

Consistent Bibliographic Data Formats with the BabbleNewt Project

S. Koby Taswell, Aniruddh Anand, Max Montes-Soza, and Carl Taswell

Brain Health Alliance, Ladera Ranch, CA, USA

BHAVI Symposium online 9 October 2023



A Need for Reference Citation Management

- Proper citation of references should be required for every scholarly research report
- Bibliographic citation reference tools allow for storage and management of bibliographic metadata
- Citation analyses evaluate reports and their metadata to calculate a variety of metrics most importantly for plagiarism detection
- Automated plagiarism detection will rely on natural language processing (NLP) and artificial intelligence (AI) methods
- But AI is susceptible to 'Garbage In, Garbage Out' (GIGO)

Welcome to TeX

- Technical STEM fields use TeX and LaTeX for document typesetting instead of Microsoft Word, Apple Pages, or other word processors
- Authors who use LaTeX often also use an integrated writing environment, such as [TeXstudio](#), for preparing manuscripts and formatting reference citations with the data format BibTeX and executable utilities bibtex or bibtex8 as parsing tools to process the source *.bib reference files
- As an extended version of BibTeX, a more recent format BibLaTeX which pairs with the tool biber, now compete with the older versions but many publishers still depend on those older versions
- This presentation was written with LaTeX and BibLaTeX, but our work has been guided by the principle of maintaining both backward and forward compatibility with diverse bibliographic data formats.

General Work Pipeline with LaTeX

- LaTeX with manual bibliography formatting
 - ① Author writes the paper in LaTeX markup language
 - ② LaTeX compiler parses the LaTeX markup and returns a typeset document
- LaTeX with automated bibliographic formatting
 - ① Author writes paper in LaTeX markup language referencing citations in a BibTeX or BibLaTeX file
 - ② LaTeX compiler parses the LaTeX markup returning a first pass on typesetting, also returning a temporary file with cited references in a temporary file
 - ③ bibtex, bibtex8, or biber parses the BibTeX or BibLaTeX file returning the cited references formatted as defined by the bibliography style (i.e. APA, MLA, etc)
 - ④ LaTeX compiler parses again, including the style-formatted references from the temporary file, returning a complete typeset document

Anatomy of a BibTeX Entry

```
@article{Patashnik1998biblatex,  
  author   = "Oren Patashnik",  
  journal  = "TUGboat",  
  number   = "2",  
  pages    = "204-207",  
  title    = "BIBTEX 101",  
  volume   = "19",  
  year     = "1998",  
}
```

A Brief History of LaTeX

- TeX originated in 1978, written by Donald Knuth (Knuth [1984](#))
- LaTeX built on foundation of TeX in 1984, written by Leslie Lamport (Lamport [1986](#))
- Later that year, Oren Patashnik worked with Lamport to create BibTeX, a reference management system and processor for use in typesetting with LaTeX (Patashnik [1984](#))
- Several re-implementations of BibTeX now exist such as BibTeXu, CL-BibTeX, MLBibTeX, Bibulous, and BibLaTeX
- Each variation has its own changes differing from the original BibTeX
- BibLaTeX may also refer to the data format paired with a parsing utility called biber, created in 2006 by Philipp Lehman, maintained by Philip Kime and Moritz Wemheuer (Lehman et al. [2020](#))

Current Challenges with BibTeX

- Inconsistencies in acceptable format and parsing across re-implementations such as:
 - some implementations require for comma after last attribute name-value pair, some do not
 - some implementations accept both “attribute-value” and {attribute-value}
- Variable tolerance of BibTeX and LaTeX reserved characters (i.e. %, , \$, &, and others)
- Commas in attribute fields can cause issues when not formatted correctly
 - double commas: “, ,”
 - improperly spaced commas: “and ,”
 - double commas specifically before or after the word “and”
- Many more issues can be found on Stack Overflow threads and Git error requests

Format Design Considerations

- Simplify the data format to make it easier and faster to read a data record for both computers and humans with fewer errors
- Increase interoperability with other data formats such as MARC, BibFrame, json, XML, and more
- Multiple perspectives of **defensive design** and **defensive coding** to prevent read-parse-write errors:
 - Simplify data structures: minimize probability of errors caused by poorly and/or inconsistently formatted data
 - Simplify parsing algorithms: minimize probability of errors caused by mistakes in regular expressions, lexers, parsers, and flow control structures in programming languages

Project BabbleNewt

- Project BabbleNewt supports interoperability and progressive transition between 5 related data formats:
 - ① PdpBibtex (*.pbtx) similar to the original definition of BibTeX by Patashnik
 - ② PdpBibtexgen (*.pbtg) generalized version of BibTeX with some changes to avoid issues with certain characters
 - ③ PdpBiblatex (*.pbtl) similar to the original definition of BibLaTeX by Kime, Wemheuer, Lehman
 - ④ PdpBiblatexgen (*.pblg) generalized version of BibLaTeX with some changes to avoid issues with certain characters
 - ⑤ PdpBabblenewt (*.pbbn) simplified consistent data format designed to be interoperable with all other bibliographic data formats
- Project BabbleNewt introduces unconstrained lists of entity-attribute names which can be used to interoperate between BibTeX, BibLaTeX, and other formats such as BibFrame which may have arbitrarily defined lists

Comparison of Format Definitions

Format	Opener	Name-Value Pair	Closer	Attribute List
pbtx	@Etyp{Ekey,	Anam= "Aval",	}	specified
pbtg	@Etyp{Ekey,	Anam= {Aval},	}	unconstrained
pblt	@Etyp{Ekey,	Anam= {Aval},	}	specified
pblg	@Etyp{Ekey,	Anam= [Aval],	}	unconstrained
pbbn	@{	Anam= [Aval],	}@	unconstrained

Comparison of Format Timings

	Bibtex	Bibtexgen	Biblatex	Biblatexgen	Babblenewt
Initialization	.37s	.36s	.33s	.40s	.32s
8 records	.38s	.38s	.35s	.42s	.33s
80 records	.61s	.55s	.57s	.61s	.52s
800 records	2.62s	2.07s	2.35s	2.36s	1.95s
8000 records	20.99s	16.52s	18.90s	16.95s	14.75s


Future Directions


- Implement an efficient bibliographic parsing tool for the PdpBabblenewt format analogous to biber for BibLaTeX
- Integrate PdpBabblenewt tools into our PDP-DREAM software for the NPDS cyberinfrastructure platform
- Build out the FAIR Metrics for Plagiarism Detection with NLP and PdpBabblenewt in PDP-DREAM and NPDS.


Contact Info


- ktaswell@bhavi.us
- www.BHAVI.us
- www.BrainHealthAlliance.org

References

 Knuth, Donald E. (1984). *The T_EXbook*. Addison-Wesley.

 Lamport, Leslie (1986). *L^AT_EX: A Document Preparation System*. Addison-Wesley.

 Lehman, Philipp et al. (Dec. 31, 2020). "The biblatex Package". In.

 Patashnik, Oren (1984). "BIBTEX 101". In: *TUGboat*.