Grant Agreement ECP-2007-DILI-527003

ARROW

Guidelines for Technical Interoperability

Deliverable number/name	*D4.2*
Dissemination level	*Public*
Delivery date	*31 July 2009 (updated April 2010)*
Status	*Final*
Author(s)	*Françoise Bourdon (BNF), Linda Bennett, Sally Chambers (KB), Giulia Marangoni, Paola Mazzucchi (AIE)*

e*Content*plus

This project is funded under the *e*Content*plus* programme[1],
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

LIST OF TABLES

I. EXECUTIVE SUMMARY

According to the IEEE definition, interoperability is *the ability of two or more systems or components to exchange information and to use the information that has been exchanged*. However, interoperability per se, is nothing, without a valid purpose. For ARROW, the purpose of enabling technical interoperability is to support libraries wishing to digitise a book and use it. The Rights information Infrastructure will facilitate the process of carrying out a diligent search by means of a comprehensive system. The system will therefore underpin the objectives of the i2010 Digital Library initiative, in accordance with the principles defined by the High Level Expert Group on Digital Libraries as far as Out of Print and Orphan Works are concerned.

The aim of *D4.2 Guidelines on technical interoperability* is to report back on the findings of the information gathering phase of ARROW to assess the status of technical interoperability in three significant domains identified by ARROW:

- **Library domain**: in the context of ARROW, *National Libraries*, who are "designated and funded by a national government to serve the nation by maintaining a comprehensive collection of the published and unpublished literary output of the nation as a whole"
- **Books in Print domain:** organisations managing information about books for commercial purposes.
- **RRO domain:** the collective management organisations (CMO) operating throughout Europe in charge of managing and licensing of reprographic rights on behalf of individual rights holders

This deliverable incorporates the results of desk analysis, the design and undertaking of the WP4 – WP5 ARROW questionnaire and the feedback gained from face to face meetings at national level with key stakeholders in order to assess the current state of technical interoperability. It proposes practical solutions to enhance this interoperability and to provide a scalable roadmap for the development of the requirements for the ARROW system.

Following the results achieved in D5.2 (Design of system architecture), the definition of ARROW workflow and the development of the Rights Information Infrastructure (D6.1), this deliverable has been revised and updated in April 2010. At the request of some Books in Print and Reprographic Rights Organisations which have been surveyed, the confidential information collected about their databases have been gathered in a distinct document, with restricted dissemination level.

II. METHODOLOGY

The current section outlines the methodology adopted for D4.2 with the purpose of illustrating the main steps of the work done.

The Guidelines on technical interoperability presented here (see Section III) result from the analysis and examination of complementary sources of information which have been collected throughout the first phase of the project:

- Desk analysis and processing of WP 4 – WP 5 questionnaire findings[2]: the analysis provided extensive data for the assessment of existing databases in Europe in terms of quality and coverage of the information, relevance for ARROW objectives, standard identifiers, metadata and technical protocols implemented thus enabling (1) the identification of main *technical areas* to be addressed by the D4.2 guidelines in order to enhance actual interoperability among the different resources involved in the ARROW information infrastructure; (2) the identification of standards applicable for the effective management and exchange of bibliographic, rights and rights holders information within the ARROW framework, as reported in D4.1; (3) the identification of the best sources of information available at country and cross-country level according to the type of information needed to fulfil ARROW purposes, as reported in D5.1 [3]

- Face to face meetings at national level with key stakeholders: the purpose of the face to face meetings was to follow up on any issues that needed further clarification from the questionnaire results and from the examination of the intermediary reports per country. As a result, a subset of organisations to be involved in the first phase of ARROW piloting has been selected. In addition, meetings with the representatives of the relevant databases which will be involved in ARROW pilot have been arranged to discuss technical and business issues related to the integration of individual databases. The national meetings enabled (1) the refinement of the D5.1 analysis in terms of domains at cross-country level to be involved into ARROW information infrastructure (2) the refinement of technical areas to be approached by D4.2 analysis thus facilitating the identification of key requirements that individual databases need to meet in order to achieve actual interoperability within the ARROW network.

[2] WP4 – WP5 questionnaire aimed at gathering background information for WP4-5 activities. It was acknowledged that the types of information needed for the two WPs were complementary and they could be collected within the same target group of professionals dealing with the management of information resources in stakeholders organisations. WP4 and WP5 leaders agreed on working together to harmonize the WP4 and WP5 set of questions related to individual work packages in a single questionnaire thus avoiding the duplication of efforts.

[3] The methodology illustrated here is shared by also D5.1 as far as the preparation of the questionnaire and the first reworking of results are concerned.

Here follows a step by step synthesis of the methodology adopted:

Outline of ARROW objectives and preliminary identification of relevant information resources to be involved within the system infrastructure: a preliminary outline of the ARROW system was drafted in order to identify, for each task to be performed, the type of data needed and which of the organisations' databases should be considered as relevant information resources to be investigated and possibly integrated into the ARROW information infrastructure. A pre-questionnaire has been circulated among partners to provide a first overview of the resources available from partners in the first place and guidance for a further refinement of the questionnaire. The outline of the ARROW objectives was also intended to provide guidance for project partners in order to identify the relevant databases in each country, so that the relevant organisations could be contacted and interviewed within the framework of WP4-5 questionnaire[4].

Furthermore, the document provided interviewers with clear criteria to evaluate the relevance of the information resources investigated during the survey and was annexed to Guidelines for the Interviewers[5].

WP4-WP5 questionnaire design[6]: The ARROW system outline provided the basis for designing the WP4-WP5 questionnaire. For each type of resource (bibliographic, rights, parties information) within the ARROW information infrastructure, key issues to be investigated through the questionnaire were identified and appropriate questions prepared accordingly[7].

To address interoperability issues, specific sections aimed at gathering key information for the technical assessment of existing information resources have been developed, focusing in particular on:

- Use of standard identification systems, with particular attention to standard identifiers for textual works and manifestations (respectively, ISTC and ISBN)
- Implementation of standard metadata schema for managing different types of information (bibliographic, rights, rights holders data)
- Technical protocols implemented to access and exchange metadata among the different stakeholders' databases (libraries, publishers, bibliographic agencies, commercial databases, RROs) in the content value chain

[4] For the complete list of the stakeholders interviewed, please see D5.1
[5] See D5.1 Annex II– *Guidelines for the Interviewer and Glossary*

[7] See the table illustrating ARROW objectives in D5.1 Annex II – *Guidelines for the interviewer and Glossary*. For each tasks to be performed by ARROW system, relevant questions included in WP4-WP5 Questionnaire are tracked.

Survey on existing bibliographic resources in Europe and standards implemented: In order to collect exhaustive information, the WP 4 - WP 5 questionnaire has been conceived and designed as a structured grid for conducting face to face discussions with representatives of the organisations selected for the interviews.

The management of the survey was based on the "ARROW groupings" approach, aimed at facilitating the coordination of the survey on a geographical/regional basis and effectively involving both project partners and supporters in the interviews. Six ARROW groupings were established, either at national level (e.g. Spain, UK, France) or at regional level (e.g. Nordic countries), each one with a National Contact Point, i.e. the project partner working as the single point of content for WP leaders and the Coordinator[8].

Analysis of the questionnaire results: data collected through the WP4-WP5 questionnaire has been further processed in order to obtain more structured reports, tailored for the specific WP4 and WP5 objectives:

1. Summary of individual stakeholders' databases: organisations in charge of conducting the interviews in each country provided a summary of results for each database investigated using a template structured in four thematic areas:
    - Relevance of the database(s) in relation to ARROW objectives
    - Data quality
    - Technical conditions to access and query database(s)
    - Business conditions to access and query database(s)
2. Outline of National scenarios: questionnaires and related summaries[9] have been subsequently collected by National Contact Points, who were in charge of a further aggregation of results to provide a synthesis of the scenario which emerged at national/geographic level[10], aimed at identifying the relationships among the main databases interacting in the book value chain at a national level (national libraries, ISBN Agencies, publishers databases, Books in Prints, RROs)
3. Examination of the country reports: on the basis of the information collected in the previous steps of the analysis, detailed reports were prepared for each country involved in the survey. These country reports were internally used as a basis for identifying the key databases to be involved in piloting the ARROW infrastructure.

8 The rationale underlying ARROW groupings is illustrated in D1.1 *ARROW Project Operative Work pl*an
9 All the questionnaires collected and related summaries are available on ARROW website in Partners Only area; WP4 and WP5 sections: http://www.ARROW-net.eu
10 See methodology enclosed in D5.1 ANNEX II - *Guidelines for the interviewer and Glossary*.

National meetings: once the key databases to be involved in the pilot phase had been identified, following the assessment of the national scenarios emerging from the detailed country reports, national meetings, organised by the National Contact Point and the Project Coordinator, with key stakeholders from the countries involved in the ARROW pilot[11] were held. Experts with specific knowledge about the structure and content of the databases selected for the pilot were also invited to attend these meetings. The main goal of these national meetings was to gather clear indications for the ARROW system design in terms of:

- requirements (definition of users groups and expectations)
- resources (which databases will be queried by ARROW, in which order/hierarchy)
- use-cases (actual test bed for the system test phase)

Refinement of the methodology and structure of the analysis: input received from the national meetings enabled:

- the identification of three cross-country *domains* to organise D5.1 Analysis: National Library Domain, Books in Print Domain, RRO Domain
- the refinement of *technical areas* for D4.2 analysis on the basis of domain-specific interoperability issues within the ARROW infrastructure

D4.2 Analysis: for each domain identified (National Library, Books in Print, RRO) the analysis focused on technical areas which are relevant for the achievement of interoperability within the ARROW infrastructure.

In order to enable the analysis of the WP4-WP5 questionnaire results according to the specific D4.2 objectives, the information relevant to each technical area has been gathered by grouping the relevant questions from the original questionnaire, thus identifying macro-criteria for the analysis of domain specific interoperability issues.

For each technical area, an assessment of the strengths and weaknesses for each domain-specific issue, with respect to ARROW objectives, is provided. In addition, recommendations as to how interoperability could be improved have been made.

The analysis of each technical area enabled the identification of guidelines for technical interoperability within ARROW, based on current practices along the book value chain. Ideally, in the long term each specific database of each specific domain should adopt these guidelines in order to achieve interoperability within the ARROW infrastructure. However, one of ARROW's main

[11] Reports on National Meetings are available on ARROW website in Partners Only area; WP4 and WP5 sections: http://www.ARROW-net.eu

objectives is to base its system on existing databases and practices, therefore guidelines reflect this approach so that the organisations have minimum changes to operate to integrate ARROW.

III. GUIDELINES ON TECHNICAL INTEROPERABILITY

1. Introduction

According to the IEEE definition, interoperability is *the ability of two or more systems or components to exchange information and to use the information that has been exchanged*[12]. This definition presupposes that the information exchanged is used to serve a clearly defined purpose and that the information exchanged is relevant for that purpose. Moreover it is essential that the information exchanged is processed and semantically interpreted in the correct way by the systems using it.

The fulfilment of a shared purpose is what requires two or more systems to be interoperable, not a vague need for interoperability in itself. Before starting building any interoperability framework, it is therefore necessary to answer the following three questions: *What is the purpose? What is the information? What are the systems?*

With regard to interoperability in the ARROW context, in the first place it means translating these basic principles into the environment that ARROW is intended to operate in. Following this translation process, it is necessary to identify where interoperability is needed and then to define ways in which such interoperability can be facilitated.

The *purpose* is to support libraries wishing to digitise a book and use it. The Rights information Infrastructure will facilitate the process of carrying out a diligent search by means of a comprehensive system. This system will be able to provide the libraries with answers as to whether they are allowed to digitise a book or not. In the case of a positive response, the system will provide details of whom they have to approach to ask for the permission to digitise the book. The system will therefore underpin the objectives of the i2010 Digital Library initiative[13], in accordance to the principles defined by the High Level Expert Group on Digital Libraries[14] as far as Out of Print and Orphan Works are concerned.

The *information* needed to fulfil the purpose includes[15]:

---

[12] *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries,* IEEE, New York, 1990, ISBN 9781559370790)

[13] http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm

[14] http://ec.europa.eu/information_society/activities/digital_libraries/experts/hleg/index_en.htm

[15] For the analysis of the information resources involved in the ARROW context, see D5.1 Analysis of bibliographic resources and clearing mechanisms in Europe

- the rights status of the work: if it is in public domain or copyrighted, if it is in print or out of print.

- the rights holders of the work: publishers, authors and other contributors, or their agents, that is organisations entitled to manage rights on behalf of rights holders.

- terms and conditions under which digitisation and use are allowed

However, according to the responses to the questionnaire and to the subsequent analysis carried out in D5.1, which are the fundamental basis for this deliverable, (see Part II – METHODOLOGY) it must be noted that the above mentioned pieces of information are very much interconnected and in many cases are the result of additional processing of other data. For example to determine if a work is in the public domain or not, supporting data concerning all manifestations belonging to the work is to be collected, matched and consolidated, as well as information about contributors.

The *systems* that need to be interoperable are the organisations in the book value chain that manage bibliographic, parties and rights databases for their own business purposes: from the responses to the questionnaire underpinning this deliverable it has emerged that in each European country considered, there are many types of organisations holding information relevant to ARROW, (e.g. Books in Print, E-tailers, Publishers, Aggregators, Collecting Societies, ISBN Agencies, ISTC Agencies, Libraries, etc.). There is also much overlap in terms of type of data and many different levels in terms of data quality and quantity.

With the benefit of this analysis, it has been concluded that, in the majority of cases, the most relevant data that can be obtained and processed for ARROW purposes, can be retrieved from a combination of a National Library catalogue, a Books in Print catalogue (BiP) or similar commercial database, and a Reproduction Rights Organisation (RRO) database.

The combination of the above mentioned components – purpose, information and system – can be further explored in the analysis of the logical steps of a diligent search. In this way, each step of the search is the purpose of an ARROW task. The data that ARROW needs to be collected for each step is a piece of information and the resource, where the information is stored, is a system, as shown in Table 1.

Each of the sources of information, however, serves its own specific purpose. By itself these information are not, without the support of a framework technological infrastructure, interoperable with the other sources of information needed for ARROW purposes. Such an infrastructure will provide tools and crosswalks to overcome the existing interoperability gaps in terms of metadata formats and identifiers; access protocols; formats for metadata export; terms and conditions to access and query databases; etc.

The role of ARROW is therefore that of "interoperability facilitator"[16] to access and query different systems in the ARROW environment and to retrieve the relevant data. It will then process and exchange this data with other systems in the same environment and add complementary data from other systems that is relevant to ARROW in order to produce new meaningful information.

Having identified that the National Library, the Books in Print and the Reproduction Rights Organisations are the three main domains making up the ARROW environment, the present document *D4.2 Guidelines on technical interoperability* focuses on the analysis of the strengths and weaknesses in each domain (Library Domain, BiP Domain, RRO Domain) in relation to specific core issues (technical areas) emerged in the questionnaires. It proposes practical solutions to enhance interoperability and to provide a scalable roadmap.

