is the computer-readable text representation of the scanned text content. (3) The enhancement layer allows additional annotations such as personal markers and notes, links, highlights, and nano-publications (*"a set of annotations that refer to the same statement and contains a minimum set of (community) agreed upon annotations"* [7]). (4) The communication layer is responsible for social interactions such as content sharing, social tagging, discussions, reading history visualizations, and search agents.

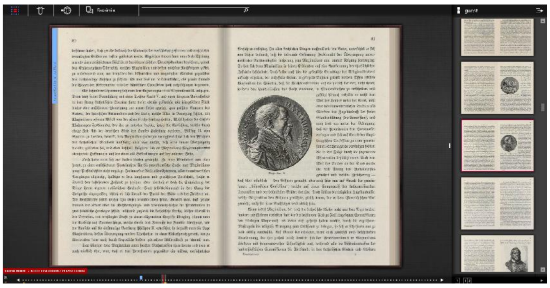


Figure 2 Reading in an interactive web book

This layered structure with some simple interactivities and the connection to the Austria-Forum of our web book library offers the following main features:

(i) Bidirectional behavior of Information Integration.

While most internet books only link into one direction, the integration into the encyclopedia Austria-Forum allows several further features, such as bidirectional linking. Knowledge content from Austria-Forum can be enhanced by web book content and web books can be enriched by Austria-Forum content. This is supported by the two major principles links and transclusions.

Links represent references to other sources. The web book reader does not only support the link to internal and external URLs, but also to different media types, such as images, audio, or videos. (See Figure 3)

Ted Nelson's idea of transclusion (inserting and embedding the original content through shared instancing) [13] is applicable to the library. While the encyclopedia pages already support the inclusion of referenced pages (using 'insert page' commands), the same principle should be possible with book content. Transclusion has different advantages compared to linking. Users have the possibility of exploring the original content instead of looking for the context in linked pages, the linked content always uses the original and updated information, and less disk space is required [11]. More on the integration of transclusion in the context of Austria-Forum can be found in [8] in this issue of this Journal.

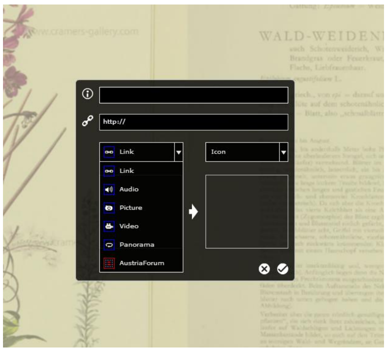


Figure 3 Linking content and media into the web book

(ii) Interactive and Engaging

The web book library supports different interactive functionalities. Users can highlight content, add comments, add bookmarks, or share content. Switching between the original facsimile representation and an OCR-processes text version makes even historic books available to all user groups (e.g. without knowledge how to read fraktur). Users can view the content such as images in original or zoomed size. The search functionality of the book gives the users an overview of the search results on each relevant page. This makes it easy to find context sensitive content (see Figure 4). Not only the results on the page {1}, but also in the entire book {2, 3} are displayed.

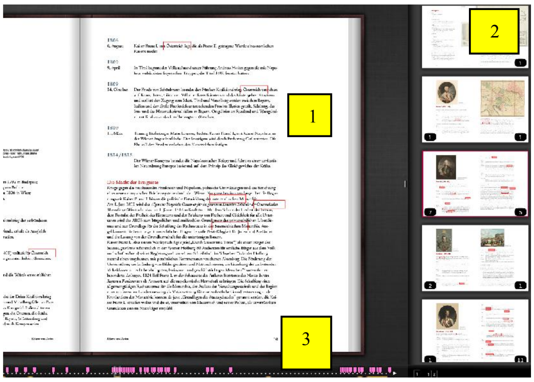


Figure 4 Context sensitive searching in web books

(iii) Social

Other important features which have become more and more relevant in the last years are social and community-based interactions. Users don't only want to read content, they also want to share and discuss content immediately within a social network. The web books support group-intern social interactions, which means that marked, shared, or commented content is only visible to a particular working group. This allows different forms of online collaboration (e.g. collaborative research), community projects (e.g. the manual correction of OCR content), or collaborative e-learning (e.g. team projects for school classes).

