Adding multilingual capabilities to bibliography-formatting software

Marius L. Jøhndal
mariuslj@ifi.uio.no
Department of Informatics
University of Oslo
Norway

Abstract

Bibliography-formatting software has been widely available for a long time, and can be successfully used, even in a primitive form, by authors and editors to assist in formatting citations and reference lists for use in scientific texts. The available tools are, however, still not easily used with non-English languages, and mixing multiple languages or using several scripts concurrently might give unsatisfactory results.

This paper describes some of the difficulties that arise in multilingual bibliography-formatting software, and briefly describes a programmable bibliography processor that aims to tackle these problems.

1 Introduction

Citations, reference lists and bibliographies are an integral part of most forms of scientific publication. Unfortunately for authors and editors, producing bibliographies can be a time-consuming and error-prone process. Being consistent about citations can be hard, an existing bibliography is not easily converted to a different style, and it is difficult to maintain sets of references that are usable in multiple documents [8].

To assist authors and editors in composing bibliographies, various kinds of bibliography software have been developed. Already in 1994, more than forty programs, predominantly commercial, existed for the purpose of formatting bibliographies [11]. Many of these still exist, in addition to a growing number of open-source projects.

Bibliography tools primarily serve two purposes: *Bibliography managers* are designed for the management of bibliography databases, and they ordinarily allow the user to do advanced querying and editing, to share bibliography databases on a network, and to import references from library catalogues or on-line databases like PubMed.

A *bibliography processor*, on the other hand, translates citations and references into a format ready for presentation in publications. Bibliography managers and processors can be combined, and, indeed, most of the commercially available programs, like Reference Manager and ProCite, aim to provide the combined functionality of managers and processors.

2 Bibliography styles

At first glance, a bibliography processor appears to be a simple and straightforward piece of software. The apparent simplicity diffuses when one realises the diversity of bibliographic data, the plethora of bibliography styles, and the range of linguistic and typographic conventions that a bibliography processor is expected to handle.

Both the techniques used for inserting citations in the text and tying these to reference information vary considerably. The details of bibliography styles are usually specified in instructions to authors by publishers or editors, but these instructions rarely diverge significantly from the general style used in publications within the discipline. Most style guides and standards (e.g. [2–3, 7, 10]) use a threefold division of styles, albeit with varying names.

In the *key systems*, citations are inserted parenthetically in the text and consist of one or more keys, usually numbers, that identify a particular reference in an accompanying reference list (example taken from [13]):

> ”Such a proof is unavailable for gapped local alignments, but computational experiments strongly suggest that the same type of distribution applies[10].”

The reference list itself is sorted by increasing keys:

> **10** Altschul, S.F. and Gish, W. (1996), *Methods Enzymol.* 266, 460–480

The assignment of keys to references is done by first ordering the references either by author or by order of citation, and then assigning the keys sequentially.

The *author-date systems* are similar, but slightly more complex, since the citations are made up of the author's name and the year of publication, and thus introduce data from the references in the citations (example from [15]):

> ”[...] anti-anxiety drugs become less potent the longer you take them, and they probably are addictive (Tinklenberg, 1977; Olivieri, Cantopher, and Edwards, 1986; Nagy 1987; Roache, 1990).”

Reference lists are sorted by author's name and year of publication.

The most challenging group of systems are the *author-title systems*. Citations are placed either in the running text or in notes, and they may or may not rely on separate lists of references. In case no reference list is used, the first citation has to contain the full reference to the work. In other cases, each citation is a brief form of the reference, built around the author's name and the title of the work (example from [14]):

> 5. *The Collected Works of William Morris*, London, 1915, xxiii, p. 147.
> 6. Ibid., 1914, xxii, p. 26.
> 7. Mackail, op. cit., ii, p. 99.
> 8. *Coll. Works.*, xxii, p. 42; xxiii, p. 173.
> 9. Ibid., xxii, pp. 47, 50, 58, 73, 80, etc.

When deciding on the implementation details of a bibliography processor, further classification of the bibliography systems is required, since there is considerable variation within each of these systems in layout and typographic conventions. A classification using five systems is often used [1, 9].

