Use of Authorities Open Data in the ARROW Rights Infrastructure

Nuno Freire
The European Library,
Europeana Foundation
Netherlands
nfreire@gmail.com

Markus Muhr
The European Library,
Europeana Foundation
Netherlands
markus.muhr@kb.nl

Abstract

The ARROW rights infrastructure provides the means to support mass digitisation projects by finding automated ways to clear the rights situation of books to be digitised. ARROW provides seamless interoperability across a distributed network of national data sources, which contain essential information for determining the rights status of works, including national bibliographies from national libraries, books-in-print databases, and rights-holders databases. This paper presents how open data about authors, from the Virtual International Authority File (VIAF) is being used in ARROW to support the data interoperability across ARROW data sources, and how it is being used for the outputs of the rights clearance process.

Keywords: open data, entities, copyright, rights information infrastructures.

1. Introduction

The process of clearing the rights situation of a work for digitisation is very difficult and expensive to be performed for works that have been published during the 20th and 21st centuries, due to the fragmentation of rights information across multiple data sources. ARROW stands for Accessible Registries of Rights Information and Orphan Works, and it provides an infrastructure to support mass digitisation projects by finding automated ways to clear the rights of the books to be digitised. ARROW provides seamless interoperability across a distributed network of national data sources, which contain essential information for determining the rights status of works, including national bibliographies from national libraries, books-in-print databases, and rights-holders databases. ARROW provides automatized ways to support the identification of a work, the clarification of its rights status and the identification of the rights holders (ARROW, 2011).

The rights clearing process depends on the availability of existing bibliographic and rights data. There are already-established information sources for printed material in national bibliographies, books-in-print and the databases of rights organisations. This paper presents how open data about authors, from the Virtual International Authority File, or VIAF (Bennett at al., 2006), is being used in ARROW to support the interoperability of these data sources and the outputs of the rights clearance process.

VIAF provides a consolidated data set 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 and is actively maintained and synchronized with the local authority files of national libraries.
This paper will follow with an overview of the complete ARROW workflow, and will describe how the matching of contributors with VIAF is made, the data from VIAF that is used in ARROW, and the potential uses and benefits for the rights clearance process. The final section will conclude and present future work.

2. The ARROW rights clearance workflow

Figure 1 shows an overview of the general workflow of ARROW. It starts from a library as a potential user that wishes to digitise a book, and shows the process that the ARROW system

supports to provide a response containing the requested rights information. The process depends on data from several sources:

- National bibliographies' data aggregated in The European Library[1] (TEL);
- Author data from the Virtual International Authority File (VIAF);
- Publishing data from Books In Print databases (BIP);
- Rights-holders data from Reprographic Rights Organisations (RRO).

The initial steps of the workflow depend on The European Library's system to fulfil the information requirements of the process regarding national bibliography data. Three tasks are carried out:

- Identification of the exact record, from the national bibliography, of the book that the library intends to digitise;
- Identification of other records of books which share the same intellectual work and, therefore, are essential for the rights clearance process;
- Improvement of the data about the contributors of the work.

