The HyperView Approach to the Integration of Semistructured Data

Lukas C. Faulstich[1]

Dissertation
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Eingereicht am:
16. November 1999
Verteidigt am:
15. Februar 2000

Gutachter:
Prof. Dr. Heinz Schweppe
Prof. Dr. Herbert Weber (TU Berlin)
Prof. Dr. Hartmut Ehrig (TU Berlin)

Betreuer:
Prof. Dr. Heinz Schweppe
Prof. Dr. Herbert Weber (TU Berlin)
Prof. Dr. Hartmut Ehrig (TU Berlin)
Dr. Ralf-Detlef Kutsche (TU Berlin)
Dr. Gabriele Taentzer (TU Berlin)

[1] Supported by the German Research Society, Berlin-Brandenburg Graduate School on Distributed Information Systems (DFG grant no. GRK 316)

to *Myra*

Abstract

In order to use the World Wide Web to answer a specific question, one often has to collect and combine information from multiple Web sites. This task is aggravated by the structural and semantic heterogeneity of the Web. *Virtual Web sites* are a promising approach to solve this problem for particular, focused application domains.

A virtual Web site is a Web site that serves pages containing concentrated information that has been *extracted*, *homogenized*, and *combined* from several underlying Web sites. The goal is to save the user from tediously searching and browsing multiple pages at all these sites.

The HyperView approach to the integration of semistructured data sources presented in this thesis provides a methodology, a formal framework, and a software environment for building such virtual Web sites.

To achieve this kind of integration, data has to be *extracted* from external Web documents, *integrated* into a common representation, and then *presented* to the user in form of Web documents.

The HyperView approach treats these three steps of extraction, integration, and presentation uniformly as consecutive views that map between different levels of abstraction. Each of these levels is modeled by schemata and corresponds to an architectural layer of the HyperView System. The HyperView methodology provides a guideline for modeling each of these layers and defining views between them in order to establish a virtual Web site.

In HyperView, the contents of Web sites as well as the consecutive results of the views are represented as graphs. A special graph-based data model (CGDM) has been developed to this purpose. The view mechanism of HyperView supports mappings between graphs. Views are defined by sets of graph transformation rules. Since it is in general not feasible to materialize views over Web sites in advance, a demand-driven rule activation mechanism has been formally described and implemented in the HyperView System. This mechanism incrementally materializes views in response to queries issued against them.

The HyperView System has been implemented in Prolog. Graph transformation rules are compiled into Prolog predicates that can be executed efficiently. Web documents are loaded into the HyperView System using a standard HTTP client. HyperView based virtual Web sites are supported using the Java servlet technology.

The case studies in the fields of Digital Libraries and of Town Information Systems included in this thesis demonstrate the applicability of HyperView for integrating semistructured information sources and for building virtual Web sites. An explorative study shows how the HyperView approach can be applied in the context of the emerging standards related to XML.

The main contributions of this thesis are:

1. the key idea of applying the same view mechanism uniformly to solve the problems of extraction, integration, and presentation,
2. the HyperView *methodology* for modeling and integrating Web sites,
3. the *formal framework* defining the data model, rule concept, and in particular the *demand-driven view materialization* mechanism of HyperView,
4. the HyperView System *prototype* providing a platform for building virtual integrated Web sites
5. the *validation* of the HyperView methodology and system in the mentioned case studies.

In short, this thesis covers the whole problem of building virtual Web sites including methodology, formal foundation, and software support.

Contents

Bibliography

[Abiteboul *et al.*, 1993] Serge Abiteboul, Sophie Cluet, and Tova Milo. Querying and updating the file. In *19th Intl. Conference on VLDB*, volume 19, pages 73–85, 8 1993.

[Abiteboul *et al.*, 1997] Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, and Janet L. Wiener. The lorel query language for semistructured data. *Journal of Digital Libraries*, 1(1), 1997.

