hdblp.xml, a short documentation

Oliver Hoffmann[0000−0002−3808−9042], Florian Reitz[0000−0001−6114−3388]

Schloss Dagstuhl - LZI, dblp group, Wadern, Germany
{oliver.hoffmann, florian.reitz}@dagstuhl.de

Abstract. This document is a short description of `hdblp`, a file which contains the historical data of the dblp collection. `hdblp` can be used to study the historical development of dblp up to December 17 2018.

1 Introduction

The dblp computer science bibliography[1] collects metadata for publications in computer science and related fields[1]. The project was founded in 1993. Since then, dblp has gathered (as of December 2018) data on more than 4.2 million publications written by about 2 million different authors. For each publication and each author, the collection maintains a simple metadata record which contains all descriptive information. The set of all records is available as daily snapshot as well as stable monthly releases. The records are modified frequently to add additional information or to correct defects. In this document, we describe the `hdblp` collection which contains the historical content of each metadata record. For every time that a record is modified, `hdblp` contains a revision of the record's content. Within certain limits, `hdblp` can be used to reconstruct the state of dblp for each day between July 1999 and December 2018. This data can be used to study the development of the collection over a period of more than 19 years. Like the dblp snapshots themselves, `hdblp` is published under the open ODC-BY 1.0 license[2].

2 hdblp

`hdblp` is created from nightly backups of the primary dblp collection. If the content of a record deviates from the content of the last backup, a revision (i.e., a full copy of the content) is created and stored in `hdblp`. As `hdblp` is created nightly, there can only be a single revision of a record per day. If there are multiple modification of a record during the same day, the resulting revision will combine them. The earliest available revision in `hdblp` is from 1995-10-08. However, due to a misconfiguration of the backup system, all dates before 1999-06-02 are unreliable.

[1] `https://dblp.org`
[2] `https://opendatacommons.org/licenses/by/1.0`

dblp provides metadata records for publications and for authors. An example for publication [2]:

```
<article key="journals/jsyml/NewmanT42" mdate="2017-05-28">
  <author>M. H. A. Newman</author>
  <author>Alan M. Turing</author>
  <title>A Formal Theorem in Church's Theory of Types.</title>
  <pages>28-33</pages>
  <year>1942</year>
  <volume>7</volume>
  <journal>J. Symb. Log.</journal>
  <number>1</number>
  <url>db/journals/jsyml/jsyml7.html#NewmanT42</url>
  <ee>https://doi.org/10.2307/2267552</ee>
  <ee>http://projecteuclid.org/euclid.jsl/1183389307</ee>
</article>
```

For each author, dblp maintains a person record. The person record lists synonyms used by the author as well as weblinks, external IDs (such as Wikidata or ORCID) and affiliation data. Person records are handled like publication records. The person record for *Alan M. Turing* is:

```
<www key="homepages/t/AlanMTuring" mdate="2018-01-08">
  <author>Alan M. Turing</author>
  <title>Home Page</title>
  <url>http://www.turingarchive.org/</url>
  <url>https://en.wikipedia.org/wiki/Alan_Turing</url>
  <url>https://www.wikidata.org/wiki/Q7251</url>
  <url>http://dl.acm.org/author_page.cfm?id=81100339134</url>
  ...
  <url>http://www.genealogy.ams.org/id.php?id=8014</url>
  <url>https://zbmath.org/authors/?q=ai:turing.alan-m</url>
</www>
```

Person records for all authors were introduced in June 2009. Before this date, person records were only created in special cases. The structure of both record types is defined in a DTD[3]. A short description can be found in[1].

