BibTeXML:
An XML Representation of BibTeX

Erik Wilde
ETH Zürich

Outline of the Talk

- Motivation
  - BibTeX as De-facto Standard for Bibliographies
  - XML as De-facto Standard for Structured Data
- Short Introduction to BibTeX
- BibTeXML: what it is and how to get there
  - Problems and Lessons learned
- `BibTeXML.org`: Web-based Service
- XLinkbase as target for BibTeXML data
- Conclusions, Further Work, and Ideas

Motivation

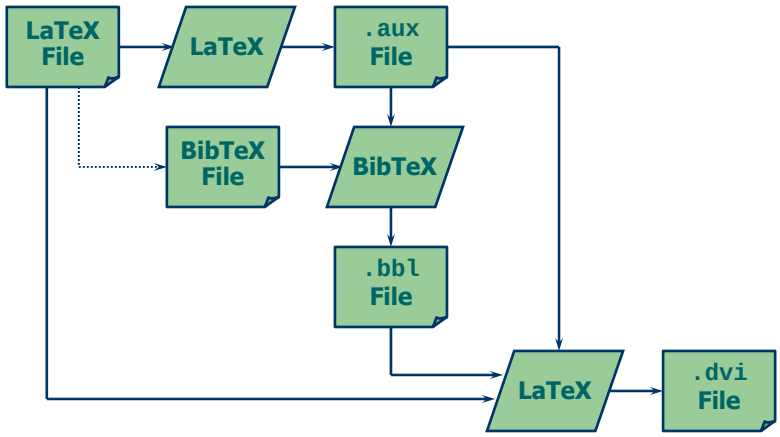
- LaTeX is widely used in the scientific community
  - superior formatting quality for complex typography
  - cross-platform, free, and stable
- BibTeX is a small add-on to LaTeX
  - pros:
    - format and program for bibliographies
    - automagically compiles references for a paper from a "database"
    - cross-platform, free, and stable
  - cons:
    - hard to manage, hard to manipulate
- test platform for our XLinkbase (similar to topic maps)

Why XML?

"XML is ASCII for the 21st Century"

- well-suited format for structured data
  - cross-platform, free, and stable
  - much easier and cheaper than (R)DBMS
  - XSLT as ideal tool for manipulation and extraction
- XML is better than a home-grown syntax
  - well-defined, no ambiguities
  - available tools and well-trained users
  - powerful query languages (XPath/XSLT/XQuery)

A Short Introduction to BibTeX



BibTeX: What it is

- very simple format for bibliographic references
  - small number of predefined entry types
    - book, proceedings, technical report, article, ...
  - small number of predefined fields
    - author, title, editor, page, month, year, publisher, ...
  - undefined fields are allowed, but ignored
- very simple macro mechanism
  - usage is not enforced in any way
  - typically, cut&paste is easier and more widely used

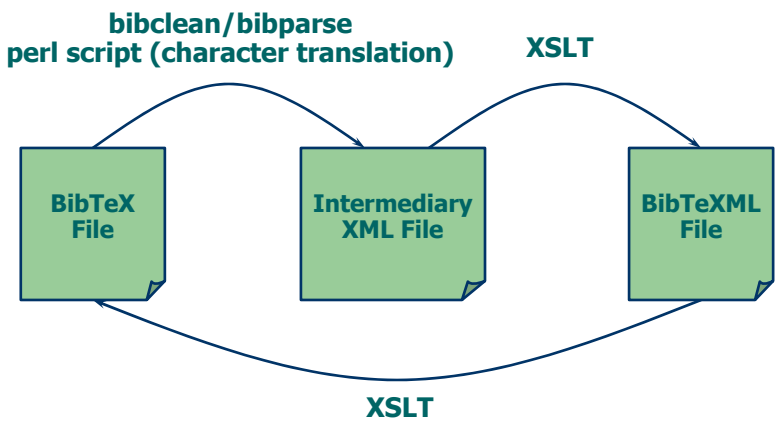
A BibTeX Sample Entry

```
@Book{lamport:86,
  author =    "Leslie Lamport",
  title =     "{\LaTeX}: A Document
               Preparation System",
  publisher = "Addison-Wesley",
  year =      "1986",
  language =  "en-US",
  index =     "LaTeX typesetting" }
```


A BibTeX Sample Entry (with Macro)

```
@Book{lamport:86,
  author =    "Leslie Lamport",
  title =     "{\LaTeX}: A Document
               Preparation System",
  publisher = awl,
  year =      "1986",
  language =  "en-US",
  index =     "LaTeX typesetting" }
```

BibTeXML: How to Get There



A BibTeXML Sample Entry

```
<book id="lamport:86">
  <authors>
    <name>
      <prename>Leslie</prename>
      <surname>Lamport</surname>
    </name>
  </authors>
  <title><tex code="{\LaTeX}">LaTeX</tex>:
  A Document Preparation System</title>
  <publisher>Addison-Wesley</publisher>
  <year>1986</year>
  <language>en-US</language>
  <index>LaTeX typesetting</index>
</book>
```


Problem#1: Undefined Syntax

- BibTeXML syntax is very simple, but undefined
  - defined by the implementation
  - BibTeX never got past 0.99c
- Nelson Beebe created a hypothetical syntax
  - based on *trial and error* with the program
  - established as standard for many applications
  - but still has some strange things in it
- parsing BibTeX can be an adventure

