FORWARD

Project Number:	**325135**
Project Acronym:	**FORWARD**
Project Title:	**Framework for a EU-wide Audiovisual Orphan Works Registry**
Instrument:	Pilot B
Thematic Priority:	CIP-ICT-PSP-2012-6

D2.6.2 Standardisation Report – Year 2

Due Date:	**30/09/2015**
Submission Date:	**13/10/2015**
Start Date of Project:	**01/07/2013**
Duration of Project:	**36 months**
Partner in Charge of Deliverable	**CRB**
Version Status	**Final**
Dissemination Level	**PU**
File Name:	FORWARD_D2.6.2_standardisation report

Revision History

Revision	Date	Author	Organisation	Description
N 1	15/09/2015	G.Scipione	CINECA	Content
N°2	10/10/2015	Nicola Mazzanti	CRB	Content, Final

Table of Contents

1. EXECUTIVE SUMMARY

Standardization is a horizontal, ongoing task across the project, part of WP2 and responsibility of the partner CRB.

Scope of this deliverable is to report about the actions undertaken in Year2 on Standardization.

Standardization activities within the FORWARD project aim at:

- Assessing whether standardization actions are required, and which areas,
- Initiating and coordinating actions to facilitate standardization.

The type of actions to be initiated depend obviously on the specific areas in which standardization is required. As several filmographic metadata standards already exist, as well as schemata for recording rights information, any standardization actions will be aimed at expanding, refining or improving existing standards rather than at creating (yet) another metadata standard.

The project's partnership is obviously well aware of all the existing metadata standards that are relevant (in fact most of them were directly involved in defining the "CEN Standard for Cinematographic Works EN 15744 and EN 15907").

As planned, activity in Year 2 focused on evaluating in details existing schemata and standards for all metadata necessary to the development of the project, namely descriptive (filmographic) and administrative (i.e. rights).

The analysis of the existing metadata standards and schemata will be completed in the phase of implementing the FORWARD system. In other words the existing standards will be evaluated in a "real life" environment, that of the FORWARD system.

Furthermore, the FORWARD System is developing and internal language to communicate among its different components. The result of this work might lead to the desire to standardize this internal language, if different enough from existing standards and schemata.

In Year 3 this "real life" analysis will be completed and standardization actions will be initiated where and if necessary.

2. INTRODUCTION

The work around surveying and evaluating current metadata standards and schemata was carried out by the project's partners with the support of the subcontractor EDItEUR, who has been asked to work with CINECA to provide advice and guidance in the areas of metadata and identifiers for key entities. This request is based upon EDItEUR's established expertise and familiarity with these areas, as well as its previous involvement with the ARROW project that explored concepts parallel with those of FORWARD (diligent search for rights holders, requests for licenses, determination of possible orphan status) but for the domain of libraries and books.

Specifically, FORWARD carried out an initial review of a number of leading metadata schemes in the overlapping domains of Film, Audio-Visual (AV) and Broadcasting, which appear most relevant to the aims of FORWARD. Similarly, it has reviewed several standard and proprietary identifiers intended for use with the AV resources that are the subject of FORWARD services.

The work is ongoing and it will be developed and refined around specific areas that require further in depth development in the implementation phase of the FORWARD system. These areas include:

- Who will be the main players in FORWARD, both as customers/clients and as partners in the provision of the service? The likely modes of interaction with the central FORWARD service, as well as the identities and capabilities of each constituency may influence the most effective approaches.

- Which will be the key decision criteria, particularly in determining whether or not an AV work is In Copyright, In the Public Domain or potentially Orphan? Knowledge of these criteria will be valuable, so that we can "home in" on relevant subsets of information, rather than replicating full sets of cataloguing data, some of which may have no material impact on the decisions.

- What is the uptake in the relevant domains of each of the main metadata schemes and identifiers discussed? To encourage uptake of the service, FORWARD should ideally communicate with its various constituencies in formats with which they are already familiar. Interaction with the database of Orphan Works implemented by OHIM in Alicante is and remains a key concern and objective of the project and of its standardization efforts.

3. Metadata Schemes in the Film/AV/Broadcast Domain

In this section, we briefly review six established metadata schemes that appear directly relevant to the aims of FORWARD. An initial mapping between these standards is underway to discover any critical gaps or advantages and the results will be circulated shortly.

