ACE’s New TOC Editor

Presentation for ACE community Call March 2024

Presenters:
Ravit H David, ACE service lead
Siena Smith, GSLA
Milica Ivetic, GSLA

The Problem:

- Table of Contents functionality (the ability to download individual chapters) helps students a lot
- Large PDFs (especially high quality scans) can be slow on certain computers
- As physical books don’t have this issue, chapter downloads hold value as an accessibility aid
- Our ToC coverage is low, and that coverage is low-quality
- Our target: all requests received after March 1st being checked for ToC quality

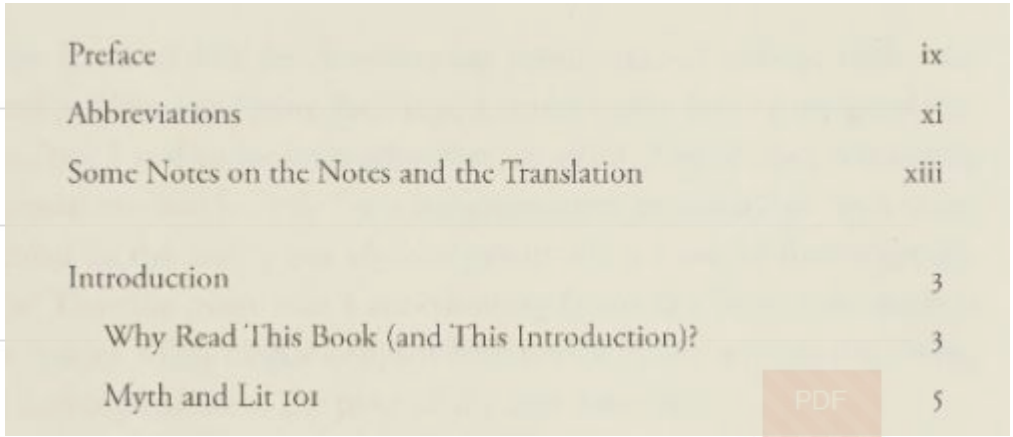
The Challenge

The majority of books we process (nearly 60%) display no ToC to the user; about 30% of them have dysfunctional ToC with garbled chapter titles

- Internet Archive generates ToC data automatically based on the OCR-generated text of scanned books (we convert to our own metadata format and generate chapter downloads)
- Layout context is lost
- Multilevel chunking cannot be handled by the algorithm

Solutions previously considered: editing the raw XML and potentially cancelling all ToC generation in future

What good OCR looks like



Preface

Abbreviations

Some Notes on the Notes and the Translation

Introduction

Why Read This Book (and This Introduction)?

Myth and Lit ror

PDF

PDF

What bad OCR looks like

BRIEF Series Editor's Introduction Acknowledgments Preface About the Authors Chapter 1 Introduction Chapter 2 Describing and Visualizing Sequences Chapter 3 Comparing Sequences Chapter 4 Identifying Groups in Data Analyses Based on Dissimilarities Between Sequences Chapter 5 Multidimensional Sequence Analysis Chapter 6 Examining Group Differences Without Cluster Analysis Chapter 7 Combining Sequence Analysis With Other Explanatory Methods Chapter 8 Conclusions References Index 154 167 DETAILED Series Editor's Introduction Acknowledgments Preface About the Authors Chapter 1 Introduction 1.1 Sequence Analysis in the Social Sciences 1.2 Organization of the Book 1.3 Software Data and Companion Webpage Chapter 2 Describing and Visualizing Sequences 2.1 Basic Concepts and Terminology 2.1.1 Sequences With Recurrent States 2.1.2 Episodes and Transitions 2.1.3 Subsequences 2.2 Defining Sequences 2.2.1 The Alphabet 2.2.2 Sequence Length and Granularity 2.2.3 Sequences of Unequal Length Censoring and Missing Data 2.3 Description of Sequence Data

The Basics 2.3.1 Time Spent in Different States and Occurrence of Episodes 2.3.2 Transition Rates 2.3.3 State Distribution and Shannon Entropy at Different Positions 2.3.4 Modal and Representative Sequences 2.4 Visualization of Sequences 2.4.1 Data Summarization Graphs 2.4.2 Data Representation Graphs xii 2.5 Description Sequences

