

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

The Functional Extension Parser (FEP) A Document Understanding Platform

Günter Mühlberger

University of Innsbruck

Department for German Language and Literature Studies

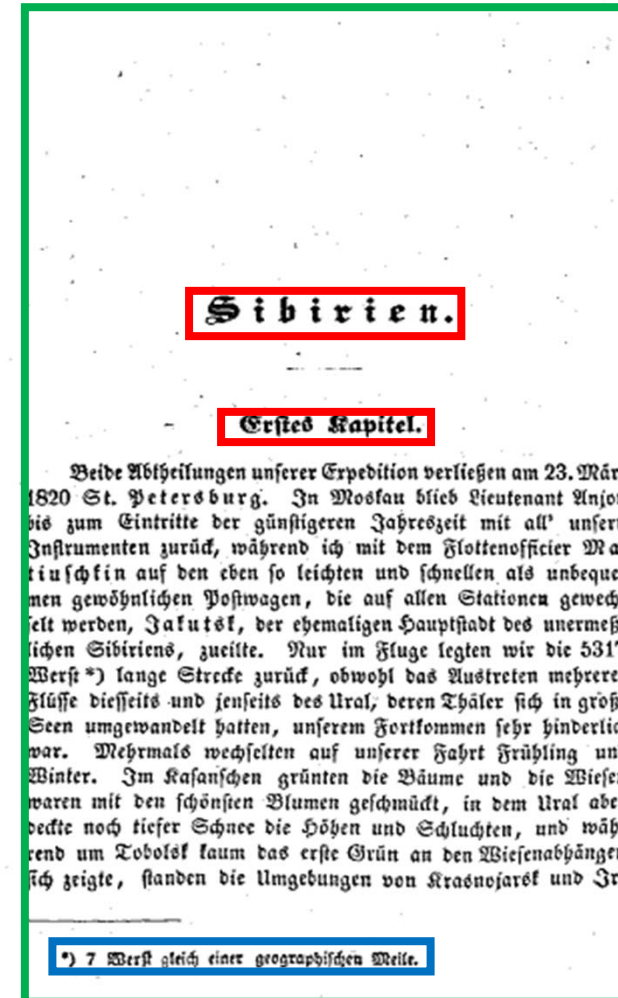


Introduction

- A book is more than just pure text – it contains a lot of structural metadata
- These metadata are (often) encoded in the layout of a document
- Size of characters, position on page, distance to other lines, etc. is used to express structural meaning
- FEP is a platform to process digitised or born digital documents and to “understand” the meaning of the layout by using a rules engine
- FEP was developed within the IMPACT project by Lukas Gander, Raphael Unterweger, Sebastian Colutto, Cornelia Lezuo (about 80 person months were invested)



- Headlines
- Footnotes
- Print space (cropping)





- Running title
- Page number
- Signature mark

russisch-amerikanische Handelsgesellschaft. Auf viele Tausende von Werst im Umkreise strömen während der wenigen Sommerwochen hieher die köstlichsten wie die gemeinen Pelzwaaren aller Art, sowie Walrosszähne und die riesigen Knochenüberreste des vorweltlichen, dem asiatischen Elephanten verwandten Mammuts, dessen Stoßzähne besonders als fossiles Elfenbein in den Handel kommen. In dieser Zeit treffen hier auch die Kaufleute aus dem Süden ein, die dieser an Allem armen Gegend fast Alles, was zum Leben nöthig ist, zuführen. Die Hauptverkaufszeit ist im Monat August, weicht aber von dem, was wir unter einem Jahrmarkt verstehen, sehr ab, indem die Kaufleute ihre Waaren in den Häusern und Höfen gleichsam verstecken, um ihre Preise, sowie die Namen ihrer Abnehmer vor einander möglichst geheim zu halten. Die Bewohner (gegen 4000) stehen noch auf einer sehr niedrigen Stufe geistiger Bildung. Der Heiligenkalender (Swäpy) ist fast ihr einziges Buch und die Erziehung ist die mangelhafteste. Die Kinder werden gewöhnlich bald nach der Geburt einer Jakutin übergeben, die sie nach 2 bis 3 Jahren fast als kleine Jakuten den Eltern wiederbringt. Später lernen sie etwas lesen und schreiben und werden dann mit dem Pelzhandel vertraut gemacht, der die Einwohner ausschließlich beschäftigt. Doch leben sie sehr gesellig, wiewohl Essen und Trinken bei ihren lärmenden Versammlungen die Hauptrolle spielen. Die Herren sitzen bei dem Punschglas und die Damen in einem bei uns längst veralteten Staate um den Theekessel, während die Jugend nach den Klängen der Gushli, einer Art liegenden Harfe mit Metallsaiten, ein Tänzchen macht.

Lieutenant Anjou ging schon im Anfange August mit seiner Abtheilung die Lena hinunter. Ich brach am 12. September, nachdem ich vorher den Ritschmann Matiuschkin und den Steuermann Kosmin mit unsern Borräthen, um die nöthigen Vorbe- reitungen für unsern dortigen Aufenthalt zu treffen, nach Nis'hne-Kolymsk vorausgeschickt hatte, eben dahin auf.

Von Jakutsk führt keine gebahnte Straße nach dem Norden, sondern die Reise muß zu Pferde auf engen, holperigen Fußstei- gen, die durch Moräste und dichte Wälder, über steile Berge und zwischen zahlreichen Landseen dahin führen, fortgesetzt wer-

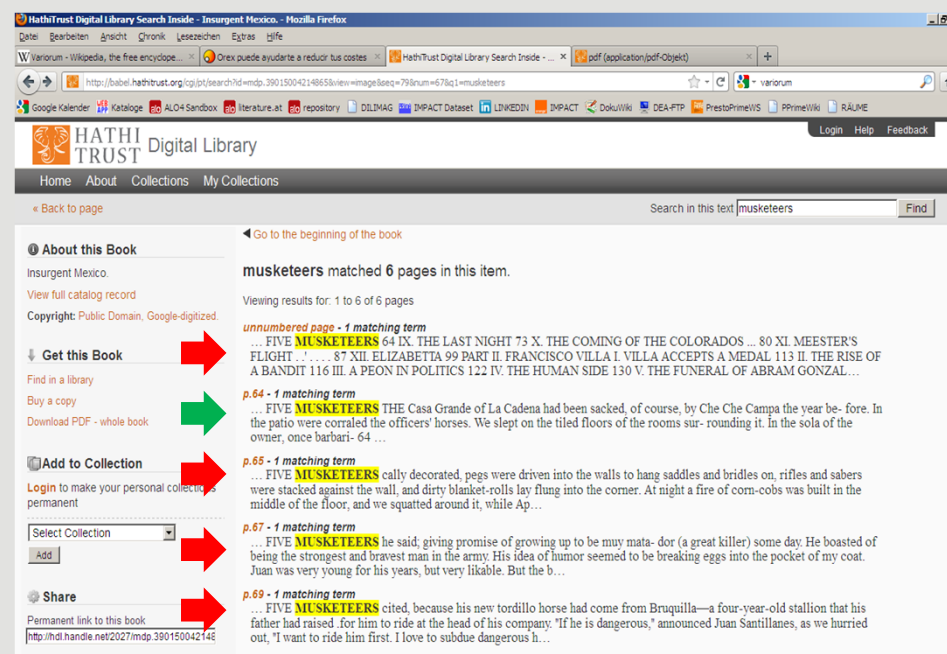
5



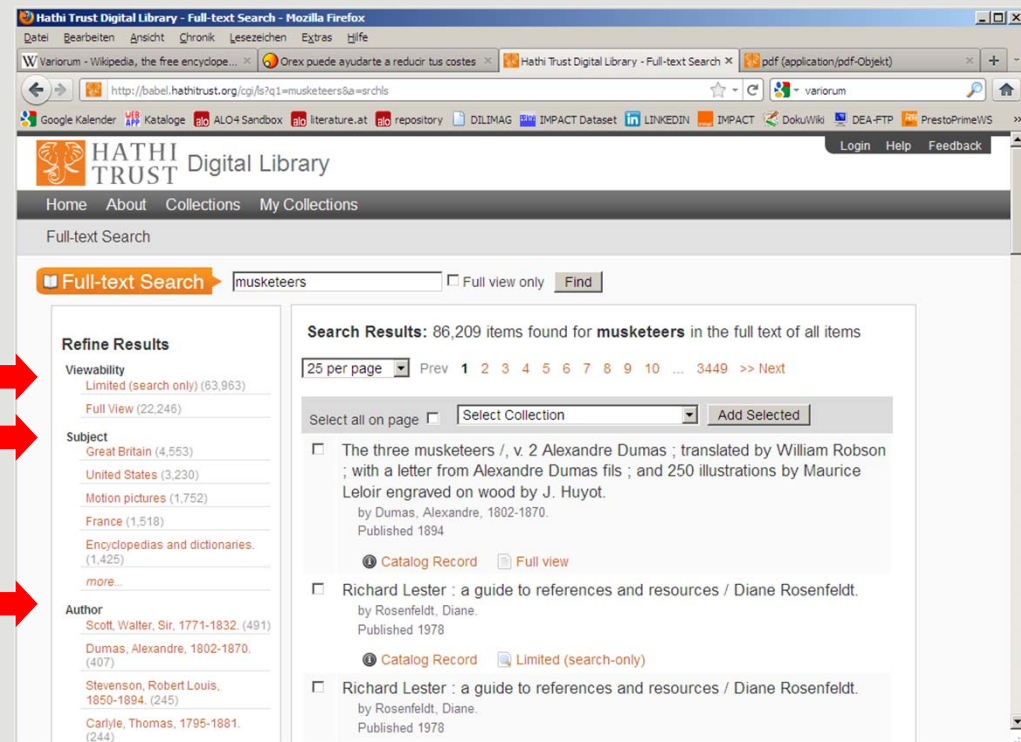
Why structural tagging is important – some examples

- Search & Retrieval
- References and links to other documents
- Reading: analogue and digital

