

How to Carry Over Historic Books into Social Networks

Heimo Müller
Medical University of Graz
Stiftingtalstraße 24
A-8010 Graz, Austria
+43-316-385-72916
heimo.mueller@mac.com

Hermann Maurer
Graz University of Technology
Inffeldgasse 16c
8010 Graz, Austria
+43-316-873-5612
hmaurer@iicm.edu

ABSTRACT

This paper describes how to make use of e-books that look like printed books in a knowledge network. After an overview of digitalization efforts and current digital library initiatives we introduce quality measures for the digitalization process. After digitalization an Interactive Internet Book (IIB) has to offer a kind of digital binding, annotation efforts and sophisticated ways for user interaction. We claim that the quality and the enhancements of an Interactive Internet Book go far beyond what is traditionally assumed: it is not enough to scan books. The scans have to be of high quality, allow good OCR to permit full text searches; books need not only be “packaged” but also need meta-data and functionalities that one can expect from a computer supported medium that go far beyond what is possible with traditional printed books. Those factors are critical for the use of e-books in social media environments, yet this is often still overlooked. Finally, we describe a working prototype and demonstrate the advantages obtained with a use case.

Categories and Subject Descriptors: H.5.0
Information Systems, INFORMATION INTERFACES AND
PRESENTATION (I.7), General

General Terms: H.5 Information Interfaces and Presentation

Keywords: e-book, Human Factors.

1. INTRODUCTION

Social media networks use a variety of digital media related to a certain knowledge space. That space usually consists of two types of material (i) specifically prepared content by a member of the network, e.g. pictures taken from some trip and (ii) existing content and knowledge objects available in general purpose information sources, such as the Wikipedia or other general or special purpose encyclopedias and databases.

One cluster of one general and over 30 special purpose encyclopedias is the Austria-Forum [1]. In contrast to other efforts it has a number of distinguishing features described in [2]. Some of the outstanding are: Contributions have a well-defined source (author or archive) with a description of the source; contributions

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BooksOnline '11, October 24, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0961-5/11/10...\$10.00.

have a time stamp, hence they can be quoted like contribution in journals or such. Special search facilities allow to restrict the search to a subset of the material and/or to use meta-data for locating information otherwise impossible to find. Further, different views of the same item (from different sources and potentially different moments in time) are welcome to allow users to form their own opinion based on pointed presentations, rather than being offered an unsatisfying compromise of opinions as it is often the case with standard encyclopedias.

We have further added so-called bookshelves to the Austria-Forum. They look similar to a real bookshelf and each bookshelf is dealing with different domains: one maybe with encyclopedias, one with books on history, another one with novels, another one with biographies, or on travel, etc. Each bookshelf contains a set of what we call Interactive Internet Books. Interactive Internet Books aim to transform traditional books, classical textbooks and (historic) encyclopedias into the digital domain and interlink them tightly with other entries in the knowledge network, in our prototype predominantly with the Austria-Forum. Lest we are misunderstood: the Austria-Forum is not some kind of interesting prototype but is (after Wikipedia) the world-wide largest Wiki system, with currently over 220.00 “objects”, and growing.

The mentioned Interactive Internet Books preserve, on the one hand, the qualities of analogue textbooks through the provision of a facsimile version supported by a dedicated viewer, and add, on the other hand, enhancement layers and cross-linking functionalities not available in a classical book. The ability to heavily cross-link a textbook with other information sources and to add (personal) remarks and links to a textbook turn Interactive Internet Books into a valuable teaching and learning tool. However, the huge amount of already digitized books, see chapter 2 for an overview, can only be exploited, if the following three prerequisites are met:

1. **Adequate quality level in digitization:** The error rate of the scanning process and optical character (OCR) recognition OCR has to be reasonable low. Defects often found are: bent-over pages, off-focus scans, errors due to moving a page during scanning, wrong language settings for the OCR, etc. A low error rate in scanning and OCR is a prerequisite for (2) and (3) and is not always easy to achieve in books with special or older fonts.
2. **The book has to be integrated into a digital knowledge space.** This implies the generation of a search index from the text version, extraction of the book’s table of contents and of indices of item/persons, extraction of images and image captions, the possibility to address a specific part in

the book by an URI and the problem of information consolidation (the same book scanned by two institutions has to be recognized as the same object).

3. **Navigation and Interaction** have to meet the standard of printed books but also integrate new communication facilities. Besides standard viewing functionality (zooming, panning and rotation) and full-text search an Interactive Internet Book should support book marking, annotations, link creation and information sharing (Facebook, Twitter) of selected parts.

The above listing is not at all complete. Yet it shows that building a usable library of Interactive Internet Books does not at all stop with large scale digitalization, but goes far beyond this, a fact not enough acknowledged in most massive digitalization efforts ongoing right now.

Of course we are aware of the fact that historic textbooks and encyclopedia are only one part of the digital book universe. However, a better integration of historic reference material will allow building on proven sources and extending the digital space by large collections of existing books that are being digitized at the moment by several institutions.

2. RELATED WORK

One definition and history of the term “e-book” is found in Armstrong [3]. He distinguishes between digitized and born-digital book. The first e-books, like those in the Gutenberg Archive, were manually typed. However, this manual digitalization approach was soon replaced by book scanning and optical character recognition (OCR). Optical character recognition is the translation of scanned images into machine-encoded text. It makes it possible to reduce the file size dramatically, to search for a word or phrase, to extract content and to display a page without scanning defects. When high quality scans are available, the accurate recognition of Latin-script text is a solved problem today. However, the recognition of texts in other scripts, e.g. blackletter typefaces or Asian language, bad (historic) prints, multilingual documents and documents with historic text corpora are still subject to research. In addition to the pure text information OCR software also should be able to recognize font size and type and simple layout information. Much information is implicitly embedded in the layout of a document. A heading, an image caption, the table of contents, an alphabetic index of terms, etc. should be treated different from ordinary text, yet this happens only rarely. Our effort involved evaluating several OCR solutions, especially the commercial product Finereader 10 (finereader.abbyy.com) and the open source software tesseract (code.google.com/p/tesseract-ocr).

