

A

<a> tag 137
AbstractParser class 175
add() method 105
agglutinative languages 117
alias 31
analyzers. *See* search engines
annotations 50
Ant build 26–27
 See also source code
Apache Droids 164–165
Apache Gora 163
Apache Hadoop 16, 50
 Bixo 206
Apache Incubator 156–157, 161
 podlings 157
Apache Jackrabbit 192–193, 195
 and ContentHandler interface 193
 and parse() method 193
 and WebDAV 194
 content repositories 192
 Content Repository for Java API 191
 nodes 192
 text extraction pool 192–193
 TextExtractionError 193
Apache Lucene 75
 Document class 159
 ecosystem 155
 Field class 159
 Lucene Core 159–160
Apache Mahout 50, 165–166
Apache Manifold Connectors. *See* ManifoldCF
Apache Nutch 15, 162, 164
 and Bixo 207
 Apache Gora 163
 Protocol plugins 162

Apache PDFBox 75, 82
Apache Solr 161–162
Apache Tika, history of 15, 17
Apache UIMA 50
 annotations 50
application programming interfaces (APIs) 9
 Java ROME API 130
 Parser API 150
 pull APIs 88
 push APIs 88
 See also Content Repository for Java API
application/* MIME type 59
*Architectural Styles and the Design of Network-based
Software Architectures* 40
audio/* MIME type 59
AutoDetectParser 47
AutoDetectParser class 82, 176

B

Babel fish 3
Behemoth 51
/bin/lis output 135
biomarkers 197
Bixo 206, 209
 and Apache Nutch 207
 and TagSoup 207
 Cascading 206
 Fetch subassembly 206
 Parse subassembly 206
 parsing documents 207, 209
 robots.txt file 207
black lists 45
BodyContentHandler class 90
BoilerpipeContentHandler class 208
BOM markers 69
Brin, Sergey 44

build.xml file 26
 byte frequency matching 70
 byte order marks. *See* BOM markers; magic bytes

C

callback functions 19
 cancer research 196, 203
 biomarkers 197
 See also Early Detection Research Network
 Cardinality property 97
 Cascading 206
 Cascading Style Sheets 10
 categorization 48
 CF. *See* Climate Forecast model
 character encodings 69–70
 BOM markers 69
 byte frequency matching 70
 statistical encoding detection 70
 See also character sets; charsets
 character sets 43, 69–70
 validating character set detection 209
 See also character encodings
 CharsetDetector class 209
 charsets 69–70
 See also character encodings; character sets
 classes
 AbstractParser 175
 AutoDetectParser 82, 176
 BodyContentHandler 90
 BoilerpipeContentHandler 208
 CharsetDetector 209
 CompositeDetector 172
 CompositeParser 82
 DelegatingParser 121
 DeploymentAreaParser 140
 Document 76, 159
 ElementMetadataHandler 174
 ExtractingRequestHandler 161
 FeedParser 128, 131
 Field 76, 159
 FileInputStream 84
 HDFParser 131, 133
 HTMLParser 82, 134
 HtmlParser 207
 IdentityHtmlMapper 92
 IndexReader 77
 IndexWriter 78
 InputStream 84, 87, 141
 java.io.Writer 120
 LanguageIdentifier 119
 LanguageIdentifierUpdateProcessor 161
 LanguageProfile 119

LinkContentHandler 90, 138
 LinkHandler 208
 LookaheadInputStream 171
 LuceneIndexer 75
 MediaType 62–63
 MediaTypeRegistry 63–64
 Metadata 19, 105–106, 199
 MimeType 201
 MimeTypes 20, 169
 MimeTypesFactory 169
 Parse 105
 ParseContext 19, 171
 ParserDecorator 176
 ParsingReader 89
 PDFParser 75, 81
 PdfParser 105
 PDFTextStripper 82
 PDSRDFParser 184
 Product 202
 ProfilingHandler 120, 207
 ProfilingWriter 120
 Property 103
 PropertyType 105
 PropertyValue 105
 Reader 34, 77
 Reference 200, 202
 SAXTransformerFactory 91
 SimpleTypeDetector 71
 TeeContentHandler 90, 122, 207
 Tika facade class 32, 34
 TikaCallable 208
 TikaInputStream 85, 87
 TransformerHandler 91
 UpdateHandler 161
 XMLParser 173
 XmlRootExtractor 171
 classpath 26
 Climate Forecast model 12
 ClimateForecast interface 105
 cloud computing 151, 153
 clustering 48
 collaborative filtering 48
 command-line interface
 –language option 114
 See also Tika CLI
 composite design pattern 82
 CompositeDetector class 172
 CompositeParser class 82
 compression 85
 content 123, 142
 how Tika extracts it 127, 141
 organization of 128, 133
 random access 131, 133
 streaming 128, 131
 types 124, 127