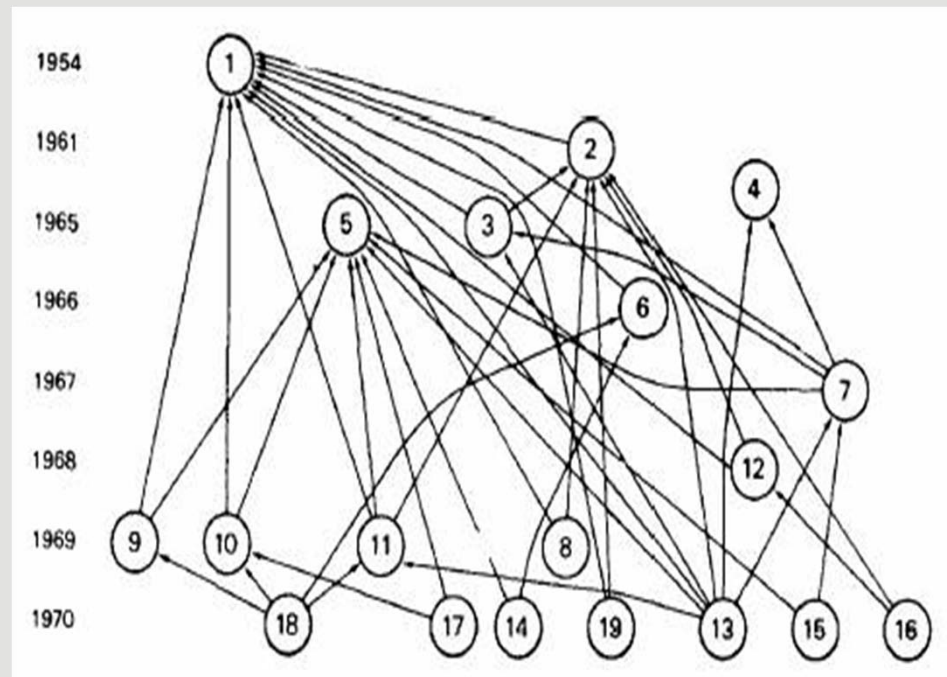
- Search & retrieval
 - Ranking and scoring, noise reduction
 - The same word appears in the running title of a journal at every page “Musketeers”
 - Front matters, such as title pages, dedications, table of contents tables, etc.
 - Back matters such as indexes, ads, etc.



- Search & retrieval
 - Facets for full-text
 - Currently facets are used for metadata such as author, year, text type, ...
 - A user might be interested in facets such as headline, footnote, index, etc...

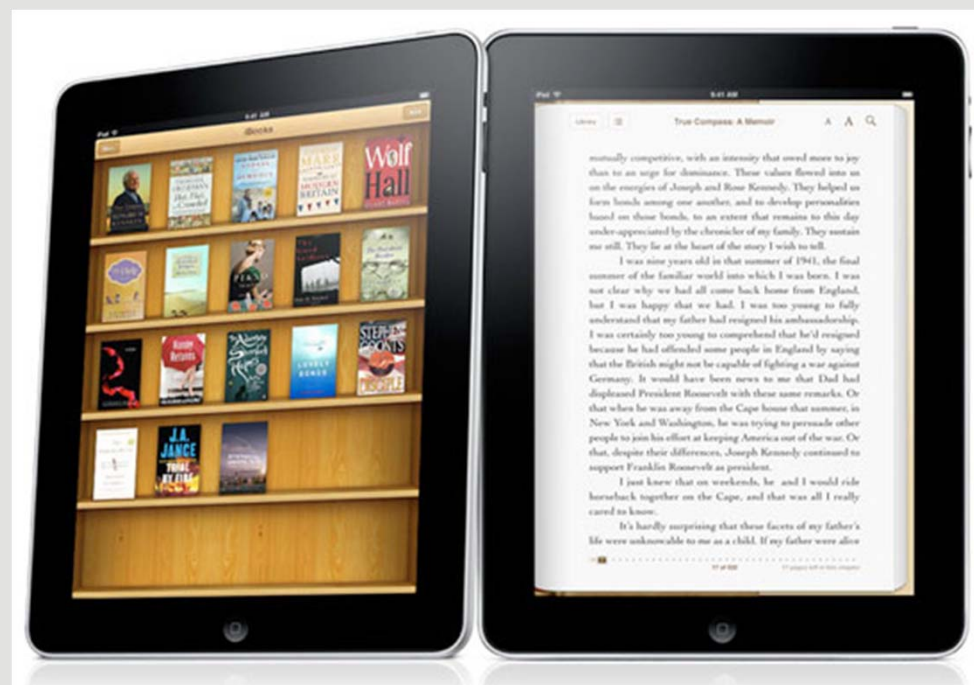


- Citations index / cloud
 - Footnotes, reference lists, citations contain bibliographic links to books, journal articles, texts, etc.
 - Structural tagging supports detection of bibliographic references
 - May also be used for catalogue enrichment

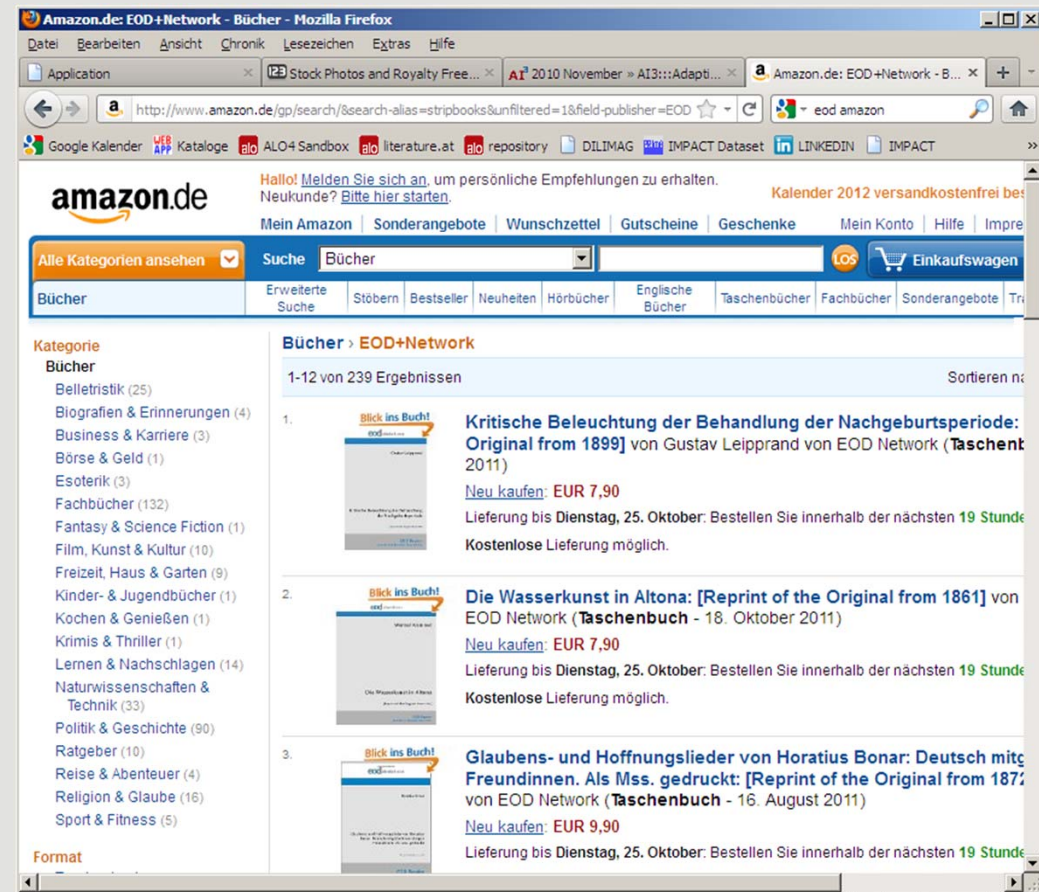


Cawkell, A. E. (1971)

- Digital reading
 - Tablet computers as alternative for reading historical books with OCR below reading quality
 - Expected features
 - Nicely cropped pages
 - Bookmarks
 - ToC page linked with headings
- Advanced reading
 - eBooks for modern texts with satisfying OCR quality
 - Structure can be encoded into ePUB etc.

