

Index

Symbols

1:*n* proximity searches, 191
360-degree view of customer, 145-146

A

access integration, 66
 performance limitations, 66-68
access probability, filtering data based on, 123
accessibility of unstructured data, 9
accessing
 data linkages, 118-122
 multiple tables, 109-110
 spreadsheets, 137
 unstructured databases, 108
 in integrated data warehouses, 182-183
accuracy, degree in semistructured data, 78-79
alerts, 38
altering source text, 52
alternate spellings, 58
alternate terms, indirect searches of, 53
analysis
 global spreadsheet analysis, 142
 of structured and unstructured data, 113-114
 intersection of data, 115-118
 in textual analytics methodology, 170-173
 unstructured databases for, 108-111
 of visualizations, 132
analytical environment, 26-31
 in structured systems, 49
analytical processing. *See* textual analytics

applications, 17
 of integrated data, 61-65
archiving emails, 144
arguments in searches, 34, 38
 enhancements to, 42-43
attributes of records, 19

B

basic word string snippet reports, 192-194
BI (Business Intelligence) technology, 26-27
blather in emails, 144-145
blobs of data, preventing, 96-97
Boolean expressions, 43
browsing queries, 118
Business Intelligence (BI) technology, 26-27
business relevance, determining, 54

C

case studies
 contract management through unstructured
 database, 209-214
 corporate taxonomy (glossary) creation, 215-218
 database for harmful chemicals, 203-208
 insurance claims case study (data warehouse),
 219-225
 textual analytics in medical research, 195-202
case-sensitivity, removing, 58
categories, external, 60-61
 in simple pointers, 106-107
CDI, 124
cell identification in spreadsheets, 141

changing source text, 52
 characteristics of unstructured data, 5-8
 chemical reports (harmful chemicals) case study, 203-208
 choosing. *See* selecting
 cluster diagrams, 43-44
 clustering text for visualizations, 130
 colloquial terms, defined, 197
 common data patterns in semistructured data, 75-76
 common values in spreadsheets, 141-142
 communication
 customer communications
 collecting, 145-146
 integration of, 62-64
 forms of
 corporate contracts, 25-26
 emails, 21-22
 legal information, 24-25
 medical records, 23-24
 spreadsheets, 22
 transcribed telephone conversations, 23
 linking, 114-115
 as unstructured data, 21, 124
 communications-based linkages between structured and unstructured databases, 177
 computers, components of, 16-17
 concatenation
 synonym, 56
 textual, 161-162
 conflicting values in semistructured data, 77-78
 context of language, 44-45
 contract management through unstructured database case study, 209-214
 contracts
 corporate contracts, 25-26
 as unstructured data, 3, 125
 corporate contracts, 25-26
 corporate functions, unstructured data in, 3-4
 corporate glossary, creating, 158
 corporate ontology, creating, 159-160
 corporate taxonomy, creating, 39-40, 159
 case study, 215-218
 cost of unstructured data, 10
 crawlers, 37
 CRM (customer relationship management) systems, 62-63, 124
 cross equivalency in spreadsheets, 141-142
 customer communications
 collecting, 145-146
 integration of, 62-64
 customer relationship management (CRM) systems, 62-63, 124

D

dangerous chemicals. *See* harmful chemicals case study
 data flow in unstructured databases, 96-97
 data integrity, need for, 86-87
 data marts, 27-28
 data models for metadata, 150
 basis for, 150-152
 corporate glossary, creating, 158
 corporate ontology, creating, 159-160
 corporate taxonomy, creating, 159
 external data model, 155-156
 generic data model, 156-158
 internal data model, 154-157
 levels of, 152-153
 relationships among components, 153-154
 role in building infrastructure, 160
 data patterns, 75-76
 data quality in unstructured environment, 161-162
 data volume. *See* volume of data
 data warehouses, 27, 86-87
 DW 2.0 architecture, 91-93
 insurance claims case study, 219-225
 integrated data warehouses, 179-180
 housing in DBMS, 180-182
 simple queries, 182-183
 unstructured databases as, 176
 database management system (DBMS), 17-19
 housing integrated data warehouses, 180-182
 databases, in harmful chemicals case study, 203-208. *See also* structured databases; unstructured databases
 decision support data, 86
 default values, in semistructured data, 77
 definitions, creating for corporate taxonomy (glossary) creation case study, 218
 degree of accuracy, in semistructured data, 78-79
 departments, 3-4
 document consolidation, 59
 document/word databases. *See* unstructured databases
 documents, moving in simple pointers, 104-105
 DW 2.0 (data warehouse architecture), 91-93
 dynamic links, 117-118