[16] For a detailed description of ARROW system workflow, technological components and system architecture, see *D5.2 Specifications of Rights Information Infrastructure*

Table 1 Combination of Purpose/Information/System in the ARROW environment			
Steps of a diligent search	**ARROW task**	**Data needed**	**Resource**
1. The librarian has a book in hand and knows some elements (possibly the ISBN and some key data: author, title, publisher, year of publication, etc.). 2. The librarian needs to know which is/are the work(s) included in that book, in order to find out its rights status.	To link a manifestation (book in hand) to the underlying work	Bibliographic metadata to identify the book in hand (including the ISBN when available or other identifier in case of books without ISBN, i.e. published before 1970) and the underlying work. The underlying work may be linked to a manifestation either through an explicit relation in the database or at manifestation metadata level; or inferred indirectly from other key metadata.	Libraries BiPs
3. The librarian needs to know if that work is in public domain	To identify the rights status of a work (1): copyrighted or in public domain work	Copyright metadata (e.g. simply describing whether the work is / is not in public domain); authors and contributors data to the determine date of death; rights holders/rights holders agents (identifier and data) to contact to clear the rights status of the work	Libraries RROs
4. In case the work is not in public domain but is in copyright, the librarian needs to know if it is *in print* or *out of print.*	To identify the rights status of a work (2): if the work is copyrighted, identify if a work is *in print* or *out of print*	Data about the availability status of the manifestation(s) of the work; rights holders/rights holders agents (ID and data) to contact to clear the availability status of the work.	BiPs RROs

Table 1 Combination of Purpose/Information/System in the ARROW environment			
5. If the work is *in print*, the librarian is likely to need to contact the publisher in order to ask for permission to digitise the book.	To link the work to the rights holders/agents	Bibliographic data, in particular contributors data, publishers' data, copyright data; authors authority files and publishers directories (to unambiguously identify the rights holders' names)	Libraries BiPs RROs
6. If the work is *out of print,* the librarian can in any case contact the publisher but sometimes the librarian will need to find the author, who retains the right to digitise the work.	To identify relevant rights holders/agents in relation to the requested use of a work	For each rights holders/rights holders agents, the rights they control (reprographic rights, digitisation rights etc.)	RROs
7. It is possible that the librarian does not need to find actual rights holder for asking permission, but the authorised (by mandate or by law) agent.	To link the rights holder/agent name to a "person" and an address where rights can be cleared and licensed (in case of out of print works, to the appropriate clearing service)	Relation between rights holder/agents' names and persons to be contacted (the link between rights holders' names in bibliographic records and rights holders' data in parties databases could be based on a rights holders identifier); parties data, in particular rights holders/agents contact details	BiPs RROs
8. To avoid duplication of effort in digitisation	To link a work to a digital collection(s), whenever the work has already been digitised	Bibliographic metadata including a link to the digital version of the work	Libraries (digital collections)

2. The Library Domain

2.1 Introduction

The first domain to be assessed in terms of interoperability for ARROW is the "Library" domain. In the context of ARROW, it is specifically *National Libraries*, those libraries who are "designated and funded by a national government to serve the nation by maintaining a comprehensive collection of the published and unpublished literary output of the nation as a whole" [17] that have been identified as including the data that is most important for ARROW. National Libraries are already key stakeholders in the ARROW project[18]. In particular, most national libraries are responsible for compiling the *National Bibliography,* "… a combination of the authoritative and comprehensive records of the national output (i.e., products of the national publishing industry) of a country…[19]". National Libraries are therefore not only collecting the publishing output of a nation, but also describing it through the creation of high quality metadata. Thus comprehensiveness and quality are two key factors for identifying National Library metadata as being crucial to ARROW.

The high quality bibliographic metadata which is created by National Libraries is stored in databases such as the national library's catalogue. National Library catalogues have therefore been identified as key systems for inclusion in the ARROW rights information infrastructure. These catalogues are therefore an excellent initial data source for gathering the necessary bibliographic information needed in order to conduct a diligent search. Bibliographic elements such as author, title, publisher and place of publication are all standard elements within a library metadata record. In addition, ***country of publication*** and ***year of publication*** can be determined from more in depth analysis of national library bibliographic records.[20] This will be

[17] According to Joan M. Reitz's Online Dictionary for Library and Information Science ODLIS, a National Library is "a library designated and funded by a national government to serve the nation by maintaining a comprehensive collection of the published and unpublished literary output of the nation as a whole, including publications of the government itself. Most national libraries are also responsible for compiling a national bibliography, and some serve as the legal depository for works protected by copyright in the country." For further information please refer to: http://lu.com/odlis/

[18] The National Libraries of the France, Germany, the Netherlands, Norway, Slovenia, Spain and the United Kingdom are contracting partners in the ARROW project. In addition, the National Library of Finland is an associated partner in the project. For further information please refer to: http://www.ARROW-net.eu/partners

[19] Please refer to the draft version of the IFLA Working Group on Guidelines for National Bibliographies: *Guidelines for National Bibliographies in the age* edited by Maja Žumer. Available online at: http://www.ifla.org/en/publications/guidelines-for-national-bibliographies-in-the-electronic-age with reference to Bell, B.L (1998). *An Annotated Guide to Current National Bibliographies.* 2nd rev. ed. München: Saur. The final version of these guidelines was published by De Gruyter in late 2009. For further information please refer to: http://www.degruyter.de/cont/fb/bb/detailEn.cfm?id=IS-9783598242878-1

[20] In library created bibliographic records, the place of publication often refers to the city where the book was published. For a human reader of the bibliographic record, if "London" is recorded in the place of publication field, s/he can automatically infer that the country of publication is therefore the United Kingdom. However, with the ARROW system, such inference would need to be done by a machine. MARC formats have various "fixed-fields"

essential to the ARROW system for determining which BiP or RRO system for querying in the next stage of the ARROW workflow. In addition to authors, information about contributors, i.e. translators, illustrators, editors etc. who may also be rights holders in the context of ARROW, is often recorded in library created bibliographic records. Indeed, MARC, which is a range of international standard digital formats for the description of bibliographic items,[21] has a specific list of over 170 "relator codes" which provide information about such contributors in a machine-readable way[22].

In addition to bibliographic control, authority control[23] is a key task often undertaken by National Libraries. The ARROW system will be able to use these authority files, which provide consistent, standard and authorised forms of personal names (e.g. authors), corporate names (e.g. publishers) and uniform titles[24] to facilitate the identification of both rights holders and the "work" underlying the particular book (manifestation) that the librarian wishes to digitise. From a National Library's authority data, the ARROW system will be able to determine, the date of death (if applicable) of authors and contributors.

Finally, National Libraries play a central role in the digitisation of a nation's cultural heritage. Often links are created from a national library catalogue to a national library's digital repository once it is available. National Libraries will therefore play a crucial role in determining whether a work has already been digitised, to prevent the duplication of effort.

---

that contain "encoded" information. For example, within MARC 21, the country of publication is included within the 008 field/ in MARC 21, see: http://www.loc.gov/marc/bibliographic/bd008a.html, the character positions 15-17 include information about regarding the *Place of publication, production, or execution.* If completed by the cataloguer, these fields will be populated with either a two or three letter country code, which will allow the ARROW system to determine the country of publication.

[21] There are a range of MARC formats, including MARC 21 (see: http://www.loc.gov/marc/) and UNIMARC (see: http://archive.ifla.org/VI/8/unimarc-concise-bibliographic-format-2008.pdf). The complete printed version of the "UNIMARC Manual" is published by K.G. Saur : http://www.degruyter.de/cont/fb/bb/detail.cfm?id=IS-9783598242847-1.

[22] For the entire list of MARC 21 relator codes, please refer to:
http://www.loc.gov/marc/sourcecode/relator/relatorlist.html#codes
For the entire list of UNIMARC relator codes, please refer to the printed edition of *Manual UNIMARC : bibliographic format* available from K.G. Saur (http://www.degruyter.de/cont/fb/bb/detail.cfm?id=IS-9783598242847-1)

[23] According to the Functional Requirements for Authority Data (IFLA model published in 2009 by K.G. Saur): "Authority control, which means both the identification of entities represented by controlled access points and the ongoing management of them, is integral to the functioning of a catalogue. Authority control is beneficial to cataloguers able to identify and distinguish between the controlled access points within a catalogue. More importantly, authority control benefits end users able to search any controlled form of an author's name or of a title to retrieve bibliographic resources within catalogues." (p.1). See:
http://www.degruyter.de/cont/fb/bb/detailEn.cfm?id=IS-9783598242823-1

[24] According to *Harrod's librarians glossary and reference book* by Raymond John Prytherch, Ashgate Publishing Ltd, 2005, a ***Uniform title*** is "The distinctive title by which a work, which has appeared under varying titles and in various versions, is more generally known, and under which catalogue entries are made. Also called 'Conventional title', 'Filing title', 'Standard title'."

As identified in Table 1 (see p 11), the Libraries domain will be most relevant for the following ARROW tasks: (1) to link a manifestation (book in hand) to the underlying work, (2) to identify the rights status of a work: copyrighted or in public domain work,(3) to link the work to the rights holders/agents (4) to link a work to a digital collection(s), whenever the work has already been digitised.

Keeping in mind the ARROW purpose and the information needed, the results of the questionnaires have been analysed according to core technical areas, relevant for interoperability among the different national libraries and with the other systems (BiP, RROs) in the ARROW environment: (1) rights holders information linked to bibliographic information (authors and other contributors); (2) bibliographic information contains information about works and links between different manifestations of the same work; (3) contains information about copyright status of work: copyrighted or in public domain; (4) identifiers used; (5) metadata formats for importing, exporting and exposing metadata; (6) protocols for importing, exporting and exposing metadata; (7) query method and criteria and (8) systems used for encoding characters.

In the European countries surveyed in this phase of the ARROW project, within the library domain the following databases have been identified as relevant for the ARROW project. These are:

Austria – Online Catalogue of the National Library of Austria [AustCat] and the Name Authority File [PND]

The Online Catalogue of the National Library of Austria is divided into three parts; (1) Online Catalogue of the Austrian National Library from 1992 onwards. This online catalogue lists all works that have been acquired since 1992: all printed items (including microforms and AV materials) appearing from 1992 onwards; all university theses that have appeared from 1989 onwards; new acquisitions with an earlier date of publication. (2) Online Catalogue of the Austrian National Library 1930-1991. This catalogue lists the printed items published in the period 1930-1991 and held by the Austrian National Library. (3) Online Catalogue of the Austrian National Library 1501-1929. This catalogue contains all printed materials that appeared during the period 1501-1929 and that were acquired and indexed by the Austrian National Library before 1992. These catalogues include approximately 8 million books and objects. In addition, there is the Personennamendatei file, which is the Name Authority File. The Austrian National Library is the name authority coordinator for all Austrian libraries.

United Kingdom - British Library Integrated Catalogue [BLIC]:

The British Library Integrated Catalogue [BLIC] is The British Library's main catalogue, which unites a number of previously separate catalogues. It contains records for well over 13

million books, serials, printed music, maps and newspapers. The British Library has several different catalogues[25] which hold bibliographic data about its collections, and is also about to go live with its Integrated Archives and Manuscripts system. For the purposes of ARROW, the British Library Integrated Catalogue [BLIC] is the most relevant. For authority control, the British Library uses the Library of Congress Name Authority File, for the identification of authors, corporate bodies, etc.

France - BnF General Catalogue – Bibliographic File [BnF Bib] and **BnF Authority File [BnF Auto]:**

The National Library of France's [BnF] General Catalogue comprises of two main files; a bibliographic file [BnF Bib] which comprises of mainly but not only monographs, serials, series of printed materials. The file contains 10.8 million records plus 2.6 million analytical records. 73% of records are of printed monographs and 6% are of printed serials. 200,000 new bibliographic records for all types of material were entered in the bibliographic file during 2008. The BnF Authority File [BnF Auto]; contains authority records covering mainly persons, bodies, titles, topics. 5m records for all types of entities: persons, corporate bodies, titles, RAMEAU (French indexing language), classification numbers. 38% are authority records with a high level of detail.

Germany – Germany National Bibliography [DNB]

The German National Library's database is called the Deutsche Nationalbibliografie [DNB], the Germany National Bibliography. Although it is described as a single database, it contains the following separate segments which appear to be interlinked but also with stand-alone capabilities: DNB (the bibliographic data file), containing 9.9 million records; PND (the persons authority file) with 3.3 million records; GKD (the corporate bodies authority file, with 1.2 million records; and SWD (the subjects authorities file) with 820,000 records.

Italy – The Central National Library of Florence (BNCF) Library Catalogue [BNCF OPAC]; the Italian National Bibliography [BNI], SBN Index and the SBN Authors Name Authority file.

BNCF is one of the two National Libraries in Italy[26]. It manages the BNCF OPAC database, i.e., the library catalogue, and is also in charge of legal deposit. The BNCF OPAC also collects records from 7 other affiliated libraries[27] and is downloaded into the SBN Index (see ICCU

[25]Other catalogues that may be relevant to the ARROW project include the Sound Archive Catalogue, the Manuscripts Catalogue, the English Short Title Catalogue [ESTC] and the Maps Catalogue.
[26] The National Central Library of Rome is the second National Library in Italy (http://www.bncrm.librari.beniculturali.it/index.php?en/1/home)
[27] The BNCF affiliated libraries are: Biblioteca Nazionale Centrale Firenze; Biblioteca dell'Archivio di Stato; Biblioteca del Conservatorio Luigi Cherubini; Biblioteca della Fondazione studi storici Filippo Turati; Biblioteca

entry below). The BNCF OPAC includes both monograph records (ca. 2.1 million records) and serial records (152.425 records). BNCF also produces the Italian National Bibliography (BNI). The BNI forms a subset of the BNCF OPAC. There are currently 620,000 entries in the Italian National Bibliography. BNCF OPAC and BNI are not actually different databases, as far as content and format are concerned, but their business and availability models are different[28].

The Italian National Library Service (SBN) is the Italian libraries network promoted by the General Direction for Library Heritage and Cultural Institutes in cooperation with the Regions and the Universities. The Central Institute for the Union Catalogue of Italian Libraries and for Bibliographic Information (ICCU) coordinates, promotes and manages the catalogue of the network of the National Library Service. The SBN Index includes approximately 10 million records. This is the national bibliographic service for Italy. ICCU also manages the SBN Authors Name Authority file, which includes approximately 100,000 controlled entries, and with an archive of over 3 million names.

Spain - BNE Catalogue and BNE Authorities

The National Library of Spain (BNE) holds two main databases. One contains the bibliographic and holdings records of all the items the library owns; its short name is BNE Catalogue. This contains approximately 3.4 million records. The other is the Authority database; its short name is BNE Authorities. It contains approximately 450,000 controlled entries from an archive of 4.5 million entries.

Slovenia - Katnuk and Conor

The National Library of Slovenia [NUK] manages two main databases, Katnuk and Conor[29]. Katnuk contains bibliographical data about all Slovenica as NUK is entitled to legal issue. It also contains data about Books in print (BIP agency within NUK) and all the ISBN numbers (ISBN agency within NUK) are also assigned in NUK (ISBN agency) and subsequently contained in the Katnuk records. It contains approximately 1 million records. The CONOR.SI database for authority control of personal and corporate names was created in December 2001. It included 20,194 authority records, automatically created by the programme, for Slovenian researchers whose works had been catalogued in the COBIB.SI database, and for

Attilio Mori dell'Istituto geografico militare; Biblioteca Marucelliana; Biblioteca Medicea Laurenziana; Biblioteca Riccardiana and Biblioteca dell'Istituto di Scienze Militari Aeronautiche.