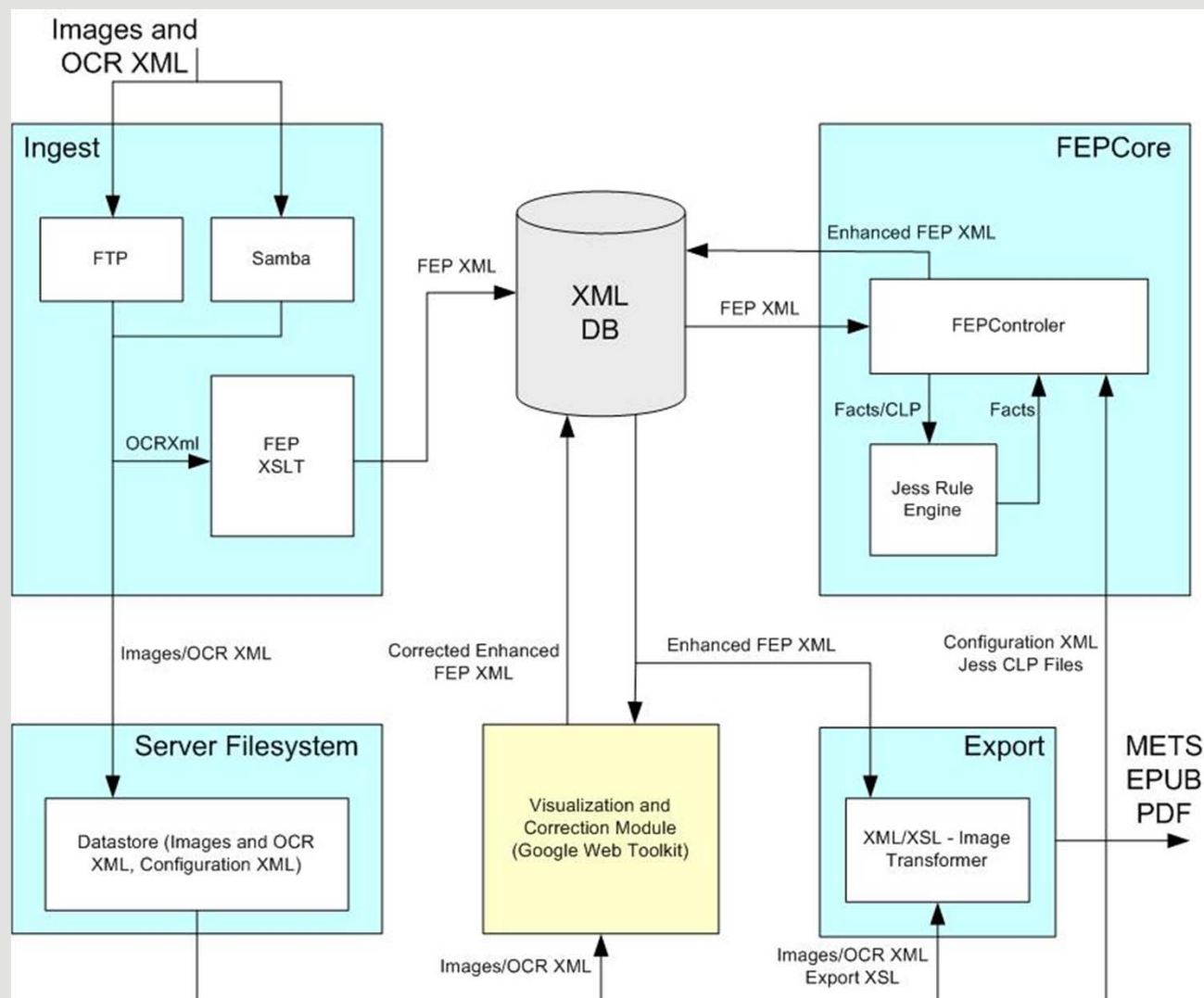


- Analogue reading
 - Print on Demand
 - Print space (exact cropping) as old concept with new benefits
 - Reconstruction helps to semi-automate the standardized production of pre-press files



Technical background

- Input
 - OCR text which needs to contain at least word coordinates
 - E.g. ALTO files, ABBYY XML or Google Books (Tesseract) HTML
- Output
 - Annotations of structural elements with coordinates, e.g. page numbers, running titles, headings, footnotes, printspace, etc.
 - Output format: METS/ALTO, XML, etc.
- FEP System
 - Images and/or OCR files are loaded via a web-service
 - OCR data are converted into internal format
 - Information is processed based on rules
 - Results are stored in a database
 - Quality control on the basis of “ground truth”, e.g. expected results
 - Rules are either manually encoded (expert knowledge) and/or based on machine learning (large document sets)



Evaluation method

- Basic rules set
 - General structural elements of books from e.g. 1700 to 2010
 - Data set: 155 books, 30.673 pages (141 training set, 41 evaluation set)
 - All pages were manually annotated (ground truth)
- Recall, Precision, F-Measure
 - 10 lines with headings in a book. We find e.g. 12 lines, 8 of them correct, 4 false:
 - Recall $= 8 \text{ of } 10 = 0,8$
 - Precision $= 8 \text{ of } 12 = 0,66$
 - F-Measure $= 2 * 0.8 * 0.66 / (0.8 + 0.66) = 0,72$
- More information
 - Important: We count lines, not structural entities!
 - E.g. if a heading has two lines one might be correct, the other one might not be recognised
 - Differences between training and evaluation set are low

Results on the evaluation set

	Recall	Precision	F-measure
Running text	0,99	0,98	0,98
Running titles	0,97	1	0,98
Page numbers	0,97	1	0,98
Footnotes	0,83	0,89	0,86
Headings	0,85	0,80	0,82
Signature marks	0,68	0,89	0,77

Table of Contents pages and entries

- Original plan was to take part in the INEX competition 2011 but time constraints made this impossible
- We got the INEX dataset for training and evaluation purposes → our results can be directly compared with the INEX results of 2011
 - the data set consists mainly of English books from the 19th century with partly really complicated ToC pages
 - Also Ground Truth was not always consistent
- Two rule sets
 - Results for table of content pages: 90% of the pages were detected correctly also the F-measure in this case is about 90%
- Evaluation of table of content entries follows XRCE link-based measure. The XRCE link-based measure permits evaluating the performance of the systems by matching ToC entries primarily based on links rather than titles.

Linking of table of content entries

Institution	Recall	Precision	F-measure
IMPACT-FEP	0.682	0.727	0.662
Microsoft Serbia	0.702	0.645	0.651
Nankai	0.674	0.676	0.632
Xerox	0.551	0.759	0.581
GREYC	0.499	0.652	0.507

Excursion: How to deal with uncertainty?

- Automated processing produces always errors!
 - OCR is a highly developed technology
 - Modern texts:
 - Clearly above 95% word accuracy
 - But for historical texts word accuracy comes down to 80% or even 50 or 50% accuracy rate
 - E.g. British newspapers from the 19th century are around 80% (Tanner)
- How to deal with these error rates?
 - Many libraries did not apply OCR since “too many errors”
 - On the other hand Google and other industrial projects certainly use OCR for enhancing their documents
- Metadata enrichment/annotation with FEP
 - The same issue: Will produce errors, how to deal with them?

How to deal with uncertainty and errors?

- **Option 1: Leave it as it is**
 - Accept the accuracy which can be provided automatically
 - Inclusion of ground truth in the database allows to exactly measure the quality of the automated processing → one knows in advance what can be expected
- **Pro**
 - Maybe the only solution for really large document sets
 - It is much cheaper to develop better rule sets than to correct large numbers of documents
 - Good results for homogenous sets are possible
 - Similar to OCR
- **Con**
 - You and your users need to accept errors
 - People want to contribute and to correct

How to deal with uncertainty and errors?

■ **Option 2: Correct it**

- Service providers or library staff needs to correct
- Manual correction with automated support

■ **Pro**

- Batch correction + off shore is relatively cheap and effective if the error rate is above a certain rate (e.g. 80%) otherwise correcting is more effort than simply do it from scratch
- Quick and standardized results
- Users are satisfied

■ **Con**

- A reasonable investment is necessary
- The complexity of the workflow may not be underestimated
- Probably it will be too expensive to correct all interesting elements, therefore you and your users still need to accept “some” errors
- Users still want to contribute but do not have a chance

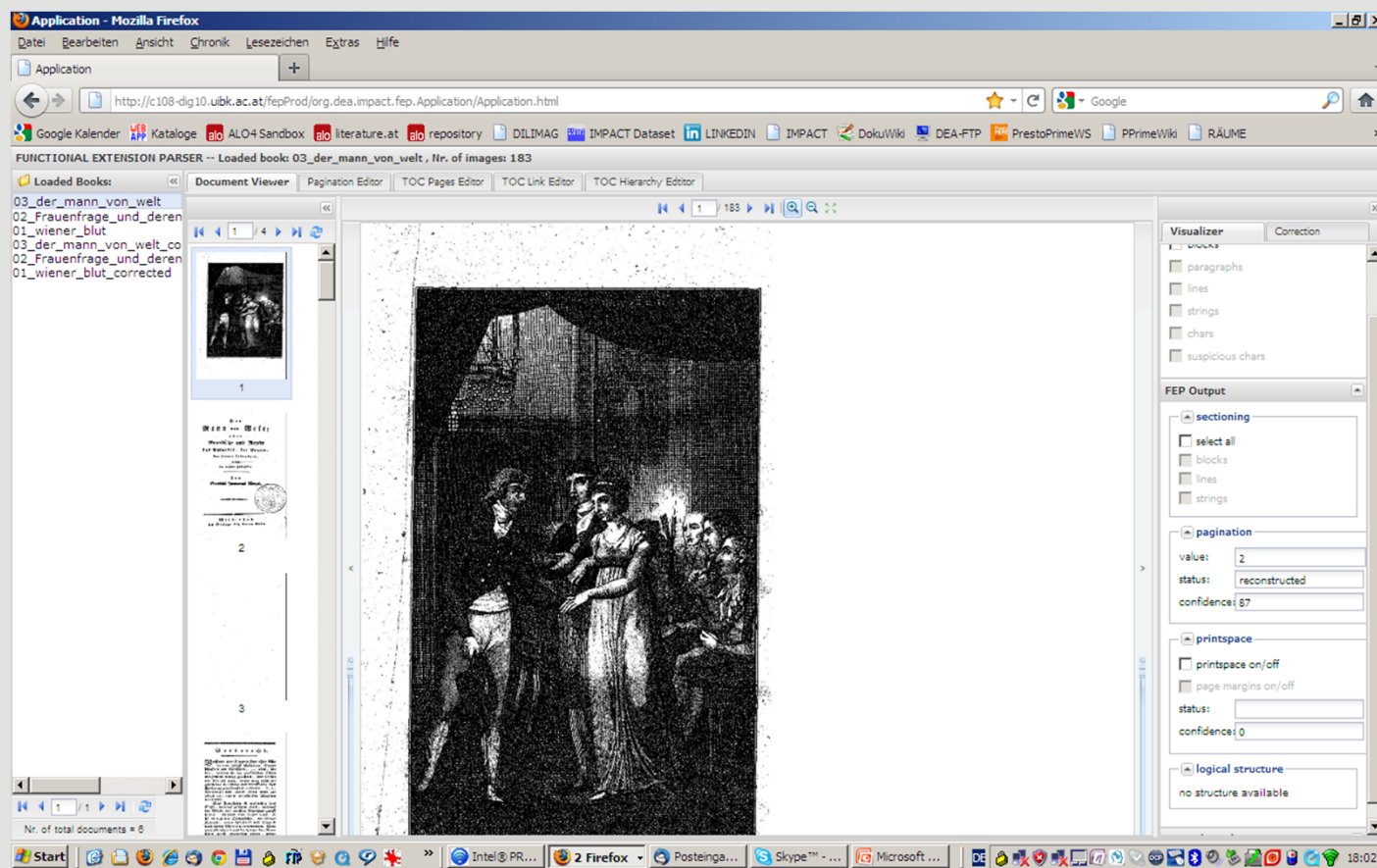
How to deal with uncertainty and errors?