[Abiteboul *et al.*, 1998] Serge Abiteboul, Sophie Cluet, and Tova Milo. A logical view of structured files. *VLDB Journal*, 7(2):96–114, 1998.

[Abiteboul, 1997] Serge Abiteboul. Querying semi-structured data. In *ICDT'97*, pages 1–18, 1997.

[Adelberg, 1998] Brad Adelberg. Nodose: A tool for semi-automatically extracting semistructured data from text documents. In *SIGMOD Conference 1998*, 1998.

[Arens *et al.*, 1993] Y. Arens, C. Y. Chee, C.-N. Hsu, and C. A. Knoblock. Retrieving and integrating data from multiple information sources. In *International Journal of Intelligent and Cooperative Information Systems, Vol. 2 No. 2*, June 1993.

[Arnaud Sahuguet and Fabien Azavant, 1999] Arnaud Sahuguet and Fabien Azavant. Wysiwyg Web Wrapper Factory (W4F). Available from the W4F WebSite http://db.cis.upenn.edu/W4F, 1999.

[Arocena and Mendelzon, 1998] G. Arocena and A. Mendelzon. WebOQL: Restructuring documents, databases and webs. In *Proc. of 14th. Intl. Conf. on Data Engineering (ICDE 98)*, 1998.

[Ashish and Knoblock, 1997] N. Ashish and C. Knoblock. Wrapper generation for semistructured internet sources. In *Proc. Workshop on Management of Semistructured Data*, Tucson, 1997.

[Atzeni and Mecca, 1997] Paolo Atzeni and Giansalvatore Mecca. Cut & paste. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 12–15, Tucson, Arizona, 1997.

[Batini *et al.*, 1986] C. Batini, M. Lenzerini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4):323–364, 1986.

[Bergholz and Freytag, 1999] A. Bergholz and J. C. Freytag. Querying semistructured data based on schema matching. In *Intl. Workshop on Database Programming Languages (DBLP)*, 1999.

[Buneman *et al.*, 1996] Peter Buneman, Susan Davidson, Gerd Hillebrand, and Dan Suciu. A query language and optimization techniques for unstructured data. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 505–516, 1996.

[Buneman *et al.*, 1997] Peter Buneman, Susan B. Davidson, Mary F. Fernandez, and Dan Suciu. Adding structure to unstructured data. In Foto N. Afrati and Phokion Kolaitis, editors, *Database Theory—ICDT'97, 6th International Conference*, volume 1186 of *Lecture Notes in Computer Science*, pages 336–350, Delphi, Greece, 8–10 January 1997. Springer.

[Buneman, 1997] Peter Buneman. Semistructured data. In *PODS'97*, 1997. Invited Tutorial.

[Chaudhuri *et al.*, 1995] Surajit Chaudhuri, Ravi Krishnamurthy, Spyros Potamianos, and Kyuseak Shim. Optimizing queries with materialized views. In *11th Int. Conference on Data Engineering*, pages 190–200, Los Alamitos, CA, 1995. IEEE Computer Soc. Press.

[Claßen and Löwe, 1995] I. Claßen and M. Löwe. Scheme evolution in object–oriented models. In *ICSE–17 Workshop on Formal Methods Application in Software Engineering Practice*, 1995.

[Cluet *et al.*, 1998] Sophie Cluet, Claude Delobel, Jérôme Siméon, and Katarzyna Smaga. Your mediators need data conversion! In *SIGMOD Conference 1998*, pages 177–188, 1998.

[Crescenzi and Mecca, 1998] V. Crescenzi and G. Mecca. Grammars have exceptions. *Information Systems*, 23(8):539–565, 1998.

[Dar *et al.*, 1996] Shaul Dar, Michael J. Franklin, Björn Thór Jónsson, Divesh Srivastava, and Michael Tan. Semantic data caching and replacement. In *VLDB'96*, pages 330–341, 1996.

