



Automatic conversion of PDF-based, layout-oriented typesetting data to DAISY

potentials and limitations

Introduction

- 2 - 3 % of new German books published every year are made available in accessible versions
- more and more print publications become available in digital form via online content delivery platforms like »libreka!«
- electronic documents might serve as data source to convert inaccessible documents into DAISY

Conversion of »libreka!« documents to DAISY



online content
delivery platform



libreka! document
(PDF format)



German Central
Library for the Blind



Digital Talking Book

Suitability of »libreka!« data for DAISY conversion

Pros:

- sample documents without Digital Rights Management: text extraction possible
- standardized file format (PDF), configuration specified by online document style guide


Cons:

- two documents created by OCR, one document still contained register marks
- no alternative texts
- no specification of primary document language
- no tag structures included

DAISY <u>markup</u>	PDF <u>markup</u>	structures in sample documents or added by experts	semantic meaning			
Part I: DTB Meta Structure (extract)						
1-2	<u><dtbook></u> , <u><book></u>	1	<Document>			root element in the Digital Talking Book
	<u>xml:lang=</u> <u>"de-DE"</u>		File > Properties > Advanced > Reading Options > Language		specification of document language	language specification
Part II(a): Major Structural Elements (extract)						
3-8	<u><level1></u> - <u><level 6></u>					major division of a publication; level 1 - 6
9-14	<u><h1></u> - <u><h6></u>	2-7	<u><H1></u> - <u><H6></u>	1-6	headlines	heading for level 1 to 6
15	<u><level></u>					major division of a publication; used for recursive structures
16	<u><hd></u>	8	<u><H></u>			text of a heading in <u><level></u> , <u><poem></u> , <u><list></u> etc.
17	<u><div></u>	9	<u><Div></u>	7	text blocks of differing visual formatting	generic container for subdivisions
18	<u><frontmatter></u>			8-10	title page, jacket text, <u>impressum</u>	guide to content of the DTB. must contain <u>doctitle</u> , may contain <u>docauthor</u> , copyright notice, table of contents, etc.
	<u><level1 class=</u> <u>"print_toc"></u>	10 - 11	<u><TOC></u> , <u><TOCI></u>	11	table of contents	table of contents

Tag structure comparison – DAISY 3

Index of Elements
DAISY 3 Structure Guidelines
Last Revised: June 4, 2008



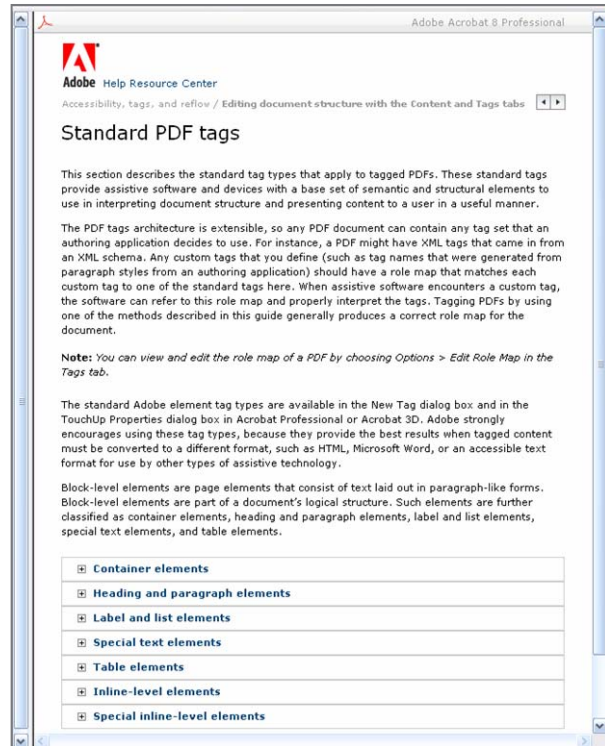
This section of the Guidelines lists all elements in alphabetical order.