The result of the OCR, optionally together with the originally scanned image, is finally stored in a computer readable format. There exists “myriads” of e-book formats each with its own advantages and limitations, which build together the Tower of e-Babel. This “is the bane of publishers, online retailers, librarians and book-lovers” [4]. We agree with the “Tower of e Babel” metaphor and therefore propose not to focus on format issues, but to make e-books online available in a Web application and to access the books through an open protocol. Nevertheless, some formats must be supported for data import and exchange. Here we recommend in addition to PDF and simple text based formats, the EPUB and DjVu document standards.

There are numerous digital libraries and book collections initiatives. The biggest is the Open Content Alliance (www.opencontentalliance.org) storing 1,6 million books hosted by the Internet Archive (www.archive.org) and describing 23 million books by the Open Library Initiative (openlibrary.org). The Million Book Project or Universal Library (www.ulib.org) from Carnegie Mellon University School of Computer Science and University Libraries with partners in India and China has digitized up to 2007 around 1 million books. It seems that the Million Book Project has been in trouble since 2007, as there is no progress and also no official statement on their web page. Some 100k books of the Million Book Project are hosted today in the Internet Archive. The European counterpart to the U.S. and English language dominated library projects is the Europeana (www.europeana.eu/portal). The Europeana is a central search portal for Europe's museums, libraries, archives and audio-visual collections including metadata for 15 millions items from over 1500 institutions. All digital objects are still stored and presented at the local institutions, e.g. the Bavarian State Library (www.bsb-muenchen-digital.de) or national portals, e.g. Austrian Literature Online (www.literature.at).

In contrast to the open access initiatives Google aims to scan several major research libraries of universities or other institutions, such as Harvard University, New York Public Library, Stanford University, University of Michigan, Columbia University, University of Oxford, Ghent University Library, National Library of Catalonia, Bavarian State Library, Austrian National Library and many more [5]. If the book copyright free copyright it can be downloaded with a watermark, read in Google's own e-book store (books.google.com/ebooks) or it can be found in further digital libraries, e.g. the Internet Archive. There is much criticism of Google's approach, concerning both the 'de facto monopoly' and copyright issues and with the missing quality control, like and poor accuracy of the book scanning and OCR. [6] [7] [8] [9].

Beyond general-purpose image classification algorithms no scientific investigation of the evaluation of overall scan quality analysis is known to the authors. For typewritten documents Cannon et al. [10] describe how to measure the small speckle factor (amount of black background speckle), white speckle factor (fattened character strokes), touching character factor (degree to which neighboring characters touch), broken character factor (degree to which individual characters are broken) and the font size factor (degradations that accompany an increase or decrease in the size of the font). Holley [11] gives some insights into analyzing and improving OCR accuracy in a large-scale historic newspaper digitalization project. He states that “The question of what is acceptable has not been answered, but in speaking to other libraries and OCR contractors, it was generally agreed for historic newspapers that good OCR is given by 98-99% accuracy, average by 90-98% and poor OCR is defined to be below 90%”.

For the automatic evaluation of OCR quality there exist several approaches. Feng and Mammatha [12] proposed a Hidden Markov Model based on a hierarchical algorithm to align OCR output and the ground truth for books. Reynaert [13] describes an automatic system for reducing the level of OCR-induced typographical variation in large collections of text. He proposes a text-induced corpus “clean up” based on high-frequency words derived from the corpus and all typographical variants for sets of words. Déjean and Meunier [14] describe how to extract logical structures from digital book collections by recognizing logical

elements (page numbers, chapter headings) and using this information for content navigation. Several e-book evaluation efforts were done, especially in educational multimedia and teaching, e.g. the JISC national e-books observatory project [15] and the Active Reading Task under the INEX book track [16].

The concept of e-books extended the limitations of paper already as it was born [17]. Armstrong and Lonsdale identified 12 specific types of added value, such as: resource links, links to reviews, author biographies or links to curricula, professors' or other educational sites, etc. In early days of library research a "Multivalent Browser" architecture was developed by Phelps and Wilensky [18], pushing the concept of extensibility and extension mechanism to the extreme. Our viewer is based on these basic concepts and supports the behavior of multipage support for scanned paper.

Today a number of additional features, e.g. the possibility to manipulate the data [3] and the integration of books into social media networks might be included. Carden classifies new paradigms for e-books covering the areas encyclopedias (databases), e-learning, academic monographs, narratives and picture-books [19]. He covers a broad range of content types, access behavior, commercial models, page layouts (intelligent re-flowable text), readings devices and formats. He only briefly mentions the specific requirements for Facsimile versions of (copyright-free) historic books, however.

Kim, Farzan and Brusilovsky have implemented in KnowledgeSea II project an e-learning system for the spatial annotation of scanned textbooks and with social navigation support from lectures to relevant online tutorials in a map-based horizontal navigation format [20]. They use annotations to enhance social navigation and guide readers through so-called cells to the most interesting pages. Pearson and Buchanan introduced very similar concepts for collaborative annotation for mobile reading devices (iPads) [21]. Their system provides a close working environment by combining portable digital annotation devices (using a gesture based interface) with a real-time digital collaboration tool.

Mazza describes how to interweave an e-Book with an interactive on-line resource for a university course on Information Visualization [22]. In this case-study he identifies as advantage that an e-book can be printed and used like any other textbook, but in addition provides basic references to eLearning materials provided by a course management systems through web-links. Ribiere, Picault and Squedin introduce the concept of sBooks, where the "s" stands for "social" [23]. They provide an extensive collection of ideas how an e-book could act as the main interface in e-learning environments and describe the sBook as a trigger for social interactions. Their concepts compromise ideas for annotations, conversations and active learning and propose new visualization techniques for sBook annotations in collaborative learning. The ideas do sound good (but like in other cases) have never been implemented.