- content extraction
 - effect of data storage on 139, 141
 - for search engines 147
 - how it works 127, 141
 - content management. *See* document management systems
 - content repositories 21, 192
 - text extraction pool 192–193
 - Content Repository for Java API 191
 - content type hints 68
 - Content-Encoding headers 208
 - ContentHandler argument 80
 - ContentHandler interface 88, 138
 - in Apache Jackrabbit 193
 - content-specific metadata standards 98–99, 101
 - compared to general standards 99
 - difficulty of comparing across standards 100
 - Content-Type header. *See* content type hints
 - context-free interaction 42
 - COO. *See* Orbiting Carbon Observatory
 - corpus 116
 - distance from 117
 - Hamshahri 158
 - OHSUMED 158
 - Tempo 158
 - Cotent-Type header 208
 - crawlers. *See* search engine
 - CSS. *See* Cascading Style Sheets
 - custom MIME types
 - custom detectors for 170, 172
 - detecting 169, 172
 - custom parsers 172, 176
 - creating 174–175
 - customizing existing parsers 173–174
-
- D**
- DAACs. *See* Distributed Active Archive Centers
 - data curation 198
 - data mining, text mining 149–150
 - data models, Planetary Data System 182, 184
 - data, linked 198
 - databases, MIME-info 61–62
 - deduplication 45
 - Definition property 97
 - DelegatingParser class 121
 - dependencies, managing 34, 36
 - DeploymentAreaParser class 140
 - design goals 17, 21
 - fast processing 19
 - flexible metadata 19
 - flexible MIME type detection 20
 - language detection 21
 - low memory footprint 19
 - MIME database 20
 - parser libraries 19
 - unified parsing interface 18
 - detect() method 34, 71, 170–171
 - detecting custom MIME types 169, 172
 - custom type detectors 170, 172
 - detecting file formats 65, 71
 - detecting MIME types 6
 - Detector interface 170
 - custom type detectors 170, 172
 - dictionary-based profiling 117
 - digital asset management 22
 - See also* document management systems
 - Distributed Active Archive Centers 187
 - document analysis 22
 - Document class 76, 159
 - document management systems 148–149
 - Content-Type headers 148
 - Document Object Model 19
 - document stream
 - InputStream 87
 - InputStream class 84
 - documents 4, 14
 - analyzing 22
 - as text 9
 - custom 92–93
 - document management systems 148–149
 - input stream 84, 87
 - language detection 113, 122
 - parsing with Bixo 207, 209
 - text mining 149–150
 - See also* files
 - DOM. *See* Document Object Model
 - downloading
 - Git 25
 - Subversion 25
 - Tika source code 25
 - drag and drop 30
 - Droids. *See* Apache Droids
 - Dublin Core 11, 98–99
-
- E**
- Early Detection Research Network 197, 203
 - data model 197
 - data sets 201
 - eCAS Curator 199
 - EDRN Catalog and Archive Service 198
 - identifying MIME types 201–202
 - linked data 198
 - metadata extraction 199, 201
 - protocols 197
 - scientific data curation 198
 - use of Tika 198, 202

Earth Science Enterprise 186, 190
 Distributed Active Archive Centers 187
 how Tika fits in 187, 190
 principal investigator 187
 Science Information Processing Systems 187

eCAS Curator 199
 Ingestor 199
 references 199
See also EDRN Catalog and Archive Service

e-commerce, useful user data 49

EDRN Catalog and Archive Service 198
 eCAS Curator 199
See also Early Detection Research Network