[28] The BNI database is hosted and managed by Licosa and access is allowed by subscription. BNCF OPAC records are extracted in UNIMARCXML format and imported into the BNI database.

[29] The National Library of Slovenia [NUK] manages the content and editing of the records stored in Conor. However, the Slovenian Institution of Information Science [IZUM] hosts and provides the database for Conor.

Slovenian authors who had appeared in the Slovenian bibliography in the preceding four years.

Norway – BIBYSIS, NORBOK, NORART and NORNOTER

The National Library of Norway is a library, a bibliographic agency and an ISBN, ISSN and ISMN agency. It manages BIBYSIS, NORBOK, NORART and NORNOTER. BIBYSIS is the catalogue for university libraries, the national library, college libraries, and a number of research libraries and institutions. It contains bibliographic records for several media types. NORBOK is the database containing the national bibliography (books). NORART is a database for indexed Norwegian and Nordic articles. NORNOTER is a national bibliography for music sheets. BIBYSIS holds 4.9m records; NORBOK holds 426,000 records; NORPER holds 26,891 records NORNOTER holds 28,202 records.

The Netherlands – KB-MDO

KB-MDO is the National Library of the Netherlands integrated metadata repository. It includes approximately 14 million metadata records describing primarily text and still image resources. The KB-MDO metadata records have been divided into many sub-sets. The KB-MDO includes the Shared Automated Cataloguing System of the Dutch Research libraries (OCLC-GGC). This shared catalogue includes the metadata records of most Dutch University libraries.

2.2 Technical area 1: rights holders information linked to bibliographic information (authors/other contributors)

The metadata records included in National library catalogues contain several bibliographic elements such as author, contributor, publisher that may be useful to determine the rights holder of a particular work. This information is enhanced by the practice of authority control, which ensures consistency in the forms of the names, including personal names and corporate bodies. With authority records themselves, additional information such as the birth and death dates of the person described, plus notes fields for including biographical and other historical data are included.[30] Extensive linking practice between bibliographic and authority data is undertaken in most National Libraries. All this controlled information, will help the ARROW system to disambiguate between rights holders and, in addition, provide additional information to help the rights holders to be located. Furthermore, the link to rights holders information, and in particular the date of death of relevant rights holders, may help to determine whether a work is in copyright or in public domain (See Technical area 3). It is worth noting that 4 out of

[30] For more information about MARC 21 Authority Records please refer to: http://www.loc.gov/marc/authority/ For more information about UNIMARC Authority Records please refer to: http://www.degruyter.de/cont/fb/bb/detailEn.cfm?id=IS-9783598242861-1 The previous edition is however available on line at: http://archive.ifla.org/VI/8/projects/UNIMARC-AuthoritiesFormat.pdf

the 7 National Library partners in ARROW, provide their authority files to the Virtual International Authority File (VIAF)[31] project. It would therefore be interesting to explore a partnership between ARROW and the VIAF project with regard to accessing national library authority data.

Austria

The Name Authority File [PND] database contains about 2.6 million personal names. Whereas in the participating libraries in Germany it is compulsory to link from the manifestation to the authority record and (in most cases) to create a new authority record if one does not yet exist for the person in question, in Austria it is optional. The authority records for books with a publishing year before 1850 are therefore comprehensive, whereas for books relevant to ARROW (publishing year 1850 - mid/end 20th century) very few authority records exist, especially if an author has published only one book and/or is unknown.

France

In the BnF Bibliographic file, entities having an intellectual, artistic or commercial interest in the resources described - authors, translators, illustrators, publishers, etc. - are listed. Public access via the web is allowed. The BnF also holds rights holder information in BnF Authority file. Details are held for the identification of corporate bodies who hold, or whose authors hold, the rights to the documents described. Addresses and contact details of individuals are kept, but not shown on the free public access to the records that is allowed via the web. The latter information is kept confidential. At the end of 2008, 3.6m authority records for persons were held, of which 1.2m were reasonably comprehensive; and 660,300 authority records were held, of which 281,000 were reasonably comprehensive. Both files have been made available to the VIAF project[32].

UK

The British Library Integrated Catalogue (BLIC) uses the Library of Congress Name Authority File, for the identification of authors, corporate bodies, etc. This file is administered by the Library of Congress, which is one of the original VIAF partners[33]. In this way, VIAF could be used in the ARROW system in combination with the BLIC data. The BLIC uses information about authors for information, rather than to establish the identity of rights holders, and does not hold rights data specifically, except for journals; the journals rights information is acquired from the CLA. (The copyright field is recorded in MARC 21 field 540). In the UK, information about rights is held exclusively by the three (four including DACS) interrelating rights organisations.

[31] For further information about the VIAF project, please refer to: http://www.oclc.org/research/projects/viaf/
[32] Further information about the VIAF project is available at: http://www.oclc.org/research/projects/viaf/
[33] The partners are The Deutsche Nationalbibliothek, the Library of Congress, the Bibliothèque Nationale de France, and OCLC.

Germany

The PND file contains 3.3m records which link persons (authors, contributors, etc.) with titles. The GKD file links corporate bodies (contributors, etc.) with titles. Full public access is allowed to both. The PND and GKD authority files collect information to enable the identification of an author from different sources, including the corresponding national authority files for foreign authors. Within the DNB (the bibliographic data file) there is a (forward and backward) link between the bibliographic record for a title and authority record (person or corporate body) using the internal record IDs[34]. The German National Library is a founding partner in the VIAF project.

Spain

The BNE Authority file is used for controlling the access points of the bibliographic records. Metadata may be freely accessed. A total of 4,537,376 records are held. It does not hold information about rights holders as such, nor does it record *in print*, *out of print* or *orphan work* status. Links are provided between the bibliographic records and the authority records. One authority record may, for example, be linked to several bibliographic records. These links are based on the identifier of each record. The field 001 of MARC is used for the identifier of a bibliographic record or for an authority record. The Spanish National Library provides their authority files to the VIAF project.

Italy

The ICCU SBN Authority file has fields for dates of birth and death and an annotation field. It links to other sources of information, the type of author (person, institute or conference), and records the update date of the authority record. There is a link between the author name in the SBN Authority File and the related bibliographic records in the SBN index. The SBN identifier for author's names is used for this purpose. ICCU provide the Italian authority files to the VIAF project.

BNCF similarly provides links between the bibliographic record in the BNCF OPAC or the BNI and the author name in the authority file. This link is made by including the Author's ID in the bibliographic record.

Netherlands

The Dutch response does not include provide any information as to whether there is a link between the rights holder information and the bibliographic information.

Norway

[34] For example, the title record http://d-nb.info/99140615X is linked to author record http://d-nb.info/gnd/111941857

The Norwegian BibSys catalogue includes links between the bibliographic records and the authority records for names of corporate bodies. One bibliographic record can be linked to several authority records. The links are based on the identifier of the authority record. In the BibSys system, there are indexes of identifiers for bibliographic records and indexes of identifiers for authority records. The links are made by making mappings between the two indexes. The National Library of Norway is considering providing the Norwegian authority data to the VIAF project.

Slovenia

In the Slovenian Katnuk database, links are made between bibliographic records from the same author. The relationship is n:1. This linking is possible using the authority files in the Conor database.

Conclusion

Linking between bibliographic records and authority files is standard practice in most national libraries, although the internal practices for this linking may differ slightly between the different libraries. The quality and extensiveness of the information included in the authority files may also differ between the various national libraries. However, several of the National Libraries involved in the ARROW project are either partners in the VIAF project, have provided their authority files for testing by the VIAF project[35] or are considering providing their authority files to the VIAF project. It is recommended that the possibility of using the VIAF system as a data provider for the ARROW system is explored further.

2.3 Technical area 2: bibliographic information contains information about works and links between different manifestation of the same work

For the ARROW system, it is important that links can be made between the manifestation that the librarian wants to digitise and the underlying work, as, according to the HLEG definition of Out of Print Works, a work is out of print when all its manifestations are out of print. It will also be useful for the ARROW system to be able to provide information about different manifestations of the same work, for example, if a particular manifestation has already been digitised by a library. Such information will help prevent duplication of effort.

Austria

The Austrian National Library records the Uniform title in the bibliographic record, for example, for translations, where the original title is mentioned. Links are also provided between bibliographic records, for example in the case of series. Links are provided to some

[35] The British Library (who use the Library of Congress / NACO authority files), French National Library, German National Library, ICCU, and the Spanish National Library are participating in the VIAF project. The National Library of Norway is considering providing the Norwegian authority data to VIAF.

digital items, in particular scanned tables of contents and digitised versions of historical newspapers.

Within the bibliographic file of the National Library of France's [BnF] General Catalogue the standard practice is that when an authority record for a work exists, it must be used as the access point to the bibliographic records describing manifestations of the work. For example, for translations, the original title is mentioned[36]. In addition, when the same work is published in different media, this is indicated in the bibliographic record: for example, 'other edition in the same medium' (Field 431 in INTERMARC) or 'other edition in another medium' (Field 432 in INTERMARC). When BnF has digitised a printed version, the printed and digital manifestations are linked to the same bibliographic record. The call number of the printed version is recorded in the holdings of the digital version. Each record established for each manifestation is also identified by an ARK (Archival Resource Key)[37] identifier (built on the 001 identifier) which is persistent and clickable.

UK

Uniform titles are used in the British Library Integrated catalogue [BLIC] and are recorded in the MARC field 240. For example, for translations, the original title is mentioned. The British Library notes that Resource Description and Access (RDA)[38] which will replace the Anglo-American Cataloguing Rules 2nd Edition (AACR2)[39] makes the relationships between these different levels of information more explicit. The British Library has already embedded a degree of these different relationships in the structure of their catalogue which will translate partially into the new RDA structure, including the use of Uniform Titles for a 'Work'. BLIC is tracing such relationships through ISBNs, SFX, theses and URLs. Bibliographic records contain ISBNs from related manifestations (paperback, hardback or e-versions). Bibliographic records for theses contain a link to the digitised version in the EThOS database. The British Library also records some data relating to digital versions in textual format – e.g., 'also available in print'.

Germany

At present, the German National Library does not provide links from the DNB (the bibliographic data file) to other manifestations, for example the digital version of a print

[36] For an example of how the BnF catalogues a French translation of an English novel, see: http://catalogue.bnf.fr/ark:/12148/cb346498783/PUBLIC

[37] For further information about Archival Resource Keys (ARKs) see: http://www.cdlib.org/inside/diglib/ark/

[38] Resource Description and Access (RDA) will be "the new standard for resource description and access designed for the digital world. Built on the foundations established by AACR2, RDA provides a comprehensive set of guidelines and instructions on resource description and access covering all types of content and media." For further information please refer to the RDA website: http://www.rdaonline.org/

[39] The Anglo-American cataloguing rules 2nd Edition (AACR2) "are designed for use in the construction of catalogues and other lists in general libraries of all sizes. The rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time." For further information please refer to the AACR2 website: http://www.aacr2.org/index.html

work. However, it does plan to do so. The German National Library does however record the Uniform title in the "Einheitssachtitel" field in the bibliographic record, for example, for translations, where the original title is mentioned.

Spain

In the BNE catalogue, the National Library of Spain provides links to other manifestations. For example, when a part of the manifestation is described the MARC21 77X field is used to express the relationship. To indicate different versions, the MARC21 77X field is used, accompanied by a textual note. The MARC21 77X fields are linked to other records using the MARC21 001 field, which identifies the records in the database. The Spanish National Library does however record the Uniform Title in the "Título uniforme" field in the bibliographic record, for example, for translations, where the original title is mentioned. For linking to the digital manifestation, a link is supplied using the 856 field of the MARC21 record.

Italy

For the BNCF OPAC, the BNI and the SBN Index, the relationship between different editions is tracked by the means of the Uniform Title (Titolo uniforme) field. Currently, the Uniform Title is only generated in cases where the title has changed over time or it is a translated title. In June 2009, the new Italian cataloguing code, REICAT (Regole Italiane di Catalogazione) was published. REICAT, is designed to meet the needs of large cooperative catalogues and is FRBR compliant[40]. In the new cataloguing code, the Uniform Title is a mandatory field. Each bibliographic record created after the REICAT implementation should therefore track all relationships between manifestations.

Since the 1990s, BNCF has been engaged in several digitisation programmes, both of modern works and works belonging to other periods in time. In the modern collection, almost 5,000 monographs, totalling 1.6m pages, have been digitised[41]. When digitised material (partial or complete) is available, the bibliographic record contains the link to the digital version (using the MARC indicator 956, instead of the usual 856 URL used for born-digital material). If the rights have been cleared or the work is in the public domain, the digitised version can be accessed via the internet.

In the ICCU SBN Index, if a digital version of a print work exists it has a separate record. The relationship is indicated using MARC field 856 or the uniform title field. If the manifestation has been digitised by one of the SBN libraries or through a national project such as InternetCulturale, the digitised copy may be referenced in the location field. Since April

[40] For further information about REICAT please refer to: http://www.ifla.org/files/hq/papers/ifla75/107-petrucciani-en.pdf

[41] The main reason for carrying out the digitisation was to preserve the material after a flood that took place in 1968 and partly destroyed the library collection.

2009 it has been possible to search both the SBN Catalogue and the InternetCulturale repository via a single search on the OPAC. ICCU is considering the implementation of the ISTC for the SBN Index once it is available.

Netherlands

In The National Library of the Netherlands' integrated metadata repository (KB-MDO), links between the different manifestations of the same work are made either by a textual reference, by a URL or pointer. For the textual references, a reference is made to the Uniform Title in a notes (annotatie) field within the bibliographic record. For example, if the manifestation is a translation of the original work into Dutch, then in the notes field it says the manifestation is a translation of the "title of the original work". These methods are also used for linking to the digital object where it is available.

Norway

In the National Library of Norway's BIBSYS system, there are notes on different printed editions that appear in the catalogue. For example, if the manifestation is a translation of the original work the original title appears in the "original title" field. There are currently no relationships created between print and digital manifestations, however, BIBSYS does provide links to the digital content in NB-digital.

Slovenia

In the National Library of Slovenia's KATNUK database, there are currently no relationships between the manifestation and the underlying work. However, there are relationships between manifestations using a textual note in the 452 CoMARC field. For links to the digital version of a manifestation, the field 856u in CoMARC is used.

Conclusion

Within national library systems there are three types of connections made between manifestations; (1) between 2 or more bibliographic records describing manifestations (e.g. manifestations referring to the same series, or "is part of", etc; (2) between bibliographic records and authority records (one uniform title authority record to all bibliographic records describing the work identified by this uniform title) (3) between a bibliographic record and the digital object described in the bibliographic record. Making these connections has been implemented in various different ways in the national libraries surveyed; ranging from providing links from the bibliographic record to a Uniform Title authority entry, through the inclusion of textual information about other manifestations in the notes fields, to no links being available at present. Many of the National Libraries surveyed provide links to the digital of the manifestations; however, there seems to be quite a range of ways that this is done in the different libraries.

2.4 Technical area 3: contains information about copyright status of work: copyrighted or in public domain

For a librarian to be able to digitise a particular book, s/he needs to know what the copyright status of the book is; is the book "in copyright" or is it in the "public domain". It was therefore important to ask the national libraries surveyed for ARROW, if they held information about the copyright status of books or could infer it from other data sources (see Technical area 1). In this way, the ARROW project team could then decide whether nation library catalogues are good for providing such information for ARROW.