E

ECM (Enterprise Content Management), 44
 editing, types of, 163

electronic format, transferring paper data to
 (harmful chemicals case study), 205

emails, 21-22
 archiving, 144
 characteristics of, 143-144
 for collecting customer communications, 145-146
 purpose of, 143
 screening, 144-145
 as unstructured data, 2

enhancements to search arguments, 42-43

Enterprise Content Management (ECM), 44

entities, in data model, 152

environments. *See* analytical environment;
 structured environment; unstructured
 environment

equivocations, creating for corporate taxonomy
 (glossary) creation case study, 218

ETL (extract/transform/load) processing, 180

external categories, 60-61, 150
 creating, 167
contract management case study, 211
harmful chemicals case study, 207
 in data model, 153
 in insurance claims case study, 222
 in simple pointers, 106-107

external data models, 155-156

external metadata, 149
 data models for, 150-151

F

false positives, avoiding, 115

federated analytical technology with integrated
 data warehouses, 182

federated queries, 41, 181

file types, selecting, 51-52

filtering data based on access probability,
 increasing performance with, 123

first generation textual analytics, 33, 89
 cluster diagrams, 43-44
 context of language, 44-45
 federated queries, 41
 hyperlinks, creating, 40
 index searches, 36-37
 integration of unstructured data, 41-42
 locations of unstructured data, 35-36
 multiple language searches, 39
 output of searches, 35
 pattern searches, 34-35
 schedulers, 38
 search argument enhancements, 42-43

search arguments, 38
 simplicity of, 34
 tagging source text, 38-39
 taxonomies, creating, 39-40

flow of data in unstructured databases, 96-97

formatting hits (in searches), 35
 as hyperlinks, 40
 as taxonomies, 39-40

foundation integration, 66
 performance limitations, 66-68

future capabilities of unstructured databases,
 124-125

G-H

general classifications, defined, 197

generic data models, 156-158

global spreadsheet analysis, 142

glossaries (corporate), creating, 158
 case study, 215-218

harmful chemicals case study, 203-208

high-level design of unstructured databases, 109

hits (in searches), 34
 formatting, 35
as hyperlinks, 40
as taxonomies, 39-40

homogeneous data, selecting for visualizations,
 128-129

homograph resolution, 57

homographic editing, 164

homographs
 defining, 166
 in simple pointers, 105-106

hot data in unstructured databases, 111

hot emails, 144-145

hyperlinks, creating, 40

I

identifier-based linkages between structured and
 unstructured databases, 178

identifiers for spreadsheets, 138

increasing performance
 filtering based on access probability, 123
 with priority rankings, 122-123

indexes. *See also* external categories
 searching, 36-37
 in unstructured databases, 99-100