Assessing Sequence Complexity and Quality 2.5.1 Unidimensional Measures 2.5.2 Composite Indices Chapter 3 Comparing Sequences 3.1 Dissimilarity Measures to Compare Sequences 3.2 Alignment Techniques 3.2.1 Optimal Matching 3.2.2 Assigning Costs to the Alignment Operations 3.2.3 Critiques of Classical OM 3.3 Alignment-Based Extensions of OM 3.4 Nonalignment Techniques 3.5 Comparing Dissimilarity Matrices 3.6 Comparing Sequences of Different Length 3.7 Beyond the Standard Full-Sample Pairwise Sequence Comparison Chapter 4 Identifying Groups in Data Analyses Based On Dissimilarities Between Sequences 4.1 Clustering Sequences to Uncover Typologies 4.1.1 The Rationale Behind Clustering Sequences 4.1.2 Crisp (or Hard) Clustering Algorithms 4.1.3 Partitional Clustering 4.1.4 Using Cluster Quality Indices to Choose the Number of Clusters 4.2 Illustrative Application 4.2.1 Hierarchical Clustering Ward's Linkage 4.2.2 Partitional Clustering Partitioning Around Medoids 4.3 "Construct Validity" for Typologies From Cluster Analysis to Sequences 4.4 Using Typologies as Dependent and Independent Variables in a Regression Framework 4.4.1 Clusters as Outcomes 4.4.2 Clusters as Predictors Chapter 5 Multidimensional Sequence Analysis 5.1 Accounting for Simultaneous Temporal Processes 5.2 Expanding the Alphabet Combining Multiple Channels Into a Single Alphabet

PDF

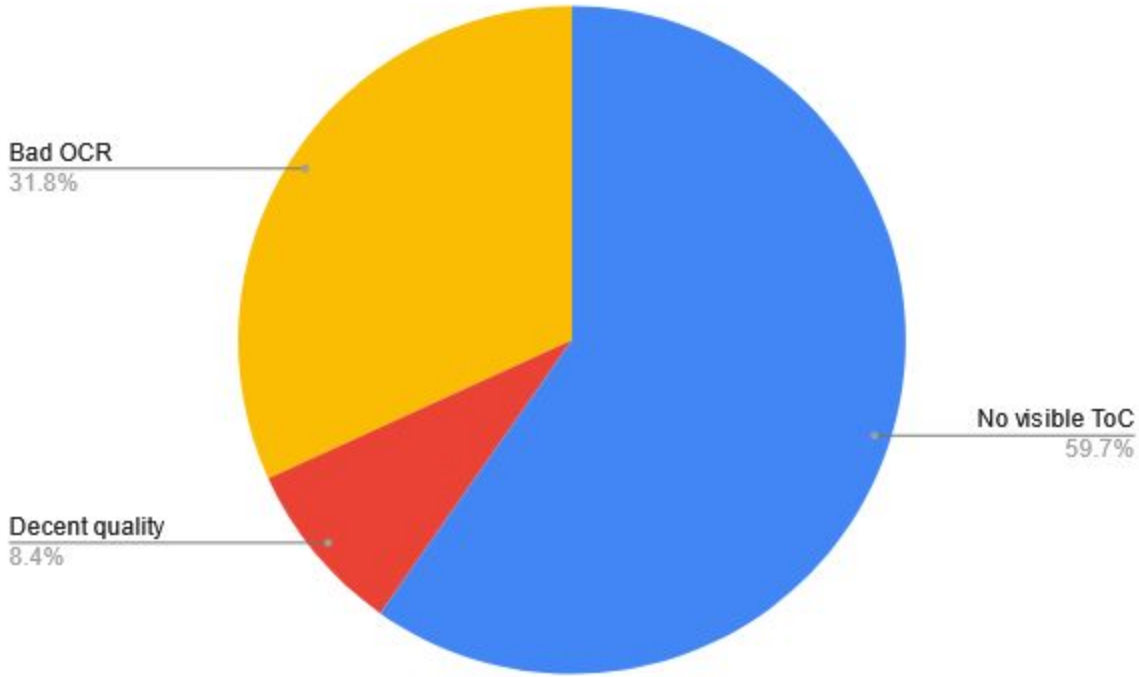
at

PDF

a3 a3

PDF

Recent ACE uploads



ACE ToCs according to Book Interchange Tag Suite (BITS) metadata

Books Without Book-Body tag in BITS: 3219 (this includes Internet Archive books)
Books Without toc.xml in File System: 2936
Bad ToC (Books without Book-Body BUT contains BAD ToC in File System): 641
Bad OCR: Unknown