Warren points out in his report on the progression of digital publishing and innovation of e-books that in reality e-books today (2010) are a "picture of book" – a book that has been digitized adds little value besides improved portability and search functionality. [24]

Warren uses the following quotation of Ray Bradbury's novel *Fahrenheit 451* to advocate a digital genesis of books which quote to conclude the introductory chapters.

"It's not books you need, it's some of the things that once were in books. Books were only one type of receptacle where we stored a lot of things we were afraid we might forget. There is nothing magical in them at all. The magic is only in what books say, how they stitched the patches of the universe together into one garment for us." (Bradbury, *Fahrenheit 451*)

3. CARRYING OVER HISTORIC BOOKS

3.1 Quality Measures

We have developed a quality classification system to describe the scan quality, OCR quality and the interface of a digital (historic) book. For each category a book is "star rated" with five stars indicating a perfect (luxury) digital version of a book.

3.1.1 Scan Quality

The rating of the scan quality is determined in the following sub-categories and defect definitions:

- **Color Reproduction:** Defect free scanning can be achieved by color calibrated scanning, preferable in 16 bit resolution per color channel. Minor defects are slight color variations. Major defects are given by clear visual differences, e.g. by an unintentional automatic color adjustment or interfering light during the scanning process. For 100% true color reproduction a "rainbow strip" is digitized with the color material. The viewer's screen has to be adjusted so that the rainbow on the screen matches the physical (paper) rainbow strip the user hold to the screen for comparison.
- **Precision of Register:** A minor defect is given by a displacement in the range of ¼ em to 1 em and a major defect by a displacement > 1 em. (1 em is the typographic measurement equivalent to the size of the letter "m" of the books standard font). In addition to the planar accuracy of fit of single pages, also the rotation of scanned pages have to be considered with a minor defects in the range of 0.5 to 3 degrees and major defects over 3 degrees of rotation.
- **Scanning Defects (mechanical):** Minor defects are small blurring, minor out of focus, minor dust spots and other small defects. Major defects are moving pages during scanning, cracked pages or quite out of focus areas.
- **Completeness:** A minor defect is given by the absence of some pages without (textual) information, e.g. the inner face of the cover. A major defect is defined by the absence of already one page with textual information.
- **Resolution:** A scan resolution below 72 dpi results in 0 points, 72-149 dpi in 1 point, 150 - 299 dpi in 2 points. 300-599 dpi in 3 points, 600– 799 dpi in 4 points and above 800 dpi in 5 points.

In each category (except the resolution) we assign points for the whole book according to the following rules:

- | | |
|----------|--|
| 0 points | very bad: Not usable for reading and/or further processing. |
| 1 point | bad: The book scan is missing major pages almost every page has defects, major defects on more then 10% of the pages. |

2 points	poor: The book scan is missing minor pages. Almost every page has defects, major defects on less than 10% of the page.
3 points	fair: The book scan is complete. No major artefacts, minor defects on more than 10% of the pages.
4 points	good: The book scan is complete. There are only tiny defects. Minor defects are on a limited number of pages (10%).
5 points	very good: The book scan is complete. There are no defects. The scan is suitable for facsimile reproduction.

Table 1 shows the mapping of the overall points (sum of the categories: color reproduction, precision of register, scanning defects, completeness and resolution) to a star rating system.

Table 1. Scan Quality

Points	Quality	Star Rating
< 10	Bad	*
10 – 14	Poor	**
15 – 19	Fair	***
20 – 22	Good	****
23 – 24	Very Good	*****
25	Reference / Facsimile	*****

3.1.2 OCR Quality

Based on our literature review [11] we recommend for the classification of the OCR quality the mapping of OCR character accuracy to a star rating system as described in table 2.

Table 2. OCR Quality

character accuracy	Quality	Star Rating
<80%	Bad	*
80%-90%	Poor	**
90%-98%	Fair	***
98%-99%	Good	****
>99%	Very Good	*****
100%	Reference / Facsimile	*****

When doing the evaluation, it is important that the numbers refer to character accuracy and not to word accuracy or some confidence level given by the OCR software. It is also important that not only pages with pure text, but also pages with a mixture of text, illustrations, graphical drawings etc. are taken into account.

3.1.3 Book Interface

The rating of the book interface is done using the following categories:

- **Page Access Speed:** When reading and browsing a book it is essential to be able access a single page fast, like when reading an paper book. In this category we award 5 points if the book interface supports a flip-through mode in full resolution, 4

points for a average latency (AL) below 20ms, 3 points for AL up to 0.8 seconds, 2 points if the AL between 0.8 and 2 seconds and 1 point for AL greater then 2 second. If the interface does not support low-resolution previews the score is reduced by 1 point.

- **Navigation and Presentation:** For each of the following features the category score is given one point. Navigation Toolbar/Buttons, Thumbnail View, Page Number Input, Table of Content, Preview for Navigation, Zoom, Rotate, Fullscreen Mode, Search, Preview of Search Results. For each underlined (main) feature missing one penalty point is subtracted
- **Personal / Group / Social Annotations:** In this category functionalities as described in the enhancement layer of Interactive Internet Books (section 3.3.3) are counted. The sum is finally normalized to a maximum of 5 points.
- **Integration into Knowledge Spaces:** In this category functionalities as described in the communication layer of Interactive Internet Books (section 3.3.4) are counted. The sum is finally normalized to a maximum of 5 points.
- **Design and Aesthetics:** This category is maybe the most difficult to judge, but as is clear from the success of Apple products, it is a very important one. However, the definition of the criteria for design / aesthetics is beyond the scope of this paper. We propose therefore either to consult experts from the design community and/or to use a public vote for a simple evaluation.

Table 3 shows the mapping of the overall points to a star rating system.

