CONTENTS

	Foreword	xvii
	Preface	xix
PART I	XML TECHNOLOGIES	1
1	HTML AND WEB PAGES	3
	Objectives	3
1.1	Hypertext and Markup Languages	3
1.2	The History of HTML	4
1.3	URLs, URIs, URNs, and IRIs	7
1.4	Survivor's Guide to HTML	8
1.5	Survivor's Guide to CSS	16
1.6	Syntax and Validation	20
1.7	Limitations of HTML	23
1.8	Unicode	25
1.9	The World Wide Web Consortium (W3C)	28
1.10	Chapter Summary	29
1.11	Further Reading	29
1.12	Online Resources	29
1.13	Exercises	30
2	XML DOCUMENTS	32
	Objectives	32
2.1	Introduction	32
2.2	Recipes in XML	33
2.3	XML Trees	35
2.4	Textual Representation of XML Documents	40
2.5	Applications of XML	44
	2.5.1 XHTML	45

2.5.2	CML	46
2.5.3	WML	47
2.5.4	ebXML	48
2.5.5	ThML	49
2.6	XML Namespaces	49
2.7	Running Example: More Recipes	53
2.8	Chapter Summary	54
2.9	Further Reading	55
2.10	Online Resources	55
2.11	Exercises	56
3	NAVIGATING XML TREES WITH XPATH	58
	Objectives	58
3.1	Pointing into XML Documents	58
3.2	Location Steps and Paths	59
3.2.1	Contexts	61
3.2.2	Axes	62
3.2.3	Node Tests	64
3.2.4	Predicates	65
3.3	Typical Location Paths	66
3.4	Abbreviations	67
3.5	General Expressions	69
3.5.1	Values and Atomization	70
3.5.2	Literal Expressions	70
3.5.3	Comments	72
3.5.4	Variable References	72
3.5.5	Arithmetic Expressions	72
3.5.6	Sequence Expressions	73
3.5.7	Path Expressions	74
3.5.8	Filter Expressions	74
3.5.9	Comparison Expressions	75
3.5.10	Boolean Expressions	78
3.5.11	Functions	78
3.5.12	For Expressions	82
3.5.13	Conditional Expressions	83
3.5.14	Quantified Expressions	83
3.5.15	Types	84
3.5.16	XPath 1.0 Restrictions	84
3.6	XPointer and XLink ★	85
3.7	Chapter Summary	89
3.8	Further Reading	89
3.9	Online Resources	89
3.10	Exercises	90

4	SCHEMA LANGUAGES	92
	Objectives	92
4.1	XML Languages and Validation	92
4.2	Regular Expressions	94
4.3	DTD – Document Type Definition	96
4.3.1	Document Type Declarations	96
4.3.2	Element Declarations	98
4.3.3	Attribute-List Declarations	101
4.3.4	Conditional Sections, Entity, and Notation Declarations	106
4.3.5	Checking Validity with DTD	109
4.3.6	Recipe Collections with DTD	110
4.3.7	Limitations of DTD	112
4.4	XML Schema	113
4.4.1	Overview	114
4.4.2	Simple Types	119
4.4.3	Complex Types	124
4.4.4	Global versus Local Descriptions	134
4.4.5	Namespaces	139
4.4.6	Annotations	141
4.4.7	Modularization	142
4.4.8	Subsumption and Substitution Groups	145
4.4.9	Defaults and Whitespace Normalization	149
4.4.10	Uniqueness, Keys, and References	150
4.4.11	Recipe Collections with XML Schema	153
4.4.12	Limitations of XML Schema	156
4.4.13	Best Practices	158
4.4.14	Other Schema Languages	159
4.5	DSD2 ★	159
4.5.1	Recipe Collections with DSD2	160
4.5.2	Rules	163
4.5.3	Boolean Expressions	165
4.5.4	Regular Expressions	166
4.5.5	Normalization	166
4.5.6	Modularization	167
4.5.7	Uniqueness and Pointers	168
4.6	RELAX NG ★	168
4.6.1	Patterns and Grammars	169
4.6.2	Datatypes	174
4.6.3	Recipe Collections with RELAX NG	177
4.6.4	Modularization	180
4.6.5	A Non-XML Syntax	181
4.7	Chapter Summary	183
4.8	Further Reading	184