Review Criteria

All of the standards discussed here, and probably others that are not directly referenced, could contribute usefully to the conceptual thinking needed to underpin FORWARD. But in terms of choosing between them – and even deciding whether such a choice is desirable or necessary – we need to consider a number of factors:

- The purpose of any standard(s) adopted by FORWARD.
- The extent to which each standard includes key information items needed for subsequent FORWARD decisions or assertions.
- How comprehensive is each standard, whilst guarding against data collection "for its own sake".
- Whether each standard is technologically sound, open and extensible in its architecture and documentation.
- The degree of uptake of each standard in the target communities that FORWARD seeks to serve.

3.1. CEN – EN 15744

The European Committee for Standardization (CEN) has developed several metadata standards relevant to the domains of interest of FORWARD. Here we focus on the standard EN 15744, whose English language title is Film Identification – Minimum Set of Metadata for Cinematographic Works.

As the name implies, EN 15744 focuses particularly on work level description. Fourteen elements (or groups thereof) convey basic information about an AV work, including several that relate specifically to the original incarnation of the work. A work identifier (recommended where possible to be the ISAN, see later) is also accommodated within the framework.

3.2. CEN – EN 15907

CEN has also produced a companion standard of relevance to FORWARD, namely EN 15907, Film Identification – Enhancing Interoperability of Metadata – Element Sets and Structures. This is significantly more complex than EN 15744, in a number of ways.

First, the scope of EN 15907 includes not only AV works but also variants, manifestations and items thereof. This results in an extended data model containing around 20 groups of elements and 8 entities. Secondly, the EN 15907 framework allows for the representation of 8 types of relationship between many

of the entities, including the cardinality of each (e.g. whether a particular relationship must be defined or whether it is optional, whether a relationship is '1 to 1' or '1 to many', etc.).

Once again the recommended AV identifier is the ISAN, but there is explicit support for multiple identifiers and identifier types. There is a published mapping available that describes how the EN 15744 elements should be represented within EN 15907. Additionally, most of the metadata model used by the European Film Gateway (EFG) is based upon EN 15907.

3.3. EBU Core

The European Broadcasting Union (EBU) has produced several standards that may be relevant to the AV domain, particularly of course for those works or resources that emerge from broadcasting organizations. EBUCore, the EBU Core Metadata Set, and CCDM, the EBU Class Conceptual Data Model, together provide a framework for descriptive and technical metadata that appears to be being adopted by a number of European broadcasters and archiving projects.

EBUCore began as an attempt to "… refine the semantics of the Dublin Core elements …" for use in describing audio archives, but it is said to have evolved to cover video and other multimedia objects and is in active use beyond specifically archival applications. The standard is well documented and appears structurally sound. It would be interesting to evaluate to what extent recent applications have really moved it beyond its audio roots (still evident in the published data model) and thus to what extent it is relevant or suitable for wider use in describing video and AV resources.

3.4. Registration Metadata for AV Identifiers

Two of the most important resource identifiers in this domain – the ISAN and the EIDR Asset ID – have well documented sets of supporting metadata. The primary purpose in these cases is to provide sufficient controlled metadata to allow for reliable identification, registration and disambiguation of the works/resources concerned, but both allow the capture of considerable amounts of additional descriptive metadata.

Given the ubiquity of these identifiers (discussed further below in Section 4), it is recommended that the content of these two metadata sets be evaluated alongside the information covered by the CEN standards, EBUCore and others.

3.5. ISAN Metadata Sets

ISAN currently maintains two interrelated metadata sets. One is ISAN Metadata for the Identification of AV Works. The other is ISAN Version Metadata for the Identification of Versions or Variants. The first of these covers both individual works and serial episodes, whilst the second deals with expressions and manifestations.

The ISAN Metadata scheme is comprehensive and well documented. The structure and metadata contained therein are expressed in XML, permitting validation at points of creation and use. For more

information, see Section 4.1 below.

3.6. EIDR Metadata

EIDR is the Entertainment Identifier Registry Association. It maintains four types of identifier, designed respectively to identify parties (usually organizations), users, AV assets and video services. Of these it is thought likely that the EIDR Asset ID and the EIDR Party ID will have most relevance to the aims of FORWARD.

The EIDR Data Fields Reference document and other sources provide extensive documentation of the underlying metadata set, as well as pointers to XML schemas, APIs defined to date and sample XML fragments. For more information, see Section 4.2 below.

3.7. MovieLabs Common Metadata

Special mention should be made of the MovieLabs Common Metadata set. This appears to be a particularly strong resource for categorizing and communicating technical aspects of AV works and assets. It is well documented and a number of its concepts or data elements are incorporated (using the md: namespace) into other metadata schemes, notably that for EIDR.