Table 3. Evaluation of the Book Interface

Points	Quality	Star Rating
< 10	Bad	*
10 – 14	Poor	**
15 – 19	Fair	***
20 – 24	Good	****
25 – 30	Very Good	*****
30	Reference Viewer	*****

3.2 Interactive Internet Books

Interactive Internet Books extend the existing E-book universe in two ways:

1. An Interactive Internet Book (IIB) augments an e-book by enhancement- and communication layers which allow personal and group annotation and editing of links. It also provides interfaces for social interaction and social formation of ontologies, see figure 1. Each page and object in an Interactive Internet Book has its unique URI, which allows to link from every webpage and/or knowledge space to a specific item in the book (and of course conversely).
2. An Interactive Internet Book captures the emotionality of a book. It gives the reader both the possibility to read a book in a cleaned OCR/text style, or as high quality facsimile view.

The publishing of an Interactive Internet Book starts just after the scanning process. All additional metadata is generated during this

process is stored in a XML description. This XML description, together with the post-processed and analyzed raw images builds the static part of an Interactive Internet Book. The dynamic part (index, user generated annotation) is stored in a relational database.

The static description of an Interactive Internet Book consists of well-known metadata for published books (e.g. the Dublin Core meta-data specification) and the definition of book layers. In our first prototype we have defined the structure of the following four layers. The implementation of the described functionality is almost finished and most facilities are offered right now. Single tools will be added to the online version of the viewer step by step within the next months.

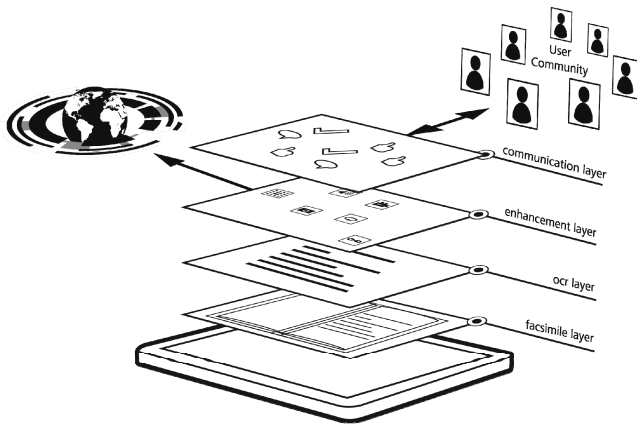


Figure 1. Interactive Internet Books Layer Architecture

3.2.1 Facsimile Layer

Within the facsimile layer the user can access the (raw) scan images. As a raw scan image is in most cases not trimmed and stored in an uncompressed format, these images have to be pre-processed. In your “digital binding” process we use the open source software scantailor to split, de-skew and trim the single scan pages, to define text content and margins and to produce a high resolution intermediate (digital master) for all further processing steps. For online reading the high resolution images are downscaled and converted to an appropriate compressed format. At the time we support the SWF format for flash based viewers and jpg format for web based viewers and mobile devices. A single page is stored in a thumbnail resolution (approx. 200 x 200 pixel, average file size 5-15 kB) and a online reading resolution of 150 dpi, which results in an average file size of 300-600kB per page for a historic book.

3.2.2 OCR / Text Layer

In the OCR/text layer the results of the optical character recognition are presented to the user. The file format for a single page can be either vector / text based, or if requested by the viewer the pages can be again converted into a raster format (jpg) in thumbnail and reading resolution. Our implementation currently supports the PDF format (native format for iOS devices), the SWF format (native format for flash based viewers), the JPEG format (web based viewer and iOS thumbnails) and text format (for the integration into the search index of the Austria-Forum wiki system). The file size of a single page in the OCR/text layer depends on the percentage of graphic and image content within a page. For text only pages the average file size can go down to 20kB. For pages with a high

amount of graphic and image areas the file size can be almost the same as in the facsimile layer.

3.2.3 Enhancement Layer

The enhancement layer provides functionality for personal and group annotation, see figure 2. Interactive Internet Books support the following basic set of tools:

- **Page Markers:** A digital “Post-it note” with optional text available in the typical post-it colors. Page Markers can be auto-aligned and manipulated on a group base, e.g. to delete all marker of certain color or hide all marker of a user group.
- **Links:** Reference to knowledge objects described as hyperlink. A link is defined by a descriptive text, a rectangle (hot area), a display group, display type (rectangle and symbol) and a predefined symbol type for audio, image, movie, panoramic images and hyperlink to the Austria-Forum.
- **Personal Notes:** Remarks and notes within a text page. Notes can be simple text, HTML / wiki formatted text areas and can display tweets. With these feature personal notes can be shared via some micro-blogging service and also collected on a user’s twitter and/or Austria-Forum profile page.
- **Highlighter:** A digital form of the well-known felt-tip pen which is used to draw attention to sections of documents by marking them with a vivid translucent color.
- **Nano-Publications:** The Concept Web Alliance has promoted the notion of nano-publications (core scientific statements with associated context). [25] The viewer will allow the definition of a nano-publication by its subject, predicate and object using elements of a book and will visualize the statements as RDF triplet.

- Page Markers
Favorites and Bookmarks
- Links
Reference to knowledge Objects
- Personal Notes
Text, Audio
- Highlighter
Transparent & fluorescent
- Nano Publications
RDF Triplets for Sematic Tagging

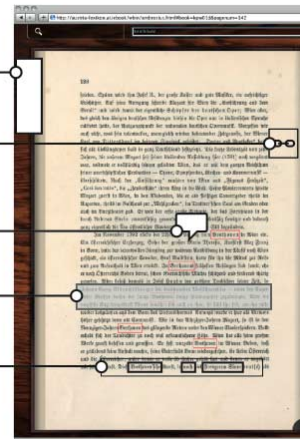


Figure 2. Enhancement Layer

3.2.4 Communication Layer

The communication layer provides functionality for group interactions and social media networks. An Interactive Internet Book supports the following basic set of tools:

- **Content Publication:** A whole book, a page or part of a page can be shared with friends using (existing) social networking services and (micro) blogging solutions.
- **Social Tagging:** Users can collectively classify, annotate and categorize parts of a book (Folksonomy [26]).

- **Discussions:** A discussion forum and chat functionality is included within an Interactive Internet Book. So the book itself becomes the interface for social interactions.
- **Reading History:** Interactive Internet Books visualize the personal and group reading history as well as annotations and links with the help of a heat-map.
- **Search Agents** support search by example and allow inductive conceptual indexing of the books.