4.9	Online Resources	185
4.10	Exercises	185
5	TRANSFORMING XML DOCUMENTS WITH XSLT	188
	Objectives	188
5.1	Transforming XML Documents	188
5.2	The Processing Model	193
5.3	Template Rules	194
5.3.1	Patterns and Matching	194
5.3.2	Names, Modes, and Priorities	196
5.4	Sequence Constructors	197
5.4.1	Element and Attribute Constructors	197
5.4.2	Text Constructors	200
5.4.3	Other Constructors	201
5.4.4	Copying Nodes	201
5.4.5	Recursive Application	202
5.4.6	Repetitions	205
5.4.7	Conditionals	206
5.4.8	Template Invocation	207
5.5	Variables and Parameters	208
5.6	Built-In Template Rules	209
5.7	Simple Traversals	210
5.8	Advanced Features	211
5.8.1	Grouping	211
5.8.2	Sorting	212
5.8.3	Numbering	214
5.8.4	Keys	215
5.8.5	Analyzing Strings	217
5.8.6	Functions	218
5.8.7	Sequence Types	218
5.8.8	Multiple Documents	219
5.8.9	Include and Import	221
5.8.10	Generating XSLT Output	224
5.9	XSLT 1.0 Restrictions	227
5.10	Stylesheets for Recipe Collections	229
5.11	XSL-FO	234
5.12	Chapter Summary	237
5.13	Further Reading	237
5.14	Online Resources	237
5.15	Exercises	237
6	QUERYING XML DOCUMENTS WITH XQUERY	240
	Objectives	240
6.1	Querying XML Documents	240

6.1.1	From Relations to Trees	241
6.1.2	Usage Scenarios	245
6.2	The XQuery Design	245
6.2.1	Relationship to XPath	246
6.2.2	Relationship to XSLT	247
6.3	The Prolog	247
6.4	Expressions	249
6.4.1	XPath Expressions	249
6.4.2	Datatype Expressions	249
6.4.3	XML Expressions	250
6.4.4	FLWOR Expressions	254
6.5	Defining Functions	259
6.6	XQuery versus XSLT ★	261
6.6.1	Emulating XSLT in XQuery	262
6.6.2	Emulating XQuery in XSLT	264
6.7	The Type System ★	266
6.7.1	Sequence Types	267
6.7.2	Validation and Dynamic Types	268
6.7.3	Type Matching	268
6.7.4	Type Annotations and Type Errors	271
6.7.5	Static Type Checking	272
6.8	XQueryX	275
6.9	XML Databases	277
6.9.1	XML Publishing	277
6.9.2	XML Shredding	279
6.10	Full-Text Searching	281
6.11	Chapter Summary	282
6.12	Further Reading	282
6.13	Online Resources	283
6.14	Exercises	283
7	XML PROGRAMMING	285
	Objectives	285
7.1	Programming with XML Documents	285
7.2	The DOM API	286
7.3	The JDOM API	289
7.3.1	Embracing Java	289
7.3.2	Data Model	290
7.3.3	Parsing, Validation, and Serializing	293
7.3.4	XPath Evaluation	297
7.3.5	XSLT Transformation	297
7.3.6	A Business Card Editor	299
7.4	XML Data Binding	304
7.4.1	Binding Compilers	305

7.4.2	The JAXB Framework	305
7.4.3	A Business Card Editor using JAXB	305
7.5	The SAX API	310
7.5.1	Streaming XML	310
7.5.2	Parsing Events	311
7.5.3	SAX Applications	313
7.5.4	SAX Filters	317
7.5.5	Streaming with XmlPull	319
7.6	Streaming Transformations with STX ★	320
7.6.1	STXPath	321
7.6.2	Transformations and Templates	321
7.6.3	Variables	325
7.6.4	Groups	326
7.6.5	Limitations of Streaming	327
7.6.6	STX for Recipes	328
7.7	Type-Safe XML Programming Languages ★	331
7.7.1	XDuce	331
7.7.2	XACT	334
7.8	Chapter Summary	338
7.9	Further Reading	338
7.10	Online Resources	339
7.11	Exercises	339
PART II	WEB TECHNOLOGIES	341
8	THE HTTP PROTOCOL	343
	Objectives	343
8.1	The Internet and HTTP	343
8.1.1	Requests	346
8.1.2	Responses	350
8.1.3	HTML Forms	353
8.1.4	Authentication	356
8.1.5	Other Advanced Features	358
8.1.6	Limitations of HTTP	361
8.2	Sessions	362
8.3	SSL and TLS	366
8.4	Web Programming with Java	368
8.4.1	TCP/IP in Java	369
8.4.2	HTTP in Java	374
8.4.3	SSL in Java (JSSE)	378
8.5	A Web Server in 145 Lines of Code	383
8.6	Chapter Summary	387
8.7	Further Reading	387