ElementMetadataHandler class 174

embedding Tika 32, 36
 Tika facade 32, 34

encoding, output encoding 31

endDocument function 19

endElement function 19

environment settings 91

errors, TextExtractionError 193
See also exceptions

events, STOP 193

example/* MIME type 59

exceptions
 IOException 75, 81
 SAXException 81
 TikaException 75, 81

extending Tika 167, 177
 adding MIME types 168–169

Extensible Hypertext Markup Language 10
 in Tika CLI 30
 structured output 87, 91

Extensible Markup Language (XML) 10, 70
 Resource Description Framework 184
See also XML files

Extensible Metadata Platform (XMP) 12, 98
 properties and property types 103

extracting text
 full text 74, 78
 with Apache Jackrabbit 192–193

ExtractingRequestHandler class 161

F

facade class 32

Facebook 49

fast processing 19

FeedParser class 128, 131

Fetch subassembly 206

Field class 76, 159

Fielding, Roy 40

file extensions 8
See also glob patterns

file formats 4–6
 combined heuristics 71
 content type hints 68
 detecting 65, 71
 filename globs 66, 68
 HDF 125–126, 131, 133
 headers 133–134
 magic bytes 68
 OLE 70
 RSS 126–128, 131
 XML 70

file headers 133–134

File Manager catalog 202

file naming conventions 134, 137

file storage. *See* storage

FileInputStream class 84

filenames, glob pattern 66, 68

files
 compression 85
 content of 123, 142
 file extensions 8
 formats 4–6
 headers 133–134
 HTML 133
 links between 137, 139
 magic bytes 8
 matrix data 126
 naming conventions 134, 137
 scalar data 126
 storage 139, 141
 text files 9
 vector data 126
 XML 8
See also documents

formatted text 30

full-text extraction 74, 78
 incremental parsing 77–78
 indexing 75, 77

full-text indexes, for large-scale systems 152

G

general metadata standards 97, 99
 compared to content-specific standards 99
 Dublin Core 99

Geographic interface 105

get() method 86

getContentHandler() method 174

getDefaultRegistry() method 63

getFile() method 86

getLanguage() method 120

getLinks() method 140

getSupertype() method. *See* MediaTypeRegistry class

getSupportedTypes() method 79, 83, 174–175
 Git 25
 glob patterns 66, 68
 graphical user interface. *See* Tika GUI

H

Hamshahri corpus 158
 handling custom documents 92–93
 hasFile() method 86
 HDF 125–126

- matrix data 126
- organization of content 131, 133
- scalar data 126
- vector data 126

See also Hierarchical Data Format
 HDFParser class 131, 133
 <head> tag 134
 headers

- Content-Encoding 208
- Content-Type 148, 208

 heuristics 71
 Hierarchical Data Format 125–126

- organization of content 131, 133

Hitchhiker's Guide to the Galaxy 3
 HTML. *See* Hypertext Markup Language
 HtmlMapper interface 92
 HTMLParser class 82, 134
 HtmlParser class 207
 Hypertext Markup Language 10, 40

- <head> tag 133
- in Tika CLI 31

I

IANA. *See* Internet Assigned Numbers Authority
 IdentityHtmlMapper class 92
 image/* MIME type 59
 implementing parsers 80, 82
 incremental language detection 120–121
 incremental parsing, streaming 77–78
 indexers

- full-text indexing 152

See also search engines
 indexing, full-text search 75, 77
 IndexReader class 77
 IndexWriter class 78
 information overload 40
 Ingester 199

- references 199

 input, standardizing 84
 InputStream argument 80
 InputStream class 84, 87, 141

- and parse() method 85

interfaces

- ClimateForecast 105
- ContentHandler 88, 138, 193
- Detector 170
- Geographic 105
- HtmlMapper 92
- org.apache.tika.parser.Parser 18
- org.xml.sax.ContentHandler 19
- Parser 20, 35, 78, 83, 130, 135
- TIFF 105

 intermediaries, promotion 42
 International Organization for Standardization 115
 internet

- complexity of 42, 44
- scale and growth of 40, 42

 Internet Assigned Numbers Authority 4, 60

- MIME type registry 6

 inverse indexes 147
 IOException 75, 81

- input error 81

 isMultiValued() method 105
 ISO 639 115
 isReasonablyCertain() method 120
 isSpecializationOf() method. *See* MediaTypeRegistry class

