

## Grant Agreement ECP-2007-DILI-527003

ARROW

# SPECIFICATION OF RIGHTS EXPRESSION METADATA

<b>Deliverable number/name</b>	<i>D4.3.1</i>
<b>Dissemination level</b>	<i>Public</i>
<b>Delivery date</b>	<i>26 April 2010</i>
<b>Status</b>	<i>Final</i>
<b>Author(s)</b>	<i>Françoise Bourdon (BnF), Tim Devenport (EDItEUR), Paola Mazzucchi (AIE)</i>



**eContentplus**

This project is funded under the eContentplus programme<sup>1</sup>,  
a multiannual Community programme to make digital content in Europe more accessible, usable and  
exploitable.

---

<sup>1</sup> OJ L 79, 24.3.2005, p. 1.

## Table of contents

<b>EXECUTIVE SUMMARY .....</b>	<b>3</b>
<b>1. RIGHTS EXPRESSION METADATA IN ARROW: CONTEXT OF APPLICATION .....</b>	<b>5</b>
<b>2. SOURCE OF WORK AND METHODOLOGY .....</b>	<b>5</b>
2.1 OUTCOMES FROM D 4.1, D4.2 AND D5.1.....	5
2.2 ARROW WORKFLOW .....	6
2.3 TECHNICAL WORKING GROUP FOR MESSAGES DEFINITION .....	7
<b>3. ARROW MESSAGES DEFINITION FOR ALPHA RELEASE.....</b>	<b>8</b>
3.1 SCHEMA 1: LIBRARY ⇔ ARROW: M1Q/M1R.....	9
<i>M1Q: Initial resource and usage request.....</i>	9
<i>M1R : Initial resource and usage response .....</i>	9
3.2 SCHEMA 2: ARROW ⇔ TEL: M2Q/M2R .....	10
<i>M2Q : Resource identification request.....</i>	10
<i>M2R : Resource identification response.....</i>	10
3.3. SCHEMA 3: ARROW ⇔ LIBRARY: M3Q/M3R.....	10
<i>M3Q: Manifestation matching review request.....</i>	10
<i>M3R: Manifestation matching review response .....</i>	10
3.4. SCHEMA 4: ARROW ⇔ TEL: M4Q/M4R .....	11
<i>M4Q: Cluster creation request .....</i>	11
<i>M4R: Cluster creation response .....</i>	11
3.5 SCHEMA 5: ARROW ⇔ LIBRARY: M5Q/M5R.....	11
<i>M5Q: Cluster review request.....</i>	11
<i>M5R: Cluster review response.....</i>	11
3.6 SCHEMA 6: ARROW ⇔ BOOKS IN PRINT: M6Q/M6R .....	12
<i>M6Q: Related BIP request .....</i>	12
<i>M6R: Related BIP response .....</i>	12
3.7 SCHEMA 7: ARROW ⇔ RRO: M7Q/M7R .....	12
<i>M7Q: Formal licence request .....</i>	12
<i>M7R: Licence proposal or refusal.....</i>	13
3.8 STRUCTURE OF THE DELIVERABLE .....	13
<b>4. MESSAGE DEFINITION: TOWARD THE BETA RELEASE .....</b>	<b>13</b>
<b>5. CONCLUSION .....</b>	<b>14</b>

## EXECUTIVE SUMMARY

One role of ARROW is to act as an “interoperability facilitator”. If it is to provide a pan-European service that will facilitate identification of book rights status and rights holders, support any “diligent search” model and create a European registry of orphan works, the system will need to enable processes which can access and query a range of different systems and retrieve relevant data: bibliographic description from National Library databases, publishing status and commercial availability from Books in Print databases, rightsholder identification from Reprographic Rights Organisation databases.