3 Bibliographic data

The archetypal examples of bibliographic data are books and journal articles. Each discipline, however, has its specialities: Engineers may cite standard documents with a standard number, geographers may cite cartographic material with a scale, and art historians may refer to paintings on a certain material, in a particular format and located in a specific museum. Furthermore, publications could be translations, reprints, unpublished or fragmentary versions, authors can be unknown or known only by pseudonym, or parts of works can be pinpointed using unusual division systems such as the number of a scene or a list of stanzas.

Further complications include the use of multiple reference lists, annotated bibliographies and reference lists divided into thematic sections [9].

4 Existing bibliography processors

Existing bibliography processors handle the conventional bibliography styles and the most frequent publication types quite well, although few programs handle all variations with equal success. Multilingual bibliographies are, however, usually not supported at all.

The file formats used by these programs are simple, text-based formats lacking formal specifications, and often vague on issues such as encoding of characters and proper separation of content from presentation.

The chief goal of this project has therefore been to study the implications of using multilingual bibliographies in a bibliography processor, without limiting the functionality of this to one particular style or a narrow choice of data.

An important implication of this is that the processor should be extensible, so that functionality required by peculiar citation styles, unforeseen bibliographic data or languages with particular needs can be added at a later stage. Also, although the focus is on mark-up languages like LaTeX or XML, the design should not preclude support for systems operating under different principles, e.g. word processors.

5 Multilingual references

In a document written in English one might find a reference such as this:

> [27] Brian W. Kernighan and Denis M. Ritchie: *The C Programming Language.* 2nd edition. Prentice Hall, 1988.

If one, on the other hand, is writing a document in French, one may choose to render the above reference in two different ways. The first alternative, the *reference-dependent* approach, is to keep the reference as it is, i.e. to use the language of the work referenced. The other alternative, the *document-dependent* approach, uses the language of the document in the reference [6]. This might look like this:

> [27] Brian W. Kernighan et Denis M. Ritchie : *The C Programming Language.* 2^{e} édition. Prentice Hall, 1988.

The reference-dependent approach seems to be unusual in English publications, but is frequent in Russian, for example. The rationale behind the approach is that the referenced work is accessible only to readers of the language in which the work is written.

Examining the French translation, we see that the keywords *and* and *edition* have been translated, and the adjectival ending of the ordinal number is both translated and written as a superscript. When translating references in this manner, the following content alterations may occur:

- **Keywords.** Any keyword that occurs in a particular style may require translation. In addition to the above examples, such keywords include *translator*, *reprint*, *volume* etc. These words often occur in one or more distinct abbreviated forms, e.g. some English-language styles use *ed.* for the keyword *editor* and *eds.* for the plural form; others use *ed.* in both numbers.

- **Names of places.** Names of places depend on language and on historical context, and at least the more common cases should be translated by the processor, like English *Florence* for Italian *Firenze.* Place names may also require qualification: The Norwegian city *Bodø* needs no further specification in a Norwegian text, but the name of the country should be appended in an English text. Furthermore, it is customary in some languages to abbreviate the names of certain cities, e.g. *СПБ.* for *Saint Petersburg* in Russian.

- **Dates.** Dates may occur in many different forms, e.g. *16. Jahrhundert* or *Spring 1576*, but even the standard dates using year, month and day cannot be translated word by word, e.g. 17th June 2000 is 17 հունիսի 2000թ in Armenian and *2000. június 17-e* in Hungarian.

- **Numerals.** Numerals (and other words indicating quantities) may require inflection of succeeding words. This is evident in English as the opposition between the singular and plural form, as in *1 volume* and *3 volumes.* Other languages distinguish further forms, such as case inflected forms in Slavic languages.

- **Transliteration and translation of names and titles.** Names and titles may be transliterated or translated if written in a language or using a script that is considered unfamiliar to the reader (example from [12]):

  Alemayehu, N. (1999). *Development of a Stemming Algorithm for Amharic Language Text Retrieval.* Ph.D. Thesis, University of Sheffield.

  Amare, G. (1990EC). ዘመናዊ የአማረኛ ሰዋሰው በቀላል አቀራረብ. አዲስ አበባ፡ ንግድ ማተሚያ ቤት. (Zemenawi yeamareNa sewasew beqelal aqerareb. Adis Abeba: ngd matemiya bEt.)