The editor

What you need for access:

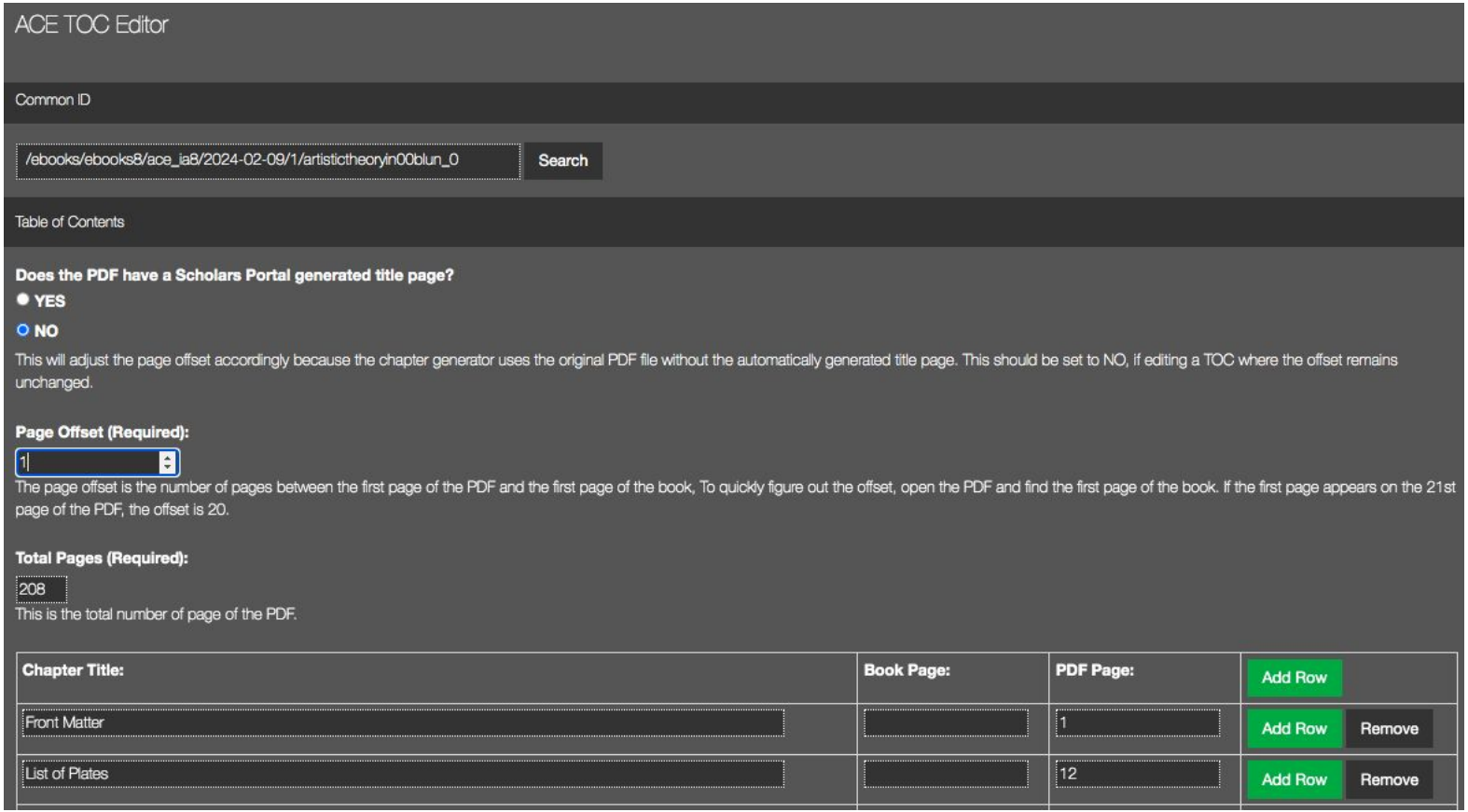
- Admin tool account with editing credentials (Link to Admin tool)

Who can edit a ToC right now?

- Anyone with an account for the editor (currently a handful Scholars Portal members)
- Accounts can be easily created and are available upon request!

