GBV VZG

Verbundzentrale des GBV

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
ELECTRICAL ENGINEERING & COMPUTER SCIENCE, UNIVERSITY OF KASSEL

Bibliographic Hash Keys

Mapping Bibliographic Records

Jakob Voß[1], Andreas Hotho[2], Robert Jäschke[2]

This poster presents a set of hash keys for bibliographic records called bibkeys. Unlike other methods of duplicate detection, bibkeys can directly be calculated from a set of basic metadata fields (title, authors/editors, year). It is shown how bibkeys are used to map similar bibliographic records in BibSonomy and among distributed library catalogs and other distributed databases.

Motivation

To check whether two citations or bibliographic records refer to the same publication, either manual work or unique identifiers or good heuristics of duplicate detection are needed. Centralized identifiers (ISBN, DOI, LCCN etc.) cannot be derived from other metadata fields but must be looked up. Other systems like OpenURL [1] and SICI [2] require detailed and clean metadata which you rarely find in normal citations. Methods of duplicate detection are common in digital libraries but they mostly build on direct comparisons of full records or multiple comparisons of minimal distances of multiple signatures [3]. Methods or FRBR work detection use similar methods or they are bound to specific bibliographic record formats or authority files [4]. In Contrast bibkeys can be applied by anyone who knows the authors (or editors), title, and year of a publication. An important feature of bibkey is that records are matched without having to directly compare them. Instead a bibkey is calculated by a simple method for each record and can directly be matched .

Examples

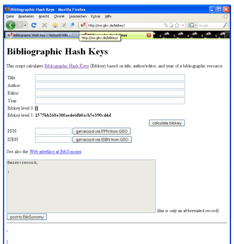
Given the book with authors "Trudi Bellardo Hahn and Charles P. Bourne", title „"A History of Online Information Services, 1963-1976", published in 2003, these metadata fields are joined to a string (bibkey level 0) and its checksum to bibkey level 1:

bibkey level 0: ahistoryofonlineinformationservices19631976 [t.hahn,c.bourne] 2003

bibkey level 1: 14ed100f75dd4459cffeb272bdbc2d1e7

Author names are abbreviated by splitting the names into tokens at white spaces. If the first and the last token are equal, this is returned once as surname. Otherwise the the first character of the first token (given name) followed by a dot is prefixed to the last token (surname). Some examples of both cases:

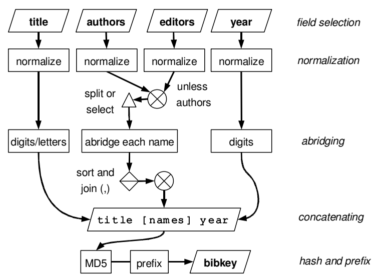
"knuth knuth"	"knuth"
"knuth"	"knuth"
"donald e. knuth"	"d.knuth"
"d.e. knuth"	"d.knuth"
"donald knuth"	"d.knuth"



check out http://ws.gbv.de/bibkey/

Specification

A general bibkey is based on four metadata fields: title, authors, editors, and year. The editors field is only used if no authors are given. First all fields are normalized. The authors/editors field is either split into single names or the first author is selected. Names are abridged, sorted and joined to a comma-separated list. Year and title are reduced to digits or digits and letters. Finally the fields are concatenated as title + " [" + names + "] " + year. Either this string is used or the MD5 Message-Digest Algorithm checksum [5] of its UTF-8 representation plus a prefix makes a hashed bibkey.



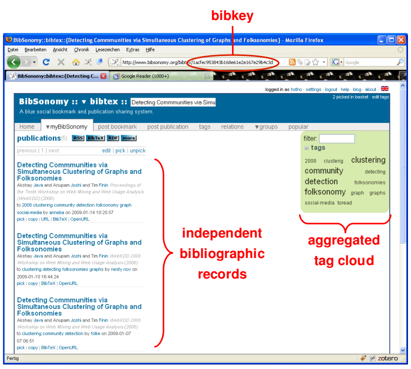
For a specific bibkey version, normalization, abridging, sorting, concatenating, and the prefix need to be defined. Bibkey level 1 uses Unicode Normalization Form Compatibility Composition (NFKC) [6], case folding to lower case, and replacement of white spaces with one space, except at the beginning and and of a string as normalization. All author names are used and the abridging method is shown in the examples-section above. The reduction to letters and digits in the title respects all Unicode letter. The prefix that is added to a MD5 checksum is "1", so in total a bibkey has 33 characters from a-z, 0-9. The latest specification is available online as well as reference implementations in Java and Perl.