Austria

The Name Authority File [PND] database is relevant to ARROW objectives, as on the one hand the authors' names are normalised and on the other hand important data such as birth and death dates are stored with the name of the person. In addition, these persons are linked to their manifestations and vice versa the manifestations in the library catalogues are linked to the PND. However, no actual information on the copyright status of a work is held.

France

The BnF does not currently include, or plan to include in the future, metadata about the copyright status of a work, as to whether the work is in copyright or in the public domain, in the BnF catalogue général – Fichier bibliographique [BnF Bib]. Such information would not be applicable to be included in the Fichier d'autorité [BnF Auto]. Internally, within the BnF, decision trees are used in order to decide whether a document is in public domain, or not.

UK

The response from the British Library does not include information about whether the BLIC includes metadata about the copyright status of works.

Germany

The response from the German National Library does not include information about whether they hold metadata about the copyright status of works.

Spain

The response from the Spanish National Library does not include information about whether the BNE catalogue includes metadata about the copyright status of works.

Italy

Both the BNCF and the ICCU responded that including metadata about the copyright status of works is not applicable to their databases.

Netherlands

The National Library of the Netherlands responded that they do include metadata about the copyright status of the work if they are in public domain. They occasionally include the metadata if the status of the work is in copyright.

Norway
The response from the Norwegian National Library does not include information about whether their databases include metadata about the copyright status of works.

Slovenia
Records from the Slovenian National Library do not include information about the copyright status of works.

Conclusion
From the responses received from the National Libraries it is clear that the majority of the national libraries surveyed do not currently store information about the copyright status of works. Indeed, some of the responses noted that this question was not relevant for their databases. From these answers received, it would be fair to conclude that National Library catalogues should not be considered to be a suitable information source for the copyright status of works. However, it is possible that the data that is included in national library catalogues could be used for pre-processing to make an automatic suggestion as to whether a particular manifestation is in copyright. For example, if the date of publication or the death date of the author, are 70 years ago, or more, an "ARROW status" could be assigned to the record, saying that this manifestation ***could*** be in the public domain. If this information is not included a National Library's data, an additional query to the VIAF system to obtain, for example, the death date of the author, could be considered. For this, further investigation would be needed. However, an initial automatic assessment of the copyright status of a work within the libraries domain could be useful for ARROW.

2.5 Technical area 4: identifiers used

The availability of identifiers in the bibliographic data that will be used by the ARROW system will be a key to its success. As we move towards the semantic web, identifiers will play an even greater role in the provision of automatically linked data. For ARROW, key identifiers are the International Standard Book Number (ISBN)[42] – the unique machine-readable book identifier; the International Standard Text Code (ISTC)[43] - a numbering system to enable the unique identification of textual works and the International Standard Name Identifier (ISNI)[44] – which is a method for uniquely identifying the public identities of contributors to media content such as books, TV programs and newspaper articles. The purpose of this part of the survey was to

[42] For further information about the ISBN please refer to: http://www.isbn-international.org/
[43] For further information about the ISTC please refer to: http://www.istc-international.org/
[44] For further information about the ISNI please refer to: http://www.isni.org/

identify which National Library catalogues currently implement such identifiers or intend to in the future when they become available.

Austria

The Austrian National Library catalogues use both ISBNs and ISSNs. It is not clear from the questionnaire response as to whether it is intended to implement the ISTC and the ISNI when they become available.

France

The current identifiers used by BnF Bib are ISBNs (2,098,102 records having at least one ISBN held in April 2009); ISSNs; and URN[45]s (ARK standard). In the future, BnF intends to hold ISTCs (2010?) and ISNIs (2010?). BnF Bib and BnF Auto also use proprietary identifiers, for the unique identification of bibliographic records (001 INTERMARC Field).

UK

Current identifiers used by the British Library are ISBNs (3,691,294 records held, with one or more ISBNs); ISSNs (184,126 records held); and ISMNs. In the future it intends to hold ISTCs (2010?) and ISNIs (2011?).

Germany

The DNB uses the following identifiers: ISBN, ISSN, DIN report number, ISMN, URN, DOI. It plans to use ISSLs (no date), ISTCs and ISNIs (both in 2010).

Spain

The BNE uses ISBNs and ISSNs as identifiers. The response does not indicate whether it intends to implement ISTC and ISNI when they become available.

Italy

BNI and ICCU both use ISBNs and ISSNs as identifiers, and both are considering using ISTCs when they become available. BNI is also considering using DOIs and URNs. Both use proprietary identifiers: BNCF OPAC, for identification of records within the SBN system, holds 2.3m records[46]; BNCF BNI, for identification of records which are included in the Italian National Bibliography, holds 1,287,926 records, including records for serials. The ICCU SBN Index and SBN Record ID are used to identify records within the SBN system. About 10m records are held. The ICCU SBN Authority file is used to identify authors' names: about 100,000 records are held.

Netherlands

[45] URN = Uniform Resource Name.

[46] It should be remembered that the SBN ID record is not a persistent identifier: it may change over time.

The National Library of the Netherlands (KB) uses ISBNs, ISSNs, URNs and ILNs in its integrated metadata repository, KB-MDO. The questionnaire response did not state how many records contained each of these types of identifiers. The Digital Author Identifier (DAI)[47] is also available in those records harvested from Dutch university libraries into the KB-MDO system. The KB also uses various proprietary identifiers for different purposes. It is not stated as to whether the use of ISTC or ISNI is planned. However, there is some overlap between the DAI and the ISNI.

Norway

The NLN databases use ISBNs, ISSNs and ISMNs as appropriate. ISTCs are being considered.

Slovenia

Katnuk uses ISBNs and ISSNs as identifiers. DLD uses URNs (it holds 239,932 records). So far implementation of ISTC or ISNI has not been planned. Katnuk also holds 800 ISMNs. NUK uses a proprietary identifier: Cobiss ID. This is a catalogue record (sometimes the same work, always the same title) identification number in the Slovenian union catalogue. Approximately 1 million of these records are held.

Conclusion

ISBNs and ISSNs are widely used in the National Libraries surveyed. At present, it appears that there is a mixed response from national libraries regarding the implementation of ISTCs and ISNIs.

2.6 Technical area 5: metadata formats for importing, exporting and exposing metadata

In order for the ARROW system to be able to interoperate with the national library databases, it is important to know what metadata formats are currently in use within the national libraries for importing, exporting and exposing their metadata.

Austria

The metadata available from the National Library of Austria is available in MAB (Maschinelles Austauschformat für Bibliotheken or Automated Library Exchange Format)[48]. For data harvesting via OAI-PMH, the metadata is available in oai_dc format.

France

[47] For further information about the DAI please refer to: http://www.surffoundation.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifierdai.aspx

[48] For further information please refer to: http://www.d-nb.de/eng/standardisierung/formate/mab.htm

For importing records: BnF is using INTERMARC/B for bibliographic records[49] and INTERMARC/A for authority records[50]. *For exporting records*: BnF is using INTERMARC/b or UNIMARC/B for bibliographic records and INTERMARC/A or UNIMARC/A for authority records. Users of BnF records can choose the format they prefer. *For exposing bibliographic records*: BnF is using Dublin Core and MPEG21-DIDL in BnF OAI-NUM repository, Dublin Core only in OAI-CAT repository. *The formats to enable bibliographic data harvesting via Z39.50* are INTERMARC/B and UNIMARC/B. *For importing records:* BnF is using INTERMARC/B for bibliographic records and INTERMARC/A for authority records. *For exporting records*: BnF is using INTERMARC/B or UNIMARC/B. The BnF is considering supplying bibliographic records in MARC 21 format. For use in the ARROW project, a sample of 1 million records has been converted to MARC 21 format.

UK

For importing, exporting and exposing bibliographic metadata about manifestations, BLIC uses Catalogue Bridge. The formats in the databases are MARC 21 and UTF-8. The formats used for data import are UK MARC, MARC 21 and UTF-8; the formats used for data export are UK MARC, MARC 21, and MARC-XML.

Germany

For internal data management, DNB has a proprietary format (ILTIS). There are interfaces to the external data formats. The formats for data import are: MAB 2; MARC 21; ONIX XML and UNIMARC. The formats for data export are: MAB 2; MARC 21; ONIX XML; UNIMARC and Dublin Core. The formats to enable data harvesting are: MAB 2; MARC 21; Dublin Core.

Spain

For importing, exporting and exposing metadata about works, manifestations and parties, both the BNE catalogue and the BNE authorities file are in MARC 21 format. The formats for data import are ISO2709 and MARC 21. The formats for data export are ISO2709 and MARC 21. There are no formats to enable data harvesting.

Italy

For importing, exporting and exposing metadata about manifestations and parties, the format in the BNI database is UNIMARC. The format for data import is UNIMARC and the formats for data export are UNIMARCXML, ISO 2709, and TXT. The format to enable data

[49] A description of the INTERMARC/B fields used in the BnF products may be found at http://bnf.fr/pages/infopro/produits/pb-INTERMARC_B.htm. A description of the UNIMARC/B fields used in the BnF products may be found at http://www.bnf.fr/pages/infopro/produits/pdf/UNIMARC(B)_conversion.pdf

[50] A description of the INTERMARC/A fields used in the BnF products may be found at : http://www.bnf.fr/pages/infopro/produits/pb-INTERMARC_A.htm
A description of the UNIMARC/A fields used in the BnF products may be found at : http://www.bnf.fr/pages/infopro/produits/pdf/UNIMARC(A)_conversion.pdf

harvesting is Dublin Core unqualified (currently being experimented with)[51]. In the ICCU SBN authority file, the format in the database is SBNMARC. The format for data import and export is UNIMARC Authority; and no formats are available to enable data harvesting.

Netherlands

In the National Library of the Netherlands' integrated metadata repository, KB-MDO, metadata is stored in Dublin Core (DC), Dublin Core Extended (DCX) and partially in MARC21 and PICA+ formats. For data import, KB-MDO accepts DC, MARC 21, PICA+ and various legacy formats. For data export, including harvesting, DC, DCX and MARC 21 are available. It is noted that all the above mentioned formats are implemented, however, are sometimes only available for parts of the collection.

Norway

For importing, exporting and exposing metadata about manifestations, works and parties, there are many formats in the NLN databases. For import and export, they include MARC 21, DC and proprietary formats. MODS is also used for data export. Formats to enable data harvesting are DC, MODS and MARC 21.

Slovenia

For importing, exporting and exposing metadata about manifestations, works and parties, the format in the Katnuk database is Comarc. The formats for data export and to enable data harvesting are MARC 21; COMARC; SUTRS; and XML[52]. The format in the DLD database is DC. Its formats for data import (manifestations and works) are DC and UNIMARC; its formats for data export and for data harvesting is DC.

Conclusion

MARC is used in each of the national libraries surveyed, albeit in the various different flavours, where MARC 21 and UNIMARC being the most widely used. Dublin Core is also widely used. Many of the national libraries also support a range of differing proprietary formats. The choice of format is made according to use.

2.7 Technical area 6: protocols for importing, exporting and exposing metadata

In order for the ARROW system to be able to interoperate with the national library databases, it is important to know what protocols are currently in use within the national libraries for importing, exporting and exposing their metadata.

Austria

[51] See http://www.bncf.firenze.sbn.it/progetti/unimarc/slim/documentation/unimarcslim.html?rigamenu=UNIMARCXML .

[52] See http://home.izum.si/cobiss/xml/COMARC.xsd

The metadata in the Online Catalogue of the National Library of Austria [AustCat] is available for harvesting via OAI-PMH. The Name Authority File [PND] database can be queried via Z39.50.

France

For bibliographic data: The protocol for searching the BnF Catalogue Général Bibliographic file BNF is currently Z39.50. Searching for manifestations and holdings information is possible. In the future, the development of a Z39.50 web application that can be launched when a searching session is opened is being planned, along with the development of SRU web services to search bibliographic data. Records from the BnF Catalogue Général Bibliographic file are available for both data export and data harvesting using via OAI-PMH from the BnF OAI-CAT repository.

For authority data: It is not possible to search the BnF catalogue général – Fichier d'autorité using a search protocol. There are also no protocols for data import. For data export, the available protocols are FTP + (ISO-2709). No protocol to enable data harvesting: not applicable, as there is no repository.

Bibliographic and authority records are also available for data export via FTP and ISO-2709.

UK

It is possible to search the British Library Integrated Catalogue [BLIC] using Z39.50.

Germany

The DNB (the bibliographic data file), PND (the persons authority file) and GKD (the corporate bodies authority file) are searchable via HTTP, Z39.50 and SRU.

Spain

The BNE catalogue can be searched using Z39.50.

Italy

The BNCF OPAC and the BNI are both searchable via SRU. The SBN Index and the SBN Authority File are searchable by both SRU and Z39.50.

Netherlands

The National Library of the Netherlands integrated metadata repository; KB-MDO is searchable via the SRU protocol. The metadata is made available for harvesting via OAI-PMH.

Norway

All the Norwegian databases are searchable via SRU and Z39.50 in addition to via a proprietary search protocol. The metadata for the Norwegian database is made available for harvesting via OAI-PMH.

Slovenia
Katnuk and CONOR are searchable via Z39.50. dLib.si is searchable via SRU.

Conclusion
The majority of the national libraries surveyed provide searchable access to their catalogues, and in some case their authority files via Z39.50. In some case, searchable access to the authority files is not provided. In addition, searchable access via SRU is also available in several cases. Also, protocols such as FTP and ISO 2709 are available for data export in some cases, including for some authority files. OAI-PMH repositories are available to allow the harvesting of metadata from national libraries in several cases.

2.8 Technical area 7: query method and criteria

It is important to know which fields of the various national library databases are available for searching. In addition, it is important to know whether it would be possible for the ARROW system to search the various database free of charge, or if there will be licensing implications for including a particular database in the ARROW system.

Austria
The catalogue of the Austrian National Library has a simple and an advanced search interface. Via the simple search interface it is possible to search in the author, title, keyword, theme, institution, publisher, year, classification number, ISBN and ISSN. In the advanced search interface, the same search elements are also available plus other filters such as searching between particular dates, languages or for example, dissertation types. The Name Authority File [PND] is searchable via a number of fields including title, person name, number, subject, corporate body, publisher, location, coded data, classification, content table, publication year; filter functionality: kind of material, location /catalogue and classification. The business conditions for the inclusion of the database in the ARROW project should be negotiated with the German National Library.

France
For querying the BnF General Catalogue – Bibliographic File [BnF Bib] via Z39.50, about 20 Bib-1 attributes of the Z39.50 protocol are available for searching[53]. The data is available for free. Via the search interface on the web, most of the MARC elements and keywords are searchable. The data is also available for free. The search interfaces on the web are as follows: simple search; advanced search; search by equation; search by keywords of the records.

For the BnF Authority File [BnF Auto], a range of search criteria are available on the «Autorités BnF interface» of the general BnF catalogue on the web. They include type of

[53] For further information please refer to: http://www.bnf.fr/pages/infopro/produits/pb-protocole_z3950.htm

authority records: persons, bodies, titles, etc.; entry word of the name, title, etc.; key word(s) of the name, title, etc. Access to the date is available free of charge or restrictions via the web.

UK

For the British Library Integrated Catalogue [BLIC], the searchable data elements are: all elements in the MARC record, including the ISBN, title, author, publisher, date of publication, place of publication, etc. Access to these records via Z39.50 to SUTRS format is available free of charge. MARC records are exported to third parties under licence. The NACO authority file is received from the Library of Congress via overnight transfer every day. It does not say in the questionnaire response whether this could be made available to the ARROW project, however, it is assumed that a licence exists between the British Library and the Library of Congress for this data.