How it works

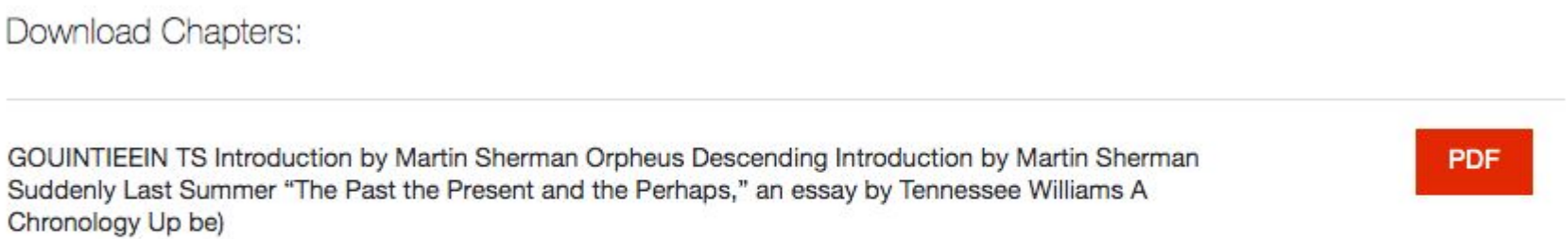
How it works


ACE TOC Editor
Common ID
/ebooks/ebooks8/ace_ia8/2024-02-09/1/artistictheoryin00blun_0
Search
Table of Contents
Does the PDF have a Scholars Portal generated title page?
YES
NO
This will adjust the page offset accordingly because the chapter generator uses the original PDF file without the automatically generated title page. This should be set to NO, if editing a TOC where the offset remains unchanged.
Page Offset (Required):
1
The page offset is the number of pages between the first page of the PDF and the first page of the book, To quickly figure out the offset, open the PDF and find the first page of the book. If the first page appears on the 21st page of the PDF, the offset is 20.
Total Pages (Required):
208
This is the total number of page of the PDF.
Chapter Title:
Book Page:
PDF Page:
Add Row
Front Matter
1
Add Row
Remove
List of Plates
12
Add Row
Remove

Example

Let’s consider two examples

1: Incomplete, lacking.



Common ID:
/ebooks/ebooks8/ace_ia8/2024-01-11/1/orpheusdescendin00will_0

Example

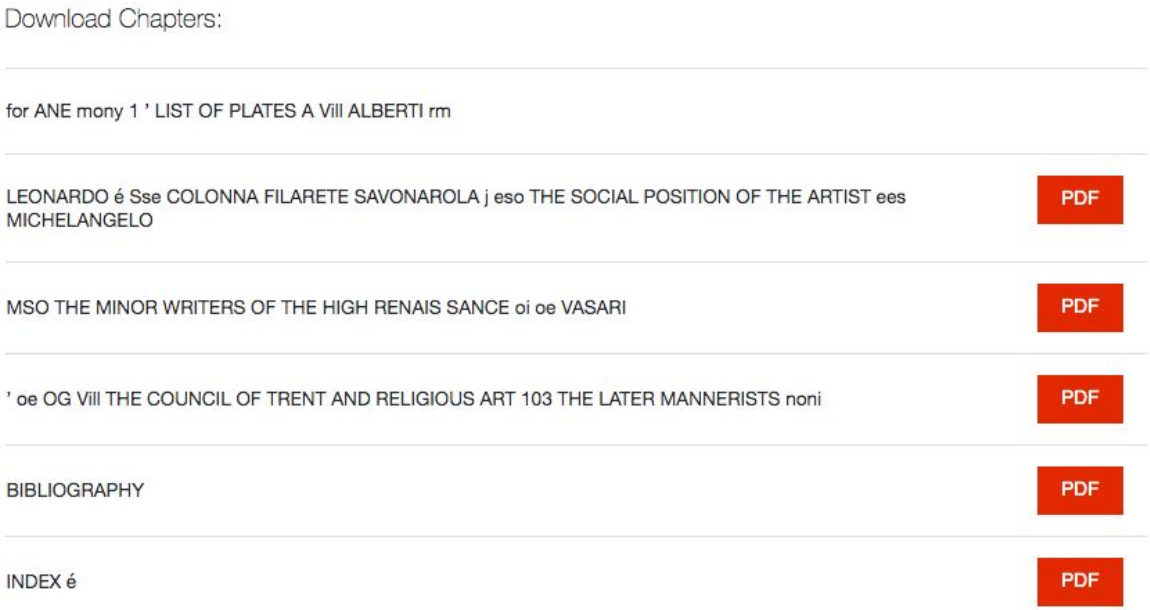
1: Incomplete, lacking. Edited!

