Grant Agreement ECP-2007-DILI-527003

ARROW

ANNEX I

Standards applicable in the framework of digitisation programmes

Deliverable number/name	*D4.1*
Dissemination level	*Public*
Delivery date	*30 July 2009*
Status	*Final*
Author(s)	*Piero Attanasio (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.

The context

Any mass digitisation programme involves very delicate issues regarding the treatment of bibliographic data and rights information. Let's imagine a book in a library collection that is candidate to enter a digitisation programme. The librarian should determine whether the book is in print or out of print and whether there is a known rightholder.

All these elements require data to be managed and exchanged, and thus there is the need for standardisation in the definition of data formats and tools for information exchange. The ARROW consortium was very aware about this since the start up, but such awareness was reinforced when we had the occasion to compare our approach on this subject to the approach contained in the Settlement agreement between Google and the American Association of Publishers and the Authors Guild, which has been made public at the end October 2008.

The coincidence between the ARROW start up and the announcement of the Settlement made this comparison natural and, in the following steps, also productive, since we had the occasion to start a discussion between ARROW and the Google staff in charge of the implementation of the Settlement bibliographic database as well as the representatives of the forthcoming "Book Right Registry", the creation of which is foreseen by the Settlement .

Such discussions, in particular that with the BRR, aimed at finding a common understanding of the problems and at envisaging common solutions. The first step was to have a shared vision of the generic context of any digitisation programme from the data management viewpoint. Then the issue was to make the discussion more concrete, and to analyse how best we might move from a general common statement of principle to more specific practical implementation plans.

It was impossible to ignore the new context that such an initiative created, in particular as far as standard issues are concerned, since – by definition – standards are something that have to consider all players acting in the same environment. Therefore we decided to enrich *D4.1 State of the art and guidelines on standard applicable* with a picture of the environment where such standards have to find the ground for application, between the two sizes of the Atlantic and possibly world wide.

There are standard issues to be managed at every point along the value chain of the digitisation programme. These are to be defined in the relationship between digitisation programmes (Europeana or Google), intermediaries like ARROW or the BRR, collecting societies, publishers and other rightholders.

To understand how best to resolve these issues, it is critical to describe the communication issues that are involved along the value chain:

1. Identifying the "books" - By definition, a digitisation is a process which creates a new "book as product" (in digital form) starting from another "book as product" (in printed form). Therefore, there are issues related to standard identification of these two classes of items.

2. Identifying the "works" - Digitisation always implies a need to deal with rights (even if the rights are in the public domain). Rights are not defined with respect to "books" (manifestations), but to "works", i.e. the content of the book. The second issue is therefore to have a mechanism for the standard identification of works themselves, as separate referents from the "books".

3. Identifying the "rightholders" - In the communication with rightholders, the first requirement is to be able unambiguously to identify them. There is a need, then, for the standard identification of rightholders, or – at least – of the names of rightholders and/or their agents, those authorised by them to clear rights for defined uses.

4. Standard bibliographic metadata - Effective communication about books, works and rightholders implies the use of standard metadata to describe them.

5. Standard rights metadata - The communication along the chain also requires the ability to exchange "Rights Information" (RI), i.e. metadata about rights, which as we have already indicated are related to works. This includes pieces of information such as their "public domain" status, their "commercial availability" status, authorisations and conditions for the different display uses, etc. Here again, we can identify a requirement for the use of standard metadata – for rights expression.

6. Standard e-content formats - If rightholders choose to enter a digitisation programme, it is possible for them to provide either digital files or printed books to be digitised. Here we can identify the issue of standard formats for these files. Later, when the file is in the database, there is also an issue related to the standard formats available for the final user.

7. Resolutions systems - Once a book is in the digitisation programme, the service may provide links to external resources. This can be seen as best managed by the adoption of a standard resolution mechanism to direct users from the book to related resources and/or from the citations or a metadata record of any book entered the programme to the "location" where to find the book itself.

8. Interoperability for common search - Finally, some discussion has focused on interoperability among the different digital collections, resulting from digitisation programmes (within private or public initiatives) and the offer provided directly by

publishers or other intermediaries, in order to provide users with richer results from their search, in particular when they start from a search engine service.

We are aware that the resolution of the issues described in this document – as it is entirely typical in standards – requires the collaboration of all stakeholders. It is not a question of a single player having to adopt standards. All must be asked to meet their individual responsibilities: ARROW, publishers, collecting societies, bibliographic agencies, libraries, Google, the BRR, etc.

The current document looks at those issues to a medium term perspective.

1. Standards for book identification

Identifying "pre-ISBN" books

The ISBN is very well established as the identifier of books, and is used all over the world. There are over 150 ISBN agencies around the world. However, when dealing with historic library collections, it is inevitable that many books do not have an ISBN; as the standard was only widely adopted in the Seventies.

There is a problem in the unambiguous identification of printed books that do not carry an ISBN. For example, in the current database available for the Google Settlement management, two other codes have been used: OCLC and LCC (Library of Congress) numbers. Nevertheless, a significant proportion of the books in the database are not identified with any (external) unique code.

European libraries have their own numbering systems for older titles, frequently shared at national level; however such national systems are not interoperable between themselves or with OCLC or LCC systems, though important efforts are made within the TEL initiative (The European Library) to enhance interoperability between the European national systems.

Possible solutions for this problem (the order of listing should not be interpreted as indicating a preference) are:

- to register ISBNs on a bulk basis for the books that do not have them. This may be done for all the records in a library catalogue or be triggered at very first steps of a digitisation programme for the individual books involved.

