

# A Deterministic Model for Semistructured and Structured Data

Peter Buneman

Affiliation:  
Department of Computer and Information Science  
University of Pennsylvania

## Contact Information

200 South 33rd Street  
Philadelphia, PA 19104-6389  
Tel: (215) 898 7703  
Fax: (215) 898-0587  
URL : <http://db.cis.upenn.edu/~peter>

## WWW PAGE

Project URLs

[www.cis.upenn.edu/~hahosoya/xduce/](http://www.cis.upenn.edu/~hahosoya/xduce/)

<http://db.cis.upenn.edu/>

<http://db.cis.upenn.edu/~wctan/DataProvenance/>

## List of Supported Students and Staff

Wang-Chiew Tan, research student

Haruo Hasoya, postdoctoral researcher

## Project Award Information

- Award Number: IIS-9977408
- Duration: 10/01/99 -- 9/30/02
- Title: A Deterministic Model for Semistructured Data

## Keywords

XML type systems, Data Models, Scientific Databases, Data Provenance

## Project Summary

This project investigates a new model for semistructured data which is more general than existing models in that the labels may be arbitrary values. They may, in fact, be pieces of semistructured data. It is less general in that the graph is constrained to be *deterministic*. For an edge-labeled graph this means that the out-edges from any vertex must have distinct labels. A number of advantages are claimed for this model: It is entirely syntactic. All operations can be understood in terms of syntax, and serialization of data is immediate. Object identifiers in particular are syntactic constructs, they have structure, and certain database transformations may be achieved by manipulation of this structure. A simple type system can be used to express both structure and constraints, and the syntax of types allows a direct representation of conceptual models such as entity-relationship diagrams. Labels are first-class values, and may be treated as such. It is possible, for example, to represent matrices in the model by using integer labels and to express operations such as matrix transposition in a query language. Finally, when data is represented in this model, there is a "persistent" data structure that allows space-efficient encoding of database versions and other forms of variant data.

In addition to supporting work on this data model, funds for this project have also been used to support Xduce, a typed XML programming language. This language is based on some earlier, related work by the PI and his colleagues on type systems for semistructured data.

## Goals, Objectives, and Targeted Activities

The first year of research on this project has clarified two important goals of the project. First, in connection with our work on data models and scientific databases, we have found it necessary to characterize *where* a piece of data comes from. In particular, suppose a database is produced by a query how does one trace back a data element back through a query? We have also found that the concept of a *key* for XML and for semistructured data, which is essential in any kind of scientific citation, has not been fully understood. We have some initial publications on both these topics, but they both need further development. Finally we are developing an efficient archiving system for scientific data. We hope to have a report on it shortly.

In connection with the work on Xduce, there are an enormous number of challenges, both theoretical and practical. In the PI's opinion there are two critical challenges. First, how is Xduce to be embedded in a conventional programming environment? One could go the SQL route and have a dynamically typed, fragile, interface. But this would negate the statically typed benefits of both the host and guest languages. One would like to see some degree of static type-checking carried across the interface. The second is the introduction of some form of (record) subtyping into the language. Anyone who has examined a number of DTDs will be painfully aware of the various highly unsatisfactory "hacks" people use to express subtyping. A subtype system that extends DTDs and that can be used in a programming language for static checking appears to be essential.

## Indication of Success

The main impact of the two aspects of the project has been the effect on the development of schemas and languages for XML. Xduce is the *first* language to exploit the DTD as a static type system. The first validator for XML-schema was written in Xduce. To see something of the effect of Xduce on query systems for XML, please refer to the XML Query Algebra.

More recently the development of keys for XML (a side-effect of our work on data models) has sparked some interest in the various attempts to produce schema and constraint languages for XML (e.g. XML-schema)

## Project Impact

- One PhD has been completed under this project, and it has also supported a female research student.
- A book co-authored by the PI is being used in a number of seminar courses on Web data. The project represents a continuation of that work
- The project has been a focus of collaboration between the Department of Computer and Information Science at Penn and with the Linguistic Data Consortium (linguistic corpora) and the BioInformatics Center (genomic databases)
- There is regular interaction between the Xduce project and researchers at AT&T (now Avaya)

## References

The following papers were fully or partially supported by this project

- **On Computing Functions with Uncertainty** [\[abstract\]](#) [\[.ps\]](#) [\[.pdf\]](#) [.bib](#)  
*Proceedings of ACM Symposium on Principles of Database Systems (PODS)* (2001)  
[Sanjeev Khanna](#) [Wang-Chiew Tan](#)
- **Keys for XML** [\[.ps\]](#) [\[.pdf\]](#) [\[.html\]](#) [.bib](#)  
*WWW10* (2001)  
[Peter Buneman](#) [Susan Davidson](#) [Wenfei Fan](#) [Carmem Hara](#) [Wang-Chiew Tan](#)
- **Why and Where: A Characterization of Data Provenance** [\[abstract\]](#) [\[.ps\]](#) [\[.pdf\]](#) [.bib](#)  
*International Conference on Database Theory (ICDT)* (2001)  
[Peter Buneman](#) [Sanjeev Khanna](#) [Wang-Chiew Tan](#)
- **Reasoning about Keys for XML** [\[abstract\]](#) [\[.ps\]](#) [\[.pdf\]](#) [.bib](#)  
*Technical Report MS-CIS-00-26* (2000)  
[Peter Buneman](#) [Susan Davidson](#) [Wenfei Fan](#) [Carmem Hara](#) [Wang-Chiew Tan](#)
- **SilkRoute: Trading between Relations and XML** [\[.ps\]](#) [\[.html\]](#) [.bib](#)  
*WWW9* (2000)

Mary Fernandez [Wang-Chiew Tan](#) Dan Suciu

- **UnQL: A Query Language and Algebra for Semistructured Data Based on Structural Recursion** [\[abstract\]](#) [\[.ps\]](#) [\[.pdf\]](#) [.bib](#)  
*VLDB Journal* (2000)  
[Peter Buneman](#) Mary Fernandez Dan Suciu
- **Data Provenance: Some Basic Issues** [\[abstract\]](#) [\[.ps\]](#) [\[.pdf\]](#) [.bib](#)  
*Foundations of Software Technology and Theoretical Computer Science* (2000)  
[Peter Buneman](#) [Sanjeev Khanna](#) [Wang-Chiew Tan](#)
- [Regular Expression Types for XML](#). Haruo Hosoya. PhD Thesis at *The University of Tokyo*, 2001.
- [Regular Expression Pattern Matching for XML](#). Haruo Hosoya and Benjamin C. Pierce. In *The 25th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, 2001.
- [Regular Expression Types for XML](#). Haruo Hosoya and Jérôme Vouillon and Benjamin C. Pierce. In *Proceedings of the International Conference on Functional Programming (ICFP)*, 2000.
- [XDuce: A Typed XML Processing Language](#). Haruo Hosoya and Benjamin C. Pierce. In *Proceedings of Third International Workshop on the Web and Databases (WebDB2000)*, 2000.

## Area Background

Please visit the [Data Provenance](#) web site and look at the introduction to the [project description](#). It contains a general description of the problem for which this project provides the technical support.

## Area References

Please see the URLs in the previous section for some interesting references concerning the interdisciplinary nature of this project.

## Potential Related Projects

If you have ideas for projects or collaborations with other researchers within the scope of the IDM program or in other related areas, describe some of them briefly here.

I have several such ideas, but the National Science foundation has not seen fit to fund them!