4. Domain-specific Identifiers

4.1. ISAN

The ISAN is a numbering system for the identification of AV content, both works and versions thereof. ISAN is an ISO standard, published as ISO 15706 and ISO 15706-2. The ISAN International Agency (ISAN-IA) manages an international network of 18 or more ISAN Registration Agencies, governs and assists in the implementation of the standard, and manages a central repository containing all ISANs and associated metadata so as to ensure the uniqueness and availability of the information.

The format of the ISAN consists of 24 main hex digits; when printed for human reading this is extended by hyphens between each block of four digits and the addition of two check digits to help identify transcription errors, like this:

ISAN 1881-66C7-3420-0000-7-9F3A-O245-U

Root Episode Version

In this example, the ISAN is assigned to a work that is not an episode or part of a serial AV work (hence the '0000' block), and the characters '7' and 'U' are check characters.

Structurally, as shown above, each ISAN has three segments:

- A root segment permanently assigned to a core work
- An episode (or part) segment permanently assigned e.g. to a television episode or film part
- A version segment that differs for each version identified.

ISANs may be assigned to single AV works, to episodes or parts (as mentioned above) and to composite works whose components may or may not have their own ISANs. To build hierarchies and to minimize data input duplication, ISAN supports 'parent–child' relationships and users may be prompted to declare such relationships at the time of ISAN registration.

To allow for the assignment of identifiers early in the process of creation and production, in-development or 'InDev ISANs' can be registered, based upon a smaller metadata set: these may not be used in the same way as an official ISAN but can be confirmed and transformed into a regular ISAN by the registrant once the production process is complete and the requisite additional metadata has been supplied.

Comprehensive documentation is available and ISAN, together with its associated metadata scheme, can be represented and communicated in XML using a published XML schema. Interactions with the ISAN database are supported via web applications, web services, XML file processing and Oracle data dumps.

4.2. EIDR Asset and Party IDs

As mentioned earlier (in Section 3.7), EIDR maintains four types of interrelated IDs. Of these, the EIDR Asset ID is directly aimed at the unique identification of AV assets whilst the EIDR Party ID may also be of interest to FORWARD as a primary or alternate identifier for the various parties or organizations associated with particular AV works or versions.

EIDR IDs are expressed as Digital Object Identifiers (DOIs), conferring useful additional properties in terms of resource discovery, linking and resolution. Structurally, each EIDR ID consists of a DOI and registry prefix, followed by a slash, followed by 20 hex digits and a check digit, separated by hyphens, like this:

10.5240/XXXX-XXXX-XXXX-XXXX-XXXX-C

DOI and registry prefix — DOI suffix

In this example, the string '5240' denotes the EIDR registry for Asset IDs. Colours are used here only to aid legibility. EIDR Party IDs follow the same format, but include the string '5237' instead of '5240'.

Unlike the situation with the ISAN, no particular meaning or significance is associated with any part of the DOI suffix. The status or kind of the AV asset, as well as any relationships that may exist with other assets or parties, is defined in the metadata that is stored along with the ID.

As shown in the EIDR Data Fields Reference, an extensive set of asset-related metadata can be declared and stored. However it is noteworthy that only a rather small subset of this metadata is mandatory. Two statements from EIDR's website FAQs are worth repeating here:

"EIDR is designed to be cost-effective for large-scale use and is intended to provide an inexpensive mechanism for tracking even micro-assets all the way down to clips and encodings and combinations. To this end, the metadata required and stored by the Registry is restricted to those core elements that help uniquely identify the object that is being registered."

"The system is not intended to replace or compete with commercial metadata providers. EIDR will not provide:

- Metadata intended for consumers,
- Extended or non-factual metadata (e.g., cast & crew, synopses, artwork, ratings, etc.),
- A rights repository.

Rather, EIDR is a B2B service designed to facilitate and support these and other forms of value-added services."

Having noted these deliberate scope restrictions on the part of EIDR itself, it is still evident that the EIDR metadata model is rich and well formed. Each EIDR asset is referred is essentially an 'object'. FRBR-like categorization is possible but should be thought through carefully since concepts like Work,

Manifestation and Version would be modelled via interrelations between EIDR objects, rather than being simply declared as such.

4.3. Other Identifiers

To paraphrase the standard disclaimer, “other identifiers are available”!