Download Chapters:

Front Matter	PDF
Introduction by Martin Sherman (1)	PDF
Orpheus Descending	PDF
Introduction by Martin Sherman (2)	PDF
Suddenly Last Summer	PDF
"The Past, the Present, and the Perhaps," an essay by Tennessee Williams	PDF
A Chronology	PDF

Example

2: Gibberish.

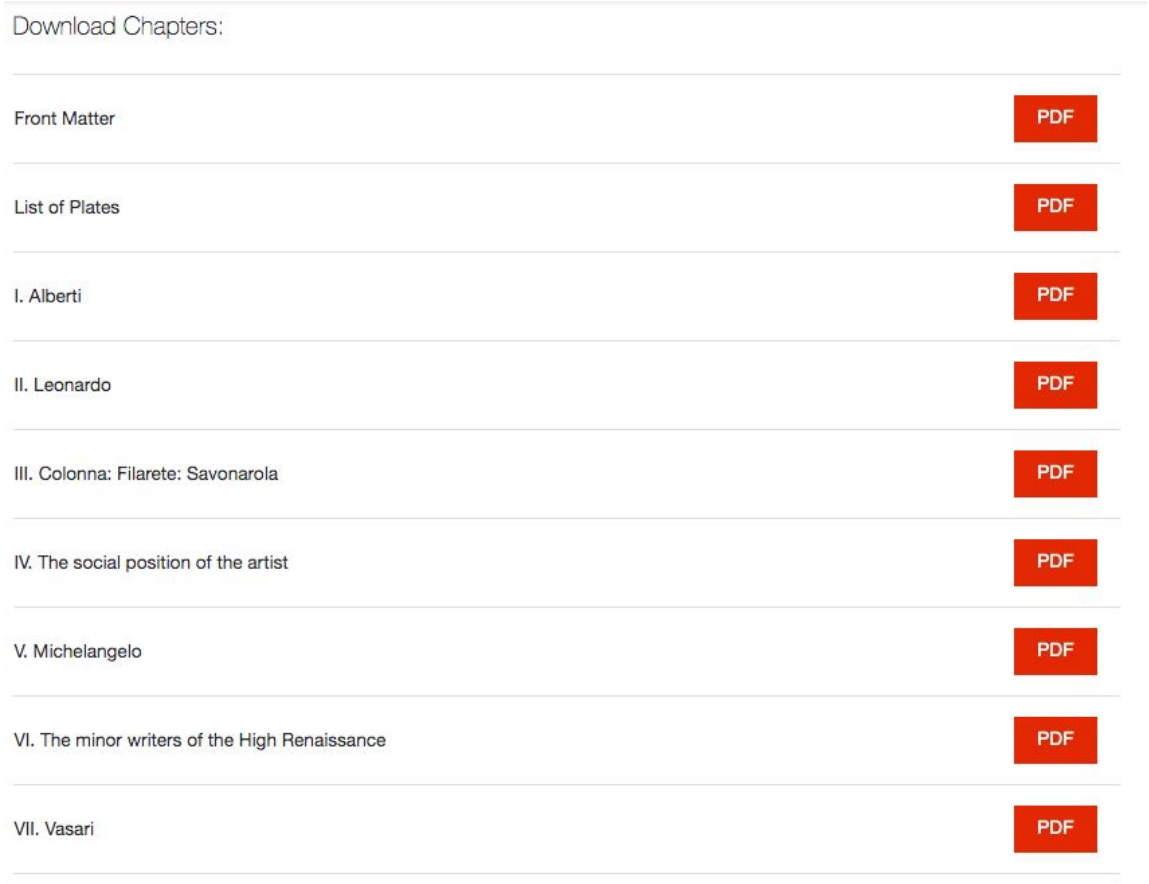


Common ID: /ebooks/ebooks8/ace_ia8/2024-02-09/1/artistictheoryin00blun_0

Example

2: Gibberish.

Edited!



Special Thanks

Bart Kawula, Scholars Portal

Annie Thomas Selvarajan, Scholars Portal

Thank you for listening!

Want to complete a TOC or two?

Email us at ace@scholarsportal.info