Specification
http://www.gbv.de/wikis/cls/Bibliographic_Hash_Key

Reference implementations
http://ws.gbv.de/bibkey/
http://www.bibsonomy.org/help/doc/inside.html

Usage

Bibkey version level 1 is used as "interhash" by the social bookmarking application BibSonomy [7] to detect if the same publication has been entered by different users.



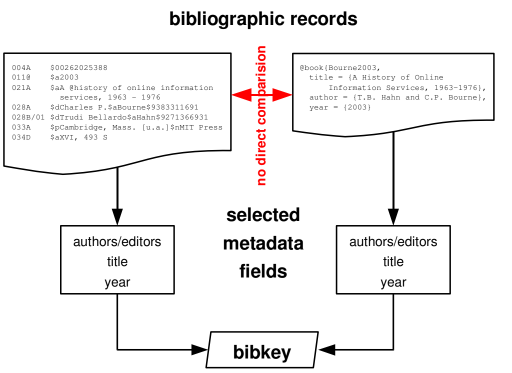
The bag-model of social tagging allows each user to manage its own bibliographic records that then can be aggregated. Other applications can quickly look up by bibkey, whether a given publication has already been entered in BibSonomy. For this purpose BibSonomy provides a JSON API and VZG provides a wrapper to the SeeAlso Linkserver protocol [8]. The Kölner Universitäts Gesamtkatalog (KUG) indexes its records with bibkey and uses it to link BibSonomy. Lookup of records via bibkey in other library catalogs and in the Wikipedia project is planned.

Usage of BibSonomy JSON API
http://www.bibsonomy.org/help/addons/integration

SeeAlso Services at VZG
http://ws.gbv.de/seealso/services/

Use of bibkey in OpenBib/KUG
http://blog.openbib.org/2008/07/15/neue-version-von-kug-und-openbib/

Planned bibliographic record store for Wikipedia using bibkeys
http://de.wikipedia.org/wiki/Benutzer:Duesentrieb/Biblio



Evaluation and Outlook

Each bibkey method defines a binary classifier for duplicate detection of bibliographic records. Thereby two kinds of error exist: first, same publications could be mapped to different keys (false negative) and second, different publications could be mapped to one key (false positive). It turned out that the first error highly depends on quality of the metadata and the definition of "same publication". Sensitivity can further be increased by improvement of the normalization step and by selecting only the first author/editor. A next version of bibkey should for instance normalize by removing all diacritics. Improvement in the abridging step can also help, especially with organizations as authors. Abridging could also include usage of authority files but this would limit the ease of bibkey usage. The second error only occurs in special cases like anonymous works, works without known year and for articles with standard titles like "Introduction", "Book Reviews" or "News" which are frequently found in journals. Further development of bibkeys will aim on testing its benefit for FRBR work detection by removing the year field and on usage of bibkeys as link targets on the Semantic Web.

References

[1] ANSI/NISO. The OpenURL Framework for Context-Sensitive Services. 2004 (Z39.88).

[2] ANSI/NISO. Serial Item and Contribution Identifier. 1996 (Z39.56).

[3] L. Padmasree, V. Ambati, J.A. Chandulal and M.S. Rao. Signature Based Duplication Detection in Digital Libraries. In *Proceedings of the ICUDL*, Alexandria, 2006. http://tera-3.ul.cs.cmu.edu/conference/2006/25.pdf

[4] T. Hickey, J. Toves. FRBR Work-Set Algorithm. OCLC, 2005. http://www.oclc.org/research/software/frbr/frbr_workset_algorithm.pdf

[5] R. Rivest. The MD5 Message-Digest Algorithm. 1992 (RFC 1321).

[6] M. Davis, M. Dürst. Unicode Normalization Forms. Revision 29, 2008-03-28 (Unicode Standard Annex #15). http://www.unicode.org/reports/tr15/tr15-29.html

[7] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proceedings of the CS-TIW*, p 87–102, Aalborg, 2006. http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006bibsonomy.pdf

[8] J. Voß. SeeAlso: A Simple Linkserver Protocol. Ariadne 57, 2008. http://www.ariadne.ac.uk/issue57/voss/