DTBook Elements, in Alphabetical Order

- [a](#)
- [abbr](#)
- [acronym](#)
- [address](#)
- [annoref](#)
- [annotation](#)
- [author](#)
- [bdo](#)
- [blockquote](#)
- [bodymatter](#)
- [book](#)
- [br](#)
- [bridgehead](#)
- [byline](#)
- [caption](#)

- 82 pre-defined structures in DAISY 3, including
- 37 main block level and inline elements
- four additional tag sets for tables, images, poetry and mathematics
- tag set sufficient for production of accessible material on various publishing channels

Tag structure comparison – Adobe PDF



- 36 standard PDF tags, five of which describe containers, and six others substructures of tables and ordered lists
- only 25 tags remain to mark up semantic structures within the document itself
- the PDF tag set is an almost complete subset of the DAISY 3 tag set

Block elements missing in PDF Standard Tag Set

DAISY markup	PDF markup	semantic meaning
Part I: DTB Meta Structure (extract)		
1-2	<dtbook>, <book>	1 <Document> root element in the Digital Talking Book
	xml:lang="de-DE"	File > Properties > Advanced > Reading Options > Language language specification
Part II(a): Major Structural Elements (extract)		
3-8	<level1> - <level 6>	major division of a publication; level 1 - 6
9-14	<h1> - <h6>	2-7 <H1> - <H6> heading for level 1 to 6
15	<level>	major division of a publication; used for recursive structures
Part II(b): Block Elements		
25	<address>	address
26	<author>	author
27	<bridgehead>	free-floating heading not associated with hierarchical structure
28	<byline>	information about creator of, or contributor to, a work
29	<code>	17 <Code> fragment of computer code
30	<dateline>	creation date and/or place
31	<epigraph>	epigraph

in PDF, no block elements available for

- addresses and authors
- datelines
- epigraphs
- linegroups
- producer's notes
- information in a sidebar

Inline elements missing in PDF Standard Tag Set

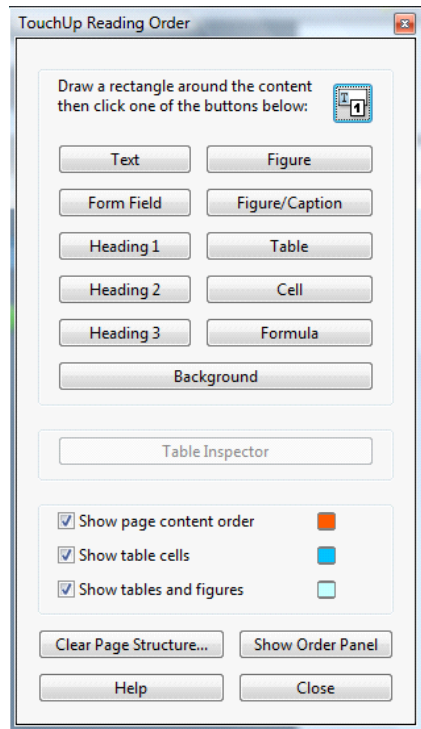
no elements to mark up

- abbreviations
- acronyms
- keywords or definitions
- emphasis
(``, ``, `` or `<i>`)
- super- or subscript characters
- page or line numbers

no elements to mark up
substructures in tables like

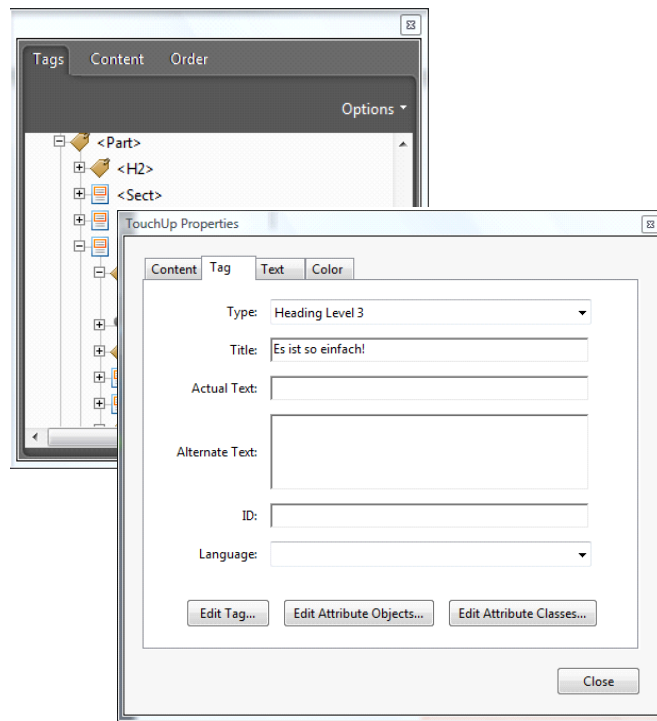
- table columns or column groups
- table header information
- table footer