There are, in addition to the language used for content, several linguistic and typographic parameters that may change at any point within a reference:

- **Punctuation.** Usage and choice of punctuation varies between languages. Note for example the subtle change in punctuational conventions in the French reference example above: The space surrounding the colon is greater than in the English version. Another type of variation is the choice of the actual symbol used for a particular purpose (example taken from [5]):

American English	“Brand of the Werewolf”
British English	’Brand of the Werewolf’
French	« La marque de la bête »
German	„Im Zeichen des Werwolfs”

- **Hyphenation rules.** In the English reference example above, English hyphenation rules are used for the entire reference. In the French translation, it is necessary to alternate between English hyphenation rules for the names and the title, and French hyphenation rules for the rest of the reference.

- **Script.** Different languages employ different writing systems, and thus the script may change in the middle of a reference (example from [16]):

  > **89.** Baade W *Evolution of Stars and Galaxies* (Ed. C Payne-Gaposchkin) (Cambridge: Harvard Univ. Press, 1963) [Бааде В *Эволюция звезд и галактик* (М.: Мир, 1966); 2-е изд. (М.: УРСС, 2002)]

  The document-processing system may require special instructions for this to work (e.g. change of fonts or font encoding).

- **Ligatures.** Related to the choice of script is the use of aesthetic ligatures.[1] An aesthetic ligature like *ﬂ* may be used in English, but should never be used in for example a Portuguese text. In most Germanic languages, English included, ligatures are prohibited in particular cases, usually at the joining point of word components, like in the German word *Auflage*.

- **Typefaces.** There may also be preferences related to the choice of typeface or type style, e.g. Chinese characters are used both in mainland China and in Taiwan, but the former uses simplified characters (简体字 jiǎntǐzì) and the latter traditional characters (繁体字 fántǐzì).

- **Writing direction.** Writing direction becomes an issue when using languages like Arabic or Hebrew. Switching between left-to-right and right-to-left languages is unfortunately not always a simple matter and may lead to complex cases.

There are also a few more general language-related issues:

- **Sorting order.** Reference lists may be sorted, but the rules of alphabetising depend on language. Not only does the order of individual letters differ from language to language, but there may be letter sequences that count as single symbols during alphabetising, e.g. the digraph *dž* in Serbo-Croatian. Also, several letters may sort as one and the same symbol, e.g. the letters *u* and *ü* in German. There may even be several competing sorting systems, and some may require dictionaries. Conventions regarding the treatment of spacing and punctuation in sorted strings also become a factor, and there may be words or prefixes that are commonly ignored when sorting. Finally, multilingual bibliographies frequently lead to lists in which multiple scripts occur, and it is not always obvious how such mixes should be treated.

- **Personal names.** Personal names (anthroponyms) are particularly important in bibliographies, and since names appear in several forms in both citations and reference lists, a bibliography processor has to be able to distinguish the parts of a name. There are, however, a large number of name systems, and even

[1] In contrast to linguistic ligatures, which are essential to the writing of a particular language, or contextual ligatures, which are part of the script itself [4].

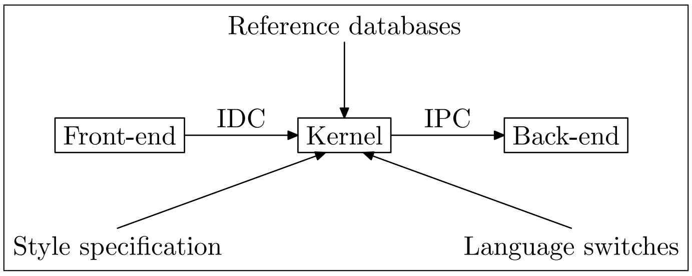


Figure 1: The principal components of the bibliography processor. IDC and IPC are the intermediate document and presentation codes.

> "simple" systems like Western names can prove troublesome. Particles like *de* and *von* are usually treated as part of the surname in English, but in most of the languages in which these particles originate, they are not. Other examples include the traditional Spanish family names, which consist of both the father's and the mother's family name. Sorting is, however, done exclusively using the father's family name. In a similar fashion, Icelandic patronymics count as surnames, but alphabetising is done using the first name. In many Asian languages, the family name precedes the personal name, and when used with a bibliography style that presents names on an inverted form, such names need special treatment, since the parts of the names should not be transposed.