indexing
 all unstructured data, 102
 selective unstructured data, 103

- semistructured data, 100
 - with simple pointers, 104-107
- indirect indexes. *See* external categories
- indirect searches of alternate terms, 53
- indirect searches of related terms, 53
- indirect SQL queries, 189-191
- industries, unstructured data in, 4-5
- information systems
 - DW 2.0 (data warehouse architecture), 91-93
 - growth of, 83-87
 - data integrity*, 86-87
 - online processing*, 83-84
 - personal computers*, 85
 - unstructured data, need for inclusion, 87-91
- insurance claims case study (data warehouse), 219-225
- integrated data warehouses. *See* data warehouses
- integration. *See also* methodology for textual analytics
 - of text for visualizations, 129
 - of unstructured and structured databases, 175-176
 - communications-based linkages*, 177
 - housing in DBMS*, 180-182
 - identifier-based linkages*, 178
 - as integrated data warehouse*, 179-180
 - simple queries*, 182-183
 - word linkages*, 176-177
 - of unstructured data, 41-42, 49-50, 89-91
 - external categorization*, 60-61
 - importance of*, 53-54
 - issues to address*, 54-60
 - reading raw text*, 50-52
 - selecting file type*, 51-52
 - selecting type of integration*, 65-70
 - usage examples*, 61-65
- integrity of data, 86-87
- internal data models, 154-157
 - corporate glossary, creating, 158
 - corporate ontology, creating, 159-160
 - corporate taxonomy, creating, 159
- internal metadata, 148-149
 - data models for, 152
- intersecting structured and unstructured data, 115-118. *See also* linking
- iterative development
 - contract management case study, 211-212
 - of visualizations, 131

J-K-L

- joining linkage data, 120-122
- keys
 - in records, 18
 - for relational tables, 98-99
- KPIs (key performance indicators), 26
- languages
 - context of, 44-45
 - multiple languages, searching, 39
 - in unstructured data, 10
- lassos, 132
- legal information, 24-25
- linkages (of structured and unstructured data)
 - accessing, 118-122
 - monitoring, 119-120
 - multi-database linkages, 179-180
 - real-time online queries, 122-124
- linking. *See also* intersecting structured and unstructured data
 - communications, 114-115
 - with dynamic links, 117-118
 - with probabilistic links, 116-117
 - with semistructured data, 116
 - with static links, 117-118
 - unstructured and structured databases
 - communications-based linkages*, 177
 - identifier-based linkages*, 178
 - word linkages*, 176-177
- loading unstructured databases, 110

M

- m:n* proximity searches, 191
- many-to-many joins, 121-122
- medical field, unstructured data in, 124
- medical records, 23-24
 - as unstructured data, 3
- medical research case study (textual analytics), 195-202
- metadata, 62
 - data models for, 150
 - basis for data model*, 150-152
 - corporate glossary, creating*, 158
 - corporate ontology, creating*, 159-160
 - corporate taxonomy, creating*, 159
 - external data model*, 155-156
 - generic data model*, 156-158
 - internal data model*, 154-157
 - levels of data model*, 152-153
 - relationships among components*, 153-154
 - role in building infrastructure*, 160

- definition of, 147
- in spreadsheets, 138-139
- in structured environment, 147-148
- in unstructured environment, 148
 - external metadata*, 149
 - internal metadata*, 148-149
- methodology for textual analytics, 163
 - analysis stage, 170-173
 - preparation stage, 163-167
 - processing stage, 167-168
 - source data organization stage, 168-170
- mid-level data model, 153
- misspellings, 58
- monitoring linkages, 119-120
- moving documents in simple pointers, 104-105
- multi-database linkages, 179-180
- multiple languages, searching, 39
- multiple tables, accessing, 109-110

N–O

- natural language processing (nlp), 44-45
- negativity exclusion, 58
- network managers, 16
- non-textual unstructured data, 2
- numeric data in spreadsheets, 140
- OCR (optical character recognition), 50, 205
- one-to-many joins, 121
- online processing, 83-84
- ontology, creating corporate ontology, 159-160
- operating systems, 16
- opportunities of unstructured data, 11-12
- optical character recognition (OCR), 50, 205
- organization of source data in textual analytics
 - methodology, 168-170
- organizational functions, 3-4
- output of searches in first generation textual analytics, formatting, 35, 39-40
- output types for preprocessing semistructured data, 80-81

P

- paper, converting to electronic data, 50, 205
- partial searches, 124
- pattern searches, in first generation textual analytics, 34-35
- patterns of data in semistructured data, 75-76
- performance, increasing
 - filtering based on access probability, 123
 - with priority rankings, 122-123