ARROW is being developed in a challenging context which is well illustrated by this cross-domain complexity. The three domains - the Library, Books in Print (BIPs) and Reprographic Rights Organisation (RROs) Domains - have distinct professional practices, goals and history. In response to this legacy, their databases are extremely heterogeneous in terms of metadata content, formats, query and access protocols. There is also the dimension of international complexity: partner countries have different structural, legal and technical environments. Finally, synergies have to be found with other relevant initiatives such as The European Library, the TELplus project, the Virtual International Authority File (VIAF), the International Standard for Textual Code (ISTC), the International Standard for Name Identifier (ISNI), etc.

The aim of WP 4 *Interoperability* is to enhance the deployment of standards along the digital library value chain for identifying and describing content and for web resolution, to promote interoperability in the digital library value net, from rights clearance to content search, to specify and implement standard message formats for rights expression within digital library initiatives, and for metadata messaging and exchange between rightsholders and rightsholder representatives, intermediaries, libraries and e-retailers.

Since ARROW is being designed primarily as a machine to machine system, standard messages are needed to support communication between stakeholders. Within the framework of WP4, the aim of the deliverable D4.3 *Specification of rights expression metadata* is to provide guidance in defining which metadata fields, out of the existing metadata schemas used in the different domains under consideration, are relevant for rights expressions, to foster interoperability and data exchange between the domains by defining metadata and messaging format specifications based on the existing implemented standards and relevant initiatives: among these, we need to consider the MARC family of standards for the library domain, the ONIX for Books family of standards for the BIP and RRO domains, ONIX for ISTC for “work” description, the ONIX for RRO framework for the RRO domain, and the ONIX for Licensing Terms framework for the formalisation of

permission and rights expression. The standard licence models for usages of Out of Print Works and Orphan Works defined by the HLEG have been taken into account as well.

This was done under the supervision of EDItEUR, the standards setting organisation responsible for the ONIX family of standards ([www.editeur.org](http://www.editeur.org)), so as to ensure high level integration between the project results and general international development.

The present first version of the deliverable (D4.3.1) includes a presentation of the methodology used to define the messages, results of face to face meetings of the Technical Working Group appointed to work on these messages, the functional requirements and context of application for the ARROW messages, the ARROW Messages Version 0.1 Release for the Alpha version of the ARROW pilot, plus the related documentation, and a first assessment on messages enhancements towards the Beta release of the ARROW pilot.

## 1. RIGHTS EXPRESSION METADATA IN ARROW: CONTEXT OF APPLICATION

Reference to “rights” in the ARROW Rights Information Infrastructure context means taking into account two different perspectives: first, defining the rights status and identifying the rightsholders of a work starting from its manifestations (those specific manifestations the library wants to digitise and use); and secondly defining the terms and conditions under which that work may be used by classes of users for specific purposes.

The first one relates to metadata, the second one refers to licences and licensing terms.

To achieve its goal, the ARROW system has to access and query different systems and retrieve relevant data: bibliographic description from National Library databases, publishing status and commercial availability from Books in Print (BIP) databases, rightsholder identification from Reprographic Rights Organisation (RRO) databases. The stakeholders involved, libraries, BIPs and RROs, belong to different cultural and economic domains, each of them having distinct professional practices, goals and histories. Before defining standard messages to support machine to machine communication between the databases in these communities, preliminary studies were necessary to cope with the heterogeneity of these databases and to design the workflow within which the messages have to function.

## 2. SOURCE OF WORK AND METHODOLOGY

### 2.1 Outcomes from D 4.1, D4.2 and D5.1

The outcomes of D 4.1, D4.2 and D5.1 (previously issued) have proved very valuable.

*State of the art and guidelines for standards applicable* (D 4.1) is an analysis of a wide range of standards selected on the basis of their potential application, to ARROW;

*Guidelines for technical interoperability* (D 4.2) and *Analysis of bibliographic resources and clearing mechanisms existing in Europe* (D 5.1) analyse the existing databases of the three domains (Libraries, BIPs and RROs).

These three deliverables provided key information about the metadata content of each domain’s databases, the structure and organisation of this metadata in the different databases, the existing exchanges between these databases, the standards and protocols used for these exchanges and an overview of the local interconnections between the 3 domains in each partner country.