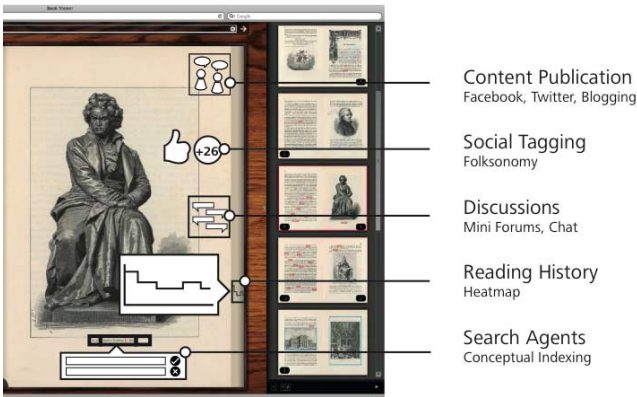


Figure 3. Communication Layer

Conceptual indexing [27] is performed by the following steps:

- In the first step a ‘topic filter’, e.g. find terms associated with a topic such as ‘church music’ or ‘Beethoven’ or ‘war’ is defined.
- A clustering algorithm on a collection of relevant texts weighted by the topic filter (i.e. terms associated with the topic get a higher weight in clustering than non-related terms) is performed.
- The reader identifies the clusters, i.e. assigns a short label, and an additional construct so that a classifier can select items in this cluster from the universe.

The combination of a classifier, a cluster of text and a label together produce a concept that can be used to structure the relevant information concerning a topic in a collection of Interactive Internet Books.

Interactive Internet Books focus strongly on communication and social interaction of communities. They incorporate the link between actors (readers of a book) and concepts embedded in a book. By keeping track of what readers, when and in what context produced, edited or consulted a social ontology can be generated. Such bookkeeping of facts and readers together provides the basis for collaborative filtering (recommending) and for collaborative moderation (evaluation of relevance and quality). An Interactive Internet Book can even generate requests, e.g. for tagging of nano-publications, annotation or evaluation tasks, to readers in such a way as to engage rather than irritate them.

4. USE CASE

In our hypothetical use-case a scholar (music student) writes an essay about the relationship between the composer Ludwig van Beethoven and the Viennese society at the end of the 19th century, in particular the relation of Beethoven and Franz Ludwig von Hatzfeldt.

For this undertaking he needs reliable sources, which provides in-depth information preferential from primary literature. He already knows that the Austria-Forum contains an online version of the encyclopaedia „Die österreichisch-ungarische Monarchie in Wort und Bild“ published between 1884 and 1902. In volume 1 „Wien und Niederösterreich“ he searches for „Beethoven“ and gets all pages, which contain the text Beethoven, see supplemental material No 1.* The search term is marked in the pages and the heat-map and the thumbnail view show the number of occurrences within a page. As the student cannot read fluently text in black-letter fonts he switches from the facsimile view to the OCR/text layer. He marks the page with a personal bookmark for further reading. At the paragraph of Beethoven Viennese time, he puts a link on the name of “Johannes Georg Albrechtsberger” to introduce this composer to his colleagues.

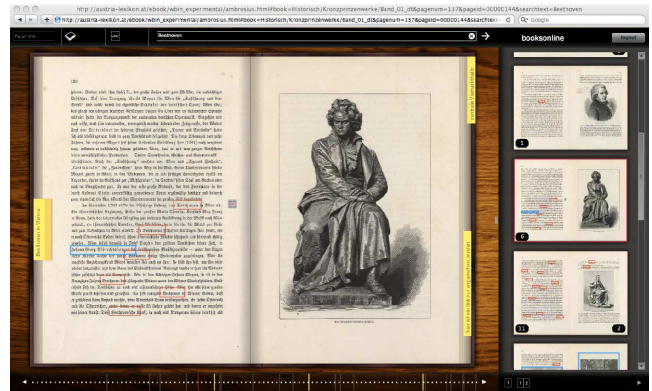


Figure 4. Search for “Beethoven” in an Interactive Internet Book

The student follows the link to the Beethoven article in the Austria-Forum, supplemental material No. 2, and finds the fact, that the first scientific biography was written in 1864, 1867, 1877 by Ludwig Nohl in a 3 volumes edition.

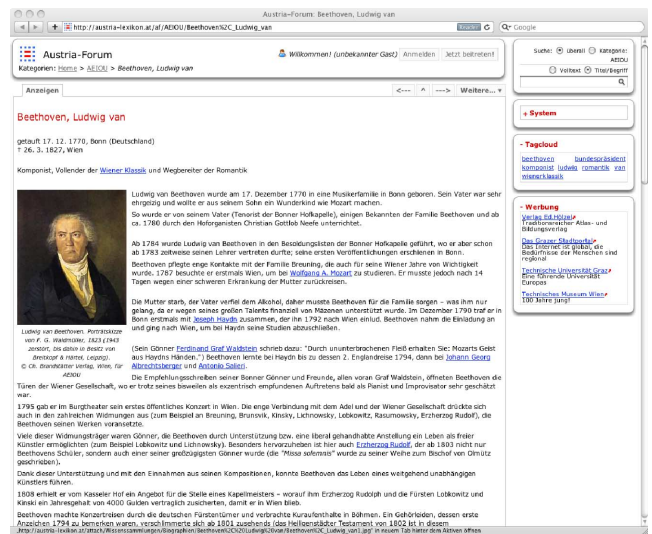


Figure 5. The “Beethoven” Article in the Austria-Forum

* the number (x) indicates the link in the supplemental material, <http://www.austria-lexikon.at/af/Austria-Forum/Supplement> Please login to the Interactive Internet Book with the username bookonline and the password 2011.

This is exactly the kind of primary literature as indicated by his professor. The student is happy, that all 3 volumes of Nohl's Beethoven biography are already digitized. The student looks at the Internet Archive, supplemental material No. 3, and finds several digitized books of Ludwig Nohl, amongst other 9 versions of the Beethoven biography.