Germany

The DNB (the bibliographic data file), the PND (the persons authority file) the GKD (the corporate bodies authority file) and the SWD (the subjects authorities file) are all searchable via a range of data elements, including: title, person name, number, subject, corporate body, publisher, location, coded data, classification, content table, publication year; filter functionality: kind of material, location /catalogue and classification. Free access to the bibliographic catalogue data is available via the OPAC website[54]. Free access is also provided for data harvesting via OAI-PMH[55].

Spain

It is possible to search in almost any field of the records of the BNE catalogue. Access is provided free of charge. For the BNE authorities file, the main fields of the records can be searched alphabetically. Access is provided free of charge or with restrictions.

Italy

The web interface of the BNCF OPAC is searchable via Author; Title; Subject; Series; Class (CDD number); Class (CDD keyword); Place of publication; Publisher/Printer; Standard ID (ISBN, BNI, CUBI); Printer's device; Collocation; Keyword; Title identifier; Author identifier; Subject identifier; Inventory. Via SRU, the searchable data elements are dc.creator; dc.date; dc.description; dc.identifier; dc.language; dc.publisher; dc.subject; dc.title. Free access is provided via the web, via SRU and via OAI-PMH. Access to the BNI is via subscription.

Access to the ICCU SBN Index is available via the SBN OPAC. The searchable data elements are author; title; series; uniform title; place of publication; publisher; subject; Dewey

[54] For further information please refer to: https://portal.d-nb.de/opac.htm?method=showOptions#top
[55] For further information please refer to: http://www.d-nb.de/netzpub/ablief/pdf/automatisierte_ablieferung.pdf (available only in German)

classification; Dewey description; keywords; ISBN; ISSN; BNI identifier; SBN record ID; + filters. All records are available free of charge or restrictions. For the ICCU SBN Authority file, the searchable data elements are name; notes, SBN identifier; all fields; date of birth / beginning from; date of death / end by; author type. All records are available free of charge or with restrictions.

Netherlands

The bibliographic data within the National Library of the Netherlands' integrated metadata repository is searchable via the Dublin Core, plus some additional elements. Access to the data is free, except for the licensed material that is stored in the KB's e-Depot. Controlled access is provided to this material. Only the data that can be made free of charge is accessible for data harvesting via OAI-PMH.

Norway

For the National Library of Norway's databases, BIBYSIS, NORBOK, NORART and NORNOTER, all data elements are searchable. The data is available for free or under specific licensing terms. This includes the data that can be harvested via OAI-PMH.

Slovenia

For the Katnuk database, all the Comarc fields are searchable. The data in the Katnuk database is also available free of licensing terms. The Conor authorities database is searchable via Personal name, researcher's code, institutional code, corporate body and language. The data is also available for free for searching. However, for harvesting the data negotiation would be required.

Conclusion

It is possible to search the catalogues and the other bibliographic databases using an extensive range of fields. At this stage it is anticipated that all the fields that the ARROW system would want to search would be possible. Regarding access to this data, the majority of the bibliographic databases are available free of charge, or with some restrictions. The nature of these restrictions would need to be clarified at a later phase in the project. Some of the value-added authorities' data have licensing restrictions. In this revised version of *D4.2 Guidelines for Technical Interoperability* the decision has been made by the ARROW Management Board to use The European Library as the central metadata repository for the library-domain in the ARROW system. The use of the bibliographic records from the National Libraries who are partners in The European Library is already permitted.[56] The permission to use the National Library authority data may need to be considered separately.

[56] In spring 2010, a new partnership agreement between the owners of The European Library, the Conference of European National Librarians (CENL) and their partner National Libraries will be signed. The use of National Library metadata, for non-commercial exploitation, as part of the ARROW system is permitted within the bounds of this

2.9 Technical area 8: systems used for encoding characters

The ARROW system will need to be able to handle a wide range of character sets and encodings. It is therefore important for the ARROW team to be aware of what character encodings are already in use in the national library databases.

Austria

Unfortunately, it is not clear from the response from the National Library of Austria what character encoding scheme is used.

France

The system used by the BnF for encoding characters when metadata about manifestations, works or parties are imported, exported or exposed for the BnF Bib file is as follows: the format in the database is UTF-8. There is no format for data import. The formats for data export are ISO 5426 and UTF8 + Latin 1 for Z39.50. For BnF Auto, the format in the database is UTF-8. To import data: UTF8. The formats for data export are ISO 5426 or UTF8; the user has the choice. For the sample of 1 million bibliographic records in MARC 21 format, only MARC 8 character encoding is available.

UK

The system used by the British Library for encoding characters when metadata about manifestations is imported, exported or exposed is Unicode UTF-8. The formats for data import are UTF-8 and MARC-8; the formats for data export are UTF-8 or MARC-8 (if required).

Germany

The systems used by the DNB for encoding characters when metadata about manifestations or parties are imported or exported are as follows: MARC 8; UTF8; ISO 8859-1; ISO 646 (IRV) and ISO 5426; CP 850 (Ext. ASCII).

Spain

The system used by both BNE files for encoding characters when metadata about manifestations or parties are imported or exported is UNICODE-UTF 8.

Italy

The system used by the BNI for encoding characters when metadata about manifestations or parties are imported or exported is UTF-8, which only began to be used in 2009[57]. The ICCU ISBN Index also uses UTF-8, for both manifestations and parties data.

Netherlands

partnership agreement. The use of National Library authority data has not been explicitly stated in the partnership agreement and may need to be considered separately.

[57] Previously the proprietary encoding SW was used.

The National Library of the Netherlands uses UTF8 character encoding for storing, importing and exporting its data.

Norway

The National Library of Norway uses UTF8 for character encoding for both data import and data export for all of its databases. This applies to metadata about both works and manifestations.

Slovenia

The formats in Katnuk for encoding metadata about works or parties are UTF-8 and WIN-1250. The format for data import and export is UNICODE. The format in the DLD database is WIN-1250. Its formats for data imports are UNICODE, UTF8, and WIN-1250; for data export, they are UTF8 and WIN-1250.

Conclusion

UTF-8 is used in all of the databases of the surveyed national libraries[58]. In the libraries domain, it is recommended that UTF-8 is used for the character encoding standard for the ARROW system.

2.10 Conclusion for Library Domain

The Library domain is a key domain to enable the development of a digital rights infrastructure for Europe. In particular, the bibliographic and authority data included in (national) library databases, will contribute to enabling the ARROW system to (1) to link a manifestation (book in hand) to the underlying work (2) to identify the rights status of a work: copyrighted or in public domain work (3) to link the work to the rights holders/agents and (4) to link a work to a digital collection(s), whenever the work has already been digitised.

Using the questionnaires responses, the current state of interoperability for the ARROW system in the Library domain is as follows:

- For the ARROW system, it is important to be able to make links between rights holders' information and the bibliographic information about manifestations that a librarian wishes to digitise. The bibliographic data managed by the national libraries surveyed, contains important bibliographic elements for determining the rights holder of a bibliographic work. This bibliographic data is further enhanced by the authorised forms of names and additional information about rights holders provided in authority records. Linking between bibliographic and authority data is common practice in the national libraries, even if there are several ways that this is currently been done. It may be valuable for the ARROW to recommend the adoption of standard practices in this area, in particular, the use of (clickable) URIs

[58] Although the BnF uses UTF-8 internally, the sample of 1 million MARC 21 records provided is only currently available in MARC 8 format.

- For the ARROW system, it is important that links can be made between the manifestation that the librarian wants to digitise and the underlying work, as, according to the HLEG definition of Out of Print Works, a work is out of print when all its manifestations are out of print. Currently, linking between manifestations and the underlying work has been done using various different methods, some of which lend themselves more to machine processing (such as the use of a uniform title entry) than others (for example, including textual notes about the original title of a translated work). A machine processing method would be important for inclusion in the ARROW system. It would be worth exploring whether the processing of bibliographic records, using techniques based on the Functional Requirements for Bibliographic Records (FRBR) conceptual model would be useful to the ARROW project. Work on the processing of national library bibliographic records according to the FRBR model is already being carried out as part of the TELplus project[59]. Within this task, which involves 11 national libraries, including 4 national libraries who are involved in the ARROW project[60], works by Nobel Prize Winners of Literature and the Bible are being clustered. It is recommended that the ARROW project further explores the synergies of this work with the TELplus project team. At the time of this revised version (April 2010) of *D4.2 Guidelines for Technical Interoperability*, considerable work has been undertaken on the clustering and matching of bibliographic records by the ARROW project team. This work is described in detail in *D6.1 Rights Information Infrastructure – Release 1.*

- Many national libraries provide links from their national library catalogues to the digitised versions of manifestations. It would be useful for the ARROW system to be able to tell the librarian wanting to digitise a book whether the manifestation of the book that s/he wishes to digitise has already been digitised. In this way, it may not be necessary to digitise the same book again. From the national libraries surveyed, it seems that currently national libraries provide links to the digitised versions of manifestations in various different ways. This may be an area which would benefit from further standardisation.

- For the librarian to be able to digitise a particular book s/he needs to know what the copyright status of the book is; it is “in copyright” or is it in the “public domain”. From the questionnaire responses received from national libraries it is clear that the majority of national libraries do not hold information about copyright status. However, it may be possible to infer a “preliminary copyright status” from bibliographic and linked authority data held by national libraries, using in particular, the death date of the author and/or other rights holders and the date of publication of the manifestation concerned. If the relevant information is not available in the National Library authority data, an additional query to the VIAF system could be used to

[59] For further information about the TELplus project please refer to: http://www.theeuropeanlibrary.org/telplus/
[60] The ARROW partners who are taking part in the FRBR Task in TELplus are the National Libraries of France, Germany, Spain and the United Kingdom.

obtain this information. This "preliminary copyright status" could then be cross-checked and confirmed using the data available in the BiP and RROs domains.

- Libraries have used standard identifiers, such as ISBNs and ISSNs, since the standards were developed. However, implementation of newer standard identifiers such as the ISTC and ISNI are less prevalent in the library domain as yet, even though many librarians are involved in the development of such standards. ARROW has a role to play here in advocating the implementation of these new standards in the Library and indeed in the BiP and RROs domain as soon as they are available. In the meantime, the identifiers in the VIAF system[61] could perhaps be used until the ISNIs are available. The possibility of this could be explored further with the VIAF team.

- The family of MARC formats is a widely used standard in the Library Domain, however, from the national libraries surveyed, a wide variety of different "flavours" of MARC are used, with MARC 21 and UNIMARC being the most widely used. For exposing the data for external portals, such as The European Library[62] and Europeana[63], often Dublin Core-based formats are used. There are cross-walks available between the various different MARC formats, for example UNIMARC to MARC21[64] and there are also open-source tools available for making metadata cross-walks[65]. However, it may be useful for ARROW to consider the possibility of building on the interoperability already in place in existing services. The European Library[66] may be a good starting point for this exploration. At the time of writing this revised version of *D4.2 Guidelines for Technical Interoperability* (April 2010) the decision has been made by the ARROW Management Board to use The European Library as the central metadata repository for the library-domain in the ARROW system. In order to extract the work information from the bibliographic records, as described in *D6.1 Rights Information Infrastructure – Release 1,* the full MARC record is required. To reduce the risk of non-delivery of the system, by increasing its complexity during the pilot phase, the ARROW Technical Working Group recommended that MARC 21, which is more widely used in the potential ARROW pilot countries, is to be the only metadata format used in the library-domain for the ARROW pilot. However, the ARROW

[61] The VIAF system uses VIAF IDs to identify the linked authority records for particular people. For example, the VIAF ID for Lorcan Dempsey is: http://www.viaf.org/viaf/36978042

[62] For further information about The European Library please refer to: http://www.theeuropeanlibrary.org

[63] For further information about Europeana please refer to: http://www.europeana.eu/portal/

[64] For further information about the UNIMARC to MARC 21 conversion specifications please refer to: http://www.loc.gov/marc/unimarctomarc21.html

[65] There are various tools for making metadata crosswalks, including REPOX: http://repox.ist.utl.pt/ and the Universal MARC Record Converter: http://www.bl.uk/bibliographic/usemarcon.html

[66] Launched in March 2005, The European Library is a free service that offers a single point of access to the bibliographic and digital collections of the National Libraries of Europe. By the end of 2009, 46 out of 48 national libraries in Europe – including the 7 national libraries who are partners in ARROW – had included their collections in The European Library. For further information about The European Library please refer to: http://www.theeuropeanlibrary.org

Technical Working Group recognise that as UNIMARC is widely used within Europe's National Libraries[67], technical interoperability between UNIMARC and the ARROW system may need to be investigated further at a later stage in the project. Z39.50, and to a lesser extent SRU, are protocols that are widely used in the Library Domain, for the querying of both bibliographic and authority databases. Other protocols are also supported, but these would be of less interest to the ARROW project. Data harvesting into portals, using OAI-PMH is becoming more common place. Several of the National Libraries surveyed, already have functioning OAI-PMH repositories in place. To prevent the need to query each of the national library databases individually, it would be worth exploring the possibility of querying The European Library as a single point of access to the bibliographic and digital collections of the national libraries of Europe. Not only are all ARROW national library partners in The European Library, but all 48 National Libraries in Europe. This would prepare the ARROW system for the future extensibility.

- As the querying of bibliographic data is a core business in the Library Domain it is anticipated that the query fields offered would be sufficient for ARROW purposes. Many searching options are also available for querying authority data. Again, it is anticipated that these query fields would be sufficient for ARROW purposes. With regard to permission to access the bibliographic data of the national libraries, the majority of databases would be available free of charge to the ARROW system. However, it may be necessary to negotiate access for ARROW to some of the authority data. In this revised version of *D4.2 Guidelines for Technical Interoperability* the decision has been made by the ARROW Management Board to use The European Library as the central metadata repository for the library-domain in the ARROW system. The use of the bibliographic records from the National Libraries who are partners in The European Library is already permitted.[68] The permission to use the National Library authority data may need to be considered separately. It is recommended that it would be worth exploring a partnership between ARROW and the VIAF project[69] with regard to accessing national library authority data.

[67] As part of his MA Thesis, *Controlled Vocabularies in a Multilingual Federated Search Environment,* (November, 2007) Christoph Schmidt-Supprian discovered that among the 29 libraries that he surveyed, ten (35%) used UNIMARC or a derivative (COMARC), one used UKMARC and 17 used MARC21 or a derivative (HUNMARC, NORMARC). Two libraries (Spain and Finland) indicated that they were moving from national variations to pure MARC21. Only one library also used a non-standard local format, while six used Dublin Core application profiles." p38. As this research was undertaken in 2007, the percentage of the National Libraries in potential ARROW pilot countries using MARC 21 may have changed. It may be worth for the ARROW project team to survey the National Libraries in potential ARROW pilot countries to find out how many use UNIMARC and MARC 21 as their main metadata export format.

[68] In spring 2010, a new partnership agreement between the owners of The European Library, the Conference of European National Librarians (CENL) and their partner National Libraries will be signed. The use of National Library metadata, for non-commercial exploitation, as part of the ARROW system is permitted within the bounds of this partnership agreement. The use of National Library authority data has not been explicitly stated in the partnership agreement and may need to be considered separately.