- **Option 3: Provide a user interface for the crowd**
 - Correction of OCR results may only be the start for also providing interfaces for structural annotations
 - Might be combined with some basic corrections carried out by service providers
- **Pro**
 - Satisfies the willingness of users to contribute
 - Users get immediate benefit, e.g. they are able to download structured PDFs for their iPad, or annotated full-text for further processing
 - Users are satisfied AND are able to contribute
 - Library gets correct and standardized data
- **Con**
 - An reasonable investment is necessary both for the user interface as well as for adapting the digital library application
 - User interfaces need to be powerful, self-explaining and simple
 - You and your users need to accept that there are always errors in the collection and that it will take decades to come to an end

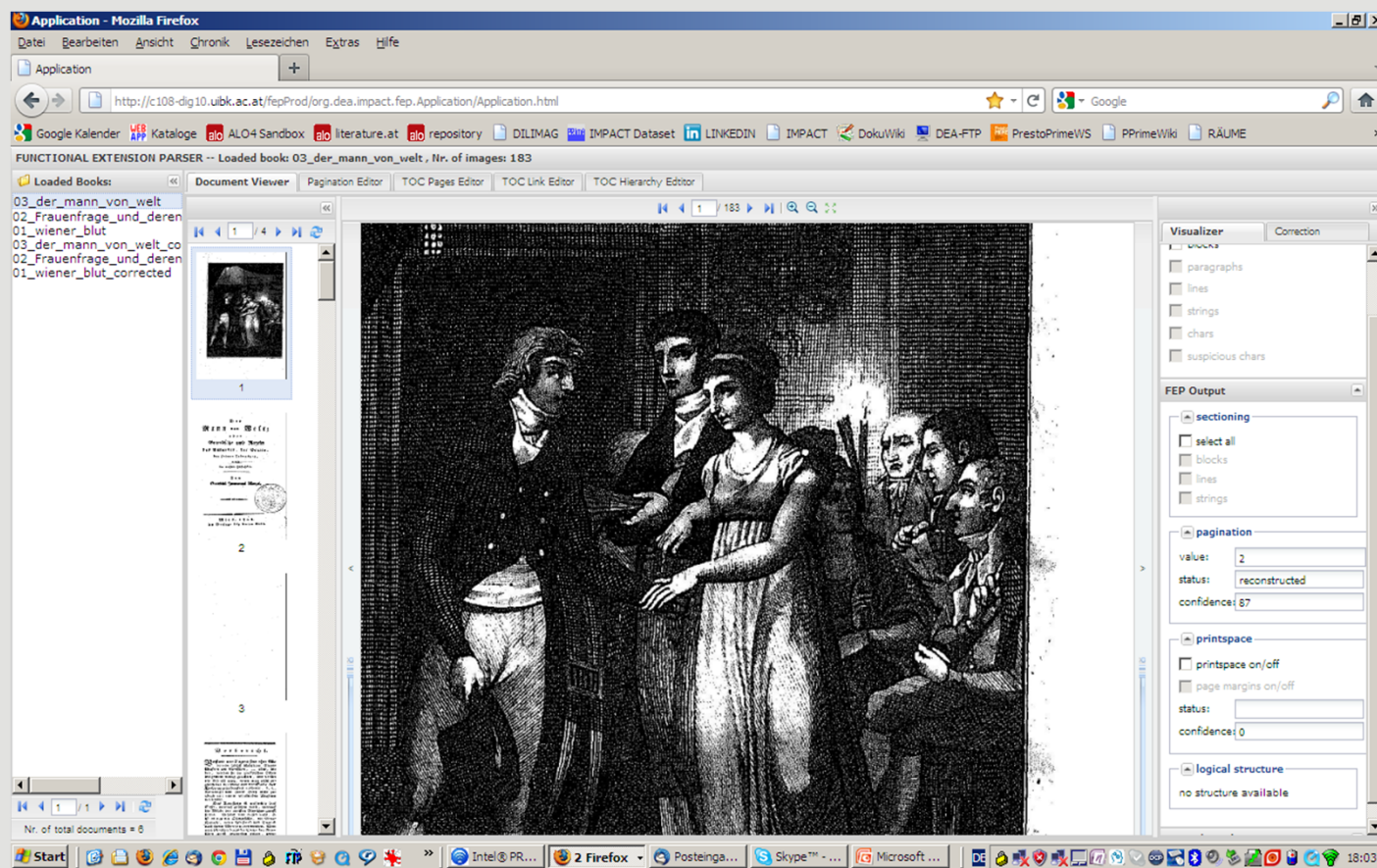
FEP User Interface

- A first attempt for a powerful, self-explaining and simple GUI
 - Currently a “general purpose interface” to display, edit and correct the structural elements of books
 - No optimisation for specific tasks and large amounts of documents
 - Has the potential to become a user interface for the crowd
 - Could look completely different!
- Based on Google Web Tool Kit (GWT)
 - Open source tool kit for complex browser based developments
 - GWT allows for features previously seen mainly in FLASH interfaces
 - Growing community
 - Good experiences: GWT allows to create interfaces in a relatively short time period

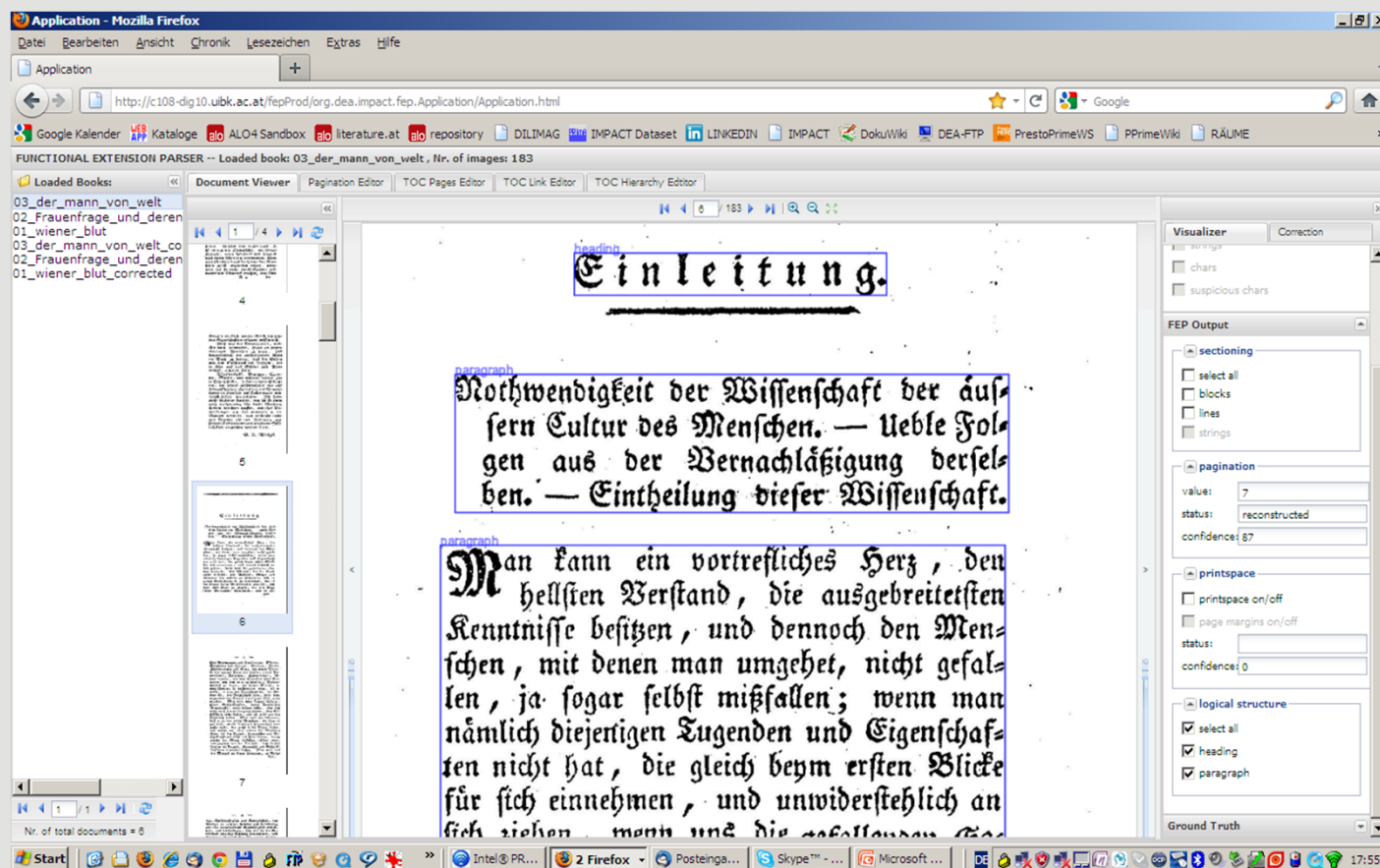
Display of results



Rich interface



Recognized elements, e.g. headings



Application - Mozilla Firefox

http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

FUNCTIONAL EXTENSION PARSER -- Loaded book: 03_der_mann_von_welt, Nr. of images: 183

Loaded Books:

- 03_der_mann_von_welt
- 02_Frauenfrage_und_derer
- 01_wiener_blut
- 03_der_mann_von_welt_co
- 02_Frauenfrage_und_derer
- 01_wiener_blut_corrected

Document Viewer

4

5

6

7

Nr. of total documents = 6

Visualizer

sectioning

- ☐ select all
- ☐ blocks
- ☐ lines
- ☐ strings

pagination

value: 7

status: reconstructed

confidence: 87

printspace

- ☐ printspace on/off
- ☐ page margins on/off

status:

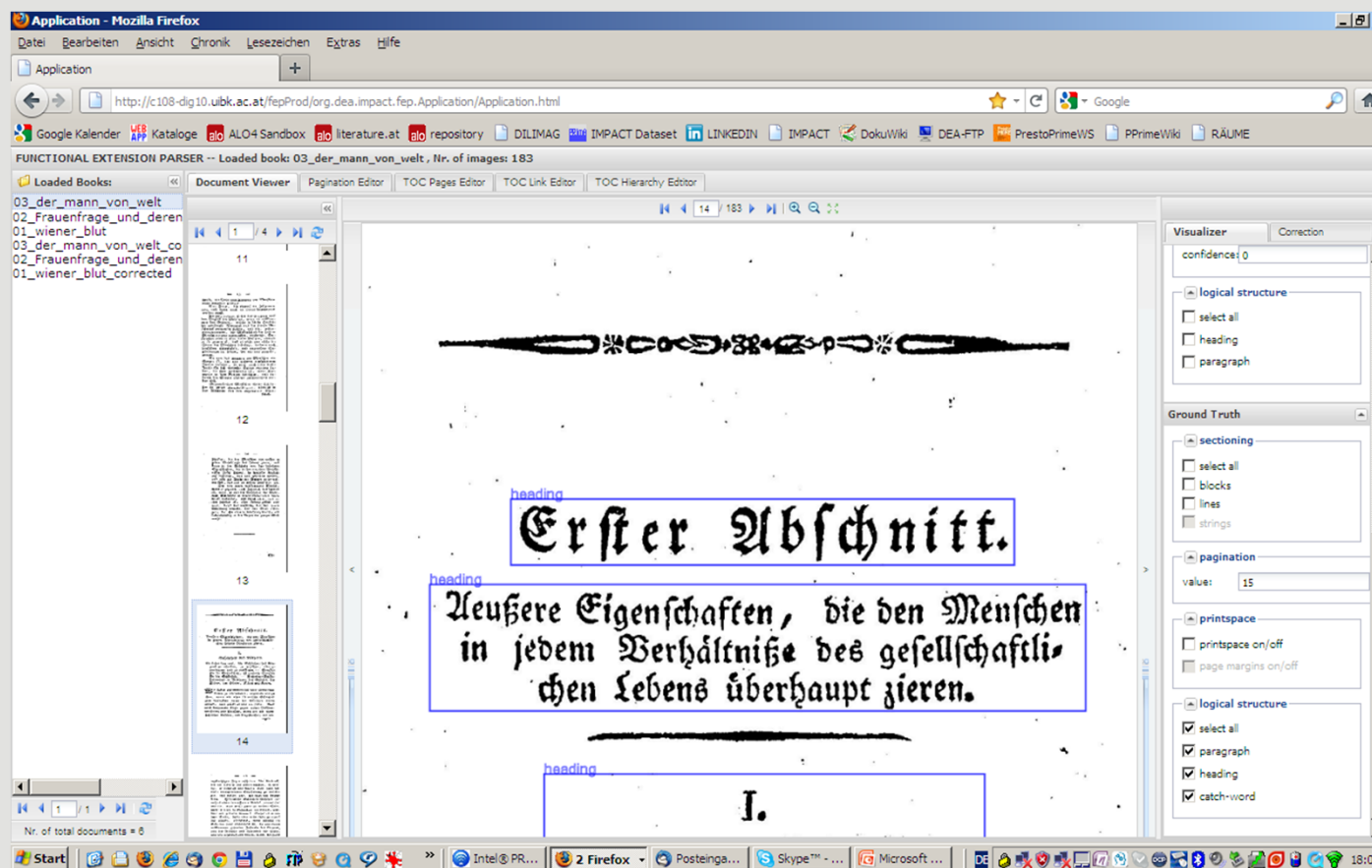
confidence: 0

logical structure

- ☒ select all
- ☒ heading
- ☒ paragraph

Ground Truth

Display of ground truth



Page numbers

Application - Mozilla Firefox

http://c-108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

Google Kalender Kataloge ALO4 Sandbox literature.at repository DILIMAG IMPACT Dataset LINKEDIN IMPACT DokuWiki DEA-FTP PrestoPrimeWS PPrimeWiki RÄUME

FUNCTIONAL EXTENSION PARSER -- Loaded book: 03_der_mann_von_welt, Nr. of images: 183

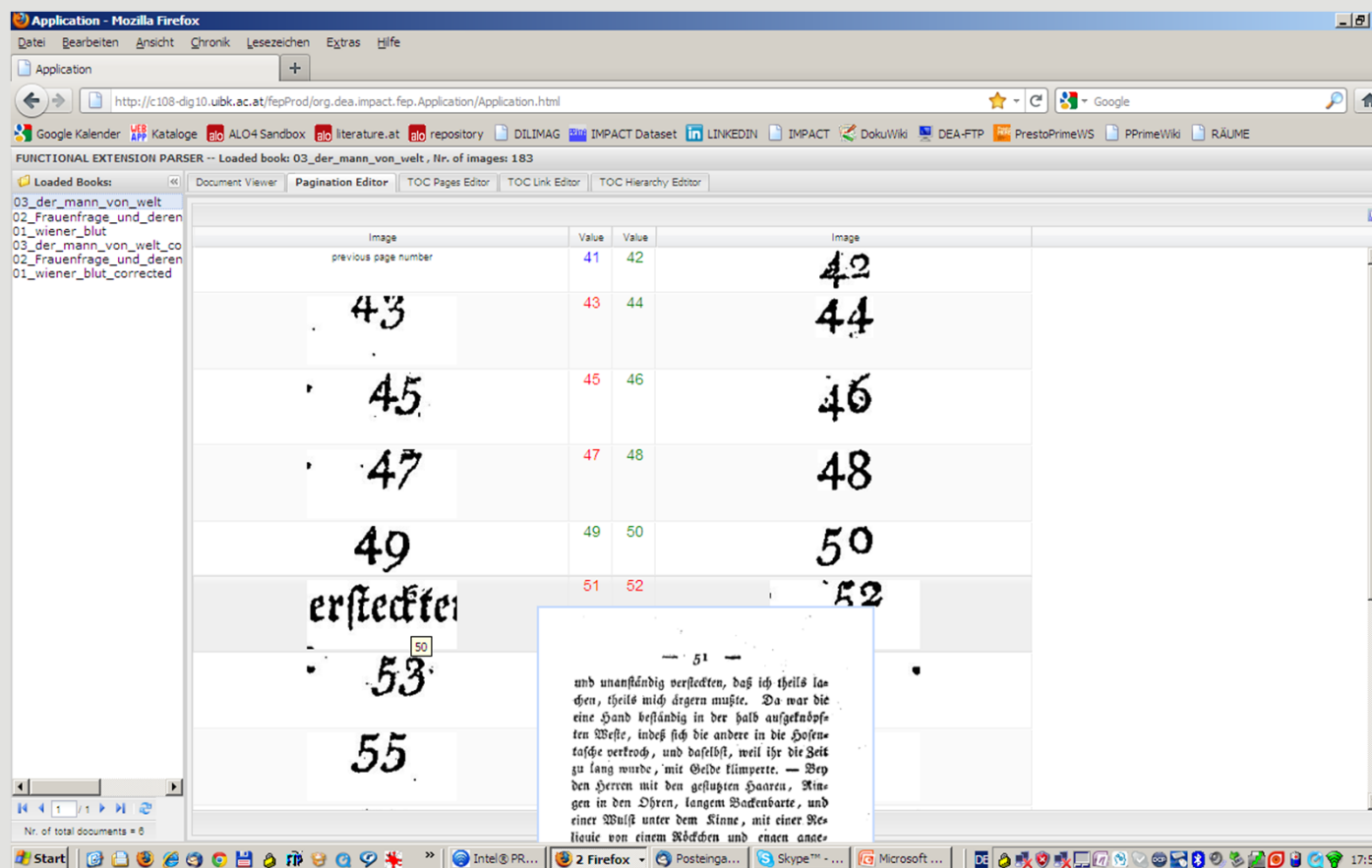
Loaded Books: Document Viewer **Pagination Editor** TOC Pages Editor TOC Link Editor TOC Hierarchy Editor

Image	Value	Value	Image
previous page number	41	42	42
43	43	44	44
45	45	46	46
47	47	48	48
49	49	50	50
erstickte	51	52	52
53	53	54	54
55	55	56	56

Nr. of total documents = 6

Start Intel® PR... 2 Firefox Postings... Skype™ - ... Microsoft ... 17:57

Page numbers control



Application - Mozilla Firefox

http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

FUNCTIONAL EXTENSION PARSER -- Loaded book: 03_der_mann_von_welt, Nr. of images: 183

Loaded Books:

- 03_der_mann_von_welt
- 02_Frauenfrage_und_deren
- 01_wiener_blut
- 03_der_mann_von_welt_co
- 02_Frauenfrage_und_deren
- 01_wiener_blut_corrected

Document Viewer | **Pagination Editor** | TOC Pages Editor | TOC Link Editor | TOC Hierarchy Editor

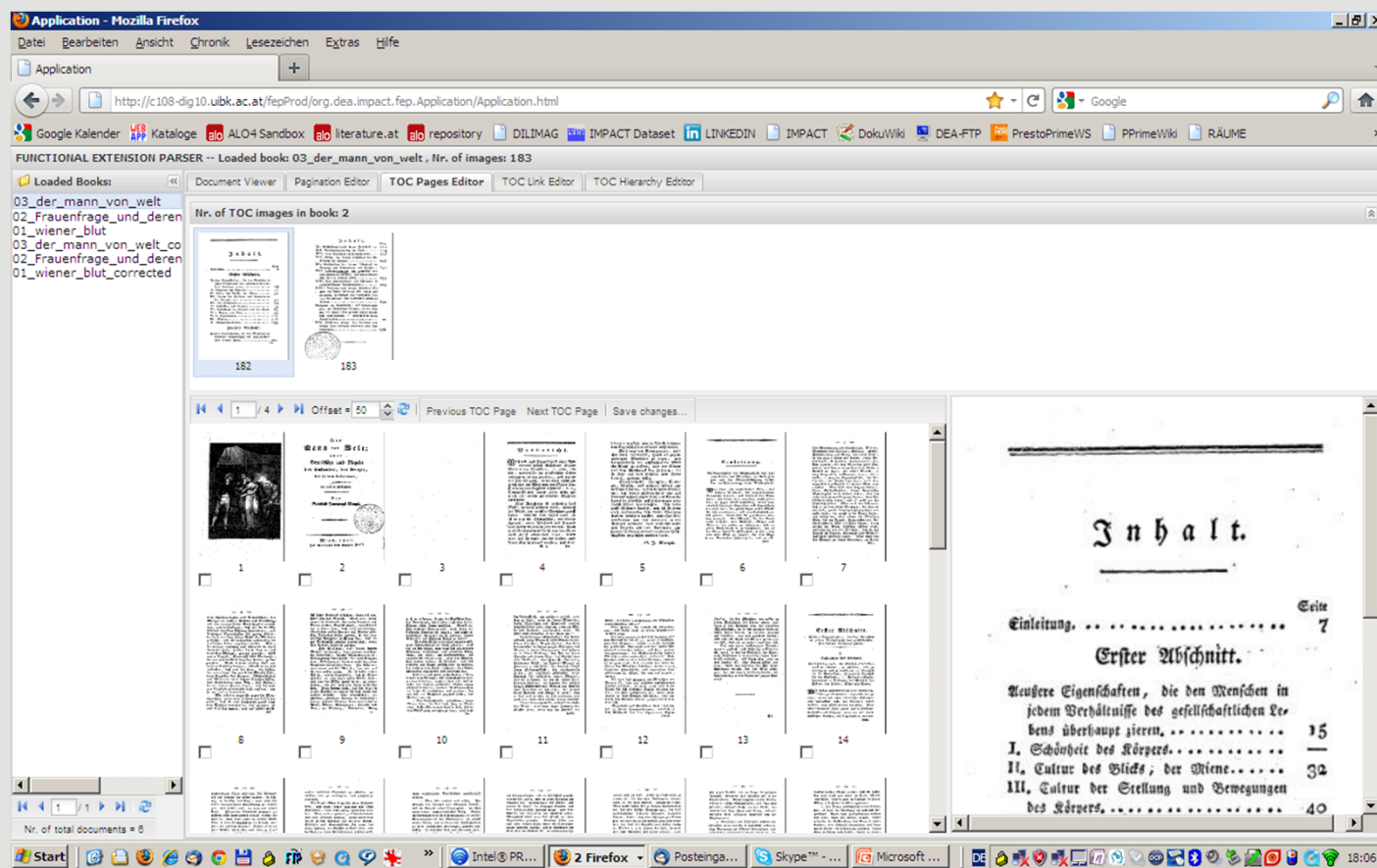
Image	Value	Value	Image
previous page number	41	42	42
43	43	44	44
45	45	46	46
47	47	48	48
49	49	50	50
ersteckten	51	52	52
53			
55			

