

Tika in Action

CHRIS A. MATTMANN
JUKKA L. ZITTING

11

MANNING
SHELTER ISLAND

brief contents

PART 1	GETTING STARTED	1
	1 • The case for the digital Babel fish	3
	2 • Getting started with Tika	24
	3 • The information landscape	38
PART 2	TIKA IN DETAIL	53
	4 • Document type detection	55
	5 • Content extraction	73
	6 • Understanding metadata	94
	7 • Language detection	113
	8 • What's in a file?	123
PART 3	INTEGRATION AND ADVANCED USE	143
	9 • The big picture	145
	10 • Tika and the Lucene search stack	154
	11 " Extending Tika	167

PART 4 CASE STUDIES.

- 12 • Powering NASA science data systems 181
- 13 • Content management with Apache Jackrabbit 191
- 14 • Curating cancer research data with Tika 196
- 15 • The classic search engine example 204

contents

foreword xv
preface xvii
acknowledgments xix
about this book xxi
about the authors xxv
about the cover illustration xxvi

PART 1 GETTING STARTED.....1

"1 The case for the digital Babel fish 3

- *- 1.1 Understanding digital documents 4
A taxonomy of file formats 5 • Parser libraries 6
Structured text as the universal language 9 • Universal metadata 10 • The program that understands everything 13
- 1.2 What is Apache Tika? 15
A bit of history 15 • Key design goals 17 • When and where to use Tika 21
- 1.3 Summary 22

Getting started with Tika 24

- 2.1 Working with Tika source code 25
Getting the source code 25 • The Maven build 26
Including Tika in Ant projects 26

- 2.2 The Tika application 27
 - Drag-and-drop text extraction: the Tika GUI* 29 • *Tika on the command line* 30
- 2.3 Tika as an embedded library 32
 - Using the Tika facade* 32 • *Managing dependencies* 34
- 2.4 Summary 36

The information landscape 38

- 3.1 Measuring information overload 40
 - Scale and growth* 40 • *Complexity* 42
- 3.2 I'm feeling lucky—searching the information landscape 44
 - Just click it: the modern search engine* 44 • *Tika's role in search* 46
- 3.3 Beyond lucky: machine learning 47
 - Your likes and dislikes* 48 • *Real-world machine learning* 50
- 3.4 Summary 52

PART 2 TIKA IN DETAIL 53

yjj Document type detection 55

- * 4.1 Internet media types 56
 - The parlance of media type names* 58 • *Categories of media types* 58 • *IANA and other type registries* 60
- AL.2 Media types in Tika 60
 - The shared MIME-info database* 61 • *The MediaType class* 62
 - The MediaTypeRegistry class* 63 • *Type hierarchies* 64
- 4.3 File format diagnostics 65
 - Filename globs* 66 • *Content type hints* 68 • *Magic bytes* 68
 - Character encodings* 69 • *Other mechanisms* 70
- ¹ 4.4 Tika, the type inspector 71
- 4.5 Summary 72

Content extraction 73

- 5.1 Full-text extraction 74
 - Abstracting the parsing process* 74 • *Full-text indexing* 75
 - Incremental parsing* 77

5.2	The Parser interface	78	.
	<i>Who knew parsing could be so easy?</i>	78	«. <i>The parse() method</i>
	<i>Parser implementations</i>	80	• <i>Parser selection</i>
5.3	Document input stream	84	
	<i>Standardizing input to Tika</i>	84	• <i>The TikaInputStream class</i>
5.4	Structured XHTML output	87	
	<i>Semantic structure of text</i>	87	•• <i>Structured output via SAX events</i>
	<i>Marking up structure with XHTML</i>	89	
5.5	Context-sensitive parsing	91	
	<i>Environment settings</i>	91	• <i>Custom document handling</i>
5.6	Summary	93	
	<i>Understanding metadata</i>	94	
6.1	The standards of metadata	96	
	<i>Metadata models'</i>	96	• <i>General metadata standards</i>
	<i>Content-specific metadata standards</i>	99	
6.2	Metadata quality	101	
	<i>Challenges/Problems</i>	101	• <i>Unifying heterogeneous standards</i>
6.3	Metadata in Tika	104	
	<i>Keys and multiple values</i>	, 105	• <i>Transformations and views</i>
6.4	Practical uses of metadata	, 107	
	<i>Common metadata for the Lucene indexer</i>	108	• <i>Give me my metadata in my schema!</i>
6.5	Summary	111	
	<i>Language detection</i>	113	
7.1	The most translated document in the world	114	
7.2	Sounds Greek to me—theory of language detection	115	
	<i>Language profiles</i>	116	• <i>Profiling algorithms</i>
	<i>The N-gram algorithm</i>	118	• <i>Advanced profiling algorithms</i>
7.3	Language detection in Tika	119	
	<i>Incremental language detection</i>	120	• <i>Putting it all together</i>
7.4	Summary	122	