Tag structure comparison – Adobe PDF



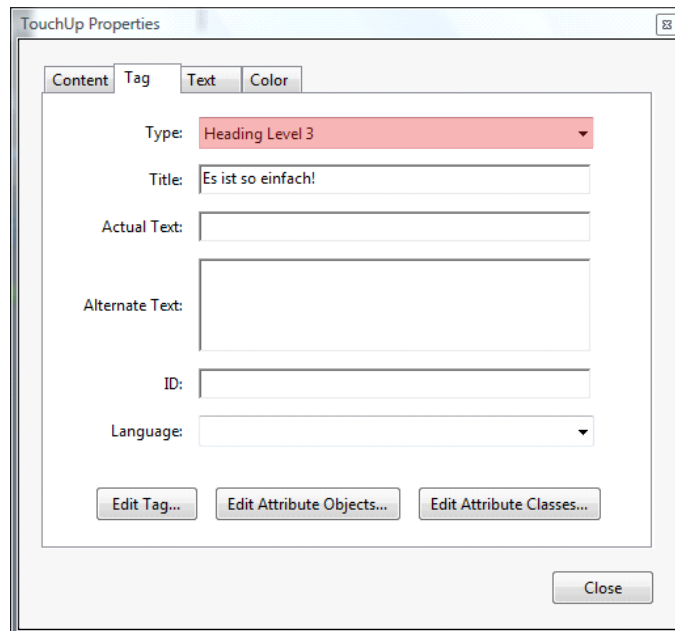
- only ten can be manipulated directly via the »Touch Up Reading Order« Tool
- other elements can only be accessed by using the »TouchUp Properties«-Feature of the »Tags« palette

Tag structure comparison – Adobe PDF



- only ten can be manipulated directly via the »Touch Up Reading Order« Tool
- other elements can only be accessed by using the »TouchUp Properties«-Feature of the »Tags« palette

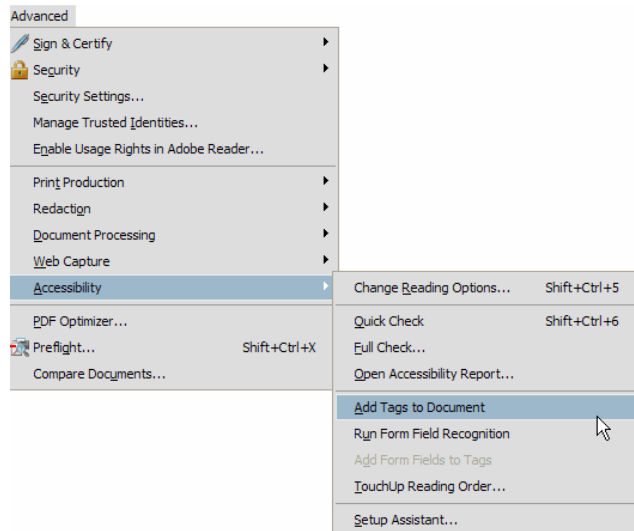
Tag structure comparison – Adobe PDF



alternative tag sets can be defined by the user, but

- each custom tag has to be mapped to an existing standard tag (by using a »role map«)
- custom tag sets can not be used for automatic tag structure generation