[Davis and Weyuker, 1983] Martin D. Davis and Elaine J. Weyuker. *Computability, Complexity, and Languages*. Academic Press, 1983.

[Ehrig and Taentzer, 1996] H. Ehrig and G. Taentzer. Computing by graph transformation. a survey and annotated bibliography. Technical Report 96-15, TU Berlin, 1996.

[Ehrig *et al.*, 1991] H. Ehrig, M. Korff, and M. Löwe. Tutorial introduction to the algebraic approach of graph grammars based on double and single pushouts. *Lecture Notes in Computer Science*, 532:24–??, 1991.

[Faulstich and Spiliopoulou, 1998] Lukas C. Faulstich and Myra Spiliopoulou. Building HyperNavigation wrappers for publisher web-sites. In *Second European Conference on Digital Libraries*, number 1513 in LNCS, pages 115–134, 1998.

[Faulstich and Spiliopoulou, 1999] Lukas C. Faulstich and Myra Spiliopoulou. Building HyperNavigation wrappers for publisher web-sites. *International Journal on Digital Libraries*, 1999. Extended version of [Faulstich and Spiliopoulou, 1998], to appear.

[Faulstich *et al.*, 1997] Lukas C. Faulstich, Myra Spiliopoulou, and Volker Linnemann. WIND: A warehouse for internet data. In *Advances in Databases – Proceedings BNCOD 15*, number 1271 in LNCS, pages 169–183. Springer, 1997.

[Faulstich, 1998] Lukas C. Faulstich. Using graph transformation techniques for integrating information from the WWW. In *Theory and Application of Graph Transformations (TAGT'98)*, pages –, 1998. To appear.

[Faulstich, 1999a] Lukas C. Faulstich. Integrating town information servers. In *FDBS-99*, 1999. Accepted for publication.

[Faulstich, 1999b] Lukas C. Faulstich. Using graph transformation techniques for integrating information from the WWW. In *TAGT'98 final proceedings*, LNCS, 1999. Extended version of [Faulstich, 1998], to appear.

[Fed, 1997] Federal Committtee on Statistical Methodology. *Record Linkage Techniques – Intl. Workshop*, 1997.

[Fernandez *et al.*, 1997] M. Fernandez, D. Florescu, A. Levy, and D. Suciu. A query language and processor for a web-site management system. In *Workshop on Management of Semistructured Data*, 1997.

[Fernandez *et al.*, 1998] Mary Fernandez, Daniela Florescu, Jaewoo Kang, Alon Levy, and Dan Suciu. Catching the boat with Strudel: experiences with a web-site management system. In *SIGMOD*, pages 414–425, 1998.

[G. Rozenberg Montanari *et al.*, 1997] U. G. Rozenberg Montanari, H. Ehrig, and H.-J. Kreowski, editors. *Handbook of Graph Grammars and Computing by Graph Transformation, Volumes 1–3*. World Scientific Publishing, 1997.

[Garcia-Molina *et al.*, 1995] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. Integrating and accessing heterogeneous information sources in tsimmi. In *Proc. of AAAI Symposium on Information Gathering*, pages pp. 61–64, 1995. <ftp://db.stanford.edu/pub/papers/tsimmis-abstract-aaai.ps>.

[Goldman and Widom, 1997] R. Goldman and J. Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In *Twenty-Third International Conference on Very Large Data Bases*, 1997.

[Goldman *et al.*, 1999] R. Goldman, J. McHugh, and J. Widom. From semistructured data to xml: Migrating the lore data model and query language. In *Workshop on the Web and Databases (WebDB '99)*, pages 25–30, 1999.

[Hammer *et al.*, 1997] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting semistructured information from the web. In *Proc. Workshop on Management of Semistructured Data*, Tucson, 1997.

[Heckel *et al.*, 1995] Reiko Heckel, Jürgen Müller, Gabriele Taentzer, and Annika Wagner. Attributed graph transformations with controlled application of rules. In *Proc. Colloquium on Graph Transformation and its Application in Computer Science*, 1995.

