# BibTeXML: An XML Representation of BibTeX

Luca Previtali, Brenno Lurati, Erik Wilde
ETH Zürich (Swiss Federal Institute of Technology), Switzerland
luca@bibtexml.org, brenno@bibtexml.org, dret@bibtexml.org

## Abstract

BibTeXML is an XML representation of BibTeX data. It can be used to represent bibliographic data in XML. The advantage of BibTeXML over BibTeX's native syntax is that it can be easily managed using standard XML tools (in particular, XSLT style sheets), while native BibTeX data can only be manipulated using specialized tools.

**Keywords:** TeX, LaTeX, BibTeX, XML, bibliography

## 1 Introduction

LaTeX [3] is an advanced text formatter which is widely used for typesetting in mathematics, physics and engineering. BibTeX [4] is a program and file format designed by Patashnik and Lamport in 1985 which provides an easy way to manage bibliographic references for LaTeX. The BibTeX format is character and field (tag) based.

Many BibTeX bibliographies are available online. These collections contain huge amounts of references (for an example see the size of `http://liinwww.ira.uka.de/bibliography` shown in Table 1). The main problem of these collections is that BibTeX is a simple format, which does not allow complex queries and data manipulation.

| | |
|---|---|
| References | > 1.1 million |
| Amount of data | 660 MBytes |
| Cross references | 16'000 |
| Number of URLs | 100'000 |

Table 1: Example of bibliographic database size

The goal of BibTeXML is to develop an XML environment for representing and structuring BibTeX bibliographies, which makes the management of bibliographic data easier, and to build an online database which allows upload and download of bibliographic entries.

## 2 Proposed Solution

XML provides an efficient way to structure data using an easy to edit and readable format. With XSLT it is possible to easily convert XML data into various formats.

The basic idea is that the whole bibliography should be managed using XML. `.bib` files are no longer edited: they are generated only for BibTeX use. To enable a reuse of all existing BibTeX collections, we need a conversion tool for transforming BibTeX bibliography into BibTeXML format (as described in Section 3.2). Once the BibTeXML representation is obtained, it is possible to store the BibTeXML entries in an online database (see Section 4). This database provides complex queries and data navigation which help the user to fetch the required references (see Figure 1).

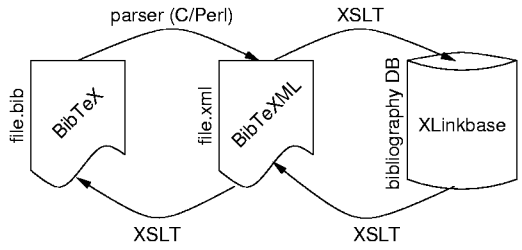


Figure 1: BibTeXML architecture

## 3 Implementation

### 3.1 BibTeX to BibTeXML Parser

BibTeX does not have a formally defined grammar, which makes the implementation of a parser difficult. Beebe [1], who works on BibTeX tools since 1990, created a prototype BibTeX grammar based on the available documentation and various experimental tests. Our parser is built on top of the *bibparse* tool (a *lex/yacc* generated lexical analyzer) developed by Beebe.

BibTeX entries may contain LaTeX-coded special characters or commands (eg, `\'e`, `\LaTeX{}`) which are (in this LaTeX-representation) not meaningful in XML. To solve this problem, we implemented a *Perl* converter which translates Unicode compliant characters into XML entities (eg, `\'e` → `&#x00E8;`) and inserts a special XML element for not Unicode-coded strings (eg, `\LaTeX{}` → `<tex code="\LaTeX{}">LaTeX</tex>`). This solution provides full LaTeX backwards compatibility and the possibility to have an XML-compliant representation.

### 3.2 BibTeXML Format

BibTeX ignores unknown fields, which makes it possible to extend the set of usable tags. BibTeXML provides two different *XML schemas*. The first one defines all standard BibTeX fields (eg, *author*, *title*, *editor*, ...), while the second defines non-standard extensions (eg, *ISBN*, *URL*, *abstract*, *language*, ...). This allows us to expand the BibTeXML capabilities maintaining full compatibility with the original format.

