

# Lecture Notes in Computer Science

1293

Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Advisory Board: W. Brauer D. Gries J. Stoer

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Barcelona*

*Budapest*

*Hong Kong*

*London*

*Milan*

*Paris*

*Santa Clara*

*Singapore*

*Tokyo*

Charles Nicholas Derick Wood (Eds.)

# Principles of Document Processing

Third International Workshop, PODP'96  
Palo Alto, California, USA, September 23, 1996  
Proceedings



Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany

Juris Hartmanis, Cornell University, NY, USA

Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Charles Nicholas

University of Maryland Baltimore County

Department of Computer Science and Electrical Engineering

1000 Hilltop Circle, Baltimore, MD 21250, USA

E-mail: nicholas@cs.umbc.edu

Derick Wood

The Hong Kong University of Science and Technology

Department of Computer Science

Clear Water Bay, Kowloon, Hong Kong

E-mail: dwood@cs.ust.hk

Cataloging-in-Publication data applied for

**Die Deutsche Bibliothek - CIP-Einheitsaufnahme**

**Principles of document processing : third international workshop ; proceedings / PODP '96, Palo Alto, California, USA, September 23, 1996 / Charles Nicholas ; Derick Wood (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Budapest ; Hong Kong ; London ; Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1997**

**(Lecture notes in computer science ; Vol. 1293)**

**ISBN 3-540-63620-X**

CR Subject Classification (1991): I.7, H.5, I.3.7, I.4

ISSN 0302-9743

ISBN 3-540-63620-X Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1997

Printed in Germany

Typesetting: Camera-ready by author

SPIN 10546383 06/3142 - 5 4 3 2 1 0 Printed on acid-free paper

# Preface

The Third International Workshop on Principles of Document Processing took place in Palo Alto, California, on September 23, 1996. PODP'96 was the third in a series of international workshops that provide forums to discuss the modeling of document processing systems using theories and techniques from, for example, computer science, mathematics, and psychology. PODP'96 took place in conjunction with EP'96 at Xerox Corporation's conference center in Palo Alto, California.

The charter of the PODP workshops is deliberately ambitious and its scope broad. The current state of electronic document processing can be characterized as a plethora of tools without a clear articulation of unifying principles and concepts underlying them. The practical and commercial impact of these tools, which include formatters, composition systems, word processing systems, structured editors, and document management systems, among others, is too pervasive and obvious to require further elaboration and emphasis. However, with the rapid development in hardware technology (processors, memory, and especially high bandwidth networks) the notion of a document and of document processing itself is undergoing a profound change. It is imperative that this change be fueled, not only by enabling technologies and tools, but also by precise, computational, and conceptual models of documents and document processing. To this end, we hope to bring to bear theories and techniques developed by researchers in other areas of science, mathematics, engineering, and the humanities (such as databases, formal specification languages and methodologies, optimization, work-flow analysis, and user interface design.)

The PODP workshops are intended to promote a happy marriage between documents and document processing, and theories and techniques. PODP provides an ideal opportunity for discussion and information exchange between researchers who are grappling with problems in *any* area of document processing.

We invited researchers to submit papers with a good balance between theory and practice in document processing. Papers that address both on a somewhat equal basis were preferred. Each paper was subjected to rigorous peer review. We extend our thanks to the other members of this Workshop's Program Committee, for their hard work under tight time pressure: Howard Blair (USA), Heather Brown (UK), Anne Brueggemann-Klein (Germany), Richard Furuta (USA), Heikki Mannila (Finland), Ethan Munson (USA), Makoto Murata (Japan), and James Sasaki (USA).

There was considerable discussion about the workshop name during and after PODP'96. As a result of this discussion, and to more accurately reflect the workshop's focus on the processing of documents in digital form, it was decided to change the workshop's name to Principles of Digital Docu-

ment Processing. The next workshop in this series, PODDP'98, is planned for April 1998, in Saint Malo, France, in conjunction with EP'98, the seventh International Conference on Electronic Publishing, Document Manipulation and Typography.

July 1997

Charles Nicholas

Derick Wood

# Table of Contents

<b>Toward an Operational Theory of Media</b>	
Ethan V. Munson .....	1
1. Introduction .....	1
2. The Data Type Model .....	2
3. An Operational Model of Media .....	3
3.1 Primitive Data Types .....	4
3.2 Dimensions .....	4
3.3 Formatting Operations .....	4
4. Applications of the Model .....	5
4.1 Comparing Media .....	5
4.2 Configuring a Presentation Specification System .....	6
5. Issues for Further Study .....	8
6. Related Work .....	10
6.1 The AHV Model .....	11
7. Conclusion .....	12
 <b>First Steps to Cross Media Publishing and Multimodal Documents</b>	
Kurt Sandkuhl .....	15
1. Introduction .....	15
2. Cross Media Publishing .....	16
3. COMPACT Approach to Cross Media Publishing .....	19
3.1 Information Processing Viewpoint .....	19
3.2 Workflow Viewpoint .....	22
4. Examples, Experience and Conclusion .....	23
 <b>Disambiguation of SGML Content Models</b>	
Helena Ahonen .....	27
1. Introduction .....	27
2. 1-unambiguity .....	29
3. Disambiguation .....	32
4. Conversion into a Content Model .....	35
5. Experimental Results .....	36