The deliverables D4.1, D4.2, D5.1 provided the real background to the message definition. In the light of their content, practical solutions to ensure interoperability and scalability were made: the choice of international standards including standard identifiers, XML structure, MARC21 format in the Library Domain and ONIX based formats in the BIP and RRO domains, the choice of The

European Library (TEL) as unique access point for the National Libraries, and the selection of countries which have the most suitable infrastructure to engage with ARROW prototyping: Germany, UK, Spain, France. It was identified, for example, that different BIPs already support different web-service-based request/response protocols for data exchange, and that ARROW may need to adopt different protocols for different BIPs, in which case the development of the standard message becomes part of the "reference architecture" rather than being an implemented message. Moreover, to comply with specific national context, 2 variants of the workflow have been designed: following Workflow A, ARROW queries the BIP and the RRO separately, while following Workflow B, ARROW only queries the RRO (which also manages the BIP data). These early decisions have been reflected in the message choreography and structure.

These 3 deliverables were essential for defining the workflow and designing the architecture as described in D 5.2 *Specification of right information infrastructure* issued in December 2009.

## **2.2 ARROW workflow**

The ARROW workflow has been defined through an iterative process and will still be adjusted and optimized during the development phase and in accordance with experience in the pilot phase. It reflects the consensus found in each country by national stakeholders.

Here is a short description of the main steps in workflow A as implemented in the Alpha release<sup>2</sup> and to which the messages described below conform:

Step 1: Library submits a query to ARROW which acknowledges receipt;

Step 2: ARROW forwards the query to The European Library for the matching process;

Step 3: The European Library processes the request, identifying the book(s) by matching the metadata query with the records present in its Central Index and returns the results to ARROW;

Step 4: ARROW forwards the results to the Requesting library for validation;

Step 5: ARROW asks The European Library to perform a clustering process for each exact match identified;

Step 6: The European Library performs the clustering process and returns the cluster(s) to ARROW;

Step 7: ARROW forwards the results to the Requesting library for validation;

Step 8: ARROW forwards the validated clusters to the BIP to match each manifestation, define the publishing status of each work, and provide publishing information;

Step 9: BIP returns the results to ARROW;

---

<sup>2</sup> Germany, the first country to pilot ARROW in its Alpha release, has adopted workflow A

Step 10: ARROW forwards the information gathered to the RRO to match with its own records, advise upon available licences and provide information about the rights holders;

Step 11: RRO returns the results to ARROW, proposing or denying a license (or stating that no license is required) and offering further information about rightsholders;

Step 12: ARROW sends feedback to the Requesting library about the rights status of each work (in print/out of print/public domain or orphan), the offer or denial of a license, details of an alternative RRO to contact for a licence where applicable and rightsholder information if available.

