Author Consolidation across European National Bibliographies and Academic Digital Repositories

Nuno Freire, Rene Wiermer, Markus Muhr, Andreas Juffinger, Chiara Latronico, Valentine Charles

The European Library, National Library of the Netherlands

Summary

This paper presents an ongoing work on the consolidation of authors across the national bibliographies of Europe and of academic digital repositories. We are studying this problem in the Driver data sets and using the Virtual Union Authority File as the consolidation data set for persons. These first results have shown an overlap in author names between the two data sets, but 41% of the names had ambiguous matches. We also present results on other existing information that may be exploited to perform disambiguation of these cases. We present future work as well, and discuss the application within The European Library and for external systems.

1 Introduction

The European Library holds several million bibliographic records from 48 national libraries of Europe. The National Bibliographies are one of the main bibliographic data sources in each country, listing every publication, under the auspices of a national library or other government agency. Depending on the country, all publishers will need to send a copy of every published work to the national legal deposit, or a national organisation will need to collect all publications.

Given that the publisher domain is very heterogeneous and that thousands of publishers might exist in a country, National Bibliographies are effectively the single point of reference with which to comprehensively identify all the publications in a country.

In Europe, National Bibliographies are typically created and maintained by national libraries. Whenever a book is published in a country, it is recorded in the corresponding national library catalogue from where the national bibliography is derived.

Currently, The European Library holds approximately 75 million bibliographic records in its centralised repository. This number is constantly increasing, as more national libraries' catalogues are included in the centralised repository. By the end of 2012, the total bibliographic universe of The European Library is expected to be approximately 200 million records.

The centralisation of European bibliographic data in The European Library is creating new possibilities for the exploitation of this data in order to improve existing services, enable the development of new ones, or provide a good setting for research.

The centralisation of the bibliographic data enables the automatic linkage of the national bibliographies across countries, through the use of data mining technologies. In this paper, we present the current status of our work on the consolidation of authors across the national bibliographies of Europe and freely-accessible, academic digital repositories holding content across several academic disciplines.

When complete, it will allow the exploration of researchers' publications across these two types of bibliographic data sources. In addition, these consolidated bibliographies will be made available for interoperability with research information systems according to the CERIF data model[1].

This paper will proceed in Section 2 with an introduction to our approach for author consolidation. Section 3 describes how references to persons are present in VIAF, in the bibliographic metadata of academic repositories, and includes other information that can be exploited for author consolidation. Section 4 presents a preliminary study aimed at identifying if author consolidation between national bibliographies and open academic repositories can be achieved, and also if there is an overlap of the sets of persons described in both data sets. Section 5 presents how we plan to implement the consolidation system, and Section 6 presents how the data with the consolidated authors can be applied. Section 7 concludes.

2 Approach for Author Consolidation

The problem of author consolidation consists in determining if two or more references correspond to the same real-world person (Elmagarmid 2007). In bibliographic data, a person may have multiple different representations (typically names and birth/death dates), and each representation might match the multiple persons (i.e., reference and referent ambiguity). The variations found in the representations of the persons may have multiple origins, such as misspellings, typing errors, abbreviations, names varying over time, etc.

Our approach leverages two key aspects of the National Bibliographies and consolidation work on authors carried out by libraries. The first aspect is that national libraries already individually perform a manual consolidation of authors through their ongoing work to maintain National Bibliographies. The second aspect is that some European national libraries actively work on the construction of the Virtual International Authority File, or VIAF (Bennett 2006). VIAF is a joint project of several national libraries from all continents. It hosts a consolidated data set containing data that national libraries have gathered for many years about the authors of the bibliographic resources held at the libraries. It is available as open data.

Using VIAF, we can already consolidate the authors across the VIAF-participating countries and soon we will exploit this resource to consolidate authors in other countries. By extracting statistics about authors consolidated in VIAF, specifically from the National Bibliographies of VIAF participants, we expect to be able to derive a probabilistic model that will allow us to consolidate the authors present in the bibliographic data available from academic digital repositories.