Nr. of total documents = 6

51

und unabhängig verflochten, daß ich theils la-
gen, theils mich ärgern mußte. Da war die
eine Hand beständig in der halb aufgeklop-
ften Weste, insofern sich die andere in die Hos-
entasche verfracht, und daselbst, weil ihr die Zeit
zu lang wurde, mit Gelde klapperte. — Wep-
den Herren mit den geklugten Haaren, Kin-
gen in den Ohren, langem Badenbarte, und
einer Wulst unter dem Kinn, mit einer Res-
lieue von einem Ködchen und einem anse-

ToC pages

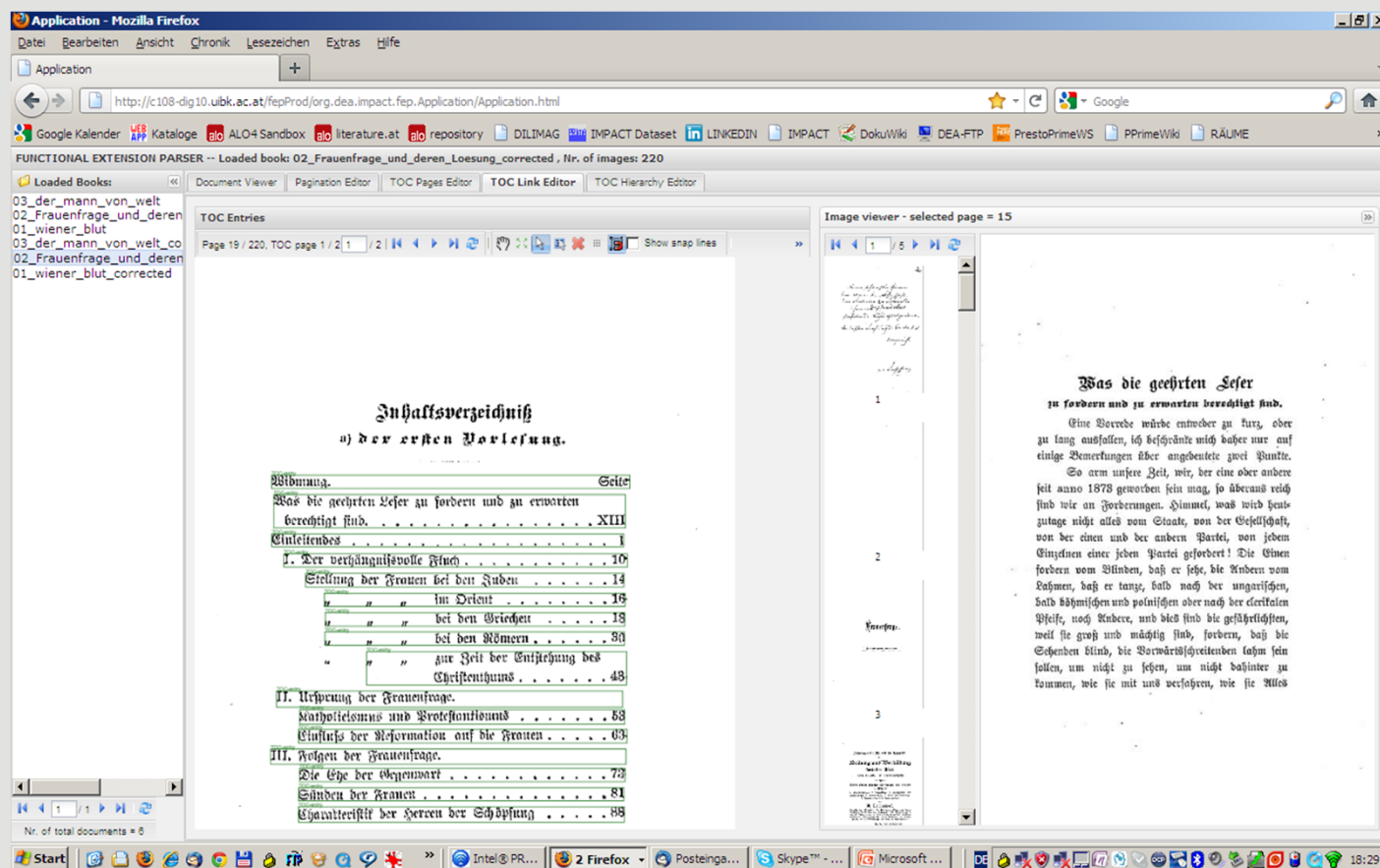


The screenshot displays the IMPACT application running in Mozilla Firefox. The browser's address bar shows the URL: `http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html`. The application interface includes a top menu bar with options like 'Datei', 'Bearbeiten', 'Ansicht', 'Chronik', 'Lesezeichen', 'Extras', and 'Hilfe'. Below the menu is a toolbar with various icons. The main content area is divided into several sections:

- Loaded Books:** A list of books on the left side, including '03_der_mann_von_welt', '02_Frauenfrage_und_deren', '01_wiener_blut', '03_der_mann_von_welt_co', '02_Frauenfrage_und_deren', and '01_wiener_blut_corrected'.
- FUNCTIONAL EXTENSION PARSER:** A status bar indicating 'Loaded book: 03_der_mann_von_welt, Nr. of images: 183'.
- Main View:** A grid of document thumbnails, each labeled with a number (1-14). The thumbnails show various pages from the loaded book, including text and images.
- Inhalt. (Table of Contents):** A large, detailed table of contents on the right side, listing sections and their corresponding page numbers. The sections include:
 - Einleitung, ... 7
 - Erster Abschnitt.
 - Neuere Eigenschaften, die den Menschen in jedem Verhältnisse des gesellschaftlichen Lebens überhaupt zielen, ... 15
 - I. Schönheit des Körpers, ... 32
 - II. Kultur des Blicks; der Miene, ... 32
 - III. Kultur der Stellung und Bewegungen des Körpers, ... 40

The bottom of the screenshot shows the Windows taskbar with various application icons and the system clock indicating 18:06.

ToC entries



Application - Mozilla Firefox

http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

FUNCTIONAL EXTENSION PARSER -- Loaded book: 02_Frauenfrage_und_deren_Loesung_corrected, Nr. of images: 220

Loaded Books: Document Viewer | Pagination Editor | TOC Pages Editor | TOC Link Editor | TOC Hierarchy Editor

TOC Entries

Page 19 / 220, TOC page 1 / 2

TOC Entries

03_der_mann_von_welt
02_Frauenfrage_und_deren
01_wiener_blut
03_der_mann_von_welt_co
02_Frauenfrage_und_deren
01_wiener_blut_corrected

Inhaltsverzeichnis

a) der ersten Vorlesung.