8.8 Online Resources 387
8.9 Exercises 388

9 PROGRAMMING WEB APPLICATIONS WITH SERVLETS 390

Objectives 390
9.1 Writing Web Applications 390
9.2 The Servlet API 391
9.2.1 The Life Cycle of Servlets 392
9.2.2 Requests 394
9.2.3 Responses 396
9.2.4 Servlet Contexts and Shared State 398
9.2.5 Sessions 403
9.3 Running Web Applications 406
9.3.1 Web Applications and Deployment 407
9.3.2 Running Servlets with Tomcat 410
9.4 Advanced Features 411
9.4.1 Listeners 411
9.4.2 Filters 413
9.4.3 Request Dispatchers 418
9.4.4 Security 419
9.5 Limitations of Servlets 422
9.6 Web Applications with JWIG ★ 423
9.7 Chapter Summary 427
9.8 Further Reading 427
9.9 Online Resources 427
9.10 Exercises 427

10 PROGRAMMING WEB APPLICATIONS WITH JSP 429

Objectives 429
10.1 The JSP Framework 429
10.1.1 Templates 430
10.1.2 Expressions 431
10.1.3 Statements 432
10.1.4 Declarations 432
10.1.5 Directives 433
10.1.6 Translation into Servlets 435
10.1.7 XML Version of JSP 438
10.1.8 The Expression Language 440
10.1.9 A JSP Shopping Cart 441
10.2 Tags 442
10.2.1 Tag Files 442
10.2.2 Tags for Quick Polls 444
10.2.3 Tag Libraries and JSTL 449

	10.2.4 Limitations of Tags	453
10.3	The Model–View–Controller Pattern	455
	10.3.1 A Model 1 Business Card Server	455
	10.3.2 A Model 2 Business Card Server	458
10.4	Chapter Summary	463
10.5	Further Reading	464
10.6	Online Resources	464
10.7	Exercises	464
11	WEB SERVICES	466
	Objectives	466
11.1	Distributed Systems and Web Services	466
	11.1.1 A Recipe Server with XML and HTTP	468
11.2	Web Service Standards	469
11.3	SOAP	471
	11.3.1 The Processing Model	471
	11.3.2 Faults	474
	11.3.3 Data Representation and RPC	476
	11.3.4 Protocol Binding	479
11.4	WSDL	480
	11.4.1 Interface Descriptions	485
	11.4.2 Binding Descriptions	486
	11.4.3 Service Descriptions	489
11.5	UDDI	489
	11.5.1 Descriptions	490
	11.5.2 Discovery	492
11.6	Chapter Summary	493
11.7	Further Reading	494
11.8	Online Resources	494
11.9	Exercises	495
12	A COMPLETE APPLICATION	496
	Objectives	496
12.1	The Web of Jokes	496
12.2	The Joke Language	497
	12.2.1 Representing Jokes	498
	12.2.2 XML Schema for Jokes	498
	12.2.3 XSLT for Jokes	500
12.3	The Joke Server	502
	12.3.1 State	503
	12.3.2 Operations	503
	12.3.3 Implementation	503
12.4	The Joke Metaserver	508
	12.4.1 State	508

12.4.2 Operations 509
12.4.3 Implementation 509
12.5 The Joke Client 512
12.5.1 State 513
12.5.2 Operations 513
12.5.3 Implementation 513
12.6 Deployment 522
12.7 Robustness 526
12.8 Reflections 527
12.9 Exercises 528

Bibliography 529
Index 535

FOREWORD

The Web has revolutionized the way the world accesses and shares information. Early on, the Web was a primarily passive, unidirectional platform for serving static HTML pages on demand to people. In no time, business users discovered that the Web not only enabled them to present unified views of products and services to customers around the clock, but that with some programming effort, the Web could support transactional customer-to-business interactions. Today, major corporations world wide increasingly depend on the Web for deploying widely distributed business-to-business processes both within and across the corporate boundary. These applications typically provide services that integrate data from Web and non-Web sources and coordinate interactions with existing business processes – a long way from one-page-at-a-time HTML programming.

The Web's core technologies – HTML and HTTP – were strained to the limit by this rapid evolution. In response, an alphabet soup of new technologies emerged to do what HTML and HTTP could not. XML, XSLT, DTD, JSP, SOAP, WSDL, and UDDI are just a few of the myriad technologies that promise to make the Web application developer's job easier. Developers, however, sit uncomfortably between the promise and the reality of Web technologies. The XML and Web technologies at their disposal are immature. Many are just emerging from standardization, have few robust or interoperable implementations, and lack the application development environments available for more mature technologies.