For each time that a record is modified, `hdblp` contains a full copy of the record's content (a revision). Revisions of the same record are listed consecutively, ordered by date of modification starting with the latest revision. The following example shows three revisions of record `journals/jsyml/NewmanT42`:

```
<article key="journals/jsyml/NewmanT42" mdate="2017-05-28">
  <author>M. H. A. Newman</author>
  <author>Alan M. Turing</author>
  <title>A Formal Theorem in Church's Theory of Types.</title>
```

[3] See `https://dblp.org/xml/dblp.dtd` for the current dtd. A version fully compatible with this data file is provided with the data publication.

```
  ...
  <ee>https://doi.org/10.2307/2267552</ee>
  <ee>http://projecteuclid.org/euclid.jsl/1183389307</ee>
</article>

<article key="journals/jsyml/NewmanT42" mdate="2014-08-05">
  <author>M. H. A. Newman</author>
  <author>Alan M. Turing</author>
  <title>A Formal Theorem in Church's Theory of Types.</title>
  ...
  <ee>http://dx.doi.org/10.2307/2267552</ee>
  <ee>http://projecteuclid.org/euclid.jsl/1183389307</ee>
</article>

....

<article key="journals/jsyml/NewmanT42" mdate="2003-10-13">
  <author>M. H. A. Newman</author>
  <author>A. M. Turing</author>
  <title>A Formal Theorem in Church's Theory of Types.</title>
  ...
  <url>db/journals/jsyml/jsyml7.html#NewmanT42</url>
</article>
```

The `mdate` attribute is the data of the first day where this content was observed. In the example, the `ee` field was modified multiple times. This is by far the most common modification (see below).

Sometimes dblp deletes records. The three main reasons are:

- A duplicate record is removed from the collection.
- An person record is removed after the profile is merged into another profile.
- The publication is withdrawn. Deleting withdrawn publications has been discontinued as of 2016. Since than, publications are tagged as withdrawn instead of deleting them.

Deleted records are denoted with an empty revision. An example of a record that was removed as a duplicate (created on 2004-11-24, deleted on 2011-01-22):

```
<inproceedings key="conf/hicss/LimGC02a" mdate="2011-01-22">
</inproceedings>

<inproceedings key="conf/hicss/LimGC02a" mdate="2004-11-24">
  <author>John Lim</author>
  <author>Binnie Gan</author>
  <author>Ting-Ting Chang</author>
  <title>A Survey on NSS Adoption Intention.</title>
  ...
  <url>db/conf/hicss/hicss2002-1.html#LimGC02a</url>
</inproceedings>
```

3 Simple Statistics

The figures presented here are extacted from a hdblp version from March 2018

Quick facts

- Publication record with the maximal number of revisions (24 revisions): journals/arobots/Hashimoto02
- Publication records with a single revision (never modified): 865.488 (21% of all publication records that ever existed)
- Person record with most revisions (16 Revisions): *Reinhard Wilhelm* homepages/w/ReinhardWilhelm
- Person records with a singe revision (never modified): 2.068.738 (95% of all person records that ever existed)
- Deleted publication records: 6.390
- Deleted person records: 117.922

Not all data fields in the records are edited with the same frequency. By far the most modified field is `ee` which holds a weblink to the publication (usually, on the publisher's site). `ee` also stores DOIs. Many modifications to `ee` are batch corrections that change the representation of DOIs.

Many modifications to the `author` field are corrections of name disambiguation related defects. See [3] for a testcollection for name defects buildt on `hdblp`.

4 Changes between dataset versions

Version 2 (May 20, 2019):

- Updated data set to contain record revisions up to December 17 2018.
- Included dtd file compatible with this version of hdblp.

Version 3 (May 20, 2019):

- Added this document. The document was omitted by accident in the previous version.

5 Contributors

The data in dblp is a joint project of the University of Trier, Germany and the LZI Schloss Dagstuhl, Wadern, Germany. See `https://dblp.org/db/about/team.html` for a list of current and past members of the dblp team. `hdblp` is created by Oliver Hoffmann. This document is primarily compiled by Florian Reitz.

References

1. M. Ley. DBLP - some lessons learned. *PVLDB*, 2(2):1493–1500, 2009.
2. M. H. A. Newman and A. M. Turing. A formal theorem in church's theory of types. *J. Symb. Log.*, 7(1):28–33, 1942.
3. F. Reitz. Two Test Collections for the Author Name Disambiguation Problem based on DBLP, Mar. 2018.