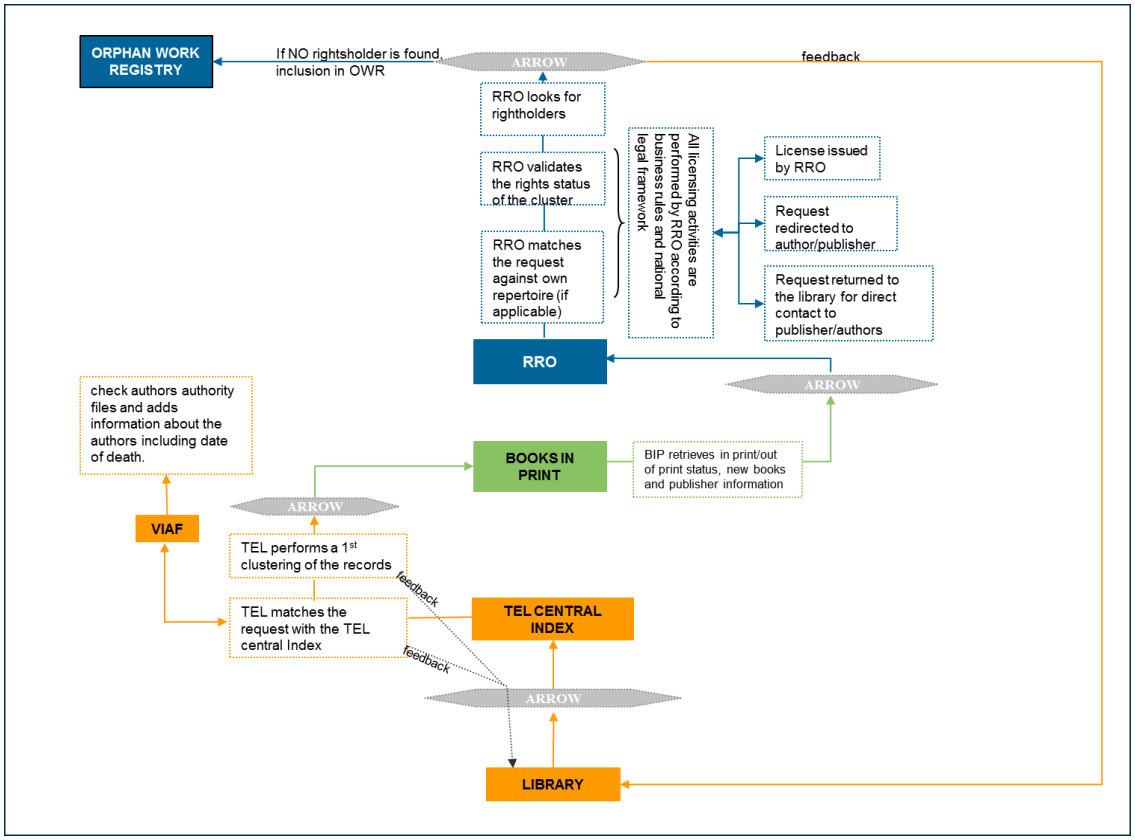

FIG. 1. The ARROW rights clearance workflow

After all publication and contributor data is identified at The European Library, the ARROW workflow proceeds with the determination of the copyright status, the identification of publications still in commerce, the identification of rights-holder(s), and finally a search for the appropriate permission from the rights-holders. In those cases where it is not possible to find information about the rights-holders, the work is registered as orphan in the ARROW Orphan

[1] http://www.theeuropeanlibrary.org

Works Registry. Additional details on the ARROW workflow and its underlying systems are available in Freire et al., 2013.

3. Authorities Data in ARROW

In the initial steps of the workflow, The European Library's system matches the data about the contributors of a work, which is present in the bibliographic records of national bibliographies, in VIAF. This section describes how the matching against VIAF is made, the data from VIAF that is used in the workflow, and its potential uses and benefits.

The identification of the VIAF record concerning a particular contributor of a work is performed by two alternative mechanisms: by authority record identifier, or by matching of available data. The first mechanism is always favoured, but it is only possible to apply it when the national bibliography records belong to a library participating in VIAF, and also if the authority record identifier of the contributor is present in the bibliographic record. When matching by identifier is not possible, other data that is available in the bibliographic record is matched against data available in the VIAF records.

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 contains also additional data which can support the matching of the contributors. FIG. 2 presents a simplified view of the data model of VIAF, representing only the data that we are exploring for matching contributors.

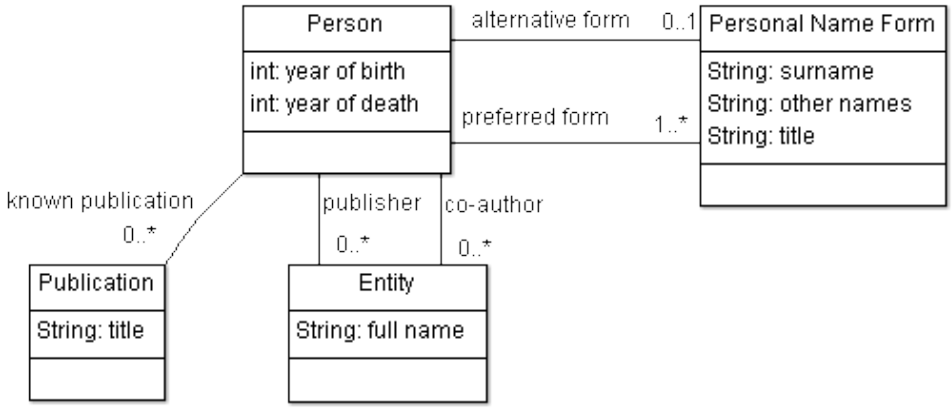

FIG. 2. Partial view of the representation of persons in the data model of VIAF

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 ARROW, 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 organizations, which have published works by the person. The publishers are represented by their name as character strings.
- Co-authors: contains names of persons or organizations, which have co-authored works with the person. The co-authors are represented by their name as character strings.