In 'An Introduction to XML and Web Technologies', Anders Møller and Michael Schwartzbach expertly decipher the alphabet soup of Web technologies. They lucidly describe the technical features of each technology, compare closely related technologies, and show how to apply them in various application scenarios. In Part I, they describe XML's genesis from SGML to become the standard data format on the Web, compare various XML schema languages including DTD and XML Schema, contrast the XML query languages XPath, XSLT, and XQuery, and most importantly, explain how XML technologies can co-exist with general purpose programming languages. In Part II, they describe technologies for building stateful Web servers – Servlets and JSP – as well as the standards for defining, publishing, and connecting to Web services.

Anders and Michael hold their readers in high regard. They write in a clear and plain-spoken voice that disguises their deep understanding of their subject. Each carefully crafted example teaches several concepts at once, and when put together, the examples yield

non-trivial (and entertaining!) results. Anders and Michael also describe the cutting-edge research that will effect the way we program for the Web in the future, which is invaluable to the reader who wants to develop a deep and lasting understanding of these technologies.

This book is a genuine pleasure to read – I learned a lot, and I learned it fast! I hope that Anders and Michael continue to track the Web's rapid changes and provide us with more valuable information in future editions.

Mary Fernández
AT&T Labs Research

PREFACE

XML and Web Technologies

In the early 1990s, the World Wide Web was defined by a triumvirate consisting of the HTML language for writing hypertext documents, the HTTP communication protocol, and the URL notation for addressing resources. Today, new Web technologies are being developed and deployed at amazing rates, building on top of the early foundations. This book offers a comprehensive introduction to the area.

There are two main threads of development, corresponding to the two parts of this book. *XML technologies* generalize the notion of data on the Web from hypertext documents to arbitrary data, including those that have traditionally been the realm of databases. In this book we cover the basic XML technology and the supporting technologies of XPath, DTD, XML Schema, DSD2, RELAX NG, XSLT, XQuery, DOM, JDOM, JAXB, SAX, STX, XDuce, and XACT. *Web technologies* build on top of the HTTP protocol to provide richer languages for constructing applications and services. In this book we cover the basic HTTP protocol and the increasingly abstract technologies of Servlets, JSP, JWIG, WSDL, SOAP, and UDDI.

These are, for better or worse, core technologies that will exist for many years or provide the foundation for future developments.

Aims of This Book

The topics covered by this book are, of course, all richly described in free online standards documents, totaling several thousand pages, but those are wholly unsuited for a self-contained course. This book is unique in providing a coherent overview of the most important XML and Web technologies. It goes into great detail but still aims for conciseness, thereby enabling the reader to see the big picture and yet obtain practical experience with the technologies and supporting tools.

The book also contains critical analyses and discussions of the technologies, in contrast to standards documents and technical manuals, which mainly present and exemplify features. It also provides a uniform terminology that is familiar to readers with a standard computer science background.

The book describes the newest technologies, including XML 1.1, XPath 2.0, XSLT 2.0, XQuery 1.0, JDOM 1.0, SAX2, JAXB 1.0, STX 1.0, HTTP/1.1, Java Servlet 2.4, JSP 2.0, JSTL 1.1, SOAP 1.2, WSDL 2.0, and UDDI 3.0. It focuses on concepts and technologies, rather than on vendor-specific tools. Moreover, the book presents selected research projects – DSD2, RELAX NG, STX, XDuce, XACT, and JWIG – that may influence future technologies.

How to Use This Book

The intended audience of this book includes computer science students, computer professionals, and researchers that want an overview of the area. Preliminary versions of this book have been used several times for undergraduate courses at the University of Aarhus and at the IT University of Copenhagen, and have been the basis for numerous industrial courses. The book can be the complete curriculum for an XML/Web course, or it may be used as a supplement for database or programming courses.

Each chapter contains carefully selected links to the essential online resources, references to further reading, and exercises that help the readers test their understanding and gain familiarity and practical experience with the most important technical specifications and tools.