After some browsing through the digitized books, the student realizes a big fluctuation in scan quality and completeness of the 9 versions, supplemental material No 4. The student makes an evaluation on the book quality, in order to base his work on reliable sources, see Table 4.

Table 4. Scan Quality "Beethoven's Leben" Internet Archive

Book id	Vol.	pages	comment	Scan quality
A	1+2	1086	2 volumes together	**
B	1	471	good scan	***
C	1	474		**
D	2	394	first chapters missing	*
E	2	616	book has 669! downloads	**
F	3/2	548	yellowed pages	*
G	3	470		*
H	3	1005		**
I	3	396		*

The student discovers further digital editions of "Beethoven's Leben" in the Europeana, supplemental material No. 5 and Table 5. As the Europeana is limited to catalogs of national libraries it only refers to books in its original context, in our case the "Bayrische Staatsbibliothek" (national library of Bavaria).

Table 5. Scan Quality "Beethoven's Leben" Bayrische Staatsbibliothek

Book id	Vol.	pages	comment	Scan quality
J	1	465	damaged cover	***
K	1	465	nice cover	***
L	2	610	complete scan	****
M	2	612	nice cover	****

Europeana's approach to aggregate only the metadata and access the books from the local library sites allows, on the one hand, the local library to brand the content with their identity, but results, on the other hand, in a great deal of book readings interfaces. One point of the criticism is that Europeana expects users to search first at the Europeana site, rather than go to Google and be redirected [28]. Our small test case has shown that the Europeana should not only aggregate everything, but also take care of being aggregated and interlinked in order to achieve a good visibility for its content items.

For the evaluation of the OCR quality we determined the character accuracy for a random sample of 10 pages for each book. The impact of the OCR quality for full text search can be clearly seen in the comparison of search results for the word "Wien".

Table 6. OCR Quality "Beethoven's Leben"

Book id	Vol.	character accuracy	"Wien" search	Wien search %	Star Rating
A	1	<80%	275	63%	*
B	1	92%	107	90%	***
C	1	<80%	12	10%	*
J	1	>99%		100%	****
K	2	no OCR available			
D	2	<80%	14	4%	*
E	2	84%	197	62%	**
L	2	>99%		100%	****
M	2	no OCR available			

Finally we have done an evaluation of the 3 different book-reading interfaces. The Internet Archive reader, see supplemental material No. 7, provides a (book like) two-page view, a well-structured interface and a fast response time. In particular the preview functionality for full text search and the book sharing functionality are of great use. Room for improvements are in the book marking and annotation section and in page zooming.

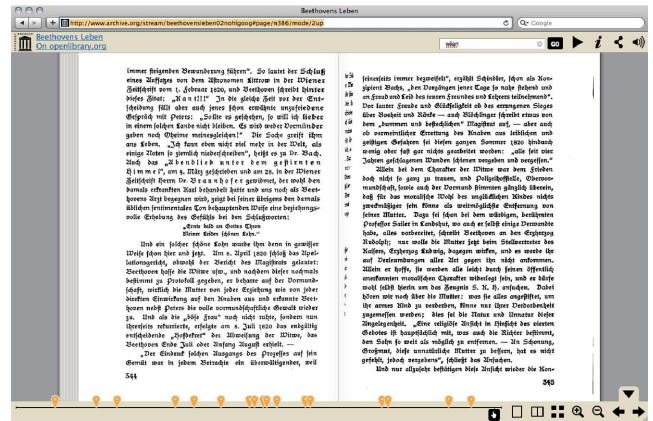


Figure 6. Interface of the Internet Archive

Figure 7 shows a screenshot of the reader of the national library of Bavaria, supplemental material No. 8.

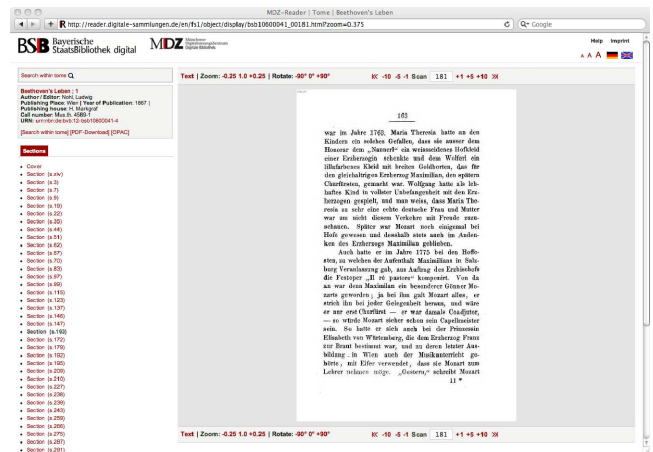


Figure 7. Interface "Bayrische Staatsbibliothek"

The reader of the national library of Bavaria, (Müncher Digitalisierungs Zentrum Digitale Bibliothek, **MDZ**), is missing most of the functionalities of a basic book reader. The sections in the navigation bar on the left side seem to be randomly chosen and do not refer to the book structure, but are related to chunks of pages of the scanning process. A user will mostly download the PDF version. However, in the PDF file the search functionality is not included which again makes the book not suitable for many application scenarios.

Figure 8 shows the interface of our Interactive Internet Book implementation, supplemental material No. 9a.

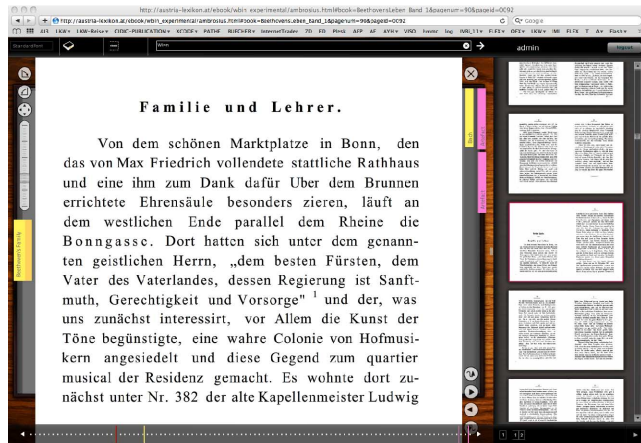


Figure 8. Interactive Internet Book OCR/text Layer

An Interactive Internet Book provides a facsimile and OCR/text viewing mode. The user can switch with one click between the two modes, e.g. to evaluate the scanning and OCR quality. In our example we have marked two typical scanning errors at the end of the book with a pink post-it, see supplemental material No. 9b and Figure 9.