[Heckel *et al.*, 1996] R. Heckel, A. Corradini, H. Ehrig, and M. Löwe. Horizontal and vertical structuring of typed graph transformation systems. *Mathematical Structures in Computer Science*, 6(6):613–648, 1996.

[Huck *et al.*, 1998] Gerald Huck, Peter Fankhauser, Karl Aberer, and Erich J. Neuhold. Jedi: Extracting and synthesizing information from the web. In *CoopIS 1998*, pages 32–43, 1998.

[Imieliński and Lipski, Jr., 1984] Tomasz Imieliński and Witold Lipski, Jr. Incomplete information in relational databases. *Journal of the ACM*, 31(4):761–791, October 1984.

[Jahnke *et al.*, 1996] J. Jahnke, W. Schäfer, and A. Zündorf. A design environment for migrating relational to object-oriented database systems. In *Int. Conf. on Software Maintenance, ICSM '96*, 1996.

[Jerome Simeon, 1998] Sophie Cluet Jerome Simeon. Using YAT to build a web server. In *Intl. Workshop on the Web and Databases (WebDB)*, 1998.

[Keller and Basu, 1994] Arthur M. Keller and Julie Basu. A predicate-based caching scheme for client-server database architectures. *The VLDB Journal*, 5(1), 1994.

[Kifer *et al.*, 1995] Michael Kifer, Georg Lausen, and James Wu. Logical foundations of object-oriented and frame-based languages. *Journal of the ACM*, 42(4):741–843, July 1995.

[Knoblock *et al.*, 1998] Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Pragnesh Jay Modi Naveen Ashish, Ion Muslea, Andrew G. Philpot, and Sheila Tejada. Modeling web sources for information integration. In *Proc. Fifteenth National Conference on Artificial Intelligence*, 1998.

[Koch, 1999] M. Koch. *Integration of Graph Transformation and Temporal Logic for the Specification of Distributed Systems*. PhD thesis, Technische Universität Berlin, FB 13, 1999. To defend.

[Konopnicki and Shmueli, 1997] D. Konopnicki and O. Shmueli. W3QS : A system for WWW querying. In *Proceedings of the 13th International Conference on Data Engineering (ICDE'97)*, pages 586–586, Washington - Brussels - Tokyo, April 1997. IEEE.

[Öksüz, 1999a] Ahter Öksüz. HyperDiscoverer – ein werkzeug für die struktur-analyse von semistrukturierten daten. Master's thesis, TU Berlin, 1999. Advisor: L.C. Faulstich. To appear.

[Öksüz, 1999b] Mürüvet Öksüz. HyperDesigner – eine entwurfs-umgebung für die konstruktion von HyperViews. Master's thesis, TU Berlin, 1999. Advisor: L.C. Faulstich. To appear.

[Lakshmanan *et al.*, 1996] L. V. S. Lakshmanan, F. Sadri, and I. N. Subramanian. A declarative language for querying and restructuring the Web. In IEEE, editor, *Sixth International Workshop on Research Issues in Data Engineering: interoperability of nontraditional database systems: proceedings, February 26–27, 1996, New Orleans, Louisiana*, pages 12–21, 1109 Spring Street, Suite 300, Silver Spring, MD 20910, USA, 1996. IEEE Computer Society Press.

[Ley, 1997] M. Ley. Die Trierer Informatik-Bibliographie DBLP. In *GI Jahrestagung 1997*, pages 257–266, 1997. <http://dblp.uni-trier.de>.

[Ludäscher *et al.*, 1998a] B. Ludäscher, R. Himmeröder, G. Lausen, W. May, and C. Schlepphorst. Managing semistructured data with florid: A deductive object-oriented perspective. *Information Systems*, 23(8), 1998. To appear.