[69] For further information about the VIAF project, please refer to: http://www.oclc.org/research/projects/viaf/

- As the ARROW system will need to deal with world-wide languages, the implementation of character encoding standards is essential. Within the Library Domain, UTF-8 is a fore-runner in this area. The character encoding implemented in the BiP and the RRO domains will also need to be taken into consideration for making recommendations for the ARROW system as a whole.

Following this analysis of the current state of technical interoperability within the library-domain, a set of guidelines, outlining the preferred standards for use by National Libraries wishing to participate within the ARROW system, have been defined The ARROW Technical Working Group do not rule out the possibility of further extending this set of preferred standards with other international standards widely used in the library-domain, such as UNIMARC. However, it has been recognised that, at this critical stage in the project, that the risk of non-delivery of the system needs to be reduced by decreasing the complexity of the system during the pilot phase of the project.

The following table provides a set of guidelines, outlining the preferred standards for use by National Libraries wishing to participate within the ARROW system, in relation to their respective technical area:

Table 2 - ARROW interoperability guidelines for Library Domain		
NL1	is a partner in The European Library	Technical Area 5 Technical Area 6 Technical Area 7 Technical Area 8
NL2	uses a MARC-based metadata format, MARC21 is preferred	Technical Area 5
NL3	uses OAI-PMH	Technical Area 6 Technical Area 7
NL4	uses the uniform title	Technical Area 2
NL5	implements FRBR	Technical Area 2
NL6	has author authority file	Technical Area 1 Technical Area 3
NL7	authority files are available through VIAF	Technical Area 1 Technical Area 3
NL8	implements ISTC	Technical Area 2 Technical Area 4
NL9	implements ISBN	Technical Area 4
NL10	implements ISNI	Technical Area 4
NL11	Uses U-TF8 character encoding	Technical Area 8

3. The BiP Domain

3.1 Introduction

The second domain to be assessed in terms of interoperability for ARROW is that of "Books in Print", organisations managing information about books for commercial purposes. Books in Print are key players in the publishing value chain in many countries as they act as *trait d'union* between the demand (the reader) and the offer (the publisher) by providing commercial information that allows sales channels and distributors to manage an efficient supply chain and have books available on the market. BiPs core business is therefore aggregating information about books from different sources (publishers in the first place), disseminating the aggregated information to third parties (booksellers, e-tailes, web portals, etc.) and maintaining it constantly updated in terms of new books – or new editions or formats - published or books no longer available on the market. Whereas in the Library Domain the driver of management of information about books is bibliographic control, in the BiP Domain the driver is the time to market: as long as information about a book is not available to sales channels, and thus to readers, that book is likely to have a limited commercial lifecycle.

Given these characteristics, from the ARROW perspective BiPs are likely to be the most authoritative source of information on:

- The rights status of books (in print or out of print), and their commercial availability
- New editions and versions of books belonging to the same work
- Information about publishers, as many BiPs also maintain publishers databases to provide the relation between book records and publisher/imprint records.

Keeping in mind the ARROW purpose and the information needed, the results of the questionnaires have been analysed according to core technical areas, relevant for interoperability among different Books in Prints and with the other systems (Library Domain and RRO Domain) in the ARROW environment: (1) rights holders information linked to bibliographic information (Publisher/Imprint); (2) bibliographic information containing information about works and links between different manifestation of the same work, including the digital version; (3) information about rights status of a book: in print or out of print; (4) identifiers used; (5) metadata format for importing, exporting and exposing metadata; (6) protocols for importing, exporting and exposing metadata.

Confidential information collected on each BIP's database has been gathered in a distinct document. In the European countries currently surveyed in this phase of the ARROW project,

within the Books in Print domain the following databases have been identified as relevant for the ARROW project[70]. These are:

France – Electre: Electre database was set up in 1985. Today it contains 1.3m records of (mainly) books, other types of publication and details of publishers. It holds approximately 700,000 records of books in print, and information about forthcoming titles. Information about availability and price is updated daily, taking feeds from the Dilicom platform (teleordering service). Electre manages also, on behalf of the Agence francophone pour la numérotation international du livre (AFNIL), the ISBN Agency for France, French-speaking Belgium and French-speaking African countries that have no national ISBN agency. AFNIL database carries records of publications from these French-speaking countries. AFNIL is responsible for the Electre publisher authority file. AFNIL and Electre databases run together as a single, cross-referable unit.

United Kingdom - Nielsen BookData: following the amalgamation of Whitaker's Books in Print with BookData Limited, Nielsen BookData became the UK's only specific "Books in Print" agency. It collects book information from more than 70 countries (including the UK, Ireland, Europe, Australia, New Zealand and South Africa). Nielsen BookData provides comprehensive bibliographic data worldwide to booksellers, libraries and publishers in 110 countries around the world. More over it manages the ISBN Agency for UK and Ireland and is one of the founding organisations of the ISTC International, acting also as ISTC Registration Agency in the UK.

Germany - Verzeichnis Lieferbarer Bücher: MVB manages the VLB database (Verzeichnis Lieferbarer Bücher) the German Books in Print service for Germany, Austria and German speaking Switzerland. The German BiP is linked to the Adressbuch für den deutschsprachigen Buchhandel (ADB), is the Directory of Publishers and Author-Publishers from Germany, Austria and Switzerland and other countries represented by a distributor located in Germany, including approximately 20,100 publishers. Moreover MVB manages the German ISBN agency, is DOI Registration agency and will likely be the ISTC Registration Agency for Germany.

Austria - VLB: see Germany

Spain - DILVE: DILVE is the "official" Spanish books-in-print repository. It is managed by the Federación de Gremios de Editores de España (FGEE) and was opened in October 2006, therefore its coverage with respect to the Spanish book market is quite small, particularly as far as out of print books (usually older books) are concerned. As of April

[70] For more information on the Slovenian Books in Print, please refer to *D 5.1 Analysis of bibliographic resources and clearing mechanisms existing in Europe*

2009 DILVE contains around 200,000 records from over 500 Spanish publishers. Publishers upload their catalogue information directly into DILVE. Only structural consistency checks are made before metadata is stored. No other checks are made and information submitted by publishers is not changed or modified in any way. In particular, contributors' names are stored exactly as submitted.

Italy - ALICE: Informazioni Editoriali (IE) is the Italian bibliographic agency, a private company which is the owner and publisher of the Italian Books in Print (BIP), named "ALICE", and manager of "Arianna", the Italian book tele-ordering system. IE manages two databases: the Italian books (in print, out-of-print and forthcoming) database created in 1975 and available in digital format since 1989 and containing over 1,000,000 bibliographic records and the Italian publishers' database. Within the framework of its business, IE manages several types of data: bibliographic data, selling data, data contained in EDI messages as orders, order responses, dispatch advices, invoices etc.

The Netherlands - Centraal Boekhuis: Centraal Boekhuis is the leading information and service provider for the Dutch book market, as well as leading distribution centre. Centraal Boekhuis manages CB Online a centralised database that serves as the basis for the Books in Print, the tele-ordering service, the ISBN agency and all other added value services offered to Dutch publishers and booksellers.

Norway - Norwegian Book Database: The DnBB contains metadata about all books published in Norway (201,520 bibliographic records as of April 2009). Its customers are Norwegian publishers, bookstores, Internet bookstores, other retailers, libraries and institutions. It contains information about availability, price, publisher and distribution as well as bibliographic information and publishers' information such as pictures, descriptions, look inside the book and tables of contents.

3.2 Technical area 1: rights holders information linked to bibliographic information (Publisher/Imprint)

According to the ARROW purpose, one relevant piece of information to be provided consists in the indication of the rights holder or rights holder's agent to ask for the requested permission for the use of a certain book. This implies a unequivocal identification of the rights holder, being the author, a contributor or the publisher. Once the rights holder has been identified, it is possible to provide further contact information, where available.

Given their business-orientated nature and the common connection in many countries with the national ISBN Agency, Books in Print are the most likely domain in the ARROW environment to manage a well structured and up to date publisher database, handling also the relation between the publisher (that is the factual rights holder) and its imprints (the name or brand

under which a book is published, usually contained in bibliographic records but without rights-ownership). The possibility to link from the single bibliographic record (information about the book) to the publisher (via the imprint) is a shortcut to build the relation book to rights holder and to retrieve relevant contact information in an effective way. The availability of such interconnected databases to be accessed and queried would be an added value to ARROW.

3.3 Technical area 2: bibliographic information contains information about works and links between different manifestations of the same work, including the digital version

According to the ARROW purpose, one relevant piece of information to be provided consists in the rights status of the work; namely if it is *in print* or *out of print*, as this influences the possibility and terms and conditions for obtaining the requested permissions. The rights status of the work can be determined only on the basis of the status of all the manifestations (books)[71] belonging to the same work, therefore it is important to be able to group all those manifestations, including the digital versions or e-books.

The results of our questionnaires show that Books in Print databases represent quite a comprehensive source of information on book production in the respective countries, and might therefore contain a significant amount of manifestations of the same work, where they exist. Some critical issues emerge in this area:

- The first is the relatively recent foundation of the Books in Print, at least as structured databases. This limits the coverage of older editions.

- The second is the lack of information about works as BiPs deal with “tradable items” that is books, manifestations. However, there are projects going on in this direction. As the possibility to group products by similarity and alternative product form is a very powerful marketing tool, BiPs are willing to offer their customers (booksellers in particular) added value to the information provided.

- The third critical issue is the small number of e-books listed in the BiPs. Whereas some BiPs do not manage e-books at all, because of compatibility issues with incorporating this new kind of information into old database structures, others report that they receive little input from publishers. This is because some publishers market their digital products outside the traditional sale channels covered by the BiP.

To overcome these possible hindrances to the fulfilment of the ARROW purpose, BIPs information on different manifestation may be cross-checked with the information in National Library catalogues, as national libraries are likely to also hold information on back lists and

[71] According to the HLEG definition of Out of Print Works, a work is out of print when all its manifestations are out of print

older books. Equally the adoption of special techniques to group manifestations of the same work is recommended, though its absence is not a barrier, as well as a stronger involvement in the correct management of e-books.

3.4 Technical area 3: contains information about rights status of a book: in print or out of print

According to the ARROW purpose, the clarification of the rights status of the book is a key piece of information, for the reasons explained in the previous technical area. Moreover there must be a clear definition of the meaning of the expression “out of print” in relation to the manifestation level (book) in order to avoid any misinterpretation of such important information.

According to the results of our survey, Books in Print are likely to be the principal source to retrieve information on *in print/out of print* status. As the name itself suggests, Books in Print databases contain, by their nature, information on books that are currently in print, that is, available in publishers’ catalogues and therefore also commercially available on the market. However, following the book lifecycle, today’s in print books are tomorrow’s out of print books, and BiPs find themselves also managing large amounts of book records with no more market appeal.

In the past, out of print book records used to be deleted from the BiPs. In this way, there might be a gap of this kind of information.

More recently, with the introduction of new technologies like Print on Demand and digital publishing, less and less books are declared out of print as publishers are reluctant to lose the opportunity to restart the lifecycle of their books at any time in the future.

This trend, together with the misleading use of the term “not commercially available”, makes the out of print status clarification a very sensitive issue that has to be approached with the consensus of all the actors along the publishing value chain. It is therefore necessary to come to an agreement on what is meant by “not commercially available” in order to guarantee the semantic interoperability of the systems. This is, in this case, even more crucial than ensuring the technical interoperability of the systems involved.

3.5 Technical area 4: identifiers used

According to the ARROW purpose, the ability to unequivocally identify a book, the work to which it belongs and its rights holder is a key requirement to be met before any meaningful exchange of information can be made. Identifiers are the most powerful tool enabling such ability, provided they are shared within a broad community of users and properly assigned.

As per the results of our survey, the major advantage of Books in Print in this area is the solid adoption of the ISBN standard for the identification of books, whenever the book has been assigned with an ISBN. Books not assigned with an ISBN within the Books in Print Domain are therefore a relatively small percentage. However, standard identifiers for works and rights holders have not taken off yet and BiPs have therefore adopted proprietary internal numbers, if any. However there is a wide consensus on the need to adopt emerging standard identifiers for works (ISTC) and parties (ISNI), once they have been fully deployed.

3.6 Technical area 5: metadata format for importing, exporting and exposing metadata

According to the ARROW purpose, the relevant pieces of information from different sources need to be accessed and retrieved, further processed and combined to produce new information. In this task ARROW acts as facilitator by providing common layers and crosswalks to enable the exchange of meaningful information among different systems. The metadata formats used to express the information to be exchanged are a key area in terms of technical interoperability. The higher the number of different metadata formats, the harder enhancing interoperability will be, without loosing the semantic value of the information.

Our survey confirms that the book industry has, on the whole, widely adopted ONIX for Books as the preferred format to exchange information on books throughout the whole supply chain. Equally, Books in Print, regardless of the native format of their database or that of the publisher, chose to use ONIX for Books for exchanging book information with suppliers, booksellers and distributors. Given this, ONIX for Books has been chosen as the preferred format for BiPs interaction also by ARROW. With regard to Publisher and Imprint databases, no standard metadata format exists so far and there is wide consensus on the need to develop a common layer to enhance information exchange.

3.7 Technical area 6: protocols for importing, exporting and exposing metadata

According to the ARROW purpose, the relevant pieces of information from different sources need to be accessed and retrieved, further processed and combined to produce new information. In this task ARROW acts as facilitator by providing common layers and crosswalks to enable the exchange of meaningful information among different systems. The protocols used to access the information and allow communication and information exchange among systems and between on system and ARROW, are one other key area in terms of technical interoperability.

Our survey confirms that given their nature of being information aggregators and providers, BiPs allow third parties to use the information contained in their databases, albeit according to different business models. However for the ARROW purpose, access to the information will

need to be permitted using automatic procedures; that is, B2B services possibly based on standard web protocols, what BiPs already currently implement or plan to implement for their own business purposes.

3.8 Conclusion for BIP Domain

Books in Print constitute the key Domain for ARROW to be able to retrieve information on out of print books, thus enhancing information coming from the Library Domain and preparing a valuable input to be further processed and enriched in the RRO Domain. Although the Books in Print Domain is made up of a quite consistent group of organisations, with regard to the assessment of the core technical areas which emerged in the questionnaires, still there are interoperability gaps to be filled. In addition, further guidance is needed where BiPs are not in place and additional sources of information are to be integrated into the Rights Information Infrastructure. Finally, as BiPs are commercially available according to different licensing and business models, an ARROW common strategy is needed to approach these type of organisations.

From the analysis of BiPs domain it has nonetheless emerged that:

- a proper BiP does not exist in all European countries. In some countries the BIP has been founded quite recently. The coverage of the database, in terms of backlist titles, where many Out of Print titles are to be found, is not consistent;

- ONIX for Books is the main international standard already in use in the book value chain, including in all more advanced BiPs databases. Therefore, using ONIX for Books for data access and exchange within the ARROW framework would greatly enhance both in-domain and cross-domain interoperability, thus facilitating the matching and processing of different data sources about books;

- ONIX for Books describes a book as a “product” that is, as a single manifestation of a work, therefore any information concerning commercial availability is at manifestation level. No information on the availability status of the Work can be inferred by a single ONIX for Books record. That is – according to the definition provided by the HLEG – it is not sufficient to know that a manifestation of a work is out of print to declare the work as out of print. A work is defined as out of print if all its manifestations are out of print. It is therefore anyway necessary, to identify and track all the other manifestations of a work and find out what their status is. This means that this result may be best achieved if all the ONIX for Books records describing a manifestation are correctly managed.