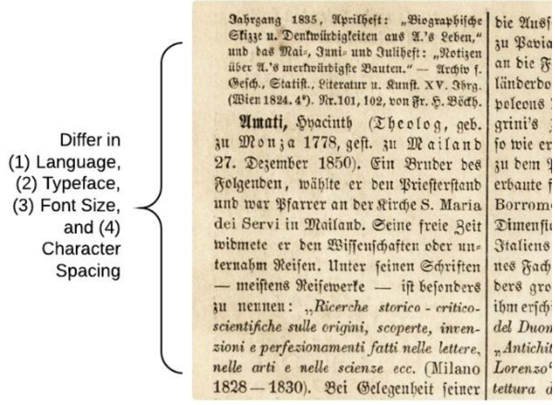
Most of the introduced features are already realized and integrated into our web book solution. Features such as transclusion and group sharing are still under development. In the next chapter we describe the process of converting historic printed books into web books.

4. Integrating Historic Books Into the Web Book Library

The web book library of the Austria-Form is in particular famous for its representation of historic books. In this chapter we discuss the single steps necessary to integrate a (historic) book into our web book library. First, the original book must be scanned and digitalized using OCR software such as Tesseract or Abbyy Finereader.

Digitalizing old books to be representable as an interactive web book, however, brings different implications. Far beyond the implications of optical character recognition of traditional books (e.g. scanning defects), old books often require special attention regarding their font style (often fraktur), their general condition (completeness, color, or damages), and the dictionary used.

An especially interesting case is the biographical 60 volume encyclopedia 'Biographisches Lexikon des Kaiserthums Oesterreich' by Constantin von Wurzbach (also referred to as 'Wurzbach') [18]. The original scans pose several challenges for OCR software and are especially hard to process with traditional tools without further modifications and enhancements. In particular different Fraktur typefaces, frequently changing font sizes and styles, arbitrary character spacing, the use of different languages (German, Latin, and Italian), and obsolete spelling and vocabulary turn Wurzbach into a challenge for common OCR algorithm (see Figure 5).


Differ in
(1) Language,
(2) Typeface,
(3) Font Size,
and (4)
Character
Spacing

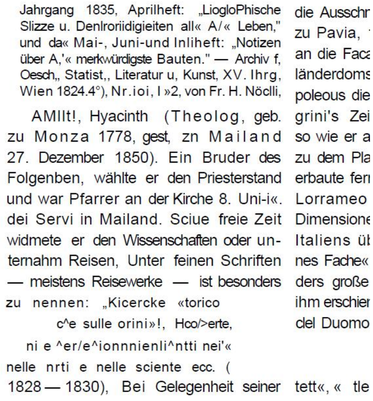
Jahrgang 1835, Aprilheft: „Biographiſche
Skizze u. Denkwürdigkeiten aus A.'s Leben,"
und das Mai-, Juni- und Juliheft: „Notizen
über A.'s merkwürdigſte Bauten." — Archiv f.
Geſch., Statiſt., Literatur u. Kunſt. XV. Jhrg.
(Wien 1824. 4°). Nr.101, 102, von Fr. H. Böckh.

Amati, Hyacinth (Theolog, geb.
zu Monza 1778, geſt. zu Mailand
27. Dezember 1850). Ein Bruder des
Folgenden, wählte er den Prieſterſtand
und war Pfarrer an der Kirche S. Maria
dei Servi in Mailand. Seine freie Zeit
widmete er den Wiſſenſchaften oder un-
ternahm Reiſen. Unter ſeinen Schriften
— meiſtens Reiſewerke — iſt beſonders
zu nennen: „*Ricerche storico - critico-
scientifiche sulle origini, scoperte, inven-
zioni e perfezionamenti fatti nelle lettere,
nelle arti e nelle scienze* ecc. (Milano
1828—1830). Bei Gelegenheit ſeiner

die Ausſ
zu Pavia
an die F
länderdo
poleons
grini's
ſo wie er
zu dem
erbaute f
Borrom
Dimenſi
Italiens
nes Fach
ders gro
ihm erſch
del Duon
„Antichit
Lorenzo
tettura d

Figure 5 Implications of OCR found in Wurzbach

For this special case traditional OCR algorithm fail to recognize the entire text in an adequate quality (see Figure 6). This level of quality would neither allow acceptable reading of the text, nor searching for specific content.