[1] http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

3 Representation of Authors

This section presents how authors are represented in the data sets we are addressing. It will also describe the data that is associated with the authors, and addresses how it can be used to provide evidence for the task of author consolidation. It starts by describing how the authors are represented in the Driver data set (Graaf 2009) and follows with how they are represented in VIAF.

The Driver data set follows a data model using mainly the general Dublin Core[2] terms for resource description, which are widely used in the Web for describing millions of resources. In the Driver data set, the authors are represented in two Dublin Core elements defined as follows (Weibel 1998):

- *creator*: an entity primarily responsible for making the resource;
- *contributor*: an entity responsible for making contributions to the resource.

The semantics associated with these data elements is very generic and does not provide a detailed data structure for the representation of the authors. The authors are represented as simple character strings, and several types of entities can be represented, such as persons, organisations, conferences, etc. These data elements often contain just the name of the author, but they may also contain other information about the author, making the consolidation difficult to perform.

The following are some examples of representation of authors in the Driver data set, which present some of the difficulties presented by this type of data:

- "Thomas Deselaers, Tobias Weyand, Daniel Keysers, Wolfgang Macherey, and Hermann Ney"
- "Fernández Mosquera, Santiago, dir."
- "Universidad de Alicante. Departamento de Prehistoria, Arqueología, Historia Antigua, Filología Griega y Filología Latina"
- "Rohrbaugh, Richard L., 1936-"
- "Žibert, Janez (avtor); Mihelič, France (mentor)"
- "Modena, Biblioteca Estense, Archivio Muratori, Filza 81, fasc. 55, lett. 140"

In the above examples, we can observe different types of entities (persons and organisations), and several authors represented in a single data element; the name and other details about the author are present in the same data element (year of birth, title, etc).

The parsing of person names from this data is also not trivial. Although some punctuation is used to separate different information about the authors, the use of the punctuation is not used in a consistent way.

In VIAF, persons are represented according to the typical data structures used in the library MARC formats (ALA 2002). The VIAF record of a person also contains additional data which

[2] http://dublincore.org/

can support the author consolidation process. Figure 1 presents a simplified view of the data model of VIAF, representing only the data that we are exploring for the author consolidation process.

The person names are structured, with separate data elements for the surname and other parts of the name (first names and middle names). Several forms of the person's name may be present, reflecting the different ways that the authors write their name on different publications. Libraries adopt one form of the name as the preferred one, and represent other forms as alternative. Since VIAF contains data from several countries, multiple preferred name forms may exist for the same person.

The birth and death dates of the person are also represented, and are often available in the VIAF records.

Additional data is available in VIAF and can be exploited to support the author consolidation process. In our work, we are exploring the following:

- Known publications: contains titles of publications that have been authored by the person. The titles are represented as character strings.
- Publishers: contains names of persons or organisations which have published works by the person. The publishers are represented by their name as character strings.
- Co-authors: contains names of persons or organisations which have co-authored works with the person. The co-authors are represented by their name as character strings.

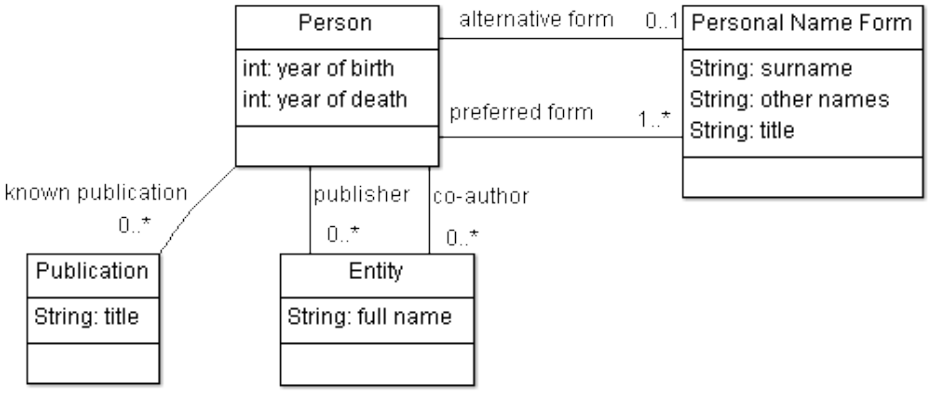

Figure 1 - Partial view of the representation of persons in the data model of VIAF