J

Jackrabbit. *See* Apache Jackrabbit
 Java

- embedding Tika 32, 36
- managing dependencies 34, 36
- ROME API 130
- service providers 83

 Java Beans 36
 Java ROME API 130
 java.io.Writer class 120
 java.util.zip package 8
 JCR. *See* Content Repository for Java API

L

language detection 21, 113, 122

- advanced algorithms 119
- agglutinative languages 117
- corpus 116
- distance 117
- in Tika 119, 122
- incremental 120–121
- ISO 639 standards 115
- language profiles 116–117
- N-gram algorithm 118
- profiling algorithms 117, 119

language detection (*continued*)
 theory 115, 119
 UDHR example 114–115
 language detection theory 115, 119
 –language option 114
 language profiles 116–117
 LanguageIdentifier class 119
 LanguageIdentifierUpdateProcessor class 161
 LanguageProfile class 119
 libraries
 Apache PDFBox 75
 Lucene Core 159–160
 parser libraries 6, 9, 19
 PDFBox 82
 Tika as embedded library 32, 36
 LinkContentHandler class 90, 138
 getLinks() method 140
 linked data 198
 LinkHandler class 208
 links, between files 137, 139
 Linnaean taxonomy. *See* taxonomy
 Linnaeus, Carl 55
 locale 91
 LookaheadInputStream class 171
 Lucene 154
 Lucene Core 159–160
 Lucene ecosystem 155
 Apache Droids 164–165
 Apache Mahout 165–166
 Apache Nutch 162, 164
 Apache Solr 161–162
 ManifoldCF 156–157
 Open Relevance 157, 159
 LuceneIndexer class 75
 and metadata 108–109
 converting metadata to RSS 109
 Luke 77

M

machine learning 47, 52
 categorization 48
 clustering 48
 collaborative filtering 48
 predicting user likes and dislikes 48, 50
 real-world examples 50, 52
 magic bytes 68
 Mahout. *See* Apache Mahout
 ManifoldCF 156–157
 mark feature 86
 mark() method 86
 matrix data 126
 Maven build 26
See also source code

Maven, memory problems 26
 media type registries 59
 MediaTypeRegistry class 63–64
 media types 56
See also MIME types
 MediaType class 62–63
See also media types
 MediaTypeRegistry class 63–64
 memory footprint 19
 message/* MIME type 59
 <meta> tag. name attribute 134
 metadata 10, 13, 30, 94, 112
 and Early Detection Research Network 199, 201
 and LuceneIndexer 108–109
 and rest 42
 and Tika CLI 31
 and Tika facade 109
 Cardinality property 97
 challenges of acquiring 101, 103
 Climate Forecast model 12
 Content-Type header 148
 converting to RSS 109, 111
 Definition property 97
 Extensible Metadata Platform 12
 flexibility 19
 how it's created 101, 103
 in Lucene Document objects 159
 instances 104
 metadata models 10
 Metadata.LANGUAGE entry 113, 122
 Name property 97
 practical uses for 107, 111
 quality of 101, 103
 Relationships property 97
 representing 107
 standards 96, 101
 transforming 107
 Valid values property 97
 Metadata argument 80
 Metadata class 19, 105–106, 199
 metadata instances 104
 representing 107
 transforming 107
 metadata models 10
 Climate Forecast model 12
 Dublin Core 11
See also metadata standards
 metadata quality 101, 103
 metadata schema 98
 metadata standards 96, 101
 content-specific standards 97–99, 101
 Dublin Core 99
 general standards 97, 99
 methods
 add() 105