6 System abstraction

As document processing and typesetting systems vary widely in their representations of in-text citations and provide different mechanisms for inclusion of foreign data, a bibliography processor would benefit from a flexible abstraction insulating it from the document processing or typesetting systems.

To accomplish this, the processor is organised as a *kernel* that interfaces the document processing system using *front-ends* and *back-ends* communicating using *intermediate document code* and *intermediate presentation code* (see figure 1). The intermediate codes are abstract representations of source documents and the presentation-ready data.

The front-ends take care of the parsing of relevant document source files and extract citations and information pertaining to reference lists. They also determine a set of language variables that are used to ensure that language dependent information is properly processed.

The kernel consumes intermediate document code and produces intermediate presentation code. In this process, references are taken from a set of *reference databases*, and *locale functions* are consulted for language-dependent information. Particular care has to be taken to ensure that the correct language attributes are emitted for the various parts of a reference list or citation.

The back-ends translate intermediate presentation code to formats usable by the target systems. This is a relatively straightforward process, but advanced typesetting systems like TeX require detailed instructions and some support code for optimal results.

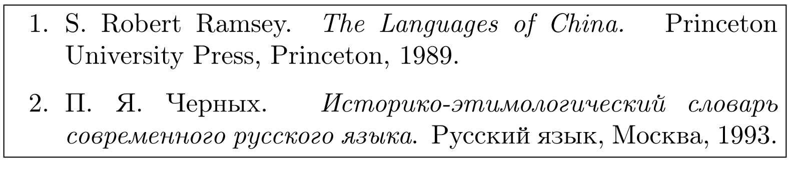
Examples of intermediate document code, intermediate presentation code for a numeric bibliography style and the resulting reference list generated by the LaTeX back-end, are shown in figures 2, 3, and 4 respectively.

```
<?xml version='1.0' encoding='utf-8'?>
<idc>
 <reflist id="rl:main" style="numeric-simple" lang="en"/>

 <citation id="cit:1" reflist="rl:main"
   reference="r:tsjernykh93" lang="en"/>
 <citation id="cit:2" reflist="rl:main"
   reference="r:ramsey89" lang="en"/>
</idc>
```

Figure 2: Example of intermediate document code.

```
<?xml version='1.0' encoding='utf-8'?>
<ipc>
 <reflist entries="2" id="dictionaries" lang="en">
  <reference key="ramsey89" lang="en">
   <label>1.</label>
   <text>
    <block>S. Robert Ramsey.</block>
    <block><i>The Languages of China</i>.</block>
    <block>Princeton University Press, Princeton, 1989.</block>
   </text>
  </reference>
  <reference key="tsjernykh93" lang="en">
   <label>2.</label>
   <text>
    <block><div lang="ru">П. Я. Черных</div>.</block>
    <block><i><div lang="ru">Историко-этимологический словарь современного
       русского языка</div></i>.</block>
    <block><div lang="ru">Русский язык</div>,
       <div lang="ru">Москва</div>, 1993.</block>
   </text>
  </reference>
 </reflist>

 <citation id="cit:1" lang="en"><super>2</super></citation>
 <citation id="cit:2" lang="en"><super>1</super></citation>
</ipc>
```

Figure 3: Example of intermediate presentation code.



1. S. Robert Ramsey. *The Languages of China.* Princeton University Press, Princeton, 1989.
2. П. Я. Черных. *Историко-этимологический словарь современного русского языка.* Русский язык, Москва, 1993.

Figure 4: Example of a reference list formatted by the LaTeX back-end.

7 Reference databases

Non of the established data formats for reference databases were designed to keep multilingual bibliographies. A new XML-based format was designed to rectify this. XML was the obvious choice for this for a number of reasons, most importantly the support for Unicode and its ability to represent information hierarchically.

The DTD of the format was designed as a superset of bibliography formats known to the author, and many rare data fields not usually used by bibliography styles, but useful in special cases, are available. Particular care has been taken to ensure that personal names can be represented properly, and that any textual information can be transliterated or translated.