- ONIX for Books messages can be used (and created) by different players along the publishing value chain for different purposes. This means that the primary source of information may not be the publisher itself in some cases, while – *stricto sensu* – the commercial availability

information (out of print according to the HLEG) should be under the control of the publisher (rights holder). However, when correctly used, ONIX for Books provides a very effective way to point out this difference: PR.20 Publishing status and dates, and copyright is to be considered a group of information “certified by the publisher” (or by other player on mandate of the publisher) and are valid for the product per se, while PR 24 Supplier, availability and prices is to be considered a group of information “certified by the supply chain” and are valid for the product in relation to a specific supplier. An agreement is needed on the use of these fields in the ARROW system;

- interoperability between BiPs and the ARROW system could be optimised using adequate standard protocols for B2B interactions such as Web Services, to be used for clearly defined purposes. Within the ARROW framework, the implementation of such a communication infrastructure would empower the access and exchange of bibliographic and commercial information and, where possible, the access and exchange of publisher/imprint information;

- e- books management is not equally efficient in every country; this might cause a loss of information regarding all the manifestations belonging to the same work. BiPs need therefore to explain to their customers (publishers) that it is possible to enter e-books titles within the same dataflow used for printed titles and provide them with guidance on the management of their metadata.

- the ISBN is recommended as primary identifier for all manifestations included in the BiP database, included digital versions of printed books or e-books. This is in line with the international trend in the e-book market, as the ISBN International itself recommends that each single version or format of e-book is assigned with a new ISBN in order to be able to track it along the supply chain;

- the ISTC is recommended to support the identification of records of books belonging to the same work, being this the purpose of the standard. It can be used within the books in print databases and across different data sources in different domains. The process of identifying out of print works is therefore made more effective. In addition, this will also support the identification of apparently orphan works and the creation of orphan works databases;

- the use of ISNI is recommended in order to assist BiPs in disambiguating contributors names and publishers names and, in particular, imprints when they change over time or are acquired by a new publisher. Furthermore, the use of a common standard identifier would enhance interoperability between the BiP domain and the other two relevant domains, where rights holders information is also managed. However adoption of ISNI depends on the timing of the standard deployment.

According to the strength and weaknesses emerged in the analysis of BiPs core technical areas, and according to actual practices in the book supply chain, a core set of guidelines has been defined to provide initial guidance to organisations on how to prepare to be interoperable, within the ARROW environment.

The following table lists the core set of ARROW guidelines for Books in Print and their relation with the respective technical area:

Table 3 - ARROW interoperability core guidelines for Books in Print Domain		
BIP1	has out of print titles metadata	Technical Area 3
BIP2	bibliographic metadata available for query by B2B service	Technical Area 6
BIP3	uses web service for B2B interactions	Technical Area 6
BIP4	implements ONIX for Books	Technical Area 5
BIP5	manages e-books	Technical Area 2
BIP6	has publishers/imprint authority data	Technical Area 1
BIP7	publishers/imprint data are available for queries by B2B services	Technical Area 6
BIP8	implements ISTC	Technical Area 2 Technical Area 3 Technical Area 4
BIP9	Implements ISNI	Technical Area 1 Technical Area 4
BIP10	Implements ISBN	Technical Area 4

4. The RRO Domain (Reprographic Rights Organisations)

4.1 Introduction

The third domain to be assessed in terms of interoperability for ARROW is that of "Reprographic Rights Organisations" (RRO), that is the collective management organisations (CMO) operating throughout Europe in charge of managing and licensing of reprographic rights on behalf of individual rights holders[72]. The key role of RROs - and more generally of CMOs - in the framework of ARROW has been clearly pointed out by the Copyright Subgroup of the High Level Expert Group (HLEG) on Digital Libraries set up by the European Commission[73]. In order to facilitate the licensing of out-of-print works and orphan works, the HLEG encouraged the establishment in each Member State of national Rights Clearance Centres (RCC), which may act as national portals and common access point for clearing rights and offer licensing agreements under the mandate of individual rights holders. It is observed that existing CMOs such as RROs could run such a service; for this purpose, RROs have been invited to observe the HLEG recommendations regarding model licences and criteria for databases and rights clearance centres for out-of-print works and orphan works, as well as the criteria for diligent search for rights holders to orphan works.

By fostering interoperability among different data sources in specific domains (Libraries, BiPs, RROs) ARROW aims at supporting libraries (and in the future, other users) to identify the rights status of a work (copyrighted, out of print, apparently orphan) and to find the appropriate rights holders (directly or through their representatives, such as RROs) and/or their representatives, so that permission to digitise a book can be requested. In this framework, RROs are expected to play a crucial role for the achievement of ARROW objectives, because:

- As rights holders' agents, RROs represent a valuable source of information as far as authors and publishers names and contact data are concerned

- RROs repertoires are a valuable source of information to be used to track the relationship between individual books and rights holders (in cases where the RROs has a title- specific mandate)

[72] With technological development, some rights holders have extended the rights and uses RROs can licence on their behalf to cover other mass uses resulting in either a physical or a digital copy (see IFFRO website: http://www.ifrro.org/)

[73] Final Report on Digital Preservation, Orphan Works, and Out of Print Works (http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/copyright/copyright_subgroup_final_report_26508-clean171.pdf); a more comprehensive analysis of the role of ARROW project in the EU legal framework is provided in D3.*1 Report on legal framework* (Edition 1)

- RROs repertoires are a valuable source of information to be used to track the relationship between rights holders and rights available for licensing (both in case of title-specific and not-title specific mandates)

- As rights holders' agents, RROs may effectively support rights holders in claiming their rights using the ARROW Registry of Orphan Works (ROW) and actively contribute to the enrichment of the ROW over time

Our survey highlighted that the current technological advancement of RROs may vary from country to country in terms of: availability of structured databases of bibliographic and rights holders metadata, use of standard metadata schema and identifiers, implementation of protocols enabling data exchange. However, Arrow system is expected to enable RROs to play a strategic role in different ways according to their own technological capabilities:

- If the RRO has structured databases for managing rights and bibliographic information, the implementation of the ARROW guidelines for technical interoperability will increasingly enhance the integration of the RRO workflow within the ARROW infrastructure, enabling the set up of automatic or semi-automatic procedures for data processing and exchange

- If the RRO does not have structured databases for managing rights and bibliographic information, the RRO interaction with ARROW will be based on alternative tools and procedures (e.g. tools enabling RROs to receive libraries requests for permission and to reply after having manually checked the request; tools enabling RROs to search the ARROW distributed network of databases of information on rights status of literary works)

In both cases, RROs are expected to effectively support the lawful exploitation of out of print works and orphan works within the ARROW framework. On users' request for permission, RROs may: (1) subject to mandates from the rights holders, provide the licence agreement, (2) if not, provide the best information available on rights holders, e.g. contact of the publisher of the last edition of the book (3) trigger the diligent search for rights holders whenever a work appears to be orphan.

Keeping in mind the ARROW purpose and the information needed, the results of the questionnaires have been analysed according to core technical areas, relevant for the interoperability among different RROs and with the other systems (Library Domain and BiP Domain) in the ARROW environment: (1) rights holders information linked to bibliographic information (Authors/Contributors, Publisher/Imprint); (2) types of usages managed on behalf of rights holders; (3) identifiers used; (4) metadata format for importing, exporting and

exposing metadata; (5) protocols for importing, exporting and exposing metadata; (6) query methods and criteria.

Confidential information collected on each RRO's database have been gathered in a distinct document.

In the European countries currently surveyed in this phase of the ARROW project, within the RRO domain the following organisations have been identified as relevant for the ARROW project[74]. These are:

France – CFC: CFC (Centre Français d'exploitation du droit de Copie) is the French Copy Centre (collecting and redistributing) and is a member of the IFFRO. CFC collects and redistributes license fees linked to the photocopying or digital distribution of books, newspapers and magazines in France. It splits its fees into two areas: on the one hand, paper photocopies of both press and print; on the other hand, digital copies of both press and print. Alongside authors' societies, the CFC has set up a group in France dedicated to researching orphan works. CFC is a member of the CSPLA (Conseil Supérieur de la Propriété Littéraire et Artistique) sub-commission, dedicated to orphan works[75].

UK – ALCS, PLS, CLA: the structure of RROs in the UK is at quite complex, but well-organised. There are three main organisations involved: the Authors' Licensing and Copyright Society [ALCS], the Publishers' Licensing Society [PLS] and the Copyright Licensing Agency [CLA]. The CLA is responsible for managing secondary copyright in situations in which the other two organisations cannot easily exercise primary copyright, for example, when dealing with photocopying or digitisation projects taking place within Higher Education Institutions (HEIs). It is co-owned by the other two organisations. All three rights holders organisations also work closely with the Design and Artists Copyright Society [DACS]. The IT systems through which these organisations operate both hold information in common and some information which is unique.

Germany - VG Wort -Verwertungsgeselleschaft Wort – is the main German rights association as far as scientific and literary works are concerned (books and journals) and is a IFRRO member. In 2008 VG Wort was acting on behalf of 390,851 individual rights holders (381,821 authors, 9,030 publishers) excluding representation via bilateral agreements with sister societies). VG Wort manages reproduction rights both in the form of printed materials and in the form of digitisation, including digitisation of older materials (out of

[74] For more information on the Slovenian RRO, SAZOR, please refer to *D 5.1 Analysis of bibliographic resources and clearing mechanisms existing in Europe*

[75] CSPLA sub-commission dedicated to orphan work: http://www.cspla.culture.gouv.fr/CONTENU/avisoo08.pdf; http://www.cspla.culture.gouv.fr/CONTENU/rapoeuvor08.pdf

print works likely belong to this category) and orphan works. VG Wort is very active in the field of digital rights management and collaborates with the German National Library and the German Publishers Association (via the MVB) to set up a integrated infrastructure to allow the timely and effective management of users requests for permissions and rights holders licensing for out of print and orphan works for German works.

Austria - Literar Mechana: Literar Mechana is the main Austrian rights association as far as scientific and literary works, as well as printed music are concerned. It is a member of IFRRO and maintains close relations with the European "sister societies" and especially to those in the German speaking countries, particularly to VG Wort. Literar Mechana was founded in 1959 and today represents 14,000 rights holders (authors and publishers); its databases are relevant information source for the identification of Austrian right holders and rights. Continuous quality checking is undertaken, in particular for the rights holder's data.

Spain – CEDRO: the Centro Español de Derechos Reprográficos is the main Spanish rights organisation for authors and publishers of printed works. Besides, foreign rights holders are managed in the framework of bilateral agreements of reciprocal representation with other Reprographic Rights Organisations. CEDRO database containing information about titles (as manifestations), rights holders (authors and publishers) and rights has been identified as a relevant data source for Arrow. Bibliographic information refers to monographs, monographic chapters and serials. Furthermore, a new service supporting the search and location of rights holders has been recently launched[76]. The service can be used for different purposes: (1) publishers may use the service to search for foreign authors in order to request translation rights (2) libraries may use the service to search for rights holders in order to request a digitisation licence.

Italy – AIDRO: the Associazione Italiana per i Diritti di Riproduzione delle Opere dell'ingegno – is one of the two main RROs in Italy managing rights as far as scientific and literary works are concerned. AIDRO manages reproduction rights outside the limits of the personal use (15% of the entire work). AIDRO currently does not hold structured databases neither for rights holders information nor for its repertoire. However, such developments have been programmed for the near future, including exploiting synergies with the Ediser Reprographic Service, described below. According to the Italian Copyright Law No. 248, 18 August 2000, a book can be photocopied for personal use only to the maximum of the 15% of the entire book and under the payment of a fee. The revenues coming from that activity are distributed by SIAE (Italian Society of Authors and Publishers) to the rights holders (contributors and publishers). Ediser, a service society of AIE (i.e. the Italian Publishers

[76] For further information, please refer to: http://www.cedro.org

Association), has been designated to provide a back office service to enable SIAE to correctly distribute the revenues of the copying activity. The service consists of the following activities: (1) to harvest, on a statistical basis, information about the photocopied books; (2) to identify all the rights holders; (3) to estimate the revenue due to each rights holder and communicate to SIAE the data needed for the distribution of revenues.

Netherlands – PRO, LIRA: VOI©E is the newly founded umbrella organization (December 2008) by Dutch RROs to manage, in a cooperative way, issues arising from the introduction of new technologies. The first achievement of this new organisation is the agreement with FOBID, representing the libraries to facilitate digitisation programmes in libraries and the use of orphan work, within the Digiti©E Committee (Digitisation of Cultural Heritage).Under the agreement, Dutch libraries, archives or museums can register their digitisation projects in a Registration Centre that has been set up by the rights holders associations. Among the RROs joining VOI©E Foundation, Stichting Lira, Stichting Reprorecht, Stichting PRO – which have been identified as relevant organisations for ARROW purposes –all use Cedar's technological infrastructure. Cedar is the Centre for Service to Authors' and related Rights (Centrum voor Dienstverlening Auteurs - en aanverwante Rechten). On Behalf of the RROs, Cedar is responsible for managing information needed for the collection of royalties for the use of material protected by copyright and the subsequent distribution of the compensation to the rightful beneficiaries, entitled to these royalties. Cedar maintains their administration and provides advice and service. The questionnaire findings gathered from PRO and LIRA have therefore enabled the identification of the relevant features of Dutch RROs technological infrastructure according to ARROW purposes.

Norway – Kopinor: Kopinor was established on 30th April 1980. Membership is in principle open to any association representing copyright holders of published works. Today Kopinor – which is also an IFFRO member - has 23 member organisations - 6 publishers' associations and 17 authors' associations. Kopinor manages reproduction rights both for printed materials and for digital materials and for digitisation purposes.

4.2 Technical area 1: rights holders information linked to bibliographic information (Authors/Contributors, Publisher/Imprint)

According to the ARROW purpose, one relevant piece of information to be provided is for the ARROW system to be able to indicate which rights holder or rights holder's agent to request permission from for the use of a certain book. This implies an unequivocal identification of the rights holder, being the author, a contributor or the publisher. Once the rights holder has been identified, it is possible to provide further contact information, where available.

The main task of an RRO is to licence reproduction rights on behalf of rights holders (according to the mandate, licenses can include both physical and digital uses), collecting remuneration

from licenses and distribute it to rights holders. Therefore, within their business activity, RROs collect and continuously update information about authors, publishers and other contributors who are eligible for remuneration (e.g. translators). In order to fully exploit such information within ARROW, to support the diligent search for rights holders, authors and publishers data should be managed within structured databases. Furthermore, if the RRO has a title-specific mandate, the existence of a structured bibliographic database linked to the rights holders database would enable the immediate retrieval of relevant rights holders (names and contact details) to ask for permission starting from the data of the book the user wishes to digitise.

4.3 Technical area 2: Type of usage managed on behalf of rights holders

In order to facilitate the licensing and use of out-of-print and orphan works in a digital environment through ARROW, RROs should have the legal mandate given by rights holders or their representatives to licence the digitisation rights and usages for out of print and orphan works[77]. Furthermore, the use of a standard communication format would enable RROs to effectively communicate this information to the other systems in the ARROW environment and in particular to the final user who asked for the permission for a certain use.