A typical BibTeX entry looks like this:

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

After the conversion, the BibTeXML entry has the following format:

```
<bibliography>
  <bibitem type="book">
  <label>lamport:86</label>
    <author>
      <firstname>Leslie</firstname>
      <lastname>Lamport</lastname>
    </author>
    <title><tex code="\LaTeX{}">LaTeX</tex>:
      A Document Preparation System</title>
    <publisher>Addison-Wesley</publisher>
    <year>1986</year>
  </bibitem>
</bibliography>
```

BibTeXML provides a macro environment helping to reduce the amount of stored data and make it more manageable. For example, each author may be stored only once, and individual author entries may reference it, enabling a simple implementation of complex queries and easy navigation within the database. Furthermore, every feature of BibTeX's macro mechanism is kept in the BibTeXML macro format. Additionally, it is possible to define more structured macros as shown in the following example:

```
<macro id="450">
  <firstname>Leslie</firstname>
  <lastname>Lamport</lastname>
  <email>lamport@pa.dec.com</email>
</macro>
...
<author id="450">
...
```

Macros can be referenced with a simple attribute inside the element.

### 3.3 BibTeXML to BibTeX Conversion

BibTeX processes BibTeX-formatted entries only, and it is therefore necessary to convert BibTeXML data into the BibTeX format. The XML structured bibliography is easily transformed into a `.bib` file using XSLT (see Figures 1 and 2). This can be locally done using the BibTeXML2BibTeX XSLT style sheet. Thanks to the XML environment, it is also possible to easily translate BibTeXML entries into various formats (eg, HTML or plain text) using easily adaptable XSLT style sheets.

### 3.4 XLinkbase

While BibTeXML is intended to be used as an exchange format and a way to represent BibTeX in XML, we see it as an intermediary format for our *XLinkbase* system, which is designed for managing large amounts of highly interlinked information, using a data model similar to *Topic Maps* [2]. XLinkbase makes it possible to easily browse and manipulate BibTeX entries, while BibTeXML is used as import and export format for the system.

Figure 2 shows the overall model of interaction, where users are working on the XLinkbase data, and only export BibTeXML if it is required for processing with the BibTeX program, or for exchange or other purposes (such as generating an HTML list of a number of bibliographic entries).

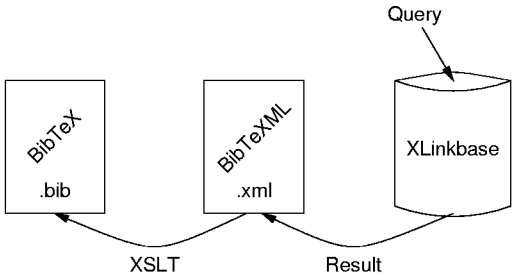


Figure 2: XLinkbase and the BibTeXML format

## 4 bibtexml.org

On `http://bibtexml.org/` it is possible to use and test some of the presented utilities and find more detailed information regarding the structure and the implementation of the BibTeXML project. In particular, it offers the following services:

- Check the correctness of a BibTeX file.
- Convert BibTeX to BibTeXML and vice versa.
- Download the conversion utilities.
- Get the BibTeXML XML Schema definitions and XML Namespace definitions.
- Browse through a sample BibTeXML XLinkbase.

We are also working on setting up a repository of BibTeXML bibliographic entries in the near future, which can then be searched using a subset of XPath expressions.

## 5 Conclusions

In this poster we describe BibTeXML, an XML syntax for BibTeXML data. The not formally defined BibTeX grammar does not allow an efficient data manipulation. Formatting bibliographies entries with XML enable us to obtain a clean and effective structure, thus making possible the creation of more powerful bibliographic databases and their manipulation using general-purpose XML-tools. The conversion tools that we have developed allow the translation of available BibTeX collections into the new format, and the BibTeXML database can also be translated to the original BibTeX-compatible format.

## 6 Acknowledgments

We would especially like to thank Nelson Beebe for the valuable feedback.

## References

[1] Nelson F. H. Beebe. Bibliography Prettyprinting and Syntax Checking. *TUGboat*, 14(4):395–419, December 1993.

[2] International Organization for Standardization. Information technology – SGML Applications – Topic Maps. ISO/IEC 13250, 2000.

[3] Leslie Lamport. *LaTeX: A Document Preparation System.* Addison-Wesley, Reading, Massachusetts, 1985.

[4] Oren Patashnik. BibTeXing. Technical report, February 1988.