Use of Authorities Open Data in the ARROW Rights Infrastructure

Nuno Freire, Markus Muhr
The European Library

DC-2013
Lisbon, September 2013
Overview

About Arrow

ARROW is a tool to facilitate rights information management in any digitisation project involving text and image based works.

ARROW infrastructure allows streamlining the process of identification of authors, publishers and other rightholders of a work, including whether it is orphan, in or out of copyright or if it is still commercially available.
Outline

- Motivation for ARROW
- The ARROW workflow
- The role of authorities metadata in ARROW
- The authorities matching process
- Conclusions and future work
Motivation for ARROW

- To support mass digitisation projects with automated ways to clear the rights of the books to be digitised.

- To identify and clear the rights associated with a book, a complex process needs to be undertaken:
  - Determine the work(s) contained within the book
  - Identify all the other expressions of the same work(s)
  - Identify the publisher(s) and contributor(s) involved
  - Determine the dates of publication at work level
  - Determine whether that work(s), and not the book itself, is still in commerce
  - If necessary, obtain any licenses from the rights holders or collective rights organizations
What is ARROW

- A rights infrastructure and system for the identification of:
  - Rights status
    - In or out of copyright
    - In or out of print / commercialised or not
  - Rights
    - Which rights are involved
  - Rightholders
    - Authors
    - Publishers
  - How and where to clear the rights
  - Orphan Works and their registration
Sources of Information in ARROW

- ARROW makes information available from several sources:
  - The European Library:
    - National bibliographies - to identify the book and to cluster it with all other books containing the same intellectual work
    - Virtual International Authority File - to better identify the authors and support the identification of in copyright works
  - Books in Print database - to know if any of the books concerned are actively commercialised by any publisher
  - Reproduction Rights Organisation – to see if they know or can trace the rightholders
The Role of Libraries

- Libraries as End User
  - Query the ARROW system to retrieve info on rights status
  - Search ARROW
  - Validate retrieved results and/or refine search
The Role of Libraries

- National Libraries as Metadata Providers
- Provide the National Bibliographies to The European Library
- Provide authority files to VIAF
The Role of The European Library (TEL)

- To match library requests with national bibliographies
- Identify all other manifestations that potentially share intellectual work with a manifestation
- To create a Work record: work metadata, manifestations, contributors, etc.
The Role of The European Library (TEL)
- To match work contributors against VIAF to gather more information for the ARROW process

check authors authority files and adds information about the authors including date of death.

TEL performs a 1st clustering of the records

TEL matches the request with the TEL central Index

License issued by RRO
Request redirected to author/publisher
Request returned to the library for direct contact to publisher/authors

if NO rightsholder is found, inclusion in OWR
The Role of Books-in-Print (BIP)

- To provide data about in print/out of print status
- To provide data about publishers
- To add new manifestation records of the work
The Role of Reproduction Rights Organisation (RRO)
- RROs as Metadata Provider
  - To provide data about authors and publishers
  - To provide data about available licenses
ONIX for Rights Information Services (ONIX-RS) used for data exchange between ARROW participants

- If no rightsholder is found, inclusion in OWR
- RRO looks for rightsholders
- RRO validates
- All legal business performed by RRO
- License issued by RRO
- License issued by RRO
- The library for direct contact to publisher/authors
- BIP retrieves in print/out of print status, new books and publisher information
- TEL performs a 1st clustering of the records
- TEL matches the request with the TEL central Index
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL
- Feedback from TEL

The ARROW Workflow
Which data from VIAF is used in ARROW

- **Name variants**
  Various forms of the name of the person or organization. May include the complete name, abbreviated names, acronyms, etc.

- **Date of birth/death**
  The dates of birth and death of the person.

- **Nationalities**
  The nationalities of a person or organization.
How data from VIAF is used in ARROW

- **Name variants**
  - For matching of names across records and data sources
  - Improves the identification of all publications of a work, the identification of publications in books-in-print databases, and the identification of the contributor in the rights-holders databases.

- **Date of birth/death**
  - Used for determining the public domain status.
  - Used for matching confirmation and disambiguation of homonyms across data sources

- **Nationalities**
  - Used, in some countries, for determining the public domain status of the work.
The matching process

- VIAF data used for matching, disambiguation, and match probability
Matching work contributors with VIAF

- Names are matched by similarity
- Confirmation of the correctness of a name match is taken from other matching data
  - The dates of birth and death
  - The title of the work is compared against the list of titles available in VIAF
  - All the contributors of the work are matched against the list of known co-authors in VIAF
  - The publisher(s) of the work are matched against the list of known publishers in VIAF
- A match is only chosen if enough supporting evidence is found
Conclusions and future work

- Currently, VIAF is the only open data source used in ARROW, and it greatly supports the outcomes of the rights clearance workflow.

- The effectiveness of this approach is currently being validated within ARROW
  - A full report is expected to be ready in the second semester of 2013.

- Future work:
  - Application of VIAF in the recognition of the contributors within the natural language text of statements of responsibility (to appear at TPDL 2013)
  - Usage of ISNI identifiers in the ARROW workflow
  - Use other sources of authority data: ISNI
Acknowledgments

www.arrow-net.eu

This work was partially supported by the ARROWplus projects, with co-funding by the European Commission programme eContentplus
Thank you

Questions or comments?

Contact:
  Nuno Freire - nfreire@gmail.com