Problem#2: Character Sets

- TeX has its own character repertoire and fonts
  - pro: everything under control of the system's author
  - con: hard or impossible to map to other formats
- TeX has many composed characters
  - u umlaut is: `\"u`, `\" u`, `\"{u}`, `"u`, `\"u{}`, ...
  - compositions can be combined: `\"\"u`
  - special cases: `\"\i{}` (or `{\"\i}`) for an ï
- not trivial to cover all possible cases
  - we need to parse a subset of TeX syntax

Problem#3: Data Normalization

- many fields contain redundant data
  - publishers typically appear very often
  - journal names (some are predefined in BibTeX)
  - authors and editors probably occur more than once
- macros can reduce or eliminate redundancies
  - problem: many people don't use them
  - solutions:
    - easy: don't try to resolve redundancies
    - hard: attempt automatic data normalization
- normalization requires intelligence

Data Normalization Approaches

- ideally, BibTeX data should be normalized
  - hardly ever used consistently
  - not enforced or even supported by BibTeX tools
- look for identical strings of identical type
  - misses people who are authors and editors
  - misses misspelled names
  - misses name variants (middle names, new names)
- long-term goal: produce normalized BibTeX

Problem#4: User-Friendliness

- conversion is a multi-step process
  - bibclean removes the most common syntax errors
  - bibparse translates into intermediary XML
  - perl script translates TeX to Unicode characters
  - XSLT transforms intermediary XML to BibTeXML
- errors may occur in several of these steps
  - reporting errors without confusing users
  - encouraging users to clean up their BibTeX

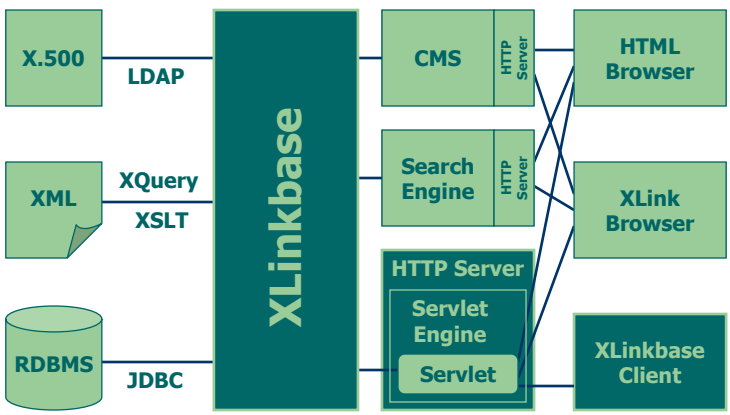
BibTeXML.org – Web-based Service

- currently under construction (goal: July 2001)
  - expect regular updates, bugfixes, and new services
- will be brought online in different phases
  1. online translation BibTeX ↔ BibTeXML
  2. locally store converted BibTeX data
     1. use for statistical analyses and problem tracking
     2. provide a query and export interface for users
  3. normalization service
     1. based on algorithms and observations
     2. based on existing bibliographic entries
- integrate into XLinkbase framework, generate GUI

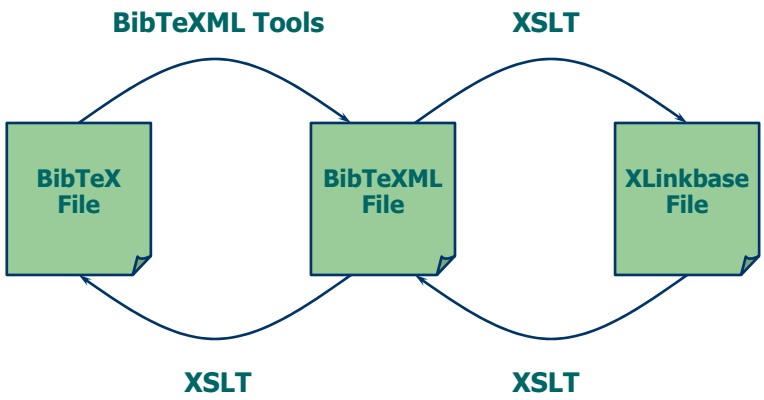
XLinkbase as BibTeXML Storage

- XLinkbase is similar to topic maps, but:
  - supports a well-defined derivation mechanism
    - topic facets are determined by derivation
    - derivation is used for supporting "typed associations"
  - supports domains for clustering topics
  - is based on XML
- XLinkbase contains references + metadata
  - references can be extracted as XLinks (or HTML...)
  - references are associated with topics
- BibTeXML is metadata about publications

XLinkbase System Architecture



BibTeXML and XLinkbase



Further Work and Ideas

- create an XML Schema for BibTeXML
  - currently only a DTD (weak typing)
  - makes extensions extremely easy (extending types)
- create import and export filters
  - for selected formats (MARC)
  - easy for XML-based formats
    - export can always use XSLT
    - import much harder for non-XML (parser required)
- integrate into XLinkbase framework

Conclusions

- converting – even seemingly trivial – "legacy" data can be surprisingly hard
  - structural problems (normalization)
  - data model problems (character repertoire)
- BibTeXML can be used
  - standalone as XML source for LaTeX/BibTeX
  - with XLinkbase for a "BibTeX topic map"
- hopefully a better way to continue using LaTeX

Contact and Further Information

- BibTeXML Web site
  - http://bibtexml.org/
  - Erik Wilde (mailto:dret@bibtexml.org)
- Converters Project Team
  - Brenno Lurati (mailto:brenno@bibtexml.org)
  - Luca Previtali (mailto:luca@bibtexml.org)
- Web Site Project Team
  - Sascha Gammaidoni (mailto:sascha@bibtexml.org)
  - Christian Mutti (mailto:christian@bibtexml.org)