From a technical point of view, the capacity of an RRO to set up automatic procedures to check the rights holders mandate and possibly provide the appropriate license on user request relies on the existence of structured databases storing information about the types of rights and usages that the RRO can license, each of them linked to the RRO repertoire (being it title specific or not - title specific). This would enable, on user request, an immediate and effective check of rights and permissions available for the requested title. The need for a standard metadata format for managing and share repertoire information has been widely recognised within the RROs community: for this purpose, RROs have promoted through IFFRO the development of ONIX for Repertoire[78]. Being a valuable tool for RROs interoperability, giving them the capacity of sharing with each other the mandates they hold from rights holders and to build specific applications for managing effectively rights and permission available[79] ONIX for Repertoire has been identified as a relevant standard within Arrow framework.
If the RRO does not have the mandate for the requested usages, it is however recommended that the RRO has the capacity to redirect users to relevant rights holders to ask for permission

[77] For a broader analysis of clearing mechanisms and licensing models for out of print and orphan works in Europe, please refer to D.3.1 Report on legal framework

[78] ONIX for Repertoire is one of the two standard message formats commissioned from EDItEUR by IFRRO, the other being ONIX for Distribution. The full package of RRO message formats is referred as "ONIX for RRO". Standards applicable within ARROW information infrastructure are presented in detail in D4.1 *State of the art and guidelines for standards applicable* (Edition 1). See also EDItEUR website www.editeur.org for further details on ONIX formats

[79] Specific types of usages to be licensed by RRO within ARROW framework will be defined according to the HLEG model agreements for out of print works (see the *Final Report on Digital Preservation, Orphan Works, and Out of Print Works*) and will be implemented into the message format used by libraries to query ARROW system.

according to the information available in the RRO database (e.g. if the work is out of print and the RRO has not the appropriate mandate, to provide users with authors names and/or contact details of the publisher of the last edition found in their repertoire). Again, the effectiveness of this process is higher when bibliographic and rights information is managed by the RRO using structured databases and appropriately linked each other.

Finally, it is worth noting that, in order to optimise the licensing process, the adoption of ONIX for PL (Publication Licence) as the standard message format for rights information exchange among publishers, RROs, and end users would further enhance the interoperability between the players involved in the whole process.

4.4 Technical area 3: identifiers used

According to the ARROW purpose, the ability to unequivocally identify a book, the work to which it belongs and its rights holders, is a key requirement to be met before any meaningful exchange of information can take place. Identifiers are the most powerful tool enabling such ability, provided they are shared within a broad community of users and are properly assigned.

With regards to the state of art in the RROs domain, it is possible to identify a common approach and solution that should be adopted without splitting the analysis into the country level: the ISBN standard is widely used for the identification of each manifestation by all the RROs involved in the project, while to identify rights holders each organisation has implemented its own proprietary identifier. However, ISNI developments are awaited, as well as an IFRRO joint project for a shared rights holders identifier. There is also a wide consensus on the need to adopt standard identification systems for work (ISTC) as soon as it will be fully deployed.

According to the ARROW tasks, three standard identification systems are recommended in order to facilitate interoperability within each specific domain as well as across all the domains involved in the rights information infrastructure: (1) the ISBN, to uniquely identify the book the user wishes to digitise/use in a digital environment; (2) the ISTC, to uniquely identify the underlying work and aggregate all the manifestations belonging to the same work; (3) the ISNI, to uniquely identify the rights holders of the work to be asked for permission[80].

With the exception of the ISBN, which has been widely adopted throughout the book value chain since the 1970's, the other relevant standard identifiers here considered are still little implemented (ISTC) or under development (ISNI). However, the adoption of these standards is supported from a disparate community of interests (including libraries, publishers, bibliographic agencies, RROs) which are also directly involved in ARROW project. Therefore ARROW piloting is expected to enforce the awareness of the importance of standards in rights

[80] Standard identifiers applicable to ARROW infrastructure are extensively treated in D4.1

management as well as to provide an infrastructure to trigger standards adoption by stakeholders involved.

4.5 Technical area 4: metadata format for importing, exporting and exposing metadata

According to the ARROW purpose, the relevant pieces of information from different sources need to be accessed and retrieved, further processed and combined to produce new information. In this task, ARROW acts as facilitator by providing common layers and crosswalks to enable the exchange of meaningful information among different systems. The metadata formats used in the RRO domain to express the information to be exchanged are a key area in terms of technical interoperability. The higher the number of different metadata formats, the harder enhancing interoperability will be, without losing the semantic value of the information.

To enable the interaction of ARROW system with the RROs domain in absence of a widely implemented standard, a common metadata format, based on existing standards, to be used by ARROW to query RROs database and retrieve information about their repertoires is to be identified. The suitable metadata format to facilitate the integration of RROs databases into the ARROW information infrastructure is ONIX for Repertoire, being it actually developed within the RRO community and currently piloted in UK by PLS and CLA. As per our survey, regarding rights holders data, at the current time no standard metadata format has been implemented in the RROs domain; nevertheless, interoperability within ARROW infrastructure should be facilitated by adopting a standard identifier system such as ISNI for rights holders' names together with appropriate redirection mechanisms to rights holders metadata hold in RROs databases.

4.6 Technical area 5: protocols for importing, exporting and exposing metadata

According to ARROW purpose, the relevant pieces of information from different sources need to be accessed and retrieved, further processed and combined to produce new information. In this task ARROW acts as facilitator by providing common layers and crosswalks to enable the exchange of meaningful information among different systems. The protocols used to access the information and allow communication and information exchange among systems and between other systems and ARROW, are one other key area in terms of technical interoperability.

So far RROs have been focused on the management of information for their internal processes and business requirements, without a strong need to "open" their system to third parties, for querying or retrieving data. This also because data managed by RROs is quite sensitive and privacy-protected. In many cases therefore data is "locked" in the databases and exchanged only for internal reasons. The development and adoption of standard formats – such as the above mentioned ONIX for RROs formats – to communicate information about repertoires and

right managed, provides strong evidence for the need of data exchange among RROs, but also between a single RRO and the rights holders. This will lead to the consequent development of tools and protocols to allow data exchange in an automatic way: ONIX for RRO champions RROs - CEDRO in Spain and the UK RROs - are in fact able to support Web Services for the exchange of metadata.

Within the RROs domain, the development of an adequate technological infrastructure enabling the effective communication between collective management organisations, thus facilitating the exchange of distribution payment information and the exchange of rights and repertoire information is recognised as a priority issue. For this purpose, IFFRO commissioned ONIX for RRO messages (ONIX for Repertoire and ONIX for Distribution) from EDItEUR[81] and is now actively supporting its members to implement these standard messages. However, in order to achieve full interoperability within the RROs domain as well as between RROs and the ARROW system, the adoption of standard communication protocols supporting metadata exchange between stakeholders' organisations is recommended as well, as it enables RROs to set up automatic procedures for data exchange and processing.

4.7 Conclusion for RRO Domain

The role of RROs within the ARROW infrastructure is essential in order to provide effective support for the due diligence search for rights holders and to facilitate the lawful exploitation of out-of print and orphan works by cultural institutions. At the same time, the RROs databases represent an highly valuable source of information to be involved in ARROW infrastructure: RROs repertoires can effectively support ARROW in identifying the existing editions of a books thus contributing to the process of matching and aggregating multiple data sources to infer the rights status of the underlying work; at the same time, rights holders databases and information about rights available for licensing can effectively support ARROW in identifying relevant rights holders to be contacted according to the requested usage of a literary work.

From the analysis of the RRO domain it has nonetheless emerged that:

- the technological development of each RROs can vary from country to country as far as the availability of structured databases of bibliographic, rights and rights holders information is concerned;

- not all RROs are entitled to manage digital rights, in particular for Out of Print and Orphan Works;

- there are several reasons to support the implementation of standard formats within the RROs domain: on the one hand, ONIX for Repertoire is the standard metadata format commissioned

[81] See also D4.1

from EDitEUR to IFFRO in order to facilitate RROs in exchanging with each other information about their repertoire; on the other, both ONIX for Repertoire and ONIX for Books - the main international standard already in use in the book value chain, including BiPs databases - rely on ONIX syntax. Therefore, using ONIX for Repertoire for data access and exchange within ARROW framework would greatly enhance both in-domain and cross-domain interoperability, thus facilitating the matching and processing of different data sources about books;

- interoperability between RROs and the ARROW system could be optimised using adequate standard protocols for B2B interactions such as web services, to be used for clearly defined purposes. Within the ARROW framework, the implementation of such a communication infrastructure would empower the access and exchange of repertoire information and, where possible, the access and exchange of rights holders information (with regard to privacy, at least publishers data could be made available to the ARROW infrastructure). Likewise, this would benefit RROs interoperability among themselves and with other domains;

- as the main standard identifier in the book value chain, the ISBN represents a valuable tool for cross-domain interoperability. In particular, in Arrow context, the ISBN is recommended in order to enable more effective searches on RROs repertoire and provide a unique key to create the relationships between books and rights holders databases. It would also allow RROs to effectively exchange with each other information about their repertoire and facilitate the distribution of revenues within the framework of bilateral agreements. At last, the use of ISBN enables the matching of metadata records in RROs repertoire with records referring to the same books in other domains (i.e. BiPs and Libraries databases) within the ARROW infrastructure;

- being specifically designed to univocally identify works, this facilitating rights management in the content value chain, the ISTC is recommended in Arrow framework in order to: enable RROs to effectively manage out of print works within their repertoire (as the out of print status refers to works and not to individual manifestations) and to facilitate data exchange about out of print works across different databases; support the identification of records of books belonging to the same work across different data sources and in different domains, thus making the process of identification of out of print works more effective; support the identification of apparently orphan works and the creation of orphan works databases;

- the use of the ISNI is recommended in order to: build a bridge between RROs internal identifiers and a cross-domain standard identification system for rights holders, facilitate RROs to disambiguate contributors names and making more effective RROs core business activities (i.e. distribution of revenues from licences to domestic and foreign rights holders); support the sharing of authority records for authors between RROs and Libraries thus enhancing the quality

of data needed for rights holders identification (according to this purpose, possible synergies between ARROW and VIAF[82] project shall be exploited).

According to the strength and weaknesses emerged in the analysis of RROs core technical areas, and according to actual practices in book rights management, a core set of guidelines has been defined to provide initial guidance to organisations on how to prepare to be interoperable within the ARROW environment.

[82] VIAF (The Virtual International Authority File, http://www.viaf.org/) is a joint project of several national libraries, implemented and hosted by OCLC aimed at matching and linking authority records for personal names using National Libraries' authority files (Deutsche Nationalbibliothek, the Library of Congress and the Bibliothèque nationale de France (BnF). The implementation of ISNI in VIAF framework would facilitate RROs to access and share authority records for authors names, thus enforcing interoperability with the Libraries domain. See also D4.1 for details on VIAF project.

The following table lists the core set of ARROW guidelines for RROs and their relation with the respective technical area:

Table 4 - ARROW interoperability guidelines for RRO Domain		
RRO1	has rights holders database	Technical Area 1
RRO2	has rights holders database connected to title database	Technical Area 1
RRO3	bibliographic metadata available for query by B2B service	Technical Area 5
RRO4	uses web service for B2B interactions	Technical Area 5
RRO5	has the legal mandate to manage digitization rights for out of print works	Technical Area 2
RRO6	has the legal mandate to manage digitization rights for orphan works	Technical Area 2
RRO7	implements ONIX for RRO	Technical Area 4
RRO8	implements ONIX for PL (Publishing Licence)	Technical Area 4
RRO9	Rights holders data are available for queries by B2B services	Technical Area 1 Technical Area 5
RRO10	Implement ISTC	Technical Area 3
RRO11	Implement ISBN	Technical Area 3
RRO12	Implement ISNI	Technical Area 3

IV CONCLUSIONS OF THE DELIVERABLE

ARROW Rights Information Infrastructure is meant in the first place to serve the diligent search process that libraries are asked to carry out before starting the digitisation of a book or of a collection of books. Such a diligent search is aimed at the identification of the relevant rights holders whom the library has to contact to ask for the permission to digitise and use a book still under copyright. The process to carry out a diligent search is very time and resource consuming, as the library is asked to search for that information in each of the information resources available for every single book the library wishes to digitise[83].

ARROW Rights Information Infrastructure is meant as well to support the growth of Europeana by facilitating the inclusion in the European digital library initiatives a number of books so far excluded from digitisation programmes because of the uncertainty of their rights status and the subsequent difficulty to obtain the permissions needed: namely Out of Print and Orphan works.

ARROW Information Infrastructure is finally meant to manage in a standard and user-friendly way the complexity of the above mentioned processes, which involve a number of different information resources and different players along the book value chain, and needs to be approached from a necessary European perspective.

Interoperability is a key word when dealing with multiple systems that need to communicate and exchange information: to fulfil ARROW purposes, the Rights Information Infrastructure needs to collect and process multiple bits of information, retrieved from multiple sources of information (databases), belonging to multiple domains, in multiple countries, acting as an "interoperability facilitator".

These guidelines provide the interoperability framework needed to unite the individual information sources and systems identified as being relevant for the ARROW purposes into a coherent system; by recommending the adoption of the main standards in the content value chain, guidelines will achieve two objectives: on one hand, to provide guidance to foster both domain and cross domain interoperability, on the other, to facilitate the creation in each domain of a common metadata layer available for internal mappings and data processing within Arrow information infrastructure.

For each domain involved in the ARROW environment, relevant technical areas have been identified and existing practices, standards in use as well as emerging standards relevant for the Arrow purposes have been analysed. This approach led to the identification of guidelines for technical interoperability within Arrow, as summarised in the following Table 5.

[83] Sector-Specific Guidelines On Due Diligence Criteria For Orphan Works (http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/orphan/guidelines.pdf)

Table 5 - ARROW interoperability core guidelines		
Library Domain	**BiP Domain**	**RRO Domain**
is a partner in The European Library	has out of print titles metadata	has rights holders database
uses a MARC-based metadata format, MARC21 is preferred	bibliographic metadata available for query by B2B service	has rights holders database connected to title database
uses OAI-PMH	uses web service for B2B interactions	bibliographic metadata available for query by B2B service
uses the uniform title	implements ONIX for Books	uses web service for B2B interactions
implements FRBR	manages e-books	has the legal mandate to manage digitisation rights for out of print works
has author authority file	has publishers / imprint authority data	has the legal mandate to manage digitization rights for orphan works
authority files are available through VIAF	publishers / imprint data are available for queries by B2B services	implements ONIX for RRO
implements ISTC	implements ISTC	implements ONIX for PL (Publishing Licence)
implements ISBN	implements ISNI	Rights holders data are available for queries by B2B services
implements ISNI	implements ISBN	iImplements ISTC
Uses UTF-8 character-encoding		implements ISBN
		implements ISNI

The assessment of the interoperability of each selected resource within each domain with the ARROW technical interoperability guidelines will provide guidance for the selection of a sub-set of countries where ARROW system will be piloted. Moreover, in order to make the piloting phase of ARROW more effective, the countries where book digitisation projects implying the diligent search for rights holders are in place or planned in the short term will be short-listed as well. Through the actual testing of the ARROW system, the adoption of technical interoperability guidelines is expected to be increasingly enhanced. Guidelines are therefore also intended as a check list or a roadmap for the project phases following the piloting, when the other countries involved in the ARROW will be brought into the Rights Information Infrastructure.