6. Conclusions .....	37
<b>SGML and Exceptions</b>	
Pekka Kilpeläinen and Derick Wood .....	39
1. Introduction .....	39
2. Extended Context-Free Grammars with Exceptions .....	41
3. Exception-Removal for DTDs .....	45
4. Concluding Remarks and Open Problems .....	47
<b>Grammar-Compatible Stylesheets</b>	
Thomas Schroff and Anne Brüggemann-Klein .....	51
1. Introduction .....	51
2. Documents and Grammars .....	52
3. Transformations and Stylesheets .....	53
3.1 Grammar Compatibility .....	54
3.2 Constructing a Partially-Bracketed Grammar from $G_{\text{source}}$ ....	54
3.3 Constructing a Nearly-Bracketed Grammar from $G_1$ .....	55
3.4 Constructing a Bracketed Grammar from $G_2$ .....	55
3.5 Constructing a Homogeneously-Bracketed Grammar from $G_2$ ..	56
4. Results .....	56
5. Conclusions .....	57
<b>Object Awareness in Multimedia Documents</b>	
M.A. Heather and B.N. Rossiter .....	59
1. Introduction .....	59
2. Background to Multimedia Document Research .....	60
3. Types of Awareness .....	64
4. Connections in Multimedia .....	66
5. Formal Modelling under Geometric Logic .....	68
5.1 Adjointness .....	71
5.2 Intension-Extension Mapping .....	73
5.3 Geometric Database Models .....	74
6. Formal Contextual Sensitivity .....	74
6.1 Limits, Colimits and Context .....	74
7. The Hypertext Lattice as a Heyting Algebra .....	78
8. Geometric Consciousness .....	79
8.1 Contextual Awareness in Hypertext .....	80
8.2 Computational Model of Consciousness .....	81
8.3 Relative and Dynamic Contexts .....	82
9. Conclusions .....	84



## **A Logic Based Formalism for Temporal Constraints in Multimedia Documents**

Peter R. King .....	87
1. Introduction .....	87
2. A Taxonomy of Constraints .....	88
2.1 Introduction .....	88
2.2 The Constraint Taxonomy .....	90
3. Interval Temporal Logic .....	91
3.1 Introduction .....	91
3.2 ITL – Atomic Operators .....	91
3.3 ITL – Examples .....	92
4. ITL as a Formalism for Multimedia Documents .....	93
4.1 Atomic Constraints – Clock Time and Duration .....	93
4.2 Relative Constraints – Two Media Items .....	94
4.3 Relations Involving More Than Two Media Items .....	95
5. Projection and Related Constructs .....	96
5.1 Multiplication .....	96
5.2 Imprecision and Adjustment .....	97
5.3 User Interaction .....	97
6. Executability and Display Forms .....	98
7. Discussion .....	99
Appendix .....	101

## **Towards Automatic Hypertextual Representation of Linear Texts**

A. Myka, H. Argenton, and U. G�ntzer .....	103
1. Introduction .....	103
2. Extraction of Nonlinear Structures .....	104
3. Statistical Evaluations .....	108
4. Natural-Language Parsing .....	110
5. High-Level Modeling .....	112
6. Monitoring User Actions .....	113
7. Link Inheritance and Accumulation .....	114
8. Comparison of Methods .....	115
9. Conclusion .....	118

## **Using Background Contextual Knowledge for Documents Representation**

Arkadi Kosmynin and Ian Davidson .....	123
1. Introduction and Motivation .....	123

2. Contextual Document Representation .....	125
2.1 Spreading Activation .....	125
2.2 Our Method .....	126
2.3 Example .....	128
3. Applications to Document Classification .....	129
4. Discussion and Future Work .....	131
5. Conclusion .....	132

## **Typed Structured Documents for Information Retrieval**

Chanda Dharap and C. Mic Bowman .....	135
1. Introduction .....	135
2. Motivation .....	136
3. Related Work .....	137
4. Model .....	138
4.1 Type .....	139
4.2 Tags .....	139
4.3 Templates .....	141
4.4 Objects .....	142
5. Tools to Implement Structured Types .....	143
6. Advantages .....	144
6.1 Multiple Inheritance .....	144
7. Precision Experiments .....	145
7.1 Measures of Usability .....	146
8. Conclusions .....	150

## **Transformation of Documents and Schemas by Patterns and Contextual Conditions**

Makoto Murata .....	153
1. Introduction .....	153
2. Transformations of Strings .....	156
2.1 Preliminaries .....	156
2.2 Transformation Rules .....	157
2.3 Applying Transformation Rules to Strings .....	157
2.4 Schema Transformation .....	158
3. Transformations of Binary Trees .....	160
3.1 Preliminaries .....	160
3.2 Transformation Rules .....	162
3.3 Applying Transformation Rules to Trees .....	163
3.4 Schema Transformation .....	166

## **Tabular Formatting Problems**

Xinxin Wang and Derick Wood .....	171
1. Introduction .....	171
2. Complexity Analysis .....	173
3. Definition of TFALN .....	174
4. A Formatting Algorithm .....	176
4.1 An Exponential-Time Algorithm .....	177
4.2 A Polynomial-Time Greedy Algorithm .....	178
4.3 An Efficient Algorithm .....	179
5. Conclusions .....	180

## **Visual Definition of Virtual Documents for the World-Wide Web**

Mark Minas and Leon Shklar .....	183
1. Introduction .....	183
2. Building Information Repositories .....	184
2.1 The Object Model .....	184
2.2 Method Sharing .....	185
3. The Visual Repository Definition Language .....	188
4. Example .....	193
5. Related Work .....	194
6. Conclusions .....	194