[Ludäscher *et al.*, 1998b] B. Ludäscher, R. Himmeröder, and W. May. Techniques and rule patterns for declaratively querying web data with FLORID. In *KI-98 Workshop "Deklarative KI-Methoden zur Implementierung und Nutzung von Systemen in Netzen"*, 1998.

[Löwe, 1993] Michael Löwe. Algebraic approach to single-pushout graph transformation. *Theoretical Computer Science*, 109:181–224, 1993.

[MathML1.01, 1999] *Mathematical Markup Language (MathML[tm]) 1.01 Specification*, july 1999. W3C Recommendation.

[Mendelzon *et al.*, 1997] A. O. Mendelzon, G. A. Mihaila, and T. Milo. Querying the World Wide Web. *International Journal on Digital Libraries*, 1(1):54–67, 1997.

[Mic, 1997] Microsoft Corp. *Channel Definition Format (CDF)*, March 1997. W3C Note, <http://www.w3.org/TR/NOTE-CDFsubmit.html>.

[Nagl, 1979] Manfred Nagl. A tutorial and bibliographical survey on graph grammars. In V. Claus, H. Ehrig, and G. Rozenberg, editors, *Graph-Grammars and Their Application to Computer Science and Biology*, volume 73 of *Lecture Notes in Computer Science*, pages 70–126, 1979.

[Paolo Atzeni, 1997] Paolo Merialdo Paolo Atzeni, Giansalvatore Mecca. To weave the web. In *VLDB '97*, pages 206–215, 1997.

[Papakonstantinou *et al.*, 1996a] Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina. Object fusion in mediator systems. In *VLDB Conference*, pages 413–424, Bombay, India, September 1996.

[Papakonstantinou *et al.*, 1996b] Y. Papakonstantinou, H. Garcia-Molina, and J. Ullman. Medmaker: A mediation system based on declarative specifications. In *ICDE'96*, pages 132–141, 1996.

[Quass *et al.*, 1996] Dallan Quass, Jennifer Widom, Roy Goldman, Kevin Haas, Qingshan Luo, Jason McHugh, Svetlozar Nestorov, Anand Rajaraman, Hugo Rivero, Serge Abiteboul, Jeffrey D. Ullman, and Janet L. Wiener. LORE: A Lightweight Object REpository for semistructured data. In H. V. Jagadish and Inderpal Singh Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, page 549, Montreal, Quebec, Canada, 4–6 1996.

[Robie *et al.*, 1998] Jonathan Robie, Joe Lapp, and David Schach. XML query language (XQL). In *QL'98 - The W3C Query Languages Workshop*, 1998. <http://www.w3.org/TandS/QL/QL98/>.

[Rozenberg *et al.*, 1999a] G. Rozenberg, U. Montanari, H. Ehrig, and H.-J. Kreowski, editors. *Handbook of Graph Grammars and Computing by Graph Transformation, Volume 2: Specification and Programming*. World Scientific, 1999. To appear.

[Rozenberg *et al.*, 1999b] G. Rozenberg, U. Montanari, H. Ehrig, and H.-J. Kreowski, editors. *Handbook of Graph Grammars and Computing by Graph Transformation, Volume 3: Concurrency and Distribution*. World Scientific, 1999. To appear.

[Rozenberg, 1996] G. Rozenberg, editor. *Handbook of Graph Grammars and Computing by Graph Transformation, Volume 1: Foundations*. World Scientific, 1996.

[Sahuguet and Azavant, 1999] Arnaud Sahuguet and Fabien Azavant. Web ecology: Recycling html pages as xml documents using w4f. In *WebDB'99*, 1999.

[SFU, 1996] Simon Fraser University Electronic Library in Computing Science. <http://fas.sfu.ca/projects/ElectronicLibrary/Collections/CMPT/>, 1996.