In addition to the name of the authors, additional information can be used from the Driver record to find a matching record with VIAF. The title of the publication described in the Driver record may be compared against the list of titles available in the VIAF records. All the authors of the resource described in the Driver record can be matched against the list of known co-authors existing in VIAF. And, similarly, the publisher of the resource described in the Driver record can be matched against the list of known publishers existing in the VIAF records.

4 Preliminary Study Results

This section presents the results we obtained in a preliminary study aimed at identifying if author consolidation between national bibliographies and open academic repositories can be achieved, given the characteristics of the data, and the possible lack of overlap of the sets of persons described in both data sets.

This study was conducted on a subset of the Driver data set. Three million records where processed, and the authors contained in them were matched against VIAF. In total, the 3 million records contained 283,114 references to authors. Approximately 3,055,000 records of persons were contained in the VIAF data set used.

This study focused on consolidating those authors which are individual persons. Although both VIAF and Driver data contain references to organisations, at this stage we addressed only the consolidation of persons.

In this study we measured five general aspects of matching data from Driver with VIAF: number of matching names; ambiguity of matching names; effect of name completeness on matching of names and ambiguity; number of matching publication titles; and number of matching co-authors.

Figure 2 presents how many names of authors from Driver matched names of persons in VIAF, and how ambiguous the name-matching can be, by showing the number of distinct VIAF matching records. In total, 283,114 matching names where found, of which 59% were not ambiguous, 26% matched two records, 10% matched three records, 3% matched four records, 1% matched five records, and the remaining 1% matched six or more records.

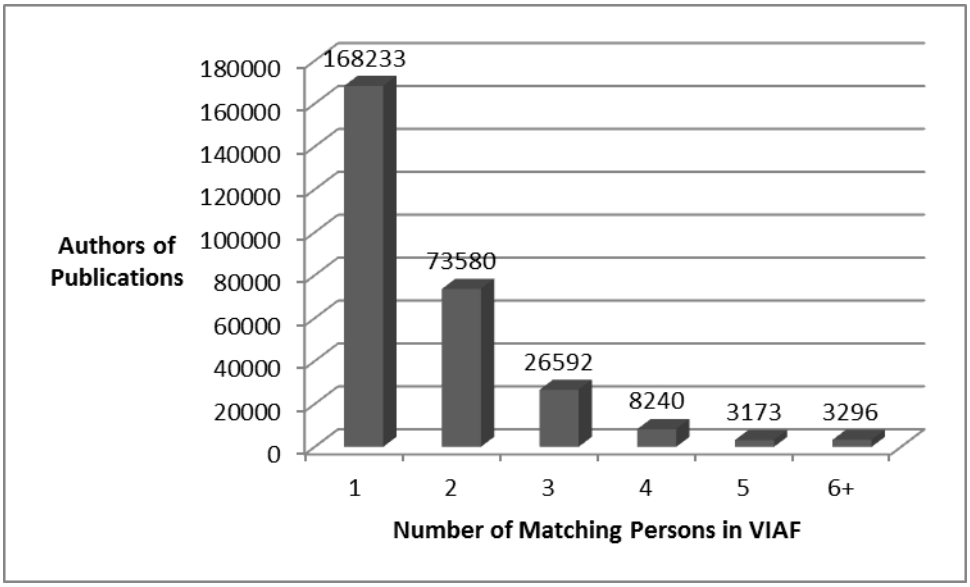

Figure 2 - Number of matching author names and their ambiguity in VIAF

We also measured the quality of the matches in names and how they influence the number of ambiguous matches in VIAF. Names of persons appear in the Driver records with different de-

grees of completeness. In some cases, the names contain first names, middle names and surnames; in other cases only the first name and surname are present, and also, only name initials may be present for first and middle names. Figure 3 presents our observations, which show how ambiguity increased with less complete name-matching. The percentage of ambiguous matches was: 47% for names with first name, middle names and surname; 66% for names with first name and surname; and 91% for names with first-name initial, middle names and surname.

