Milestone: M7.12 - A suite of test cases that will be used to test the de-duplication software

Partners: OU

Compiled by: David Morse (OU), Dauvit King (OU)

December 2011

WP7: Biodiversity literature access and data mining

Notes on M7.12 - A suite of test cases that will be used to test the de-duplication software

Due

23:59 Sunday 31 July 2011

Not delivered to schedule in the format originally anticipated. This brief report states why we were unable to release formally the test cases on schedule, in the form that was implied by the title of the milestone - 'a suite of test cases'. The report will describe the barriers to progress and the work there has been towards this milestone.

Issues hindering progress

An early concern was the changing landscape within which deliverable D7.1 - Community contributed bibliography was to be delivered. For a detailed discussion of the difficulties that were encountered in producing this deliverable, please see the report on D7.1 which is available from the ViBRANT website at http://vbrant.eu/sites/vbrant.eu/files/D7-1_report_final.pdf. In short, our time and effort has been devoted to navigating our way through the changing landscape of architectures and hosting solutions for the Bibliography of Life.

Investigating the scale of the problem

The suite of test cases is designed to support the development of reference de-duplication software as outlined in the description of 'Task 7.1 Community constructed digital bibliographies' in the ViBRANT Description of Work document:

A Develop aggregators to harvest bibliographic metadata from publically accessible repositories of biodiversity science literature.

B Develop software to identify and remove duplicate citations from the harvested metadata. In other words, de-duplicate reference lists whilst retaining links to external sources. Add stable, persistent identifiers where no identifier previously exists.

The focus of the de-duplication software is on removing duplicate citations from metadata harvested from different sources. To that end, we adopted two approaches to identifying test cases for the de-duplication software:

1. Task 7.1.E of the Description of Work was to 'Solicit community help in developing a suite of test cases to test the de-duplication software by contributing examples of duplicates that they encounter'. We have done this but the community has not been forthcoming.
2. Identify duplicates in 'publicly accessible repositories of biodiversity science literature' (Task 7.1.A, ViBRANT Description of Work). In the remainder of this section we will describe this approach to identifying a suite of test cases.

Having built a repository for the bibliographic references (RefBank, see Deliverable D7.1 report), the repository needed to be populated. To that end we have been developing import routines and aggregators that will load references from other repositories. Some of these repositories already contain duplicate references so our first task is to identify duplicates *within* a repository, before tackling the problem of finding references to the same source material that appear in *different*

repositories.

For example, the bibliographic references in both ITIS (Integrated Taxonomic Information System) and Catalogue of Life contain duplicate references.

Duplicate references in ITIS

The ITIS database is available from http://www.itis.gov/. This example is taken from from the October 2011 dump of the ITIS database.

This is a very typical example of duplicate references where one or more fields in the database are almost, but not exactly, the same. In this example the publication title appears in both abbreviated and full forms (Annot. Zool. Japan. 35 versus Annotationes Zoologicae Japonenses 35 (3)), and there are different page numbers (162-165 versus 162-165).

PUB; "2"; "Matsumoto K. 1962."; "Two new genera and a new subgenus of the family Asellidae of Japan."; "Annot. Zool. Japan. 35"; NULL; "1962-01-01"; NULL; NULL; NULL; NULL; "162-165"; NULL; "1998-04-02"

PUB; "3"; "Matsumoto K. 1962."; "Two new genera and a new subgenus of the family Asellidae of Japan"; "Annotationes Zoologicae Japonenses 35 (3)"; NULL; "1962-01-01"; NULL; NULL; NULL; NULL; "162-169"; NULL; "2000-08-01"

In the above example the second field is the record number in the database. Subsequent fields are the author(s) and date, the paper title, journal and volume, a NULL field, the date of publication, some more NULL fields, then the page numbers. The field separator is a semi-colon.

Duplicate references in Catalogue of Life

The Catalogue of Life is available from http://www.catalogueoflife.org/ and the example is taken from the 2011 edition of Catalogue of Life.

The following two references differ in language: the first reference is in Russian and the second is in French. As such, this example represents a novel angle on the concept of duplicate references – references that are expressed in different languages.

"""44445""", """Uvarov""", """1929""", """""""", """Ezhegodnik Muz. Akad. Nauk SSSR 31""", "\N"

"""45300""", """Uvarov""", """1929""", """""""", """Annuaire du Musée Zoologique de l'Acad. des Sciences de l'URSS (Ann. Mus. Zool. Acad. Sci. USSR) 31""", "\N"

In the above example the first field is the record number in the database, then there is the author name, the year of publication and the bibliographic details which are compressed into one field. The field separator is a comma.

Discussion

We have shown that two large, well-known databases that are widely used by the biodiversity community contain duplicate references. While this is not surprising, it demonstrates the scale of the problem of duplication and that de-duplication should be a two-stage process. First, to identify duplicates within a single bibliographic data source and second, to identify duplicates between

data sources. The former will probably be easier than the latter, because references within a single bibliographic database are likely to be more consistent than those between data sources.

These examples also demonstrate the challenge posed by different languages. One of the resources that we intend to develop in WP7 is a look-up service of journals, their abbreviations and variants. Ideally, as illustrated by the Catalogue of Life example, such a service should also provide language variants too.

The way forward

The above examples demonstrate that there are sufficient cases of duplicates within existing large databases that spending time and effort developing a dedicated suite of cases for testing software and approaches to de-duplication, is not justified. Rather, we intend to use existing large databases and the problems that we encounter when importing references into RefBank to expose different types of duplicate references.