[Sheth and Larson, 1990] Amit P. Sheth and James A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[Smith and Lopez, 1997] D. Smith and M. Lopez. Information extraction for semi-structured documents. In *Proc. Workshop on Management of Semistructured Data*, Tucson, 1997.

[SUL, 1998] Stanford University Libraries – Electronic Journals Collection. <http://www-sul.stanford.edu/collect/ejourns.html>, 1998.

[SVG1.0, 1999] *Scalable Vector Graphics (SVG) 1.0 Specification*, july 1999. W3C Working Draft, <http://www.w3.org/1999/08/WD-SVG-19990812/>.

[Taentzer, 1996] G. Taentzer. *Parallel and Distributed Graph Transformation: Formal Description and Application to Communication-Based Systems*. PhD thesis, TU Berlin, 1996.

[USB, 1998] Stony Brook University Libraries – electronic journals. <http://www.sunysb.edu/library/ldeljour.htm>, 1998.

[W3C, 1998] W3C DOM Working Group. *Document Object Model (DOM) Level 1 Specification*, version 1.0 edition, Oct 1998. W3C Recommendation, <http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001>.

[W3C, 1999a] W3C RDF Working Group. *Resource Description Framework (RDF) Model and Syntax Specification*, Feb 1999. W3C Recommendation, <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>.

[W3C, 1999b] W3C RDF Working Group. *Resource Description Framework (RDF) Schema Specification*, Mar 1999. W3C Proposed Recommendation, <http://www.w3.org/TR/1999/PR-rdf-schema-19990303>.

[W3C, 1999c] W3C XML Linking Group. *XML XPointer Requirements*, version 1.0 edition, Feb 1999. W3C Note, <http://www.w3.org/TR/1999/NOTE-xptr-req-19990224>.

[W3C, 1999d] W3C XML Schema Working Group. *XML Schema Part 1: Structures*, May 1999. W3C Working Draft, <http://www.w3.org/1999/05/06-xmlschema-1/>.

[W3C, 1999e] W3C XSL Working Group. *Extensible Stylesheet Language (XSL) Specification*, Apr 1999. W3C Working Draft, <http://www.w3.org/TR/WD-xsl-19990421>.

[W3C, 1999f] W3C XSL Working Group. *XSL Transformations (XSLT) Specification*, version 1.0 edition, Apr 1999. W3C Working Draft, <http://www.w3.org/TR/1999/WD-xslt-19990421>.

[Wechler, 1992] Wolfgang Wechler. *Universal Algebra for Computer Scientists*, volume 25 of *EATCS Monographs on Theoretical Computer Science.* Springer, Berlin, 1992.

[Wiederhold, 1992] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

[XHTML1.0, 1999] *XHTML[tm] 1.0: The Extensible HyperText Markup Language*, august 1999. W3C Proposed Recommendation, <http://www.w3.org/TR/1999/PR-xhtml1-19990824>.