8 Style specifications and style engines

Bibliographic styles are described in *style specifications*. These are XML files that select a particular *style engine* and supply it with a set of style parameters. The style engines constitue the programmable part of the processor, and there are style engines for each of the major bibliography style systems. These are designed to be usable with almost any bibliography style by adjusting only the style parameters. Typical parameters include such things as what delimiters to use in citations or the limit on how many authors to list in a reference.

9 Availability

The implementation of the multilingual bibliography processor described in this paper is called `ibibproc` and is available from `http://ibibproc.sourceforge.net`. The system is implemented in Perl and relies on the Unicode support found in version 5.6 or newer.

Currently included are front-ends and back-ends for DocBook XML, plain TeX and LaTeX, in addition to separate back-ends for plain text files and XHTML. There are also scripts for conversion between BibTeX and the `ibibproc` database format, style specifications that emulate the standard BibTeX styles, and a drop-in replacement script for the BibTeX program.

10 Conclusion – Future work

The project has revealed that adapting the ideas of a bibliography processor to bibliographies with multiple languages, is more complex than it seems, and several *ad hoc* solutions are required. Furthermore, it is not always possible to be consistent in the distinctions between presentation and data content. It is, however, in most cases possible to process multilingual bibliographies without manual intervention.

The implementation is still immature, and several components are missing. Support for more languages, styles and target systems, in addition to conversion utilities for additional formats, will be needed if `ibibproc` shall become a successor to current bibliography processors.

Lately, several new projects have appeared that are related to `ibibproc`, and merging `ibibproc` with these efforts may give the project new momentum. Of particular interest is the `bibx` project, which develops a DTD for multilingual reference databases, and the bibliography processor project `Bibulus` with a design similar to `ibibproc`.

References

[1] Bennett, Frank G., jr. Camel: Kicking over the bibliographic traces in BibTEX. *TUGboat*, 17(1): 22–28, March 1996.

[2] British Standards. *British Standards 5605: Recommendations for Citing and Referencing Published Materials.* BS 5605:1990, British Standards, Milton Keyns, Great Britain, 1990.

[3] *The Chicago Manual of Style.* 14th ed., University of Chicago Press, 1993.

[4] Haralambous, Yannis. Tour du monde des ligatures. *Cahiers GUTenberg*, 22: 87–99, September 1995.

[5] Hufflen, Jean-Michel. Multilingual Features for Bibliography Programs: from XML to MlBibTEX. In *EuroTEX 2002: Proceedings of the Thirteenth European TEX conference*, Bachotek, Poland, May 2002, pp. 46–59.

[6] ——. MLBIBTEX: a New Implementation of BIBTEX. In *EuroTEX 2001: Proceedings of the Twelfth European TEX conference*, Kerkrade, The Netherlands, September 2001, pp. 74–94.

[7] International Organization for Standardization. *ISO 690: Documentation — Bibliographic references — Content, form and structure.* ISO 690:1987 (E). International Organization for Standardization, Geneva, 1987.

[8] Rahtz, Sebastian. Bibliographic Tools. *Literary and Linguistic Computing*, 2(4): 231–241, October 1987.

[9] Rhead, David. The "operational requirement" (?) for support of bibliographies. *TUGboat*, 14(4): 425–432, December 1993.

[10] Ritter, R. M. *The Oxford Guide to Style.* Oxford University Press, 2002.

[11] Stigleman, Sue. Bibliography Formatting Software: An Updated Buying Guide. *Database*, 17(6): 53–65, December 1994.

Example references

[12] Alemayehu, Nega and Peter Willett. Stemming of Amharic Words for Information Retrieval. *Literary and Linguistic Computing*, 17(1): 1–17, April 2002.

[13] Altschul, Stephen F.Fundamentals of database searching. In *Trends guide to bioinformatics*, Elsevier Science, 1998, pp. 7–9.

[14] Pevsner, Nikolaus. *Pioneers of Modern Design.* Penguin Books, London, 1991.

[15] Rosenhan, David L. and Martin E. P. Seligman. *Abnormal Psychology.* 3. ed., W. W. Norton, New York, 1995.

[16] Ефремов, Ю. Н. and А. Д. Чернин. Крупномаштабное звездообразование в галактиках. *Успехи физических наук*, 173(1): 3–25, January 2003.