Following best practice in this area and similar approaches in ARROW, FORWARD should ensure that its data model and ingest/export routines allow multiple identifiers for the same object to be communicated where these are available. This can assist considerably in recognition and disambiguation of the entity concerned as well as providing “cross-walk” facilities between one identifier and another.

Thus for example it could make perfect sense to communicate and store both the ISAN and the EIDR Asset ID for a particular AV object, if these have both been registered. An interesting sidelight on this topic is the recent announcement that ISAN and EIDR are to set up a joint registration agency in the UK, where among other things customers will be offered the option of purchasing one or both of the relevant IDs for the same object.

The ISNI (International Standard Name Identifier) is worth considering as a primary or alternate party identifier for organizations or persons associated with AV assets. Other resources in this space include the EIDR Party ID mentioned earlier, the VIAF (Virtual International Authority File for names, hosted by OCLC) and perhaps in due course the ORCID (Open Researcher and Contributor ID).

Finally there are a large number of proprietary IDs in use, some entirely internally within organizations and others that are proprietary but find significant use in the wider market. Again, the same principle should apply – it can be very useful to exchange these proprietary IDs as well as standard identifiers, provided that the provenance and type of each ID is also clearly communicated.

5. High-level Requirements & Some Questions

5.1. Parties Involved in FORWARD

A number of Film Heritage Institutions (FHIs) and Commercial AV Libraries will be important contributors of information and metadata to the FORWARD service. There will be a wide user base across the EU, whether these be parties interrogating the system or posing queries or other entities actively seeking rights clearance, licensed use, etc.

We are assuming that the ultimate authorities in terms of rights information about AV resources will be the appropriate Reproduction Rights Organizations (RROs) in each country. It is not yet clear which authority files or reference sources will be used to validate or carry out searches in response to user requests.

(As a comparative example, in the books-related ARROW service, incoming requests from users are effectively matched against the pooled resources of many European national libraries, represented by TEL (The European Library), and their "in print" status (or otherwise) is checked with the Books in Print agencies (BIPs) in relevant countries.)

No similar databases exist in the AV sector.

5.2. Lessons from ARROW

Whilst not slavishly following the detailed features of the ARROW Project and the continuing service that it put in place, it appears sensible for the FORWARD team to review some of its general approaches.

This is partly because FORWARD is conceived as a rights holder information service whose objectives are broadly similar to those of ARROW, albeit in parallel domains (AV rather than books), so that some of the ideas may be a useful inspiration and avoid the reinvention of too many wheels. Also, there is the stated aim of making FORWARD and ARROW as interoperable as possible, so again it should be useful to at least begin with a conceptually similar model.

In no particular order, some key concepts worthy of review from ARROW might include the following:

Assertions. ARROW used this concept to describe statements made with some degree of authority by ARROW itself (ARROWAssertions) or other partner organizations. Simple examples include assertions that a particular book is 'In Print', 'Out of Print', 'Orphan', etc., based upon the information available and agreed algorithms having been applied.

Describing Parties Involved. ARROW used nomenclature such as Requesting Library (all of the initial set of users were libraries), Central Service (for the organization and infrastructure that operates the service, acts as a messaging hub, etc.), Reference Source (in the ARROW context, TEL and the various BIPs), RRO and Orphan Works Registry. It would probably be sensible to generalize or adapt these somewhat for FORWARD and there may be other entities that play significant roles in the process.

License/Permissions Requests. Requests for licenses or permissions to use AV content have been excluded from this part of the work and so are not considered in detail here. However, it may be worth noting that ARROW allowed for requests to include either complex machine-readable license term expressions (expressed in ONIX for Publication Licenses, ONIX-PL) or much simplified "permission sets" constructed around frequently encountered scenarios.

Modelling of FRBR-like Entities. It will be crucial to have in place (if these are not present already) robust definitions of whichever type of entity modelling is most appropriate for the various AV communities. A model constructed around the Work | Manifestation | Item characterization worked for ARROW and a similar framework needs to be in place for FORWARD to avoid ambiguities or misunderstandings.

Short Descriptions. The primary sources of input data (and requests) in the ARROW system were generally library-created MARC metadata records, whereas other players in the chain of investigation were more familiar with ONIX records. Both of these contained many of the key information items required to process ARROW requests but also a considerable amount of information (for example on cataloguing detail, marketing collateral, etc.) that was not necessary to support diligent search for rights holders. So in this instance it proved valuable to devise several types of "short description" that included information such as work identifiers and dates of birth/death of contributing authors with direct relevance to the searches and decision algorithms involved.