What's in a file? 123

8.1 Types of content 124

HDF: a format for scientific data 125 • Really Simple Syndication: a format for rapidly changing content 126

8.2 How Tika extracts content 127

Organization of content 128 • File header and naming - conventions 133 • Storage affects extraction 139

8.3 Summary 141

PART 3 INTEGRATION AND ADVANCED USE.....143

9 *The big picture* 145

9.1 Tika in search engines 146

The search use case 146 • The anatomy of a search index 146

9.2 Managing and mining information 147

Document management systems 148 • Text mining 149

9.3 Buzzword compliance 150

Modularity, Spring, and OSGi 150 • Large-scale computing 151

9.4 Summary 153

10 *Tika and the Lucene search stack* 154

10.1 Load-bearing walls 155

ManifoldCF 156 • Open Relevance 157

10.2 The steel frame 159

Lucene Core 159 • Solr 161

10.3 The finishing touches 162

Nutch 162 • Droids 164 • Mahout 165

10.4 Summary 166

11 *Extending Tika* 167

11.1 Adding type information 168

Custom media type configuration 169

- 11.2 Custom type detection 169
 - The Detector interface 170 • Building a custom type detector 170 • Plugging in new detectors 172*
- 11.3 Customized parsing 172
 - Customizing existing parsers 173- Writing a new "s" parser 174 • Plugging in new parsers 175*
 - Overriding existing parsers 176*
- 11.4 Summary 176

PART 4 CASE STUDIES.....179

Powering NASA science data systems 181

- 12.1 NASA's Planetary Data System 182
 - PDS data model 182 • The PDS search redesign 184*
- 12.2 NASA's Earth Science Enterprise 186
 - Leveraging Tika in NASA Earth Science SIPS 187*
 - Using Tika within the ground data systems 188*
- 12.3 Summary 190

13 *Content management with Apache Jackrabbit 191*

- 13.1 Introducing Apache Jackrabbit 192
- 13.2 The text extraction pool 192
- 13.3 Content-aware WebDAV 194
- 13.4 Summary 195

14 *Curating cancer research data with Tika 196*

- 14.1 The NCI Early Detection Research Network 197
 - The EDRN data model 197 • Scientific data curation 198*
- 14.2 Integrating Tika 198
 - Metadata extraction 199 • MIME type identification and classification 201*
- 14.3 Summary 203

15	<i>The classic search engine example</i>	204	
15.1	The Public Terabyte Dataset Project	205	
15.2	The Bixo web crawler	206	
	<i>Parsing fetched documents</i>	207 • <i>Validating Tika's charset detection</i>	209
15.3	Summary	210	
<i>appendix A</i>	<i>Tika quick reference</i>	211	
<i>appendix B</i>	<i>Supported metadata keys</i>	214	
	<i>index</i>	219-	

SummaryTika in Action is a hands-on guide to content mining with Apache Tika. The book's many examples and case studies offer real-world experience from domains ranging from search engines to digital asset management and scientific data processing. About the TechnologyTika is an Apache toolkit that has built into it everything you and your app need to know about file formats. Using Tika, your applications can discover and extract content from digital documents in almost any format, including exotic ones.