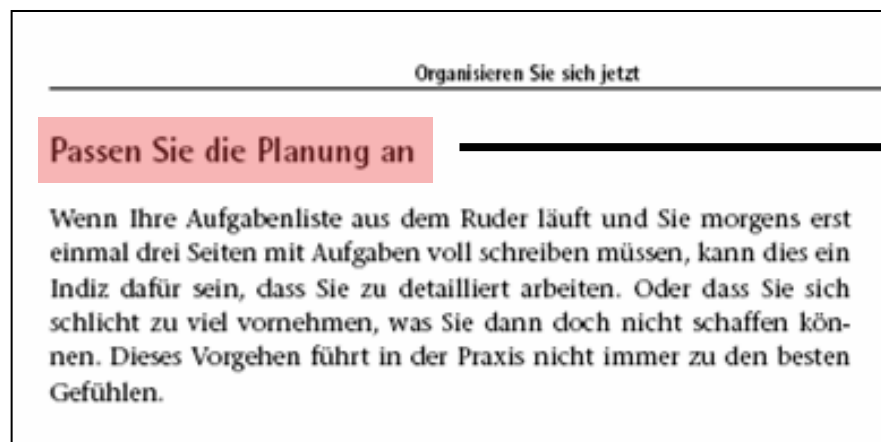
Automatic generation of PDF tag structures



- PDF Standard Tag Set too limited, but all structures can be reused in DAISY 3

- problem:
»libreka!« sample documents are non-tagged PDF
- question:
how good is automatic tag structure generation of Adobe Acrobat 8 Pro (using the »Add Tags to Document« feature)

Automatic generation of tag structures



1. identification of visual formatting
2. mapping of visual formatting to corresponding semantic element

1. identify visual formatting
font weight bold,
font size 12 pt,
increased line spacing



2. map to corresponding markup »Headline 2 «

Adobe PDF standard tags		Export to XML sample document								Export to Microsoft .Doc sample document							
Container elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
	<Document>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
divisions and subdivisions	<Div>, <Art>, <Sect>	o	-	o	x	-	x	x	x	o	-	o	x	-	x	x	x
Heading and paragraph elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
headlines	<H>, <H1> - <H6>	+	o	o	-	o	-	o	o	+	+	+	+	+	o	o	+
paragraph	<p>	-	-	+	o	+	+	+	o	-	o	+	o	+	+	+	+
Label and list elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
list	<L>, 	o	o	o	x	x	o	x	o	+	o	-	x	x	o	x	o
Special text elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
quotation	<BlockQuote>	x	x	x	x	x	-	x	-	x	x	x	x	x	-	x	+
caption	<Caption>	-	x	+	-	o	x	x	o	o	x	o	+	+	x	x	o
indexes	<Index>	o	o	o	-	o	o	x	o	+	+	o	o	+	o	x	+
table of contents	<TOC>, <TOCI>	o	o	-	o	+	o	o	-	+	+	-	+	+	o	o	-
Table elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
table	<Table>, <TR>, <TD>, <TH>	-	+	-	x	x	x	x	x	o	+	-	x	x	x	x	x
Inline-level elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
bibliography entry	<BibEntry>	x	x	x	-	x	-	x	-	x	x	x	o	x	+	x	+
quote	<Quote>	-	-	-	x	x	-	x	-	-	o	-	x	x	o	x	o
span		-	-	-	-	-	x	-	-	+	-	+	o	+	x	+	+
Special Inline-level elements		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
computer source code	<Code>	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
image	<Figure>	o	+	-	o	+	x	x	+	+	+	-	o	+	x	x	o
interactive form	<Form>	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
mathematical formula	<Formula>	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
link	<Link>	+	+	+	+	+	-	+	o	+	-	-	-	+	-	-	-
footnote	<Note>	x	x	x	-	-	x	-	o	x	x	x	o	o	x	-	o
reference	<Reference>	x	x	-	-	-	-	-	-	x	x	-	o	+	+	+	+

Results of automatic tag structure generation

Erster Schritt: Was nehmen wir wahr?

