CITEPROC

BRUCE D'ARCUS

Abstract. CiteProc is a comprehensive solution for citation coding and formatting. It uses standard XML and XSLT tools and standards to provide a rich, extensible and portable system suitable for a wide range of scholarly needs, and easy to integrate into a wide range of contexts; everything from traditional XML toolchains built around DocBook or TEI, to GUI application like Word and OpenOffice, and web applications. Providing these tools as modules that communicate over standard http, it should facilitate innovation in the realm of scholarly software, and the workflows built around them.

Formatting citations and bibliographies is an essential task for students, researchers, and scholars. Yet the software that serves this need consists of either monolithic commercial applications based on proprietary data formats, or BibTeX , which is limited to TeX.

Exploiting XML-related document formats and processing technologies like XSLT, as well as standard communication protocols like http, CiteProc offers a focused set of tools and standards to encode and format citations and bibliographies independent of specific bibliographic database solutions or document systems. By adopting a unique modular approach, this projects seeks to spur innovation in open source bibliographic software more broadly.

Existing Solutions

The Black Box. The commercial market is dominated by ISI Researchsoft, which now owns three major reference manager products: ProCite, Reference Manager, and Endnote. Beyond the virtual-monopoly position of the company and its troubling impacts on the quality of the software it releases, all of these products are proprietary black boxes. Everything from database formats, to style files, to remote connection configuration files are all proprietary binary formats that are totally opaque, and impossible to extend.

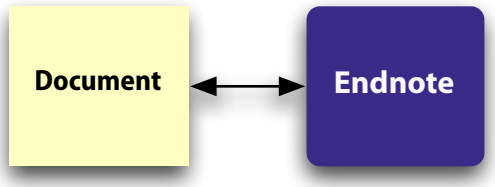

Figure 1. Processing Structure of Traditional Applications

Date: 2004.

Open But Dated. BibTeX has been the most obvious alternative to the commercial products. While quite popular, particularly in the hard sciences and math, BibTeX is limited in the following ways:

(1) it has a poor data model
(2) it lacks international support
(3) it is difficult to parse
(4) its styling language is difficult to work with
(5) is not designed for a networked world
(6) is limited to TeX

A Flexible, Modular and Free Alternative

CiteProc adopts many of the best lessons from existing solutions, but also seeks to improve on them. While broadly similar to the approach of BibTeX , CiteProc has the following advantages:

(1) has a much richer internal data model
(2) is international-ready
(3) is based on XML, so easy to process with widely available tools
(4) its styling language is an easy-to-use XML dialect
(5) communication with databases happens over http
(6) in theory, it should be able to support any text document format

CiteProc thus takes the basic principles of BibTeX but dramatically improves on it by exploiting broader advances in structured markup and text processing that have happened since the initial release of BibTeX roughly 15 years ago.

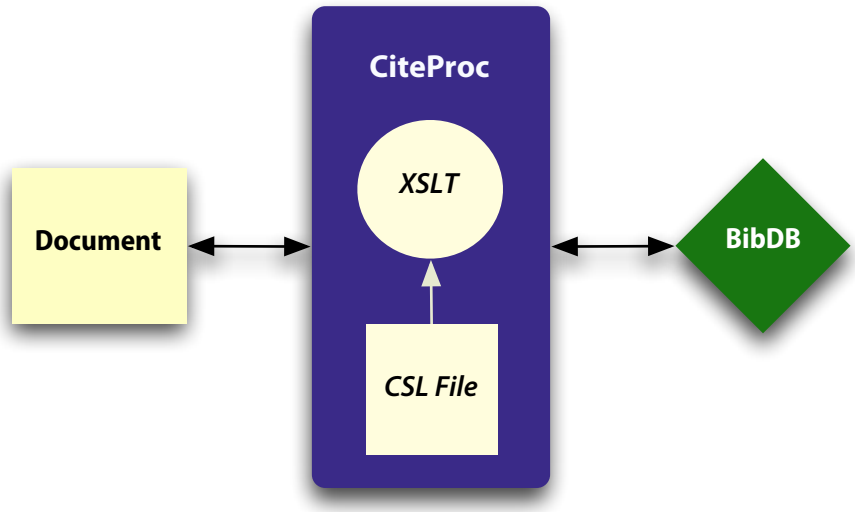

Figure 2. CiteProc Processing Structure

In contrast to commercial applications, CiteProc is free software, is modularized for integration into a wide-range of contexts, and uses open standards for virtually everything: data, processing engine, configuration files, and interapplication communication.

CiteProc Compatability

For a database project to be compatible with CiteProc, it needs primarily to be able to accept a CQL query over http and return a collection of bibliographic records in response. Those records must conform to the MODS XML Schema from the Library of Congress.