Abbildung. Seite

Was die geehrten Leser zu fordern und zu erwarten berechtigt sind. XLII

Einleitendes

I. Der verhängnisvolle Fluch 10

Stellung der Frauen bei den Juden 14

„ „ „ im Orient 16

„ „ „ bei den Griechen 18

„ „ „ bei den Römern 20

„ „ „ zur Zeit der Entstehung des Christenthums 43

II. Ursprung der Frauenfrage.

Katholizismus und Protestantismus 63

Einfluss der Reformation auf die Frauen 63

III. Folgen der Frauenfrage.

Die Lage der Gegenwart 72

Einfluss der Frauen 81

Charakteristik der Herren der Schöpfung 88

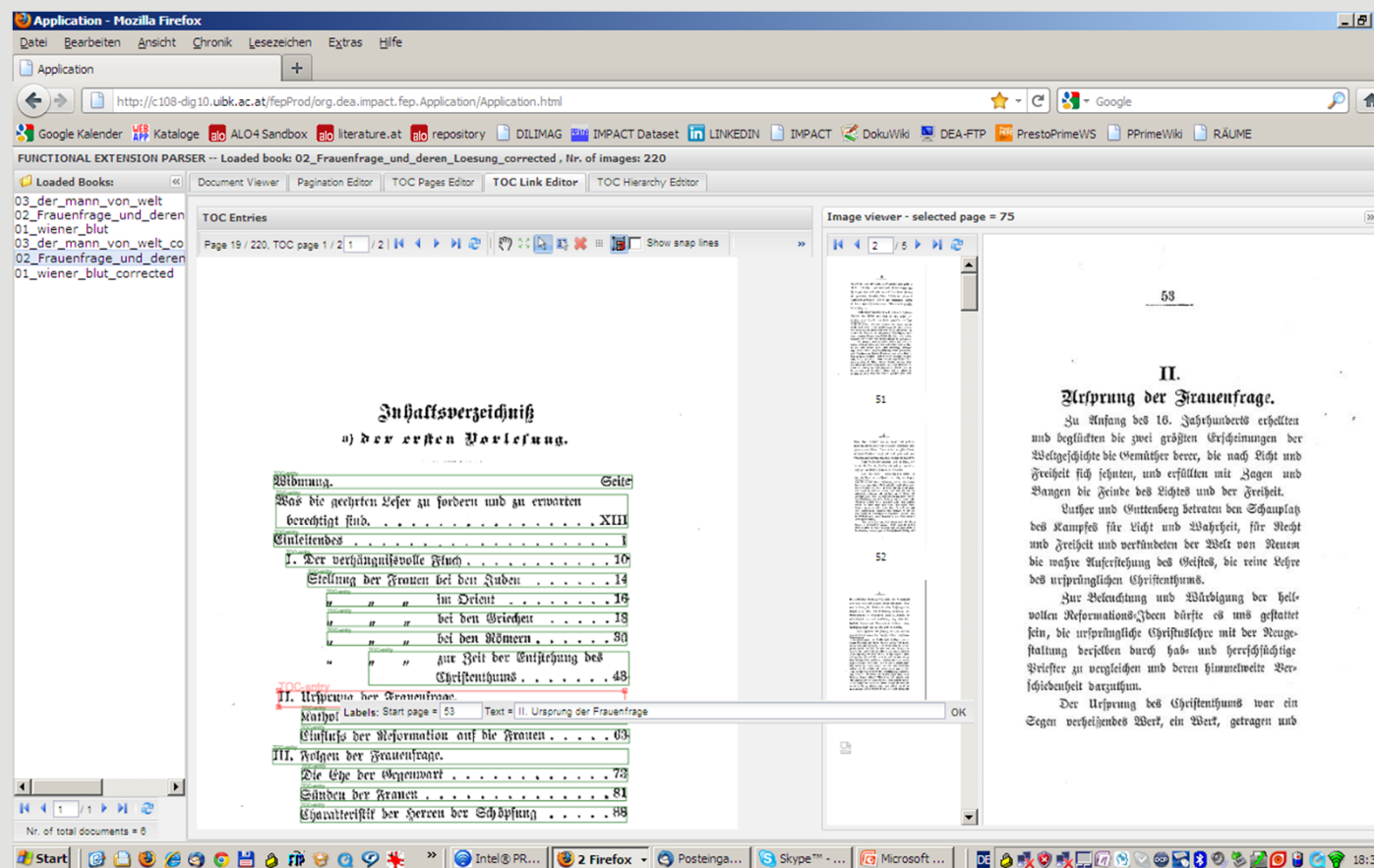
Image viewer - selected page = 15

Was die geehrten Leser zu fordern und zu erwarten berechtigt sind.

Eine Rede würde entweder zu kurz, oder zu lang ausfallen, ich beschränke mich daher nur auf einige Bemerkungen über angedeutete zwei Punkte.

So arm unsere Zeit, wir, der eine oder andere seit anno 1878 geworden sein mag, so überaus reich sind wir an Forderungen. Himmel, was wird heute zutage nicht alles vom Staate, von der Gesellschaft, von der einen und der andern Partei, von jedem Einzelnen einer jeden Partei gefordert! Die Einen fordern vom Bienen, daß er sehe, die Andern vom Eseln, daß er tanze, bald nach der ungarischen, bald böhmischen und polnischen oder nach der clericalen Pfeife, noch Andere, und dies sind die gefährlichsten, weil sie groß und mächtig sind, fordern, daß die Eselnden klug, die Vorwärtsstreichenden lahm sein sollen, um nicht zu sehen, um nicht dahinter zu kommen, wie sie mit uns verfahren, wie sie uns

Linking of entries with pages/headings



Application - Mozilla Firefox

http://c-108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

Google Kalender Kataloge ALO4 Sandbox literature.at repository DILIMAG IMPACT Dataset LINKEDIN IMPACT DokuWiki DEA-FTP PrestoPrimeWS PPrimeWiki RÄUME

FUNCTIONAL EXTENSION PARSER -- Loaded book: 02_Frauenfrage_und_deren_Loesung_corrected, Nr. of images: 220

Document Viewer | Pagination Editor | TOC Pages Editor | TOC Link Editor | TOC Hierarchy Editor

Loaded Books:

- 03_der_mann_von_welt
- 02_Frauenfrage_und_deren
- 01_wiener_blut
- 03_der_mann_von_welt_co
- 02_Frauenfrage_und_deren
- 01_wiener_blut_corrected

TOC Entries

Page 19 / 220, TOC page 1 / 2

TOC Entries	Page
Inhaltsverzeichnis	
a) der ersten Vorlesung.	
Abbildung.	Seite
Was die geachteten Leser zu fordern und zu erwarten berechtigt sind.	XIII
Einleitendes	
I. Der verhängnisvolle Kampf.	10
Stellung der Frauen bei den Juden	14
" " " " im Orient	16
" " " " bei den Griechen	18
" " " " bei den Römern	20
" " " " zur Zeit der Entstehung des Christenthums	43
II. Ursprung der Frauenfrage.	
Kapitel Labels: Start page = 53 Text = II. Ursprung der Frauenfrage	
Einfluss der Reformation auf die Frauen	63
III. Folgen der Frauenfrage.	
Die Lage der Gegenwart	73
Ständen der Frauen	81
Charakteristik der Herren der Schöpfung	88

Image viewer - selected page = 75

53

II. Ursprung der Frauenfrage.

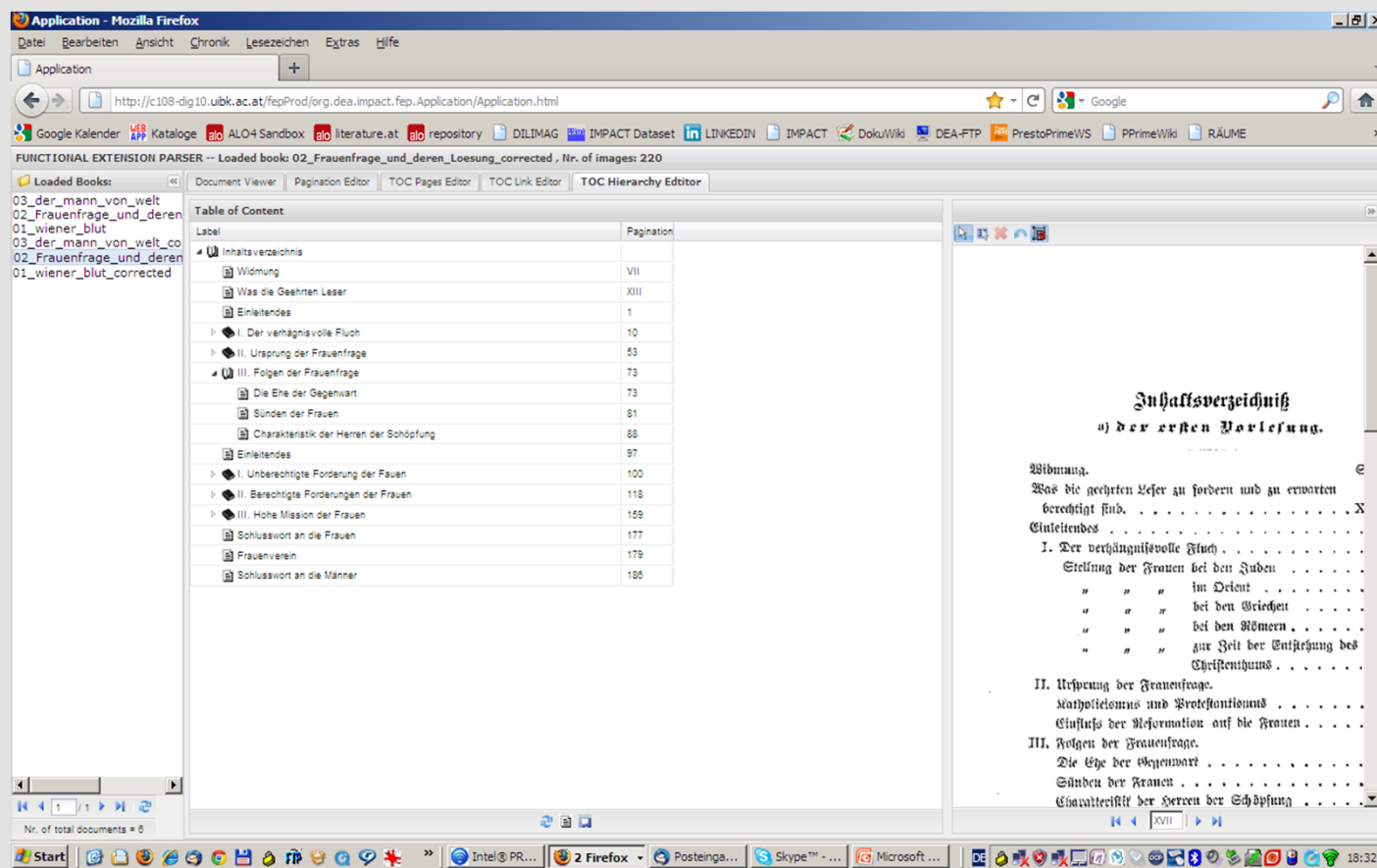
Zu Anfang des 16. Jahrhunderts erhielten und befolgten die zwei größten Christenheiten der Weltgeschichte die Vermahnung derer, die nach Licht und Freiheit sich sehnten, und erfüllten mit Hagen und Wangen die Feinde des Lichtes und der Freiheit.

Luther und Gutenberg betreten den Schauplatz des Kampfes für Licht und Wahrheit, für Recht und Freiheit und verkünden der Welt von Neuen die wahre Auferstehung des Geistes, die reine Lehre des ursprünglichen Christenthums.

Zur Beleuchtung und Würdigung der heilvollen Reformation-Ideen dürfte es uns gestattet sein, die ursprüngliche Christenlehre mit der Neugestaltung derselben durch das und herrschsüchtige Vortier zu vergleichen und deren himmelweite Verschiedenheit darzulegen.