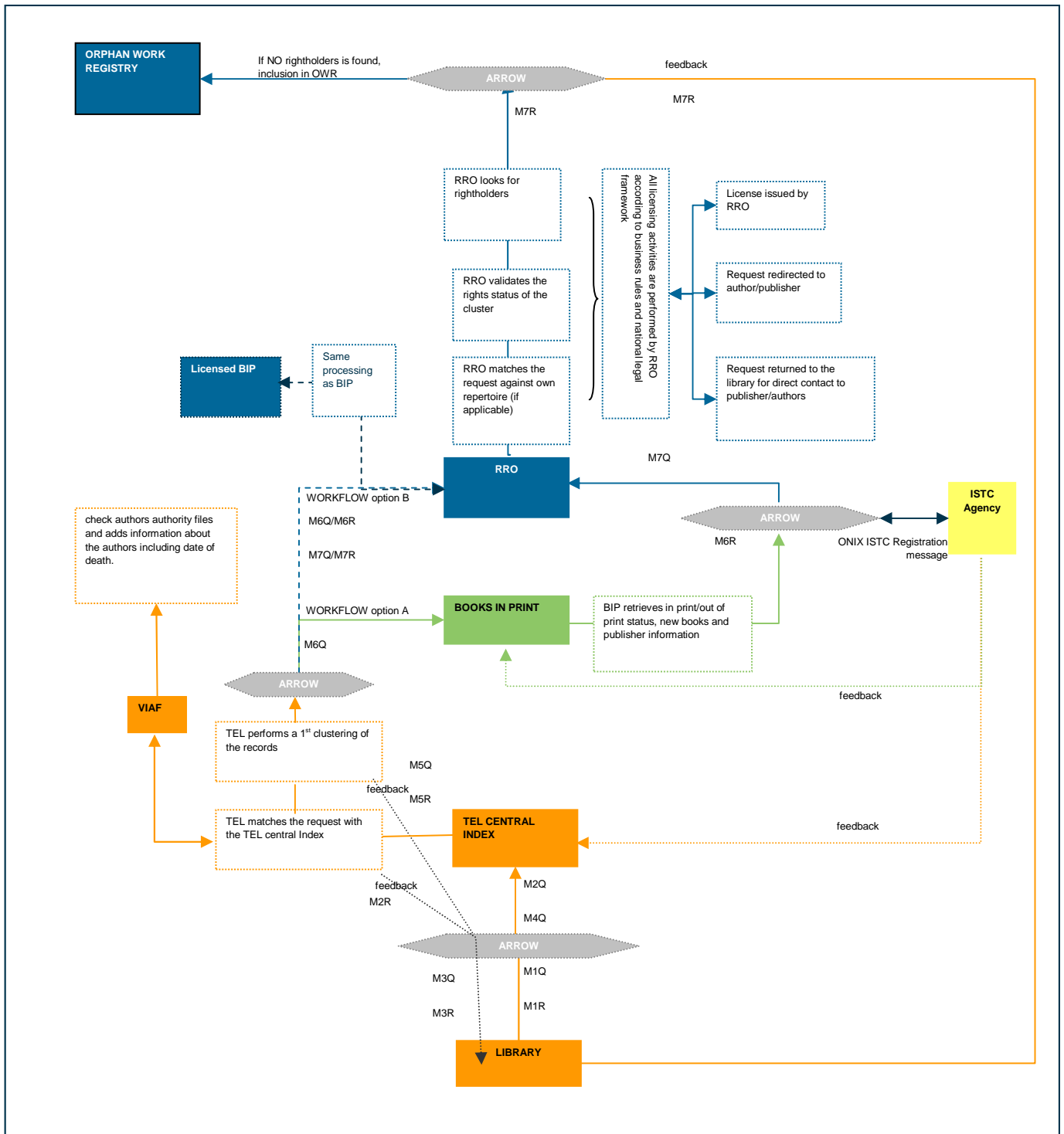
Deliverables D4.1, D4.2, D5.1 and D5.2 were used to specify the metadata messaging formats of the system and ultimately led to the implementation. The strong interconnections between WP4 *Interoperability*, WP5 *Design of system architecture* and WP6 *Implementation of the ARROW system* explain why a common methodology of work was established. The creation of technical working groups involving all the relevant stakeholders helped participants to undertake the work in an efficient way.

### **2.3 Technical working group for messages definition**

A technical working group was set up in October 2009, to define requirements for the implementation of machine readable data exchange, to serve the ARROW workflow. All relevant stakeholders are represented in the WG:

- ✓ Project coordinator (AIE)
- ✓ Leaders of WP4-5-6 (BnF, MVB and Cineca)
- ✓ Technical experts per domain: The European Library (TEL), BIPs (MVB) and RROs (ALCS) representatives
- ✓ Appointed experts: EDItEUR
- ✓ Whenever necessary guidance by standard governing organisations has been required (ISBN International Agency, ISTC international Agency for example)

Work was mainly undertaken by e-mail, but key points were discussed and validated during face to face meetings. Five meetings were held from October 2009 to February 2010 to prepare the Messages Version 0.1 Release for the Alpha version of the ARROW pilot: October 2009 (Paris), November 2009 (Frankfurt), December 2009 (The Hague), January 2010 (Milan), February 2010 (London). Minutes, including task assignment and scheduling, were prepared for each meeting. Each meeting was a milestone in the messages definition and the minutes provide a full account of how the work was achieved.