methods (*continued*)
 detect() 34, 71, 170–171
 get() 86
 getContentHandler() 174
 getFile() 86
 getLanguage() 120
 getLinks() 140
 getSupertype() 65
 getSupportedTypes() 79, 83, 174–175
 hasFile() 86
 isMultiValued 105
 isReasonablyCertain() 120
 isSpecializationOf() 65
 mark() 86
 MediaTypeRegistry 63
 parse() 34, 79–80, 122, 129, 175, 193
 parser() 105
 parseToString() 34, 74–75
 reset() 86
 set methods 105
 setMaxStringLength() 34
 setMediaTypeRegistry() 83
 MIME database 20
 MIME type identifiers 6
 MIME types 6, 56, 60
 adding new types to Tika 168–169
 adding to MIME-info database 169
 and Early Detection Research Network
 201–202
 and Parser interface 20
 application/* 59
 audio/* 59
 categories of 58
 custom 169, 172
 custom MIME type detectors 170, 172
 detecting 6, 20
 example/* 59
 identifiers 6
 image/* 59
 Internet Assigned Numbers Authority 60
 media type registries 59
 MediaType class 62–63
 MediaTypeRegistry class 63–64
 message/* 59
 MIME database 20
 MIME-info database 61–62
 model/* 59
 multipart/* 59
 parent and child types 8
 registration 6
 syntax 58
 text/* 59
 Tika MIME repository 200
 top-level 58
 video/* 59
 working with 60, 65

MIME-info database 61–62
 adding new types to 169
 MimeType class 201
 MimeTypes class 20, 169
 MimeTypesFactory class 169
 ML. *See* machine learning
 model/* MIME type 59
 modularity 150
 multipart/* MIME type 59
 Multipurpose Internet Mail Extensions. *See* MIME
 types

N

Name property 97
 NASA 181, 190
 Earth Science Enterprise 186, 190
 how they use Tika 187, 190
 National Polar-orbiting Operational Environ-
 mental Satellite System 188
 Orbiting Carbon Observatory 187
 PDS search redesign 184, 186
 Planetary Data System 182, 186
 Product Evaluation and Analysis Tool
 Element 188
 Soil Moisture Active Passive 188
 National Cancer Institute, Early Detection
 Research Network 197, 203
 National Polar-orbiting Operational Environmen-
 tal Satellite System 188
 NetCDF 131
 N-gram algorithm 118
 nodes 192
 NPOESS. *See* National Polar-orbiting Operational
 Environmental Satellite System
 Nutch. *See* Apache Nutch

O

Object Linking and Embedding 70
 OHSUMED corpus 158
 OLE format 70
 OLE. *See* Object Linking and Embedding
 OODT 198
 Open Relevance 157, 159
 Hamshahri corpus 158
 OHSUMED corpus 158
 Tempo corpus 158
 Open Services Gateway Initiative (OSGi) 36, 151
 Orbiting Carbon Observatory 187
 computing resources 189
 org.apache.tika.language package 35, 119
 org.apache.tika.metadata package 35, 105
 org.apache.tika.mime package 35
 org.apache.tika.parser.Parser interface 18, 20

- org.apache.tika.parser package 35, 174
- org.apache.tika.sax package 35
- org.xml.sax.ContentHandler interface 19
- organization of content 128, 133
- OSGI. *See* Open Services Gateway Initiative
- output
 - SAX events 88–89
 - structured XHTML 87, 91
 - XHTML 89, 91
- output serialization 91
- overriding parsers 176

P

packages

- org.apache.tika.language 35, 119
- org.apache.tika.metadata 35, 105
- org.apache.tika.mime 35
- org.apache.tika.parser 35, 174
- org.apache.tika.sax 35
- See also* java.util.zip
- Page, Lawrence 44
- Parse subassembly 206
- parse() method 34, 79–80, 105, 109, 122, 129, 175
 - and input streams 85
 - ContentHandler argument 80
 - in Apache Jackrabbit 193
 - InputStream argument 80
 - Metadata argument 80
 - ParseContext argument 80, 91
- ParseContext 91
- ParseContext argument 80
- ParseContext class 19, 171
- Parser API 150
- Parser class 105
- Parser interface 35, 78, 83, 130, 135
 - and MIME types 20
 - implementation 80, 82
 - InputStream class 84
 - SAX content handler 88
- parser libraries 6, 9, 19
- parser override 176
- parser selection 82–83
- ParserDecorator class 176
- parsers 47, 78, 83
 - as plugins 175
 - AutoDetectParser 47
 - customizing 172, 176
 - customizing existing parsers 173–174
 - DelegatingParser 121
 - DeploymentAreaParser 140
 - FeedParser class 128, 131
 - HDFParser class 131, 133
 - HTMLParser 134