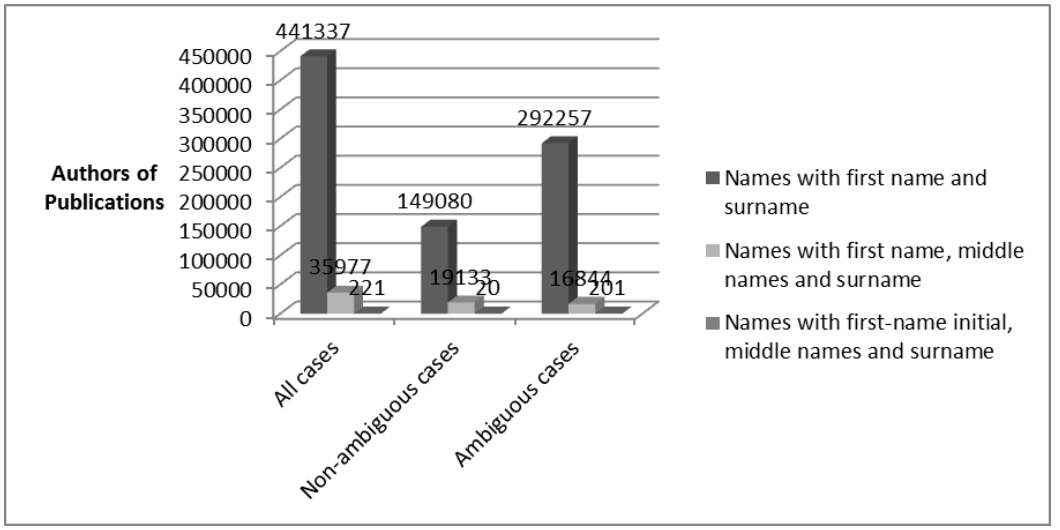

Figure 3 - Quality of the name matches their influence in ambiguous matches in VIAF

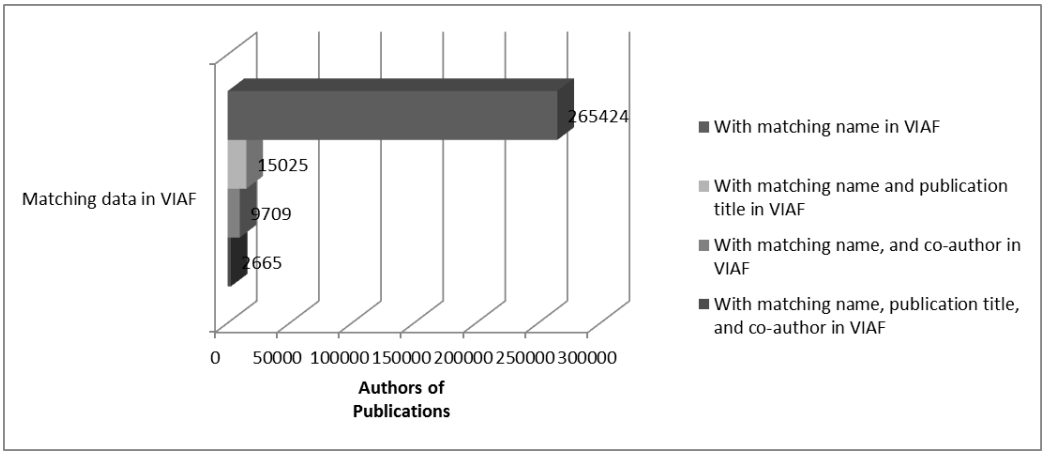

Figure 4 – Number of matches by name, publication title and co-author

We also measured the number of authors from Driver that have a match in VIAF, with just the name in common, with the name and the title of the publication in common, with the name and a co-author in common, or with the name, the title of the publication and a co-author in common. Figure 4 presents the measured values, and we can observe that, in some cases, reliable infor-

mation is available in the data to allow a correct consolidation of the references to some authors. However, for the majority of cases, only a matching name was found.