Bevor Sie Ihre Intuition aktiv in Ihrem Privatleben und Arbeitsumfeld einsetzen können, sind noch ein paar „Grundeinstellungen“ nötig, die Ihnen die Anwendung einfacher machen werden. Dieses Kapitel dient der Vorbereitung und wird Ihnen nicht nur Wissen über die Wahrnehmung an sich vermitteln, sondern Ihnen auch dabei helfen, Ihre eigenen Wahrnehmungsschwerpunkte zu erkennen.

Die verschiedenen Wahrnehmungsformen

Die verschiedenen Arten der Wahrnehmung sind mit ausschlaggebend dafür, warum Menschen intuitiv unterschiedlich entscheiden. Lernen Sie, Ihren Wahrnehmungsstärken stärker zu vertrauen und schwächere Aufmerksamkeitsformen zu trainieren.

```
<h1>Erster Schritt: was nehmen wir wahr?</h1>
<p id="para_176">Bevor Sie Ihre Intuition aktiv
<level2 id="chapter_12">
<h2>Die verschiedenen wahrnehmungsformen</h2>
<p id="para_178">Die verschiedenen Arten der w
<p id="para_179">wir nehmen unsere Umwelt mit
<p id="para_180">Aber auch wie sich etwas anfü
```

- detection of headlines, paragraphs and tables of content was ok in about half of the documents
- for other basal structures like tables or lists, there were detection problems in almost every instance

Results of automatic tag structure generation

canae Medii Aevi (Scottish Record Society, New Series, Bd. I.), Edinburgh, (1969), S. 170.	nennenswerter Erf enormes Wachstu einem der wichtig
31 Anderson, Sources, (1990), S. 143, 171.	Damit hatte D
32 Duncan, Kingdom, (1975), S. 145.	Scoticana“ bereits
33 R.B. Brooke, The Coming of the Friars, London, (1975), S. 50-57.	worfen und man gen in der päpstlic
34 Er hatte die Begegnung mit dem hl. Bernhard um Stunden nach dessen Tod verpaßt, siehe Brooke, Connoisseur, (1989), S. 325-26.	Einführung der R trieb er die kirchli

<TR> <TD/>	
<TD>31 Anderson, Sources, (1990), S. 143, 171.	
<TD>Damit hatte David I. die Grundzüge der z </TR>	
<TR> <TD/>	
<TD>hard um Stunden nach dessen Tod verpaßt,	
<TD>Einführung der Reformorden erkennen. Nach </TR>	

- complex headline hierarchies or nested tables were replicated incorrectly
- captions of tables & images were mostly ignored
- sidebars were merged with main text
- two column layouts were interpreted as tables

Results of automatic tag structure generation

34 Beginn und Entwicklung der Landwirtschaft ①

Tabelle 2.1: Weltweite Anbauflächen der wichtigsten Kulturpflanzenarten mit kommerziell genutzten transgenen Sorten (nach JAMES 2003) ②

	weltweite Anbaufläche		
	gesamt (Mio ha) ③	transgen (Mio ha)	transgen (%)
Sojabohne	76	41,4	55
Mais	140	15,5	11
Baumwolle	34	7,2	21
Raps	22	3,6	16
Gesamt	272	67,7	25

①

②

Tabelle 2.1:

	gesamt (Mio ha) ③	transgen (Mio ha)	transgen (%)
Sojabohne	76	41,4	55
Mais	140	15,5	11
Baumwolle	34	7,2	21
Raps	22	3,6	16
Gesamt	272	67,7	25

inline structures were almost completely lost in the conversion process, e.g.

- emphasis (italic or bold formatting)
- inline quotations or
- page numbers or
- column titles

Conclusion

- the PDF Standard Tag Set is still too limited for professional purposes, and the detection rate of the »Add Tags to Document« feature is poor
- page layout recognition, detection of multi-column layouts, sidebars etc. should be improved
- all text information must be retained, even if some inline element is missing in the PDF Standard Tag Set
- incorporating user intervention tag structure detection might be a good idea