# XML Retrieval: DB/IR in Theory, Web in practice

Sihem Amer-Yahia
Yahoo! Research
New York, USA
sihem@yahoo-inc.com

Ricardo Baeza-Yates
Yahoo! Research Barcelona and Latinamerica
ricardo@baeza.cl

Mariano P. Consens
University of Toronto
Canada
consens@cs.toronto.edu

Mounia Lalmas
Queen Mary University of London
UK
mounia@dcs.qmul.ac.uk

**Learning Objectives:** Tutorial attendees will learn about requirements for combined DB and IR applications that access XML content; learn about models and indexes tailored to XML and their algorithms; understand the specific problems of XML IR and XML query processing; learn about XML retrieval evaluation, INEX in particular; learn about the challenges in querying structure and content on the Web (in particular in the context of online communities).

Our main goal is to provide a clear view of the major challenges in XML Retrieval: its problems, its solutions and the pitfalls that should be avoided. We also describe the applicability of the DB and IR querying paradigms for accessing Web data.

**Topics:** XML query processing and data management, integration of text into XML and databases, and online communities on the Web.

**Relevance to VLDB 2007 Attendees:** In addition to those attendees whose main interest are the topics directly mentioned above, additional attendees will be interested in the challenges of extending retrieval techniques to search the increasing amounts of XML content accessible in the Web and the use of XML as the encoding format for data in several domains (e.g., web services, semantic web, digital libraries, enterprise intranets, application specific domains in science and industry).

**Keywords:** XML Retrieval, Database and Information Retrieval Techniques, Indexing & Query Processing, Semi-structured Data and Data Models, Retrieval Evaluation, Online Communities, Social Tagging, Web 2.0

**Target audience:** The tutorial is targeted at an audience that includes most VLDB attendees. The level of the tutorial can be considered Introductory if it is a half-day tutorial, or Intermediate if it is a full-day tutorial. In both cases, attendees should have basic introductory knowledge about standard DB and IR models and methods.


*VLDB '07,* September 23-28, 2007, Vienna, Austria.


**Tutorial Abstract:** The world of data has been developed from two main points of view: the structured relational data model and the unstructured text model. The two distinct cultures of databases and information retrieval now have a natural meeting place in the Web with its semi-structured XML model. Data in Digital Libraries and in Enterprise Environments also shares many of the semi-structured characteristics of web data. As web-style searching becomes an ubiquitous tool, the need for integrating these two viewpoints becomes even more important.

In particular, we consider the application of DB and IR research to querying Web data in the context of online communities. With Web 2.0, the question arises: how can search interfaces remain simple when users are allowed to contribute content (Wikipedia), share it (Flickr), and rate it (YouTube)? When they can decide who their friends are (del.icio.us), what they like to see, and how they want it to look like (MySpace)? While we want to keep the user interface simple (keyword search), we would like to study the applicability of querying structure and content to a context where new forms of data-driven dynamic web content (e.g. user feedback, tags, contributed multimedia) are provided.

This tutorial will provide an overview of the different issues and approaches put forward by the IR and DB communities and survey the DB-IR integration efforts as they focus in the problem of retrieval from XML content. In particular, the context of querying content in online communities is an excellent example of such an application. Both earlier proposals as well as recent ones will be discussed. A variety of application scenarios for XML Retrieval will be covered, including examples of current tools and techniques.

**Tutorial History:** This is a newly developed tutorial proposal including a substantial amount of original content and a new table of contents. The authors had separately given tutorials in related topics in DB-IR integration and XML retrieval at a number of conferences including: VLDB 2004, COMAD 2005, SIGIR 2005, ASIAN 2005, CIKM 2005, and SIGIR 2006. A related proposal was also submitted to SIGIR 2007.

**Presenter Bios:**

*Sihem Amer-Yahia* is a Senior Research Scientist at Yahoo! Research. She is interested in the interplay between structured data and search in online communities. Previously, she worked on XML search. In particular, she represented

AT&T Labs on the Full-Text Task Force within the W3C. Sihem co-chaired WebDB 2004 and XSym 2006. Her current research interest is to leverage structure when querying content, in particular, she is focusing on issues related to processing top-k queries in online shopping and community-aware ranking in online communities. Recently, she gave a keynote at WebDB 2007 on Web 2.0 search challenges.