5 The Planned Consolidation System

The author consolidation system is planned to be built as an ETL (Extraction, Transformation and Loading). The process starts with the preparation of data for consolidation processing. This step comprises tasks for selecting the relevant data from the National Bibliographies, and the academic repositories that will be used to represent an author during the consolidation process. The following data about the authors is gathered:

- Name
- Years of birth and death
- Known co-authors
- Known publishers
- Known titles of publications

The decision to match two author references is made by reasoning on the similarity scores obtained by comparing each of the above data elements. Comparison of person names is performed with the Jaro-Winkler similarity metric (Jaro 1989). Comparison of remaining data fields is based on the number of common values found in the two records versus the total number of available values. For example, similarity of co-authors is given by the number of common co-authors between two authors, and similar calculations for titles and publishers.

The solution for reasoning on the outcome of comparisons will be based on supervised machine-learned model, which determines the likelihood of two authors being the same entity. As ground truth, for building and testing the machine-learned model, we are using a data set extracted from the National Bibliographies of VIAF participants and the Driver repository.

6 Application Scenarios

In the first stage, this work will be integrated into The European Library portal, allowing the end users to explore the researchers' publications across these two types of bibliographic data sources.

The main application scenario is the search for the bibliography of a specific researcher. Through The European Library portal, users exploring a researcher's bibliography will be able to see a consolidated view of the researcher's publications across academic repositories and the traditional publications recorded in national bibliographies. For example, a user visualising the bibliography of a researcher, such as Stephen W. Hawking, will be able to see its academic publications alongside its books and all books' translations published throughout Europe.

At a later stage, we plan to make available these consolidated bibliographies for interoperability with research information systems according to the CERIF data model and interoperability services of CERIF.

7 Conclusion

This paper presented the ongoing work at The European Library, addressing the consolidation of authors across the national bibliographies of Europe and freely-accessible academic digital repositories.

The preliminary results of our studies have indicated that, although the majority of authors in the Driver data set were not found in VIAF, a considerable overlap between the authors of these two kinds of bibliographic data sets exists. We believe that the overlap between the authors represented in the two data sets is higher than we observed, since the lack of a detailed data structure to represent authors in the Driver data set poses difficulties in matching the authors' names. However, we were not able to measure the extent of the effect of this data heterogeneity.

We also observed that ambiguity of names is very high, and that it is increased by the frequent use of name initials in bibliographic references to authors. In many of the ambiguous cases, we could not find a matching title or a matching co-author; therefore, in many cases there may not be enough information in the bibliographic data records to consolidate the authors.

References

ALA; CLA; CILIP (2002): Anglo-American Cataloguing Rules. 2002 Revision.

Bennett, R.; Hengel-Dittrich, C.; O'Neill, E.; Tillett, B.B. (2006): VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files. World Library and Information Congress: 72nd IFLA General Conference and Council.

Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. (2007): Duplicate Record Detection: A Survey. IEEE Transactions on knowledge and data engineering, vol. 19, no. 1, 1-16. DOI: 10.1109/TKDE.2007.250581

Graaf, M. (2009): The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output. Amsterdam University Press. ISBN 9789053564103

Jaro, M. A. (1989): Advances in record linking methodology as applied to the 1985 census of Tampa Florida. Journal of the American Statistical Society 64: 1183-1210

Weibel, S.; Kunze, J.; Lagoze, C.; Wolf, M. (1998): Dublin Core Metadata for Resource Discovery. Network Working Group Request for Comments: 2413.

Contact Information

Nuno Freire

The European Library
National Library of the Netherlands
Willem-Alexanderhof 5
2509 LK The Hague
Netherlands

nfreire@gmail.com