Summary of the Dagstuhl Seminar 05061 (06.02.–11.02.2005) "Foundations of Semi-structured Data" organized by F. Neven, T. Schwentick, and D. Suciu

July 8, 2005

1 Motivation

As in the first seminar on this topic, the aim of the workshop was to bring together people from the areas related to semi-structured data. However, besides the presentation of recent work, this time the main goal was to identify the main lines of a common framework for future foundational work on semi-structured data. These lines of research are summarized below.

The workshop was of a very interdisciplinary nature with invitees from databases, structured documents, programming languages, information retrieval and formal language theory. Several of the lectures were presented by PhD students. We had four invited speakers and a panel on research evaluation. Due to strong connections between topics treated at this workshop, many of the participants initiated new cooperations and research projects.

2 Foundations and processing of XQuery

The invited survey talk by **Ioana Manolescu** entitled **XML Query Processing: Storage and Query Model Interplay** presented an overview of

the latest XML storage and processing techniques. There is still a long way to go to marry the relational and the XML world, and to get highly scalable XQuery engines.

Several research directions became apparent. One direction is to logically optimize XQuery-expressions themselves. **Alin Deutsch** addressed this direction in his talk **The NEXT Framework for Logical XQuery Optimization**. Another possibility is to take full advantage of existing highly reliable and scalable relational database technology. Such systems have proved their merits and strengths over the last decades. A line of research addresses how XML technology can be implemented on top of a relational database. All XML is translated into relations and all queries expressed in an XML query language are translated into SQL. In this respect, **Christopher Re** talked about **A Performant XQuery to SQL Translation**.

Although implementation and optimization is of utmost importance, it is equally important to understand the design and expressive power of XQuery. In his talk **The Complexity of Nonrecursive XQuery**, **Christoph Koch** made apparent the connections between XQuery and the complex object functional query languages that have been studied in the 90's. In his talk **Expressive power of XQuery fragments**, **Roel Vercammen** gave a detailed account of the expressiveness of the different XQuery constructs.

3 Updates

Updates have not adequately been addressed in the XML world nor in the research world. The main problem is a missing standard for updates in XML. Without doubt this topic will attract much attention in the future. **Michael Benedikt** discussed an XML update language in his talk **Semantics and Optimization of XML Updates**. **Uri Zarfaty** employed a programming language view on that topic in his presentation **Reasoning about Tree Update**.

4 Compressing

The main advantage of XML, its flexibility, is at the same time its main problem. The verbosity of an XML document is frightening. For this reason,

a lot of effort is invested in compression of XML documents. While good algorithms and techniques have been developed to compress XML for shipping, a lot of work still needs to be done to efficiently query compressed XML documents (without decompressing). Two talks gave insights into the difficulty of this problem. **Ioana Manolescu** talked about **XQueC: embedding XML Compression into Databases** while **Alberto Laender** discussed **A Word-based Query-aware Compressor for XML Documents**.

5 Streaming processing of XML

XML started out as a document format, but evolved, due to the adoption of the format by the database and semi-structured data research community, to a format for data exchange. For this kind of application it is important that XML data can be processed in a streaming fashion rather than when stored in a relational or native XML database. This poses some new problems. In this respect, **Stefanie Scherzinger** talked about **Schema-based Scheduling of Event Processors and Buffer Minimization for Queries on Structured Data Streams**. **Nicole Schweikardt** addressed the streaming problem on a theoretical level by providing and analyzing a new Turing machine model for stream based processing. The title of her talk was **Tight Lower Bounds for Query Processing on Streaming and External Memory Data**.

6 Typing

Typing is an area where database and programming language researchers meet. A number of talks addressed a variety of topics. **Helmut Seidl** discussed sound and complete typechecking of XSLT-based XML-transformations in his presentation **Type-Checking XML Transformers with Macro Tree Transducers**. **Stijn Vansummeren** addressed well-definedness of XQuery in his presentation **Deciding Well-Definedness of XQuery Fragments**. Both talks focused on sound and complete algorithms. Two programming language oriented talks were delivered by **Alain Frisch** on **Adding XML types to ML** and by **Giuseppe Castagna** on **Types and Patterns**

for Querying XML. Finally, **Wim Martens** discussed yet another interpretation of typing in the context of XML Schema languages. In his talk **Which XML Schemas Admit 1-Pass Preorder Typing?** he explained that XML Schema Definitions are not adequately modeled by unranked regular tree automata, but by a restriction thereof. This restriction is enforced by the Element Declarations Consistent constraint which is motivated by efficient processing. He also discussed a possible relaxation of the latter constraint that still allows for efficient processing.

7 Pattern languages

One of the first directions in foundational XML research was on XML pattern languages. **Christoph Koch** provided a survey of recent results in his invited talk **Logic Programming, (Automata,) and XML – or – Queries on Trees: A Journey from Dagstuhl to Vienna and Back**. **Maarten Marx** discussed **Marrying XPath to Regular Tree Queries: Looping Caterpillars** where he made correspondences with propositional dynamic logic. An analysis of XPath was done in the talks **Forward XPath-like Queries Revisited** and **Extending XMark benchmark with XPath 1.0 queries** by **Dan-Alexandru Olteanu** and **Loredana Afanasiev**, respectively. **Joachim Niehren** addressed **Queries by Tree Automata for Web Information Extraction** where he showed how learning of tree automata can be used to wrap web pages. On the foundational side, **Anca Muscholl** presented results on FO^2 over infinite alphabet strings in her talk **Boosting first-order logic with data**. She showed that FO^2 is decidable over such strings. This opens possibilities for decidable typechecking results for XML transformations that allow to compare data values. **Luc Segoufin** presented **Regular tree languages definable in FO** where he characterizes the expressive power of first-order logic over trees. **Wolfgang Thomas** talked on minimizing tree automata in his talk **Automata on Unranked Trees: Restrictions and Extensions**.

8 The future of data integration

Dan Suciu presented in his invited talk **What is next?** a new paradigm based on probabilistic databases. After a decade of research on how to integrate data, he concludes that we should look at ways to quantify the value of the integrated results. One possibility is by applying probabilities. Another new research direction that is attracting a lot of attention is the field of data exchange. **Marcelo Arenas** discussed recent results in his talk **XML Data Exchange: Consistency and Query Answering**. In his invited presentation **Active XML and distributed data management**, **Serge Abiteboul** surveyed recent work on Active XML. Closely related talks on retrieval and similarity of XML were given by **Felix Weigel** (**Node Identification Schemes for Efficient XML Retrieval**) and **Elio Masciari** (**Fast Detection of XML Structural Similarity**).