- implementation 80, 82
- overriding 176
- parser libraries 6, 9, 19
- selecting 82–83
- unified parsing interface 18
- writing a new one 174–175
- parseToString() method 34, 74–75
- parsing 78, 83
 - context-sensitive 91, 93
 - customizing 172, 176
 - customizing existing parsers 173–174
 - incrementally 77–78
 - overview 74–75
 - parser libraries 6, 9
 - PDF files 74
 - SAX events 89
 - SAX-based output 19
 - unified parsing interface 18
 - with Bixo 207, 209
 - writing a new parser 174–175
- parsing context 91, 93
 - environment settings 91
 - locale 91
- ParsingReader class 89
- PDF files, parsing 74
- PDFBox library 82
- PDFParser class 75, 81
- PdfParser class 105
- PDFTextStripper class 82
- PDS 182
- PDSRDFParser class 184
- PEATE. *See* Product Evaluation and Analysis Tool Element
- plain text 30
- Planetary Data System 182, 186
 - data model 182, 184
 - Instruments 182–183
 - labels 182
 - Missions 182–183
 - PDS Data Distribution System 184
 - products 184
 - search redesign 184, 186
 - Targets 182–183
- Planetary Data System Data Distribution System (PDS-S). *See* Planetary Data System
- plugins, parser plugins 175
- podlings 157
- principal investigator. *See* Science Information Processing Systems
- Product class 202
- Product Evaluation and Analysis Tool Element 188
 - and Tika 189
- profiling algorithms 117, 119
 - advanced 119
 - N-gram algorithm 118

ProfilingHandler class 120, 207
 ProfilingWriter class 120
 promotion of intermediaries 42
 properties, in Apache Jackrabbit 192
 Property class 103
 property types 105
 property values 103, 105
 PropertyType class 105
 PropertyType enum 103
 PropertyValue class 105
 Protocol plugins 162
 protocols 46

- in Early Detection Research Network 197

 provider configuration files 83
 Public Terabyte Dataset (PTD) 43
 Public Terabyte Dataset Project 204–205
 pull API 88
 purchase history 49
 push API 88

R

random access 85
 ratings 49
 RDF. *See* Resource Description Framework format
 Reader class 34, 77
 Really Simple Syndication (RSS) 107, 126–127

- from metadata 109, 111
- organization of content 128, 131

 Reference class 200, 202
 Relationships property 97
 Representational State Transfer. *See* REST
 reset() method 86
 Resource Description Framework format 184
 resource management, close() method 84
 REST 42

- context-free interaction 42
- principles of 42
- promotion of intermediaries 42
- use of metadata 42

 RFC 5646 115
 robots.txt file 207
 ROME 129
 root element detection, XML root detection 70
 root elements 8
 RSS 126–127

- channels 126
- organization of content 128, 131

See also Really Simple Syndication

S

SAX events 88

- parsing 89

See also Simple API for XML

SAXException 81

- output error 81

 SAXTransformerFactory class 91
 scalability 151, 153
 scalar data 126
 Science Information Processing Systems (SIPS) 187

- how Tika fits in 187, 190
- principal investigator 187

 search engine 44
 search engines 13–14, 21, 44, 47, 146–147, 204, 210

- analyzers 147
- and Tika 46–47
- Bixo 206, 209
- black lists 45
- crawlers 146
- deduplication 45
- indexers 147
- inverse indexes 147
- Public Terabyte Dataset Project 205
- structure of 146–147
- URL filtering 45
- web crawlers 45
- white lists 45

 service providers 83

- provider configuration files 83

 set methods 105
 setMaxStringLength() method 34
 setMediaTypeRegistry() method 83
 shared MIME-info database. *See* MIME-info database
 Simple API for XML 19

- callback functions 19
- parse() method ContentHandler argument 80
- structured output 88–89

 SimpleTypeDetector class 71
 SIPS. *See* Science Information Processing Systems
 SMAP. *See* Soil Moisture Active Passive
 social media 49
 Soil Moisture Active Passive 188
 Solr. *See* Apache Solr
 SolrCell 161
 source code 25, 27

- downloading 25
- Git 25
- Subversion 25

 Spring framework 151

- bean configuration 151

 startDocument function 19
 startElement function 19
 statistical encoding detection 70
 STOP event, in Apache Jackrabbit 193
 storage