### **3.1 Schema 1: Library ↔ ARROW: M1Q/M1R**

#### **M1Q: Initial resource and usage request**

This message is the one which begins the ARROW process. It involves a "Registered Library"<sup>3</sup> sending a request to ARROW. The request seeks to describe the resources the library wants to digitise and/or use and the usage(s) for which permission is being requested. For the purpose of the current project these resources are always books.

For the Alpha pilot phase it was decided that, in each request, the proposed usages will be uniform for all books included in the request. Sufficient flexibility, however has been left in the message structure for this to change in the future, to allow greater granularity of relationship between usage requests and books. It is a compound message, with 2 main distinct parts inside a single envelope: one part is composed of bibliographic records (one for each book the library wishes to digitise and use); the other is the request for permission to digitise and use the books. As the bibliographic records provided by the library are generally downloaded from the library catalogue and constitute a MARC21 XML file (see result of D4.2 and D5.1), the message allows libraries to include directly their MARC21 XML records. Permission requests from libraries may be expressed within these messages in two alternative ways. In one approach, the requesting library can define the permissions sought in terms of ONIX-PL usages, thus allowing access to the full flexibility and range of that standard. An alternative approach is to utilize one of a smaller group of predefined "bundles" of usages, assembled into a controlled "Permission Set". EDItEUR has created a short list of these Permission Sets, based upon input from the library and RRO domains: and this approach has been used exclusively during the pilot in order to simplify initial, trial implementations. The message also carries information, needed later on in the workflow, about the organisation asking for the permission (namely the Requesting library).

#### **M1R : Initial resource and usage response**

This is the initial ARROW response to the Requesting library, acknowledging that the submission has been validated (or not), and is being processed by the ARROW system. This internal acknowledgement is also used to exchange identifiers and confirm process integrity.

At this point, for valid submissions, ARROW breaks the batch request down into single lines (one for each book), and from this point forward each line is handled along the workflow as a separate record from the point of view of processing the data.

---

<sup>3</sup> A "Registered Library" is a library accessing the ARROW system as an authenticated user, i. e. after having logged in with the credentials assigned by the system.

## **3.2 Schema 2: ARROW ↔ TEL: M2Q/M2R**

### **M2Q : Resource identification request**

This message is the request from ARROW to The European Library to match each bibliographic record present in the Requesting library's original submission with an existing record in the The European Library Central MARC21 Index. Functionally it is similar to the request for BIP matching (see below M6Q).

From this point on (messages M2Q to M6R, inclusive) the messages include only the bibliographic data as the permission request information is of no practical interest for either The European Library or for the BiPs.

### **M2R : Resource identification response**

This message is the response from The European Library to ARROW conveying the results of the matching process. Functionally it is similar to the BIP matching results message (see below M6R) and to the "Select manifestation" message (see below M3Q).

The European Library response has one of three values: exact match found, partial match found or no match found, according to the results of the matching procedure. The message therefore can convey data about one or more manifestations corresponding to the library request and their matching probability percentage. Alternatively the message can indicate that no match was found by The European Library, and provide guidance for further steps in the workflow.

## **3.3. Schema 3: ARROW ↔ Library: M3Q/M3R**

### **M3Q: Manifestation matching review request**

This message is the request from ARROW to the Requesting library for validation of the matching process undertaken by The European Library. Functionally, this message is essentially identical to target resource matching response (see above M2R).

The message conveys data about all the manifestations found by The European Library as providing matches to the original request from the library, for validation. In the case of a partial match, the message sent by ARROW contains the complete set of bibliographic records found in The European Library Central Index (in the form of a subset of MARC21 XML) and asks the Requesting library to select the record corresponding (closest match) to the resource it wants to digitise and/or use.