Jahrgang 1835, Aprilheft: „LiogloPhische
Slizze u. Denlroriidigieiten all« A/« Leben,"
und da« Mai-, Juni-und Inliheft: „Notizen
über A,'« merkwürdigste Bauten." — Archiv f,
Oesch,, Statist,, Literatur u, Kunst, XV. Ihrg,
Wien 1824.4°), Nr.ioi, I »2, von Fr. H. Nöclli,

AMIlt!, Hyacinth (Theolog, geb.
zu Monza 1778, gest, zn Mailand
27. Dezember 1850). Ein Bruder des
Folgenben, wählte er den Priesterstand
und war Pfarrer an der Kirche 8. Uni-i«.
dei Servi in Mailand. Sciue freie Zeit
widmete er den Wissenschaften oder un-
ternahm Reisen, Unter feinen Schriften
— meistens Reisewerke — ist besonders
zu nennen: „Kicercke «torico
c^e sulle orini»!, Hco/>erte,
ni e ^er/e^ionnnienli^ntti nei'«
nelle nrti e nelle sciente ecc. (
1828 — 1830), Bei Gelegenheit seiner

die Ausschr
zu Pavia,
an die Fac
länderdoms
poleous die
grini's Zei
so wie er a
zu dem Pla
erbaute fer
Lorrameo
Dimensione
Italiens üt
nes Fache«
ders große
ihm erschier
clel Duomo
tett«, « tle

Figure 6 OCR processed text

Wikisource [17] provides manually created and corrected data of the first volume out of 60. To enhance the OCR algorithm we have created an own dictionary based on this data. This dictionary of the used vocabulary can be generalized to be used as a basis to train the dataset for the remaining volumes. We use the manually created volume to control the quality of our algorithm. Different quality measurements such as Levenshtein distance or text analysis are used to compare the machine-processed text with the manually corrected one, to be able to optimize the algorithm for the remaining 59 volume.

Wurzbach is a unique collection of biographies of persons (in particular of the Habsburg monarchy) living between 1750 and 1850 in the Austrian crown lands. Integrating all volumes in a machine readable way into the Austria-Forum web book library and connect it to the Austria-Forum content will allow users to search, research, and connect persons.

5. *Future of Web Books*

Our future research will focus on accessibility and interaction with knowledge embedded in web books. For this we plan to develop algorithms for automatic knowledge discovery and data mining (KDD) optimized for (encyclopedic) book collections. Such algorithms provide functionality for rule mining, subgroup discovery, graph mining, queries design, and the analysis of structured, cross-domain data sets, and in consequence allow identifying valid, novel, potentially useful, and meaningful patterns within a web book, which can be accessed as linked open data, e.g. in the nano-publication format [7].

Future user interfaces for knowledge discovery in web books will build on well-known principles of communication and social interaction in the support of comments, links, and semantic tagging of statements in nano-publications. For the manipulation of “facts”, we will develop a special editor, which will on the fly generate semantic tagging of statements in the OCR text using software agents, that can deal with open, rich and ambivalent ‘natural’ inputs, such as digitized books and engage the reader to provide their knowledge to train agents or share knowledge directly with other users. Such interactive agents can semi-automatically identify meaningful nano-publications in web books and consolidate them using the knowledge provided by readers of a web book. Finally a web books and its embedded itself becomes the interface, and embodies the link between actors (readers of a books) and concepts, which are described in the formal representation by RDF triplets. By the behavior of a large number of users (who, when, and which parts are used in a certain context) we can build a "social ontology" of the web books, and extract facts as open data which can form the base for a central knowledge repository similar Wikidata or Freebase.

6. *Conclusion*

In this article we discussed different techniques to enhance online reading experiences with focus on integrated, interactivity, and social experiences. We introduced this concept as interactive web books based on the example of a digital library, which is integrated into the online encyclopedia Austria-Forum. This article’s contribution is a model to create book experiences, which are interactive and socially attractive and integrated into a knowledge database to support not only leisure reading experiences, but also innovative and intelligent research and learning.

In future work we include research on automatic knowledge discovery and data mining algorithms for book collections and will integrate concepts such as nano-publications and will introduce user interfaces with even more social and interactive behavior. Furthermore, we focus on the integration of an advanced transclusion system, which does not only allow the integration of books pages into encyclopedia pages, but also additional inserted content into the web books.

