Standard XML Query Languages for Natural Language Processing

Ulrich Schaefer

German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab, Saarbrücken

ESSLLI 2009, 2nd week, 09:15–10:45

Why this lecture?

- More and more corpora and other linguistic resources are available in XML format
- XML often plays role of abstract syntax (having concrete syntax at the same time)
- Well implemented and established standards for querying XML are ready to be used for
  - offline access to corpora
  - online integration of NLP components
  - combination and transformation of resources
- This is a practical course introducing the main concepts and elements of W3C's XML query languages, focusing on applications in NLP

Why querying corpora or NLP XML output?

- getting statistics such as frequencies etc.
- combining resources
- pre-processing for machine learning
- visualization (answer to query is visual representation)
- interfacing to applications (QA, search)

There is often no clear distinction between query and transformation

What makes a good linguistic query language?

- not only data access, but also transformation and integration facilities
- enable composite and pipelined queries
- support relevant relationships within linguistic data
- abstract from low-level representation where possible

What will this lecture cover?

Introduction to W3C's three standard XML query languages

- XPath 1.0 and parts of 2.0
- XSLT 1.0 and parts of 2.0
- XQuery 1.0

with various NLP-related examples

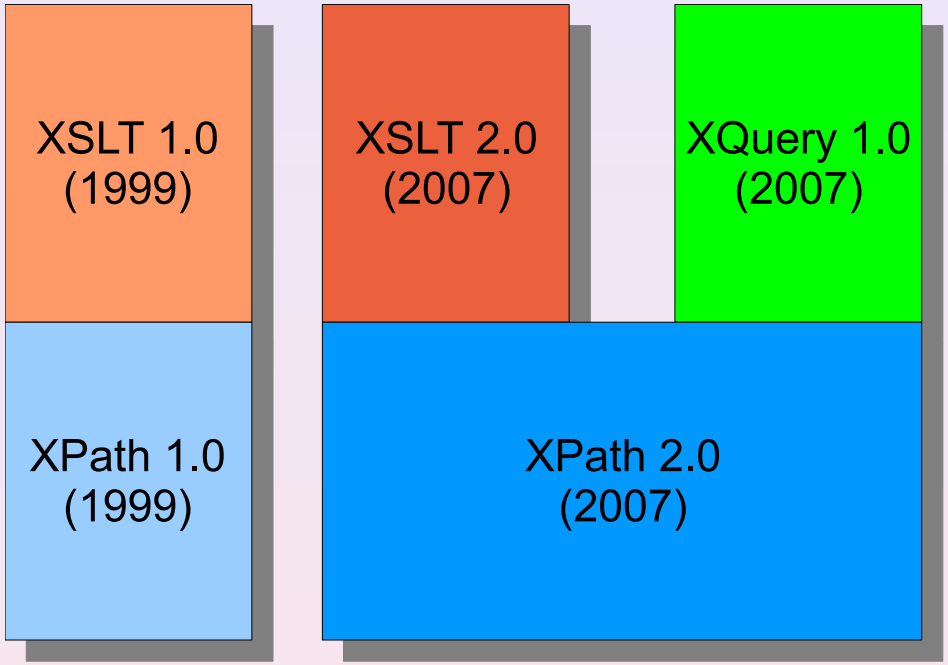

Figure: W3C XML query standards

Course material online

Course material, e.g.

- slides
- online documentation
- bibliography
- links to useful software tools

is/will be made available online at

http://www.dfki.de/~uschaefer/esslli09/