*Ricardo Baeza-Yates* is director of Yahoo! Research Barcelona and Yahoo! Research Latinamerica in Santiago, Chile. Until 2005 he was an ICREA Professor at Universitat Pompeu Fabra in Barcelona and also a professor and director of the Center for Web Research, that he founded in 2002, at the CS department of the University of Chile. His research interests include information retrieval, algorithms, and information visualization. He is co-author of the book Modern Information Retrieval, published in 1999 by Addison-Wesley. He received his Ph.D. in CS from the U. of Waterloo, Canada, in 1989.

*Mariano P. Consens* research interests are in the areas of Data Management Systems and the Web, with a focus on XML retrieval and autonomic systems. Mariano received his PhD and MSc degrees in Computer Science from the University of Toronto, and his Computer Systems Engineer degree from the Universidad de la Republica, Uruguay. Consens has been a faculty member in Information Engineering at the MIE Department, University of Toronto, since 2003. Before that, he was research faculty at the School of Computer Science, University of Waterloo, from 1994 to 1999. He has been active in the software industry as a founder and CTO of several startups.

*Mounia Lalmas* obtained a PhD in Computer Science from the University of Glasgow in 1996. Presently she is a Professor of Information Retrieval at the Department of Computer Science, at Queen Mary, University of London, which she joined as a lecturer in 1999. Her research focuses on the development and evaluation of intelligent access to interactive heterogeneous and complex information repositories. She is the co-leader of the INEX initiative, with over 50 participating organizations worldwide. She is the ACM SIGIR vice chair.

## Tutorial Full Description

### *1. Introduction*

Motivations. XML and the Web. Historical perspective on DB and IR communities and semi-structured data. Novel applications of DB and IR research to querying Web data in the context of online communities.

### *2. Conceptual Framework*

Goals. Data and Query Requirements. Sample Use Cases. DB-IR Integration issues

### *4. XML Basics and Standards*

XML Model, Schemas and Summaries. XPath and XQuery. Structured text models and query algebras.

### *5. Querying Content*

Content-only queries. Query specification and users' expectations. Nature and form of results. Ranking Measures. Comparison.

### *6. Querying Content and Structure*

User need specification in a precise query language. XQueryFT. XML IR models. How to obtain document and terms statistics. How to model relationships. How to deal with overlaps. How to interpret structural constraints.

We also discuss the case where users provide keyword queries and ranking as well as query processing makes use of structured information.

### *7. Preprocessing and Indexing Content and Structure*

Data Preparation. Indexing. XML Processing Algorithms: streaming, summaries, and indexes. Query Processing and Optimization. TopK query processing.

### *9. Evaluation*

Why we need to evaluate. Introduction to IR evaluation. Evaluation in XML IR, in particular INEX. Document collections, Topics, Tasks, Relevance, and Metrics. Lessons learned.

## Tutorial Selected Bibliography

The bibliography includes over 100 references. In addition, the tutorial will draw heavily from the following publications:


- XML, XPath, XQuery, and XQueryFT standards: World Wide Web Consortium (www.w3c.org).
- Proceedings of the ACM SIGIR Workshops on XML and Information Retrieval, edited by Yoelle Maarek *et al*., 2000 & 2002.
- Proceedings of the workshops of the Initiative for the Evaluation of XML Retrieval (INEX), edited by N. Fuhr, M. Lalmas et al, 2002-2006.
- Proceedings of the JASIST special issue on XML and IR, edited by Ricardo Baeza-Yates, David Carmel, Yoelle Maarek, and Aya Sofer, JASIST 53(6), 2002.
- Proceedings of the International Workshop on XQuery Implementation, Experience and Perspectives (XIME-P) 2004, 2005, 2006.
- Proceedings of the International XML Database Symposium (XSym), 2003-2006.
- Proceedings of the Joint Workshop on XML and DB-IR Integration, edited by Ricardo Baeza-Yates, Yoelle Maarek, Thomas Roelleke, and Arjen P. de Vries, SIGIR 2004, Sheffield, 2004.
- Report on the DB/IR panel at SIGMOD 2005, Sihem Amer-Yahia, Pat Case, Thomas Rlleke, Jayavel Shanmugasundaram, Gerhard Weikum, SIGMOD Record 34(4):71-74, 2005.
- Proceedings of the TOIS special issue on XML retrieval, edited by Ricardo Baeza-Yates, Norbert Fuhr, Yolle S. Maarek, ACM Trans. Inf. Syst. 24(4), 2006.