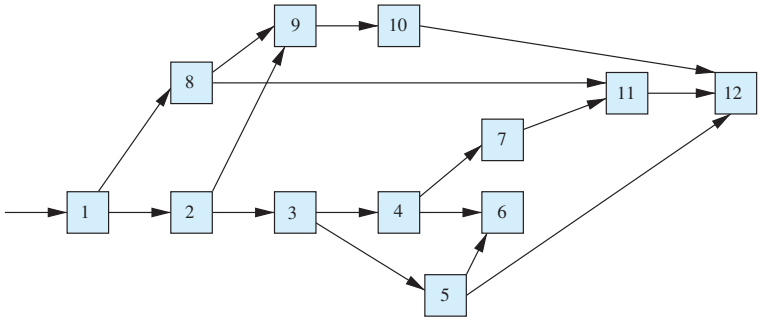
Chapter 1 contains survivor's guides to HTML and CSS, and motivates the need for XML. It also covers technical issues related to Web publication, such as URLs, URNs, URIs, and Unicode. Moreover, this chapter briefly introduces the workings of the World Wide Web Consortium, which develops many of the technologies covered by the book. In Chapter 2, we explain the XML notation, its dual nature as a textual format and a tree structure, and the namespace mechanism. Chapter 3 covers the XPath language, which is used in several other languages for pointing into XML documents and much more. Chapter 4 describes the use of schema languages. It explains the languages DTD and XML Schema, and also two less widely known alternatives, DSD2 and RELAX NG, and compares the languages. Additionally, this chapter contains an introduction to the notion of regular expressions, which is commonly used in schema languages. In Chapters 5 and 6, we show how the XSLT and XQuery languages can be used to define transformations between XML languages, and we provide a thorough comparison between the two. XSLT is mainly targeted at stylesheet transformations for presenting XML data, whereas XQuery is designed primarily for database-like queries. Chapter 7 shows how to work with XML in programming languages using DOM, JDOM, JAXB, SAX, and STX. As a running example, we develop an XML-based language for recipe collections and show how the various technologies become useful.

Chapter 8 contains an introduction to the HTTP protocol and shows how to program HTTP servers and clients. Chapter 9 describes the Servlet platform, which is a convenient API that builds upon HTTP. Closely related technologies are JSP and JSTL, which are the topics of Chapter 10. Chapter 11 explains the basic technologies related to Web services that

communicate XML data on the Web, in particular the WSDL, SOAP, and UDDI initiatives. Finally, Chapter 12 collects the essential knowledge from the other chapters and applies it to a larger project: development of an interactive Web service, *The Web of Jokes*, for sharing jokes on the Web.

All chapters contain numerous 'gold nuggets' showing concrete examples of how typical tasks can be solved with the technologies being described.

The chapter dependencies may be illustrated by the following diagram:



Chapters 7, 8, 9, 10, 11, and 12 assume a basic knowledge of the Java programming language. This language is natural to use when illustrating Web technologies because of its platform independence and native support for Unicode and migration of data and code, and furthermore, a majority of the freely available tools that exist for XML and Web development are based on Java. Except from this prerequisite, the book is self-contained.

Some sections of this book are intended for advanced studies and may safely be skipped by readers who only need to obtain basic familiarity with the presented technologies. These sections, marked with the symbol ★, contain further discussions and presentations of advanced features and research topics.

The Companion Web Site

The book has a companion Web site located at

```
http://www.brics.dk/ixwt/
```

This site contains additional material for teachers and students: an extensive collection of slides, examples from the book, additional projects, and multiple choice tests for each topic. Qualified instructors may obtain solutions to the exercises and the PowerPoint sources for the slides. Moreover, many exercises in the book refer to online data. We use *EX* as an abbreviation of the URL `http://www.brics.dk/ixwt/exercises/`.

A preliminary version of the online material has been continually updated since March 2000 and has been visited by more than 300,000 people (counted as unique IP numbers).

Acknowledgements

Thanks to Claus Brabrand, Mary Fernández, Thomas Hildebrandt, Martin Mosegaard Jensen, Lars Michael Kristensen, Henning Niss, Peter Thiemann, Phil Wadler, and the anonymous reviewers for detailed comments to the contents of this book. The participants of several XML and Web courses at the University of Aarhus and the IT University of Copenhagen have also contributed to structuring and proofreading the presented material. Thanks to the people from Pearson Education: to Sami Taalas and Kate Brewin for persuading us to enter into the project, and to Owen Knight, Mary Lince, Karen McLaren Simon Plumtree, and Simon Turner for guiding us through the process.

Anders Møller
Michael I. Schwartzbach

Publisher's Acknowledgements

Screen shots on pages 14 and 343 used with permission from Google Inc. Google™ is a trademark of Google Inc.