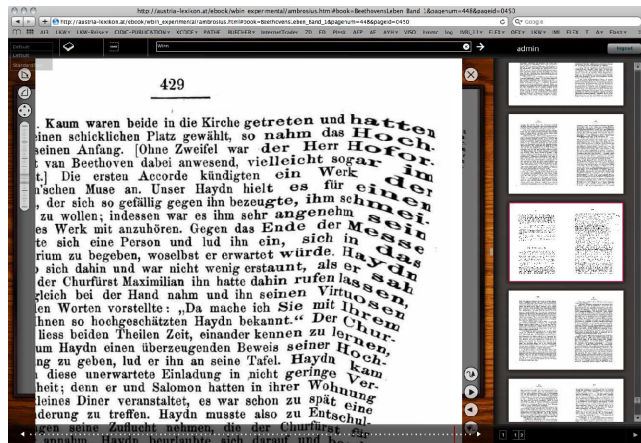


Figure 9. Interactive Internet Book Zoom in the Facsimile Layer

An Interactive Internet Book supports page rotation and continuous zoom. The zoom mode is activated by a simple mouse click into the page which can then be panned and rotated to an appropriate reading/viewing position. When the zoom mode is activated the user can still navigate through the book and switch between the facsimile and OCR/text layer.

The navigation bar in the bottom of the interface provides a small preview picture for fast navigation and a heat-map to indicate annotations and the reading history.

The thumbnail window on the right side can display 1 or 2 page style thumbnails, a hierarchical table of content (TOC) and optional also an index, see supplemental material No. 9c and Figure 10.

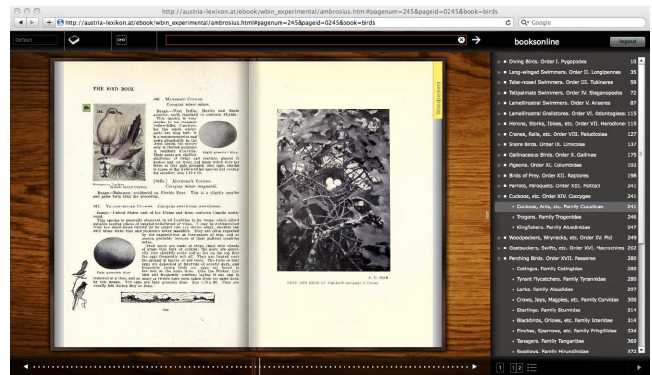


Figure 10. Interactive Internet Book - thumbnail view, table of content

Every reader can set personal bookmarks and define links and multimedia annotation, see supplemental material No. 9d and Figure 11. Bookmarks and links can be visible to all other readers, if the user has the right to the define the visible for the group "word", to selected user groups or only for the user, who generated the item.

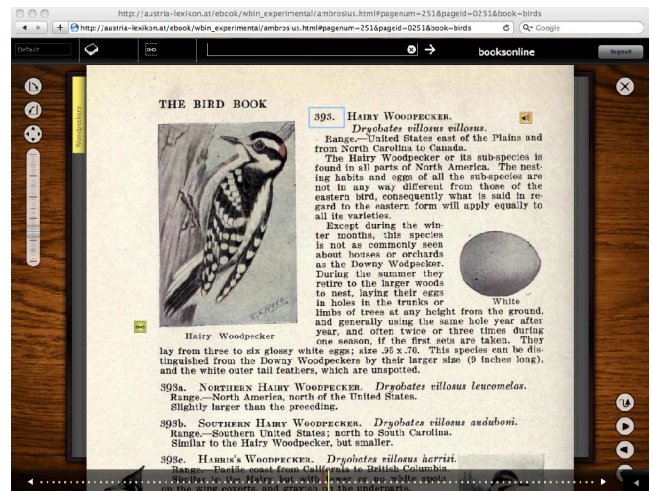


Figure 11. Interactive Internet Book - Links and multimedia annotations

Using the criteria described in section 3.1.3 we compared the interface of the Internet Archive book viewer, the viewer of the "Müncher Digitalisierung Zentrum Digitale Bibliothek" (MDZ-DB) and the Interactive Internet Book (IIB). This comparison should not be seen as a competition, but as example how a common (minimal) standard for e-book interfaces can be defined. The links in the supplemental section of the Interactive Internet Book implementation point to the development version of the viewer which is not stable in every feature, but give valuable insights into ongoing developments and project's progress.

Table 7 shows details of the interface evaluation. The category “aesthetics” was determined by a small survey (5 colleagues at each University).

Table 7. Interface Evaluation

	Internet Archive	MDZ-DB	IIB
Access Speed	4	3	4
Navigation, Presentation	8	3	8
Annotation	3	1	4
Integr. Knowledge Spaces	4	1	3
Aesthetics	4	3	4
Sum	24	11	24
Star Rating	****	**	****

5. CONCLUSION

In this paper we analyze some prerequisites for the transformation of historic books into the digital word. We argue that the quality of the book scan is essential for all later steps and that the book interface should on the one hand (technically) extend the possibility of a paper book but on the other hand also should capture the emotionality of book reading. We have further explained our approach to interweave books, in particular encyclopedias with digital knowledge spaces (Austria-Forum). Our future work will, in addition to the completion of the described features, deal with mobile viewing devices, multilingual issues, algorithms for automatic quality evaluation, and information integration (merging different digital editions of one book).