Acknowledgments

We would like to thank Hermann Maurer and the members of the Austria-Forum team. The Wurzbach project is done in collaboration with the University of Vienna (Thomas Winkelbauer), and has received much help from the digitization group of the University of Graz (Karl Lenger).

References

[1] Armstrong, C. J., “Books in a virtual worlds: the evolution of the e-books and its lexicon,” The Journal of Librarianship and Information Science 10 (3), pp. 193-206.

[2] Austria-Forum, http://www.austria-forum.org/, 7.1.2014

[3] Burstyn, J., Herriotts, M.A., “gBook: An e-Book Reader With Physical Document Navigation Techniques,” CHI 2010, pp. 4369-4374.

[4] Coyle, K., “E-Reading,” The Journal of Academic Librarianship 34 (2), 2008, pp. 160-162.

[5] Coyle, K., “Mass Digitization of Books,” The Journal of Academic Librarianship 32 (6), 2006, pp. 641-645.

[6] Coyle, K., “One World: Digital,” The Journal of Academic Librarianship 32 (2), 2006, pp. 205-207.

[7] Groth, P., Gibson, A., Velterop, J., “The Anatomy of a Nano-publication,” Information Services and Use 30(1), 2010, pp. 51-56.

[8] Maurer, H., “Austria-Forum and beyond,” Journal of Internet Research, 10 (2), 2014.

[9] Müller, H., Maurer, H., “How to Carry Over Historic Books into Social Networks” BooksOnline’11, 2011.

[10] Fenwick, Jr., J., B., Kurtz, B. L., Meznar, P., Phillips, R. Weidner, A., „Developing a highly interactive book for CS instruction,“ Proceeding of the 44th ACM Technical Symposium on Computer Science Education, 2013, pp. 135-140.

[11] Krottmaier, H., Helic, D., “Issues of Transclusions,” In M. Driscoll and T. Reeves (Eds.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2002, pp. 1730-1733.

[12] Maurer, H., Müller, H., “Can the Web turn into a digital library?,” International Journal on Digital Libraries 13(2), March 2013, pp. 65-75.

[13] Nelson, T. H., “The heart of connection: hypermedia unified by transclusion,” Communications of the ACM 38 (8), Aug. 1995, pp. 31-33.

[14] Schwartz, C., “Digital Libraries: An Overview,” The Journal of Academic Librarianship 26 (6), 2000, pp. 385-393.

[15] Thayer, A., Lee, C. P., Hwang, L. H., Sales, H., Sen, P., Dalal, N., “The Imposition and Superimposition of Digital Reading Technology: The Academic Potential of E-readers,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2011, pp.2917-2926.

[16] Wilson, R., Landoni, M., Gibb, F., “A User-Centred Approach to Ebook Design”, The Electronic Library, 20 (4), pp. 322-330.

[17] Wikimedia Foundation, “Biographisches Lexikon des Kaiserthums Oesterreich,” http://de.wikisource.org/wiki/Biographisches_Lexikon_des_Kaiserthums_Oesterreich, 08.01.2014.

[18] Wurzbach, C. v., „Biographisches Lexikon des Kaiserthums Oesterreich,“ 1891.

[19] Yi, W., Park, E., Cho, K., “E-Book Readability, Comprehensibility and Satisfaction”, Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, 2011.

Authors

Johanna Pirker studied software engineering at Graz University of Technology and concluded with a thesis on virtual worlds for physics education in a joint project with Massachusetts Institute of Technology. Currently she is a researcher at the Institute of Information Systems and Computer Media, Graz University of Technology.

Gerhard Wurzinger is currently a software engineer and researcher at the Institute for Information Systems and Computer Media, Graz University of Technology. After some years of professional experience as software engineer in the automotive industry Gerhard Wurzinger received a degree in Software Development and Business Management at Graz University of Technology. He is the technical lead of the Austria-Forum.

Heimo Müller studied mathematics in Graz and Vienna, concluding with a thesis on data space semantics. As Marie Curie fellow, Heimo Müller worked at the Faculty of Arts of the Vrije Universiteit Amsterdam in an interdisciplinary program on word and image studies and was the founding head of the Information-Design programme of the University of Applied Sciences FH Joanneum in Graz. At the Medical University of Graz he developed an interactive data exploration system and investigated large inhomogeneous clinical data collections and worked on the ICT architecture of the FET flagship proposal IT Future of Medicine (ITFoM).