- performance limitations, access integration
 - versus foundation integration, 66-68
- permutations of words, 74
- personal computers, 85
- phrases, support for, 57
- pointers, simple, 104-107
- prefaced values in semistructured data, 76-77
- preparation stage in textual analytics
 - methodology, 163-167
- preparation time for preprocessing semistructured data, 82
- preprocessing semistructured data, 79-80
 - preparation time for, 82
 - variable output types for, 80-81
- prioritization
 - increasing performance with, 122-123
 - of unstructured data, 10
- probabilistic links, 116-117
- processing stage in textual analytics methodology, 167-168
- proximity SQL queries, 191-192
- punctuation, removing, 58

Q

- quality of data in unstructured environment, 161-162
- queries. *See also* analysis
 - browsing queries, 118
 - federated queries, 41, 181
 - real-time online queries, 122-124
 - simple queries, 182-183
 - SQL queries in textual analytics, 185-187
 - basic word string snippet reports*, 192-194
 - indirect query example*, 189-191
 - proximity query example*, 191-192
 - simple query example*, 187-189
 - to unstructured and structured data
 - simultaneously, 93

R

- raw text, reading, 50-52
- reading raw text, 50-52
- real-time online queries, 122-124
- recasting visualizations, 132
- records in tables, 18-19
- reducing words to stems, 55
- related terms, indirect searches of, 53
- relational tables in unstructured databases, 97-99
- relevance, determining, 54

- removing
 - case-sensitivity, 58
 - punctuation, 58
 - stop words, 54
 - retrieval of structured data, speed of, 19-20
 - rows in tables, 18-19
- S**
- schedulers, 38
 - screening emails, 144-145
 - search arguments, 38
 - enhancements to, 42-43
 - search technologies, 33, 89
 - cluster diagrams, 43-44
 - context of language, 44-45
 - federated queries, 41
 - hyperlinks, creating, 40
 - index searches, 36-37
 - integration of unstructured data, 41-42
 - locations of unstructured data, 35-36
 - multiple language searches, 39
 - output of searches, 35
 - pattern searches, 34-35
 - schedulers, 38
 - search argument enhancements, 42-43
 - search arguments, 38
 - simplicity of, 34
 - tagging source text, 38-39
 - taxonomies, creating, 39-40
 - searchability of unstructured data, 10
 - searches
 - indirect searches, 53
 - partial (wild card) searches, 124
 - permutations of words, 54
 - simple searches, 53
 - second generation textual analytics, 50, 89-91
 - external categorization, 60-61
 - importance of integration, 53-54
 - issues to address, 54-60
 - reading raw text, 50-52
 - selecting file type, 51-52
 - selecting type of integration, 65-70
 - usage examples, 61-65
 - security of unstructured data, 10
 - selecting
 - file types, 51-52
 - homogeneous data for visualizations, 128-129
 - integration type, 65-70
 - self-organizing map (SOM), 130-131
 - analyzing, 132
 - recasting, 132
 - for semistructured data, 133
 - semantic processing, 44
 - semistructured data, 34-35, 73
 - common data patterns, 75-76
 - conflicting values, 77-78
 - default values, 77
 - degree of accuracy, 78-79
 - in emails, 144
 - examples of, 74-75
 - indexing, 100
 - linking with, 116
 - prefaced values, 76-77
 - preprocessing, 79-82
 - in spreadsheets, 140
 - visualizations of, 131-133
 - semistructured processing
 - harmful chemicals case study, 205-206
 - on unstructured data, 81
 - sensitive data, 111
 - simple pointers, 104-107
 - simple queries, 182-183, 187-189
 - simple searches, importance of integration, 53
 - simplicity of first generation textual analytics, 34
 - snippets of text, 188
 - basic word string snippet reports, 192-194
 - SOM (self-organizing map), 130-131
 - analyzing, 132
 - recasting, 132
 - for semistructured data, 133
 - source data organization stage in textual analytics
 - methodology, 168-170
 - source text
 - altering, 52
 - tagging, 38-39
 - spellings of words, 58
 - spreadsheets, 22, 135-137
 - accessing, 137
 - cell identification, 141
 - challenges of, 135-137
 - cross equivalency, 141-142
 - global spreadsheet analysis, 142
 - types of data in, 138-140
 - unique identification for, 138
 - as unstructured data, 125
 - SQL interface, accessing unstructured databases, 108