List of Mathematical Symbols

Symbol	**Meaning**	**Definition**	**Page**
$\mathbb{PG}$	category of plain graphs	Th. 3.1.1	25
$\mathbb{CG}$	category of clustered graphs	Th. 3.1.2	27
$F, G, H, \ldots$	(clustered) graphs	3.1.2	26
V_G, E_G, C_G, D_G, A_G	vertices, edges, clusters, dependencies, attribute algebra of a graph G	3.1.2	26
s_G, t_G, a_G, c_G	source, target, attribute, clustering function of a graph G	3.1.2	26
$f, g, h, \ldots$	graph morphisms	3.1.3	26
$(.)_{base}, (.)_{struct}, (.)_{attr},$ $(.)_{vertex}, (.)_{edge}, (.)_{cluster}, (.)_{dep}$	components of a (clustered) graph morphism	3.1.3	26
$dom(.)$	domain of a function or morphism	3.1.3	26
$\sqsubseteq, \sqsupseteq$	subgraph / supergraph relationship	3.1.4	27
$\hookrightarrow$	inclusion morphism	3.1.4	27
$. \cap ., . \cup .$	intersection and union of sets and graphs	3.1.5	27
$. \uplus .$	disjoint union of sets and graphs	3.1.5	27
$(.)\|_{(.)}$	restriction of a morphism or function	3.1.6	27
$.\nabla\ .$	compatibility of morphisms or functions	3.1.7	28
$. \rightsquigarrow .$	reachability relation	3.1.8	28
$\mathbb{T}$	atomic data sorts (primitive types)	3.1.9	28
$\mathbb{O}$	atomic data operations	3.1.9	28
$\Sigma = (\mathbb{T}, \mathbb{O})$	signature for atomic data algebra	3.1.9	28
$\mathbb{V}$	a variable set	3.1.9	28
$T_\Sigma(\mathbb{V})$	term algebra over Σ and $\mathbb{V}$	3.1.9	28
$type$	typing function for terms	3.1.9	28
$\mathbb{U} = T_\Sigma(\emptyset)$	universe of atomic data	3.1.9	28
S	schema graph	3.1.10	28
τ, ρ	typing / interpretation morphism	3.1.11	29
σ	a substitution	3.2.2	32
p	a rule	3.3.4	40
$Matches(., .)$	matches of a pattern graph in a data graph	3.2.5	33
Φ	an oracle	3.3.3	38
Π	a hyperview	3.3.6	42
$Apply^{(.)}(.\|.)$	the application functor	3.3.5	41
$PlanOracle^{(.)}(.\|.)$	solutions for a QEP	3.3.12	46
$Plans^{(.)}(.)$	plan functor for a hyperview	3.3.11	45
$Oracle^{(.),(.)}(., .)$	query match functor	3.3.13	46

Zusammenfassung der Ergebnisse

Um das World Wide Web zur Beantwortung konkreter Fragen zu benutzen, muß man häufig Informationen von verschiedenen Web-Sites zusammentragen und kombinieren. Diese Aufgabe wird durch die uneinheitliche Gestaltung und die inhaltliche Heterogenität der einzelnen WWW-Quellen noch erschwert. *Virtuelle Web Sites* stellen einen vielversprechenden Ansatz dar, dieses Problem zumindest für begrenzte Anwendungsbereiche zu lösen.

Ein virtueller Web Site bietet auf seinen Seiten Informationen, die aus einer Reihe von zugrundeliegenden Web Sites extrahiert, vereinheitlicht, und integriert wurden. Das Ziel ist dabei, dem Benutzer zeitaufwendiges Suchen nach möglicherweise auf alle angeschlossenen Web-Server verstreuten Seiten zu ersparen.

Der in dieser Dissertation präsentierte HyperView-Ansatz zur Integration von semistrukturierten Datenquellen besteht aus einer Methodik, einem zugrundeliegenden mathematischen Formalismus und einer Software-Umgebung, auf deren Basis virtuelle Web Sites realisiert werden können.

Ein virtueller Web-Site muß die folgenden Aufgaben erfüllen: *Extrahierung* von Daten aus den Seiten der angeschlossenen Web-Sites, *Integration* dieser Daten in einer einheitlichen Repräsentation, und schließlich die *Präsentation* der integrierten Daten im WWW, z.B. in Form von HTML-Seiten.

Im HyperView-Ansatz werden die drei genannten Schritte einheitlich als aufeinanderfolgende Sichten (Views) aufgefaßt, die Abbildungen zwischen Schichten unterschiedlicher Abstraktionsniveaus realisieren. Jede Schicht wird mittels Schemata modelliert und entspricht einer Ebene in der HyperView-Architektur. Die HyperView-Methodik stellt eine Richtlinie dar, wie diese Schichten und die Abbildungen zwischen ihnen zu modellieren sind.