Der Ursprung des Christenthums war ein Segen verheißendes Wort, ein Wort, getragen und

ToC hierarchy editor



Application - Mozilla Firefox

http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

FUNCTIONAL EXTENSION PARSER -- Loaded book: 02_Frauenfrage_und_deren_Loesung_corrected, Nr. of images: 220

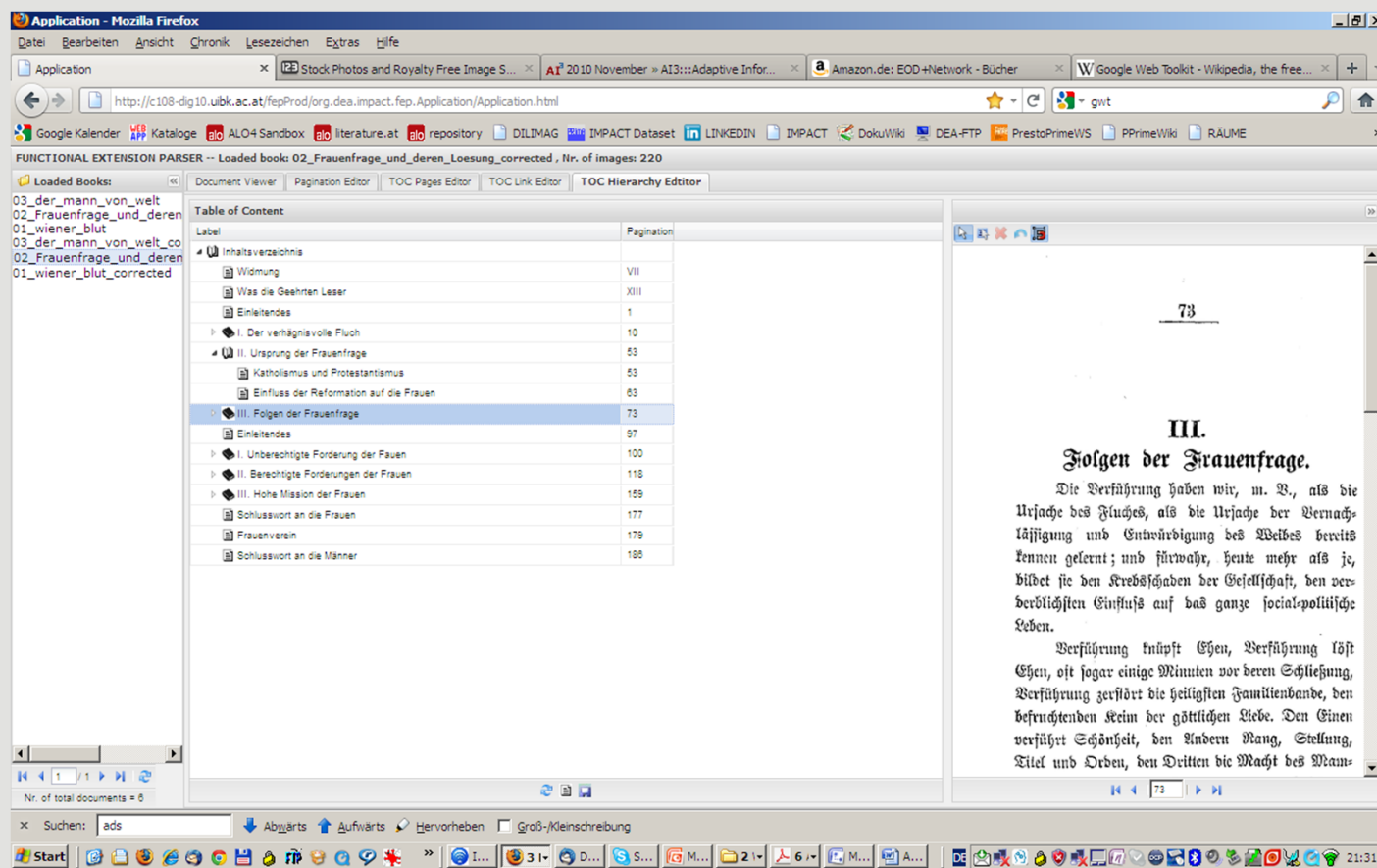
Loaded Books: Document Viewer | Pagination Editor | TOC Pages Editor | TOC Link Editor | **TOC Hierarchy Editor**

Label	Pagination
<ul style="list-style-type: none"> Inhaltsverzeichnis Widmung Was die Geehrten Leser Einleitendes I. Der verhängnisvolle Fluch II. Ursprung der Frauenfrage III. Folgen der Frauenfrage Die Ehe der Gegenwart Sünden der Frauen Charakteristik der Herren der Schöpfung Einleitendes I. Unberechtigte Forderung der Frauen II. Berechtigte Forderungen der Frauen III. Hohe Mission der Frauen Schlusswort an die Frauen Frauenverein Schlusswort an die Männer 	<ul style="list-style-type: none"> VII XIII 1 10 53 73 73 81 88 97 100 118 159 177 179 186

Inhaltsverzeichnis
a) der ersten Vorlesung.

Widmung.
Was die geehrten Leser zu fordern und zu erwarten
berechtigt sind. X
Einleitendes
I. Der verhängnisvolle Fluch.
Stellung der Frauen bei den Juden
" " " im Orient
" " " bei den Griechen
" " " bei den Römern
" " " zur Zeit der Entstehung des
Christenthums
II. Ursprung der Frauenfrage.
Matriachismus und Patriarchismus
Einfluss der Reformation auf die Frauen
III. Folgen der Frauenfrage.
Die Ehe der Gegenwart
Sünden der Frauen
Charakteristik der Herren der Schöpfung

Drag and drop of entries



Application - Mozilla Firefox

DATEI Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

Application x 123 Stock Photos and Royalty Free Image S... x A1 2010 November » A13::Adaptive Infor... x Amazon.de: EOD+Network - Bücher x Google Web Toolkit - Wikipedia, the free... x

http://c108-dig10.uibk.ac.at/fepProd/org.dea.impact.fep.Application/Application.html

Google Kalender Kataloge ALO4 Sandbox literature.at repository DILIMAG IMPACT Dataset LINKEDIN IMPACT DokuWiki DEA-FTP PrestoPrimeWS PPrimeWiki RÄUME

FUNCTIONAL EXTENSION PARSER -- Loaded book: 02_Frauenfrage_und_deren_Loesung_corrected, Nr. of images: 220

Loaded Books: Document Viewer | Pagination Editor | TOC Pages Editor | TOC Link Editor | TOC Hierarchy Editor

03_der_mann_von_welt
02_Frauenfrage_und_deren
01_wiener_blut
03_der_mann_von_welt_co
02_Frauenfrage_und_deren
01_wiener_blut_corrected

Table of Content

Label	Pagination
Inhaltsverzeichnis	VII
Widmung	XIII
Was die Geehrten Leser	1
Einleitendes	10
I. Der verhängnisvolle Fluch	53
II. Ursprung der Frauenfrage	53
Katholismus und Protestantismus	53
Einfluss der Reformation auf die Frauen	73
III. Folgen der Frauenfrage	97
Einleitendes	100
I. Unberechtigte Forderung der Frauen	118
II. Berechtigte Forderungen der Frauen	159
III. Hohe Mission der Frauen	177
Schlusswort an die Frauen	179
Frauenverein	180
Schlusswort an die Männer	

73

III.
Folgen der Frauenfrage.

Die Verführung haben wir, m. D., als die Ursache des Fluches, als die Ursache der Vernachlässigung und Entwürdigung des Weibes bereits kennen gelernt; und fürwahr, heute mehr als je, bildet sie den Krebsgeschaden der Gesellschaft, den verderblichsten Einfluss auf das ganze social-politische Leben.

Verführung knüpft Ehen, Verführung löst Ehen, oft sogar einige Minuten vor deren Schließung, Verführung zerflört die heiligsten Familienbände, den befruchtenden Keim der göttlichen Liebe. Den Einen verführt Schönheit, den Andern Rang, Stellung, Titel und Orden, den Dritten die Macht des Mann-

Nr. of total documents = 6

Suchen: ads Abwärts Aufwärts Hervorheben Groß-/Kleinschreibung

Start

Export from FEP web-interface

- METS/ALTO
 - XML Standard for digitised books and documents
- PDFs
 - Advanced PDFs for eBooks
 - [Original version](#)
 - [FEP processed version](#)
 - Pre-press files for Print on Demand
 - [FEP prepress file](#)
- ePUB
 - For modern documents with good OCR quality or corrected books (not implemented yet)

Application areas for FEP

- In the digitisation workflow
 - OCR processing + automated metadata generation on all levels
 - Quality control with specific editors
 - E.g. page numbering or detection of illustrations in books
- In the metadata workflow
 - OCR processing of specific pages, e.g. table of contents
 - Support of cataloguers
- Enriching already scanned collections
 - Option 1: Batch processing of large collections
 - Option 2: Involve end-users in the enrichment process (crowd)

Implementation of FEP

- Option 1: FEP as remote service
 - EOD Network: FEP is connected via web services
 - Image and OCR data are transferred and processed within FEP
 - Therefore the EOD Network needs not to care about technical infrastructure or support
 - E.g. a solution for crowd processing of already digitised documents
- Option 2: FEP as local service
 - E.g. for very specific processing within a digitisation workflow a local installation may be the best solution
 - Will become possible once the generic DUP will be available
- Option 3: Research partnership
 - Institutions which are interested to provide significant input to the FEP are invited to cooperate with us
 - E.g. we are in close connection with Prof. Schulz from LMU and will share the platform so that LMU can use it for their projects

Current implementations

- Europeana Newspaper Projekt (2012-2015)
 - 8 Mill. newspaper pages will be OCRed
 - For a small portion UIBK will use FEP for tracking articles in newspapers
- EOD Network
 - Pilot to implement FEP as an additional option for the workflow
 - Usual workflow is not touched: Libraries decide if they want to use the FEP and for what options
 - Simple: Correct only page numbers
 - Medium: Correct page numbers and the original table of content
 - Full: Correct print space (cropping) and create new table of content
 - We are highly interested to see the results of this pilot in the next months!
- Under negotiation: German National Library (DNB)
 - Pilot in order to use FEP to extract metadata from title pages of dissertations (author, title, year of publication, university, document type) and to provide data for the cataloguers

Future plans: a generic DUP

- Overall objective
 - Provide a contribution to the set up of European research infrastructures such as DARIAH or CLARIN
- Generic input and processing
 - Currently the “input facts” for FEP are restricted to OCR data
 - In the future we will be able to work with all kinds of data connected with text
 - Image data, e.g. for better segmentation
 - Preservation data, e.g. for providing rule based archiving
 - Morphological data, e.g. part of speech, or syntax trees
 - Semantic data, e.g. named entities such as geographical names
 - Also storage of the different data types shall become flexibel
 - E.g. simple files or Lucene index



Thank you for your attention!