SQL queries in textual analytics, 185-187
 basic word string snippet reports, 192-194
 indirect query example, 189-191
 proximity query example, 191-192
 simple query example, 187-189

static links, 117-118
 processing independently, 120

stemming, 55, 164, 167

stop words
 defining, 164
 removing, 54

stop-word editing, 163

storage of structured data, speed of, 19-20

structured data
 analyzing with unstructured data, 113-114
accessing linkages, 118-122
intersection of data, 115-118
real-time online queries, 122-124
 explained, 1
 integration of unstructured data with, 90-91
 updating, 9

structured databases, integration with unstructured databases, 175-176
 communications-based linkages, 177
 housing in DBMS, 180-182
 identifier-based linkages, 178
 as integrated data warehouse, 179-180
 simple queries, 182-183
 word linkages, 176-177

structured environment, 15-16, 18-20
 bringing unstructured data into, 30-31
 computer components in, 16-17
 DBMS in, 17-19
 integration of unstructured environment with, 49-50
external categorization, 60-61
importance of, 53-54
issues to address, 54-60
reading raw text, 50-52
selecting file type, 51-52
selecting type of integration, 65-70
usage examples, 61-65
 metadata in, 147-148
 storage and retrieval, speed of, 19-20
 storing all unstructured data in, 102
 storing selective unstructured data in, 103
 textual analytics in, versus unstructured environment, 28-30

structured systems
 analytical environment in, 49
 unstructured systems versus, 47-49

synonym concatenation, 56

synonyms
 defining, 164-165
 editing, 163
 replacing, 56
 resolving, 56-57
 in simple pointers, 105-106

T

tables
 in databases, 18-19
 multiple tables, accessing, 109-110
 relational tables in unstructured databases, 97-99

tagging source text, 38-39

taxonomies, creating, 39-40, 159
 case study, 215-218

terminology
 in corporate taxonomy (glossary) creation case study, 217-218
 in insurance claims case study (data warehouse), 220
 in medical research case study, 196-197
 in unstructured data, 9

text, visualizing. See visualizations

text clustering for visualizations, 130

text integration for visualizations, 129

textual analytics
 challenges for, 9-10
 contract management case study, 209-214
 first generation technologies, 33, 89
cluster diagrams, 43-44
context of language, 44-45
federated queries, 41
hyperlinks, creating, 40
index searches, 36-37
integration of unstructured data, 41-42
locations of unstructured data, 35-36
multiple language searches, 39
output of searches, 35
pattern searches, 34-35
schedulers, 38
search argument enhancements, 42-43
search arguments, 38
simplicity of, 34
tagging source text, 38-39
taxonomies, creating, 39-40
 harmful chemicals case study, 203-208
 medical research case study, 195-202
 methodology for, 163
analysis stage, 170-173
preparation stage, 163-167

- processing stage*, 167-168
- source data organization stage*, 168-170
- second generation textual analytics, 50, 89-91
 - external categorization*, 60-61
 - importance of integration*, 53-54
 - issues to address*, 54-60
 - reading raw text*, 50-52
 - selecting file type*, 51-52
 - selecting type of integration*, 65-70
 - usage examples*, 61-65
- SQL queries in, 185-187
 - basic word string snippet reports*, 192-194
 - indirect query example*, 189-191
 - proximity query example*, 191-192
 - simple query example*, 187-189
- in unstructured environment versus structured environment, 28-30
- visualizations, 127-128
 - analyzing*, 132
 - creating*, 128-131
 - recasting*, 132
 - for semistructured data*, 133
- textual concatenation, 161-162
- textual integration. *See* integration
- textual unstructured data. *See* unstructured data
- themes of data, 59
- transaction processing data, 86
 - unstructured data versus, 88
- transcribed telephone conversations, 23