In HyperView werden sowohl die Inhalte von Web-Sites als auch die Resultate der darauffolgenden Sichten durch Graphen repräsentiert. Zu diesem Zweck wurde ein eigenes graphbasiertes Datenmodell, CGDM, entwickelt. Der Sicht-Mechanismus von HyperView unterstützt Abbildungen zwischen diesen Graphen. Sichten werden durch Mengen von Graphtransformationsregeln definiert. Nachdem es häufig weder sinnvoll noch möglich ist, Sichten über Web Sites im vorhinein zu materialisieren, wurde für HyperView ein bedarfsgesteuerter Mechanismus zur Aktivierung von Regeln formal beschrieben und im HyperView System implementiert. Dieser Mechanismus materialisiert Sichten inkrementell, in Reaktion auf Anfragen gegen die Sichten.

Das HyperView System ist in Prolog implementiert. Graphtransformationsregeln werden in Prolog-Prädikate kompiliert und können so effizient unter Ausnutzung von Unifikation und Rückwärts-Verkettung ausgeführt werden. HTML-Seiten werden mit einer frei verfügbaren Software aus dem WWW geladen. Als Technologie für die Einbindung des HyperView Systems in einen normalen Web-Server werden Java Servlets genutzt.

In der Dissertation sind zwei Fallstudien enthalten, welche die Anwendbarkeit von HyperView in den Feldern Digitale Bibliotheken und kulturelle Stadtinformationen demonstrieren. Die Anwendbarkeit von HyperView im Rahmen der momentan entstehenden XML-bezogenen Standards wird im abschließenden Kapitel der Arbeit diskutiert.

Die Hauptergebnisse dieser Arbeit sind folgende:

1. der Nachweis, daß die Probleme der Daten-Extraktion, -Integration, und -Präsentation mit einem *einheitlichen Abbildungs-Mechanismus* gelöst werden können,

2. die HyperView-Methodik für die Modellierung und Integration von Web-Sites,

3. die *formale Definition* des Datenmodells, des Regelkonzepts und des bedarfsgesteuerten Mechanismus für die Materialisierung von Sichten,

4. die *Implementierung* des HyperView Systems als einer Plattform für die Errichtung virtueller Web-Sites, und

5. die *Validierung* der HyperView-Methodik und des HyperView Systems in den erwähnten Fallstudien.

Zusammenfassend stellt diese Dissertation somit eine durchgängige Behandlung des Problems der Erstellung von virtuellen Web Sites dar, die Entwurfs-Methodik, formale Fundierung und Software-Unterstützung umfaßt.

Lebenslauf

Lukas C. Faulstich

*21. März 1967	in Würzburg.
Eltern:	Karoline Faulstich, geb. Frink, Hausfrau, und Otmar Faulstich, Kirchenmusiker.
1973–77	Grundschule Bechtolsheimer Hof, Würzburg.
1979–86	Riemenschneider-Gymnasium, Würzburg.
1986	Abitur.
1986–87	Grundwehrdienst.
1987–1994	Studium der Informatik, Nebenfach Mathematik, Julius-Maximilians-Universität Würzburg.
1989–93	Stipendiat des Cusanuswerkes.
1991–92	Austauschstudium: State University of New York, Stony Brook, New York, USA.
1993	Diplomarbeit: Deduktive Programmierung in Eiffel.
1994	Diplomprüfung.
1994–1996	Wissenschaftlicher Mitarbeiter, Institut für Informationssysteme, Medizinische Universität Lübeck, bei Prof. Dr. V. Linnemann.
1997–Jan. 2000	Stipendiat am Graduiertenkolleg *Verteilte Informationssysteme*, Arbeitsgruppe *Datenbanken und Informationssysteme*, Freie Universität Berlin, bei Prof. Dr. H. Schweppe.

Verwendete Hilfsmittel

Hiermit erkläre ich, die vorliegende Arbeit auf Grundlage der in der Arbeit genannten Hilfsmittel selbständig verfaßt zu haben.

(Lukas C. Faulstich)