### **M3R: Manifestation matching review response**

M3R is the response from the Requesting library to ARROW, with the reviewed and validated match. This response from the library takes one of two forms: it either confirms or rejects each proposed match. Functionally this message is similar to the manifestation matching message from The European Library (M2R) -with the addition of elements needed to convey the results of the review by the Library or other agency.

### **3.4. Schema 4: ARROW ↔ TEL: M4Q/M4R**

#### **M4Q: Cluster creation request**

This message is sent once the Requesting library selects either an exact or a closest match with its target resource from the list of proposed matching manifestations. Based upon this selection, ARROW asks TEL to carry out a clustering of records from The European Library Central Index around the chosen manifestation. This message carries the information needed by The European Library to build a cluster of all records describing books that are manifestations of the same work. Functionally this message is a subset of the M2R message.

#### **M4R: Cluster creation response**

M4R is the message in which The European Library communicates to ARROW, for each match previously identified, the clusters created applying the ISTC rules in a hierarchical way. The message can carry one or more clusters providing a representation of the "work" to which the manifestation corresponding to the record submitted by the Requesting library belongs (Primary cluster) and similar or related "works" (Secondary clusters). In a cluster, there is always one work and from 1 to n related manifestations. The manifestation for which the licence has been requested is flagged in the cluster. Each "work" is accompanied by data about its related manifestations. The cluster is a combination of work and manifestation metadata. At this stage, bibliographic metadata on manifestations is still expressed in MARC21 XML, while Work metadata is expressed in an ONIX for ISTC based format; there is no way of communicating clusters of this kind following any library metadata communication standard.

### **3.5 Schema 5: ARROW ↔ Library: M5Q/M5R**

#### **M5Q: Cluster review request**

This message is the request from ARROW to the Requesting library for the review and validation of the proposed clusters. Reviewing clustering outputs appears necessary at least at the pilot stage, so as to involve libraries in the refinement of clustering algorithms. It is assumed that the Requesting library will be the party reviewing the clustering results but if not, an alternative message addressee can be easily specified. Functionally this message is almost identical to the cluster creation response message M4R.

#### **M5R: Cluster review response**

This message is the response from the party reviewing the clustering outputs to ARROW with the review and validation of the clusters, at both work and related manifestation level. Functionally this message is almost identical to the cluster creation details message M4R, and the request to review clustering outputs M5Q.

### **3.6 Schema 6: ARROW ↔ Books in print: M6Q/M6R**

#### **M6Q: Related BIP request**

This message is the request to the relevant BIP database from ARROW for each manifestation in the cluster(s) provided by The European Library to provide the publishing status (in print or out of print status), commercial availability and publisher information for each of these manifestations. Functionally this message is similar to the request for The European Library matching and identification (M2Q). For manifestations with a known ISBN the message can simply convey the ISBN.

According to standard practice in the book supply chain, an ONIX format is used from this point forward ARROW needs to map data from the MARC21 XML used to exchange messages with libraries into the ONIX format used by BIPs and RROs.

In the case of "no match" found in the The European Library process this message conveys the original bibliographic record describing the manifestation the Requesting library wants to digitise and/or use, mapped into the ONIX format needed by the BIP.

#### **M6R: Related BIP response**

This message is the response from BIP to ARROW providing data about in print/out of print status and publisher information, for each manifestation in the cluster(s) defining by The European Library that has been matched in the BIP database. The response message conveys the records retrieved by the BIP that correspond to the manifestations requested in M6Q, including records describing any additional related manifestation that is not present in M6Q (for example, forthcoming titles or e-books). Functionally this message is similar to the response with results of The European Library matching (M2R).

For the Alpha release, responses from BIP are foreseen as being in the ONIX for Books 2.1 format currently used by BIP; however, the message can easily also carry ONIX for Books 3.0 an updated format which it is anticipated BIPs will be using in the future.