- how it affects extraction 139, 141

storage (*continued*)
 logical representation 139, 141
 physical representation 141
 streaming 128, 131
 structured text 9–10, 30, 87, 91
 as SAX output 88–89
 semantic structure 87–88
 sub-class-of 64
 Subversion 25
 trunk checkout 25

T

TagSoup 207
 Taste. *See* Apache Mahout
 taxonomy 55
 TeeContentHandler class 90, 122, 207
 Tempo corpus 158
 text mining 149–150
 Text Retrieval Conference. *See* TREC standards
 text, structured 9–10
 text/* MIME type 59
 TextExtractionError 193
 TIFF interface 105
 Tika
 adding new MIME types 168–169
 and cancer research 196, 203
 and Early Detection Research Network 198,
 202
 and NASA 181, 190
 and PEATE 189
 Ant build 26–27
 as Babel fish 3
 as embedded library 32, 36
 command-line interface 30, 32
 computing resources 189
 design goals 17, 21
 extending 167, 177
 extracting full text 74, 78
 facade 32, 34
 fast processing 19
 for search engines 204, 210
 GUI 29–30
 history of 15, 17
 how it extracts content 127, 141
 how NASA uses 187, 190
 identifying files for ingestion 190
 in search engines 46–47, 146–147
 indexing across different file types 189
 introduction to 15, 22
 language detection 119, 122
 managing dependencies 34, 36
 Maven build 26
 memory footprint 19
 MIME repository 200

modularity 150
 origin of name 16
 requirements for data delivery and
 dissemination 190
 scalability 151, 153
 source code 25, 27
 standardizing input 84
 Tika application 27, 32
 Tika CLI 30, 32
 validating character set detection 209
 when to use 21–22
 working with metadata 104, 107
 working with MIME types 60, 65
 Tika Annotator. *See* Apache UIMA
 Tika application 27, 32
 documentation 28
 tika-app. *See also* Tika CLI; Tika GUI
 Tika bundle. *See* Open Services Gateway Initiative
 Tika CLI 30, 32
 HTML 31
 –language option 114
 metadata 31
 output encoding 31
 XHTML 30
 Tika CLI. *See* Tika application, Tika GUI
 Tika facade 32, 34
 and metadata 109
 detect() method 34
 parse() method 34
 parseToString() method 34, 74–75
 setMaxStringLength() method 34
 Tika GUI 29–30
 views 30
 Tika GUI. *See* Tika application; CLI
 Tika MIME repository 200
 Tika. *See* Apache Tika
 tika-app 36
 tika-bundle 36
 TikaCallable class 208
 tika-core 35
 TikaException 75, 81
 parse error 81
 TikaInputStream class 85, 87
 See also document stream
 tika-mimetypes.xml. *See* media type registries
 tika-parent 35
 tika-parsers 35
 top level project 155
 TransformerHandler class 91
 transforming metadata 107
 TREC standards 159
 Twitter 49
 type hierarchies 64–65
 media type inheritance. *See* media type
 sub-class-of 64

type hints, content type hints 68
type/subtype 58

U

UDHR. *See* Universal Declaration of Human Rights
Unicode 43
 BOM markers 69
uniform resource locators, URL filtering 45
Universal Declaration of Human Rights
 (UDHR) 114–115
Unix pipeline. *See* Tika CLI
unravelStringMet function 132
UpdateHandler class 161
updateVersion function 140
URL filtering 45
users
 characteristics 49
 item ratings 49
 purchase history 49

V

Valid values property 97
ValueType enum 103
vector data 126
video/* MIME type 59

W

web browsers 40
web crawlers 45
 protocol layer 46
web servers 40
Web-based Distributed Authoring and Versioning
 Protocol (WebDAV) 194–195
 when to use 194
white lists 45
World Wide Web
 architecture 40
 complexity of 42, 44
 scale and growth of 40, 42

X

XHTML output 89, 91
XHTML. *See* Extensible Hypertext Markup Language
XML files, root elements 8
XML. *See* Extensible Markup Language
XMLParser class 173
XmlRootExtractor class 171
XMP dynamic media 98
XMP. *See* Extensible Metadata Platform
xmpDM. *See* XMP dynamic media