The matching between the contributor in a bibliographic record and a VIAF record starts with the matching of the name of the contributor. In all bibliographic data, the names of the contributors are subjected to general data quality problems, such as typing errors, misspellings, synonyms, homonyms, abbreviations, different spellings across languages, etc (Freire et al., 2008). For this reason, the matching of names should not be performed by exact matching, since it would result in missed results in a highly heterogeneous international context like ARROW. We adopted a two phase character string similarity matching for names. The first phase restricts the matching to a subset of the VIAF names containing only records with a minimal level of similarity with the contributor name. Its implementation is based on indexing of character n-grams (Gravano et al., 2001) of the contributors' names. In the second phase a more suitable similarity metric for names is applied based on the Jaro-Winkler similarity metric (Jaro, 1989).
Typically, several VIAF records match the name of a contributor, and further evidence for choosing the correct VIAF record is searched in other data elements. The dates of birth and death are compared, if available. The title of the publication described in the bibliographic record is compared against the list of titles available in the VIAF records. All the contributors of the work are matched against the list of known co-authors existing in VIAF. And, similarly, the publisher(s) of the work are matched against the list of known publishers existing in the VIAF records. A record from VIAF is only chosen if enough supporting evidence is found.
When a match between a work contributor and VIAF is found, detailed data about the contributor in VIAF is attached to the work metadata gathered throughout the ARROW workflow (EDItEUR, 2011), and made for use in the remainder steps of the workflow to support matching in other data sources and in the processing of the outputs of the rights clearance process. A summary of the uses of the contributor detailed data in the ARROW workflow is presented in Table 1.

TABLE 1: VIAF data elements use in the rights clearance process

Data elements	Description	Use in the rights clearance process
Name variants	Various forms of the name of the person or organization. May include the complete name, abbreviated names, acronyms, etc.	Used for matching of names across records and data sources. Improves the identification of all publications of a work in the national bibliography, the identification of publications still in commerce in books-in-print databases, and the identification of the contributor in the rights-holders databases.
Date of birth and death	The dates of birth and death of the person.	Used by the public domain algorithm for determining the public domain status of the work. Used for matching confirmation and disambiguation of homonyms across data sources: national bibliography, books-in-print databases, and rights-holders databases.
Nationalities	The nationalities of a person or organization.	Used, in some countries, by the public domain algorithm for determining the public domain status of the work.

4. Conclusions and Future Work

The ARROW rights infrastructure provides automated means to clear the rights situation of a work by addressing data interoperability across a distributed network of data sources. Currently VIAF is the only open data source used in ARROW, and its worldwide coverage greatly supports the outcomes of the ARROW rights clearance workflow. The effectiveness of this approach is currently being validated within ARROW, and a full report is expected to be ready in the second semester of 2013.

The ARROW framework was designed from the start to allow open data to be found and attached to the consolidated work metadata provided by ARROW. Future work will explore further use of open data sources, and in other sources of data about contributors. The first objective, in this line of work, will be to use of the International Standard Name Identifier (ISNI)

system, which uniquely and authoritatively identifies public identities in several fields of creative activity (ISO 27729, 2012). There is also potential for exploring other sources of open data in the ARROW workflow, although there is not at the moment any concrete implementation plan in ARROW.

Acknowledgments

We would like to acknowledge the work of the participants in ARROW Plus project, who contributed to all the work described in this paper. In particular, we would like to acknowledge the contributions of Associazione Italiana Editori, the University of Innsbruck for their support in the testing and validation of this work.

This work was carried out in the ARROW Plus project, a best practice network funded under the European Commission's Competitiveness and Innovation Framework Programme, grant agreement number 270942.

References

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

ARROW (2011). ARROW Business Model. Retrieved April 4, 2013 from: http://www.arrow-net.eu/sites/default/files/ARROW_Business_Model.pdf

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.

EDItEUR (2011): ONIX for Rights Information Services. Overview. Retrieved April 4, 2013 from: http://www.editeur.org/files/ONIX-RS/ONIXRS_v1_0_PDF_Documentation.zip

Freire, N., Borbinha, J., Martins, B. (2008): Consolidation of References to Persons in Bibliographic Databases. ICADL 2008 – The 11th International Conference on Asian Digital Libraries.

Freire, N., Scipione, G., Muhr, M., Juffinger A. (2013): Supporting Rights Clearance for Digitisation Projects with the ARROW Service. LIBER Quarterly, vol. 22, no. 4, pp 265-284.

Gravano, L., Ipeirotis, P. G., Jagadish, H., Koudas, N., Muthukrishnan, S., Pietarinen, L., Srivastava, D. (2001): Using q-grams in a DBMS for approximate string processing. IEEE Data Engineering Bulletin, 24(4):28–34.

ISO 27729 (2012): Information and documentation. International standard name identifier (ISNI). International Organization for Standardization. ISBN 978-0-580-61840-6

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