### **3.7 Schema 7: ARROW ↔ RRO: M7Q/M7R**

#### **M7Q: Formal licence request**

This message is the request from ARROW to the relevant Reprographic Rights Organisation (RRO): to match the request against their own repertoire (if applicable); to establish the rights status of each work; and to identify the rightsholders. At this point ARROW has completed all that it can do in attempting to identify works corresponding to the initial manifestation the library wants to digitise and use, and in identifying the manifestations that relate to these works; and in determining their in print or out of print status. ARROW passes all that information to the RRO as an "extended cluster" of works and manifestations, including the so called "ARROW assertions" on the commercial status of the work, along with the details of the original permission request and the information about the Requesting library.

## **M7R: Licence proposal or refusal**

This message is the response from the RRO to ARROW providing data about licences and/or rightsholders, according to the “end points” defined by the RRO in relation to the permissions requested by the Requesting library. The message is designed to be able to convey a licence in the ONIX-LT format or a response stating that the RRO has or does not have a mandate, based on the ONIX for RRO format.

ARROW then provides the final answer to the Requesting library.

### **3.8 Structure of the deliverable**

The deliverable includes spreadsheet documentation for each of the messages, XSD XML schemas for each pair (request/response) of messages and HTML documentation/visualisation of the message structures.

The XSD and HTML elements can be found in the zipped file attached to this Deliverable 3.4.1., whose contents are as follows:

- ✓ ARROW Reference Schema v0.1;
- ✓ Seven individual schema files for message pairs M1 – M7;
- ✓ One schema file for all code lists (enumerated types);
- ✓ One schema file containing structures common to all message pairs;
- ✓ One schema file containing structures not common to all message pairs, but common to message pairs M2 and M4 in the The European Library domain;
- ✓ Eight HTML files containing schema documentation for each of the above eight schemas;
- ✓ Various HTML and CSS files supporting display of the documentation;
- ✓ An ‘img’ folder containing the image file diagrams for the documentation

## **4. MESSAGE DEFINITION: TOWARD THE BETA RELEASE**

The definition of messages for the Alpha release deliberately omits some capabilities that will need to be addressed in the Beta release and beyond. Messages will be defined as soon as clear requirements have emerged.

ARROW-VIAF interaction, ARROW-ISTC interaction and ARROW-ISNI interaction: these interactions in the workflow remain to be fully specified; it was decided not to take them into account in the ARROW Alpha pilot. Impacts on ARROW messaging will be evaluated subsequently. Analysis of the ARROW-ISTC interaction has been initiated, but decisions go beyond the messages domain. Standard message formats for ISTC registration are already available<sup>4</sup>, so no dedicated messages are expected to be required in the ARROW context. A study on ISNI registration in ARROW workflow should be carried out when the standard has been approved.

To date, no detailed requirements have been articulated for communication to and from the Registry of Orphan Works, nor for exchange of enhanced metadata. But, in both cases, issues are

---

<sup>4</sup> More information on ISTC and ISTC Registration message on the website of the ISTC International, [www.istc-international.org](http://www.istc-international.org)

external to the message domain; it is foreseen that the ARROW message framework could be extended relatively straightforwardly to accommodate these cases.

After testing the initial suite of messages in the Alpha pilot, enhancements and modifications to the messages are anticipated to meet both existing and new requirements.

Messages will also be tested against Workflow B where the RRO takes over the function of the BiP<sup>5</sup>, to assess their compliance with it or whether any modification is needed.

## 5. CONCLUSION

The present suite of messages has been developed to serve the ARROW workflow as it has been designed and to carry all the relevant information needed by ARROW to fulfil its goal. The structure of the messages has been designed to allow flexibility in use and implementation, according to the capabilities and requirements of the different players involved. Message constructs (syntax and semantics) are derived, whenever possible, from the standard formats used by the players in the different domains (libraries, Books in print, RROs and collective management organisations), to facilitate interoperability.

Though refinements will be needed and some work remains to be done, the goal of designing a flexible structure for the messages and for their choreography has been achieved. Collaboration between WP4-5-6 will continue, alongside future collaboration with WP7 Validation; and as a result the system will be further improved as the project continues.

---

<sup>5</sup> United Kingdom and Spain, piloting ARROW in the Beta release have adopted Workflow B