6. ACKNOWLEDGMENTS

Our thanks go to Stefan Sauer, Wilhelm Steiner, Gerhard Wurzing, Sabine Erking and Katharina Asbäck for their contributions and critical reviews, to the Internet Archive (Prelinger Library, Marcus Lucero) for the book scan of “The bird book”, to the Bayrische Staatsbibliothek for the book scans of “Beethoven’s Leben” and especially to Hans Petschar of the Austrian National Library for the book scans of the 24 volume encyclopaedia „Die österreichisch-ungarische Monarchie in Wort und Bild“.

7. REFERENCES

- [1] The Austria Forum, 2011. <http://www.austria-lexikon.at/>, last visited July 1st, 2011.
- [2] Maurer, H., Müller, H. 2011. *How to use the Web's information flood for teaching*. Proc. ED-MEDIA 2011, Lisbon, Portugal, pp. 3103-3108.
- [3] Armstrong, C.J., 2008. *Books in a virtual world: the evolution of the e-book and its lexicon*. in Journal of Librarianship and Information Science, 40 (3), pp. 193-206.
- [4] Rothman, D.H. 2006. *Razing The Tower of e-Babel, The reason e-books haven't caught on is simple: they're too complicated*. <http://new.publishersweekly.com/pw/by-topic/columns-and-blogs/soapbox/article/8355-razing-the-tower-of-e-babel.html>, last visited July 1st, 2011.
- [5] Vincent, L. 2007. *Google Book Search: Document Understanding on a Massive Scale*. in Proc. of the ninth International Conference on Document Analysis and Recognition (ICDAR), pp. 819-823.
- [6] Nunberg, G. 2009. *Google's Book Search: A Disaster for Scholars*. <http://chronicle.com/article/Googles-Book-Search-A/48245/>, last visited July 1st, 2011.
- [7] Duguid, P., 2007. *Inheritance or Loss: A Brief Survey of Google Books*. First Monday, 2007 12(8).
- [8] Darnton, R. 2009. *Google & the Future of Books*, <http://www.nybooks.com/articles/archives/2009/feb/12/google-the-future-of-books>, last visited July 1st, 2011.
- [9] Cohn, C., Hashimoto, K. 2010. *The Case for Book Privacy Parity: Google Books and the Shift from Offline to Online Reading*, <http://hlpronline.com/2010/05/the-case-for-book-privacy-parity-google-books-and-the-shift-from-offline-to-online-reading/>, last visited July 1st, 2011.
- [10] Cannon, M., Hochberg, J., Kelly, P. 1999. *Quality Assessment And Restoration of Typewritten Document Images*. in International Journal on Document Analysis and Recognition Volume 2, Numbers 2-3, 80-89.
- [11] Holley, R. 2009. *How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs*, D-Lib Magazine, March/April 2009 Volume 15 Number 3/4.
- [12] Feng, S., Manmatha, R. 2006. *A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books*, in JCDL, 2006, pp. 109–118.
- [13] Reynaert, M. 2008. *Non-interactive OCR post-correction for giga-scale digitization projects*. in CICLing'08 Proceedings of the 9th international conference on Computational linguistics and intelligent text processing.
- [14] Déjean, H., Meunier, J.L. 2010. *Document: a useful level for facing noisy data*. in Proceedings of the fourth workshop on Analytics for noisy unstructured text data (AND '10). ACM, New York, NY, USA, 3-10.
- [15] Maynard, S., 2007. *A survey on the use of different forms of scholarly output*. A report commissioned by the JISC Scholarly Communications Group.
- [16] Landoni, M. 2010. *Evaluating e-books*. in Proceedings of the third workshop on Research advances in large digital book repositories and complementary media (BooksOnline '10).
- [17] Armstrong, C.J., Lonsdale, R.E. 1998. *The publishing of electronic Scholarly Monographs and textbooks*, Report G5, London Library Information Technology Centre.
- [18] Phelps, T.A., Wilensky, R.: 2001. The multivalent browser: a platform for new ideas. In: DocEng '01: Proceedings of the 2001 ACM Symposium on Document engineering, ACM Press (2001) pp. 58–67.
- [19] Carden, M. 2008. E-Books are not books. In Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories (BooksOnline '08). ACM, New York, NY, USA, pp. 9-12.
- [20] Kim, J.-K., Farzan, R., and Brusilovsky, P. (2008) Social Navigation and Annotation for Electronic Books. In Proceeding of the 2008 ACM workshop on Research

- advances in large digital book repositories (BooksOnline '08). ACM, New York, NY, USA., 25-28.
- [21] Pearson, J. and Buchanan, G. (2010) Real-Time Document Collaboration Using iPads. In: Proceedings of BooksOnline 2010 Workshop at the 19th ACM conference on Conference on information and knowledge management: CIKM '10, Toronto, Canada, October 26, 2010, ACM Press, pp. 9-13.
- [22] Mazza, R. 2008. The integrated eBook: the convergence of ebook, companion web site, and elearning. In Proceeding of the 2008 ACM workshop on Research advances in large digital book repositories (BooksOnline '08). ACM, New York, NY, USA, 1-4.
- [23] Ribiere. M., Picault, J., Squedin, S. 2010, *The sBook: towards social and personalized learning experiences*. in Proceedings of the third workshop on Research advances in large digital book repositories and complementary media, BooksOnline '10.
- [24] Warren, J.W. 2010. *The Progression of Digital Publishing: Innovation and the Evolution of E-books*. in International Journal of the Book, Volume 7, Issue 4, pp.37-54.
- [25] Mons, B., Velterop, J. 2009. *Nano-Publication in the e-Science Era*, Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009).
- [26] Peters, I. 2009. *Folksonomies. Indexing and Retrieval in Web 2.0*. Berlin: De Gruyter Saur.
- [27] Woods, W.A. 1997. *Conceptual Indexing: a Better Way to Organize Knowledge*. Technical Report. Sun Microsystems, Inc., Mountain View, CA, USA.
- [28] Erway R., 2009. A view on Europeana from the US perspective. *Liber Quarterly* 19 (2), pp. 103–121.