5.3. Appropriate Metadata Formats for Various Purposes

Referring back to the 'review criteria' mentioned earlier, it will be important to differentiate the different purposes for which metadata formats or schemes should be selected or adopted by FORWARD. At a high level, one may distinguish between interactions with the wider communities outside of FORWARD – those supplying information, querying the service and receiving responses – and processes that can be considered as occurring within the FORWARD service itself.

For any use cases that involve the bulk loading or ingest of AV metadata – whether as a preliminary setup process or to deal with a sizeable batch of queries – then it appears to me almost paramount that FORWARD should support one or more formats that are already widely used in the AV communities or well on their way to adoption.

In other words, selection of metadata format needs to be made on the basis of both suitability and uptake in the communities concerned. This mirrors the approach in ARROW mentioned above, where the decision was taken to support both MARC records (almost ubiquitous in the library community) and ONIX Books records (very widespread in the books trade) as permitted formats.

Similar considerations may apply for interactions between FORWARD service partners, such as the central FORWARD service, the FHIs, and whichever other authority files or reference sources are used.

There may be more latitude when it comes to “internal” FORWARD processes and also perhaps in returning simple assertions about rights or orphan status to those making the original queries. In these cases, other metadata schemes or subsets thereof could come into play.

Nevertheless it is still likely to be prudent to decide upon one conceptual model – preferably based upon a public, open standard – to service these requirements too, in order to avoid the overhead and cost to FORWARD of designing and then maintaining yet another scheme.

In closing this particular section, an interesting “edge case” could be ONIX. The ONIX metadata set, covering a wide range of applications and business purposes, was used as the basis for the data model used by ARROW. It may be a step too far to envisage transposing ONIX principles and vocabularies into a service for the AV domain, even if enriched as necessary by concepts from other namespaces. But on the other hand, such an approach might significantly cut down on developmental overhead and support the aims of FORWARD/ARROW interoperability.

6. Useful Resources & Links

ARROW

Information is still available about the original ARROW and ARROW Plus projects. Responsibility for ongoing ARROW services has been taken over by the recently established ARROW Association. Messaging to and from ARROW, as well as a good deal of the internal data modeling, was developed by EDItEUR and the ARROW project team into a variant of the ONIX standard, namely ONIX for Rightsholder Information Services or ONIX-RS.

CEN

The European Committee for Standards or CEN publishes a wide range of European standards. These include the two formats most closely allied to the aims of FORWARD, namely EN 15744 and EN 15907.

EBU

The European Broadcasting Union or EBU is responsible for two pieces of work directly relevant to any broadcast content that may be the subject of FORWARD requests. These are the EBU Core Metadata Set and the EBU Class Conceptual Data Model.

EDItEUR

EDItEUR provides a wide range of standards to the publishing and related communities. In addition to the ONIX-RS suite mentioned above, it is also responsible for the widely adopted ONIX for Books, ONIX for Publication Licenses or ONIX-PL and other specialized formats.

EIDR

EIDR is the Entertainment Identifier Registration Agency. Its identifier standards are expressed as DOIs and extensive descriptions can be found in the EIDR ID Format and EIDR System Data Fields Reference documents.

ISAN

The ISAN International Agency or ISAN-IA, based in Switzerland coordinates the work of ISAN worldwide via a network of international registration agencies. Two published documents are of particular relevance to FORWARD, namely the ISAN User Guide and the ISAN Metadata Schema.

ISNI

The International Standard Name Identifier or ISNI is fast gaining traction as a widely used party identifier.

MovieLabs

The MovieLabs Common Metadata set is a valuable and extensive resource, used in its own right and also as embedded concepts within other schemes such as that of EIDR and others.

7. Conclusions

The environment in which FORWARD will have to operate is complex and, to some extent, underdeveloped when compared to other sectors like Music and Audiovisual (TV).

The reasons are obviously linked to the rather traditional (pre-digital) environment in which the works FORWARD operates with are made available, exploited etc.

As described in this document, many standards exist, and they are being led growingly to be interact with each other in a way that is as transparent as possible. Nevertheless, it is true that these standards do not fully respond to the needs of the FORWARD project as such.

The FORWARD system will integrate concepts and solutions from these various standards in order to create an 'internal language' to use in communications among different components, databases and services of the System.

Once the FORWARD System will be fully implemented and integrated, it will be possible to evaluate whether this "internal language" (based on a specific metadata schema) is a good candidate for standardization, completely or in part.