U

- unique identification for spreadsheets, 138
- unstructured data. *See also* textual analytics; unstructured databases; unstructured environment
 - analyzing with structured data, 113-114
 - accessing linkages*, 118-122
 - intersection of data*, 115-118
 - real-time online queries*, 122-124
 - bringing into structured environment, 30-31
 - characteristics of, 5-8
 - in corporate functions, 3-4
 - in DW 2.0 (data warehouse architecture), 91-93
 - emails. *See* emails
 - explained, 2-3
 - first generation textual analytics. *See* first generation textual analytics
 - inclusion in information systems, need for, 87-91
 - in industries, 4-5
 - integration of, 41-42, 89-91
 - non-textual unstructured data, 2
 - opportunities of, 11-12
 - purpose of, 21
 - semistructured processing on, 81
 - spreadsheets. *See* spreadsheets
 - transaction processing data versus, 88
 - types of, 2-3
 - updating, lack of, 9
- unstructured databases, 95. *See also* unstructured data; unstructured environment
 - accessing through SQL interface, 108
 - for analysis, 108-111
 - contract management case study, 209-214
 - as data warehouses, 176
 - flow of data, 96-97
 - future capabilities of, 124-125
 - high-level design of, 109
 - hot data in, 111
 - indexes, types of, 99-100
 - integration with structured databases, 175-176
 - communications-based linkages*, 177
 - housing in DBMS*, 180-182
 - identifier-based linkages*, 178
 - as integrated data warehouse*, 179-180
 - simple queries*, 182-183
 - word linkages*, 176-177
 - large volumes of data in, 101-102
 - multiple tables, accessing, 109-110
 - relational tables in, 97-99
 - simple pointers in, 104-107
 - storing all unstructured data in, 102
 - storing selective unstructured data in, 103
- unstructured environment, 20-26. *See also* unstructured data; unstructured databases
 - communication, forms of
 - corporate contracts*, 25-26
 - emails*, 21-22
 - legal information*, 24-25
 - medical records*, 23-24
 - spreadsheets*, 22
 - transcribed telephone conversations*, 23
 - data quality in, 161-162
 - integration with structured environment, 49-50
 - external categorization*, 60-61
 - importance of*, 53-54
 - issues to address*, 54-60
 - reading raw text*, 50-52
 - selecting file type*, 51-52

selecting type of integration, 65-70
usage examples, 61-65
metadata in, 148
external metadata, 149
internal metadata, 148-149
textual analytics in, 28-30
unstructured systems, structured systems versus,
47-49
unstructured textual data. *See* unstructured data
updating structured/unstructured data, 9

V

VCR (voice character recognition), 51-52
visualizations, 43-44, 127-128
analyzing, 132
creating, 128
iterative development, 131
selecting homogeneous data, 128-129
as SOM (self-organizing map), 130-131
text clustering, 130
text integration, 129
recasting, 132
for semistructured data, 133
voice character recognition (VCR), 51-52
voice recordings, converting to electronic data,
51-52
volume of data, 10
in insurance claims case study, 220-225
in medical research case study, 197-198
in unstructured databases, 101-102

W-Z

weak links, 177
wild cards, 42-43, 124
word linkages between structured and
unstructured databases, 176-177
word permutations, 54
word stemming, 55, 164, 167
word-to-word proximity searches, 191
worksheets. *See* spreadsheets



THIS BOOK IS SAFARI ENABLED

INCLUDES FREE 45-DAY ACCESS TO THE ONLINE EDITION

The Safari® Enabled icon on the cover of your favorite technology book means the book is available through Safari Bookshelf. When you buy this book, you get free access to the online edition for 45 days.

Safari Bookshelf is an electronic reference library that lets you easily search thousands of technical books, find code samples, download chapters, and access technical information whenever and wherever you need it.

TO GAIN 45-DAY SAFARI ENABLED ACCESS TO THIS BOOK:

- Go to <http://www.prenhallprofessional.com/safarienabled>
- Complete the brief registration form
- Enter the coupon code found in the front of this book on the "Copyright" page

If you have difficulty registering on Safari Bookshelf or accessing the online edition, please e-mail customer-service@safaribooksonline.com.



PRENTICE
HALL