- to adopt one of the alternative existing numbering systems, providing interoperability with the others through mapping. In some way, this means to attempt to transform a legacy proprietary system in a standard one, which implies – first of all – the willingness of the owner of the system to collaborate. In respect to the previous solution, this adds one level of complexity because of the need to provide interoperability between the selected numbering system and the ISBN system. The advantage, instead, can be in avoiding to use the ISBN for purposes that are not in the origin of the standard, which is mainly tailored to serve the book supply chain.

- to make the existing systems interoperable through the creation of a new numbering system developed through a combination of the legacy ones . Pros and cons are very similar to the previous solution but affording since the beginning the problem of interoperability between pre-existing systems.

Identifying digitised version of the books

There is a second problem related to the identification of new "books" created by the digitisation process. The ISBN standard (ISO 2018:2005) is very clear in defining that any new edition of a book, in either print or digital form, must be assigned a new ISBN insofar as it is "made available separately".

This could be interpreted as implying that a book does not require an ISBN if it is digitised with the sole purpose of making it searchable on the Internet (and thus is not itself "made available" using the Internet as a distribution channel). However, when the book is made available, it certainly requires a new (and different) ISBN from the printed copy that has been digitised, in particular when they are included in a supply chain .

On these themes, ARROW contacted the International ISBN Agency that decided to create a working group to better study the problem, which is one of the agenda item in the ISBN Annual General Meeting in Seoul, next September 2009.

2. Standards for work identification

As we have discussed, rights are (for the most part) in works, not in books. This creates a clear requirement for work identification and for tracking the relationships:

- between works and their different "manifestations" (e.g. the different editions, printed and digital, of the Hamlet)
- between principal works and derivate works (e.g. translations, annotated, new editions, etc.)

In both cases, the ISTC (ISO 21047:2009) is the appropriate ISO standard to be applied. This standard is right at the beginning of its deployment internationally, and decisions that are taken with respect to the management of the ARROW project (and in parallel decisions taken within the Google Settlement) will be crucial to the way in which the standard is implemented.

Therefore, any digitisation programme requires the determination of these relations. When the relations are discovered, if they are also registered in a standard way, through the ISTC, this generate a value for the whole system, which may be exploited in any other context. This is a very clear occasion to foster the start up of the new ISO standard.

The ARROW project decided to trigger the ISTC registration as far as it will be possible, through the relations with the emerging ISTC agencies around Europe . In the discussion with the representatives of the Google Settlement parties, ARROW promoted a co-ordination on this aspect. Authors and publishers associations in the US, which are the constituencies of the BRR, share the ARROW approach and clearly expressed their willingness to support the ISTC.

3. Standards for rightholders identification

Rightholders identification is another important topic in the recent standards development, specifically with the continuing work on another ISO standard, the ISNI (International Standard Name Identifier).

However, dealing with information about persons, rather than about "objects" (like books, or works), implies also sensible problems in respect to privacy issues. The concept which is crucial to deployment of the ISNI is the separation between the identity of a party (typically in this context a rightholder and a "legal person") and the identity of a public name (or often many names) used by that party. The first type of identity cannot be managed in public repositories, because of confidentiality and data protection issues; the second can, in order to facilitate the efficient exchange of information.

This is a new approach to the "name authority", a concept that is well understood in the library community, and will surely have applications within the ARROW project.

To avoid the risk that confidentiality and or data protection issues prevent the free sharing of useful information between rightholder, the forthcoming Registry of orphan works, the individual RROs and the requesting libraries, a solution to the problem should be sought in collaboration with the ISO working group for the ISNI.

4. Standards for bibliographic metadata

In the management of a complex system like the ARROW, a number of different metadata format may be involved. Once again, the analysis of what happened for the management of the claiming process within the Google Settlement provided some guidelines about the risk that lack of awareness of standardisation may create.

In the Google Settlement, publishers are encouraged to provide "ONIX" files to communicate (meta)data about books. However, publishers have no idea what "minimum data set" is required by the system in order to allow them to receive the return message containing the information that has been promised.

The use of the ONIX family of standards within the administration of the Settlement and in the longer perspective in the management of communication within any digitisation programme, and thus primarily within ARROW, has great potential. This can only happen through ad hoc implementation, to be agreed with the ONIX community, i.e. with EDItEUR and its internal governance bodies, as it has been planned for the future steps of the ARROW system.

Though ONIX is the most adopted standard for metadata exchange in the book trade, in the library world it is not used as much, and thus for ARROW to adopt ONIX for the whole process would be non-effective. When the dialogue is between the library world and ARROW, other standards have to be considered, in particular MARC and Dublin Core.

As any standard of this type, there are always pros and cons in using one or the other along a process. From the ARROW viewpoint the issue is not to select one system, but to be able to receive data in one format (e.g. MARC when they arrive from a library) and to send in another format (e.g. ONIX when we have to communicate with a Books in Print catalogue).

5. Standards for rights information

The ONIX family of standard messages is particularly useful for the exchange of rights and permissions information. EDItEUR has already implemented a framework for communicating this type of metadata, ONIX for licensing terms (ONIX-LT).

In the ARROW context, as well as in the context of the Google Settlement and the BRR, the work done for RROs within the ONIX-LT framework to facilitate the worldwide exchange of information about repertoires and to support financial distributions may be seen as particularly useful.

The requirement for rights information exchange within the system is very broad, and can include:

- Information about public domain status of a work.
- Information about the in print / out of print status of each individual title.
- Exchange of information about rights ownership and "mandates".
- Exchange information between libraries and RROs to support financial distributions, once plans for remunerated use of digitised books will be fully in place.

- Publishers and authors will need to communicate to the library – or to a private company like Google – permissions for one or more display uses for one or more works; this may also need to carry additional information (e.g. the price for the sales to final users).

The ARROW project will continue to explore each of these subjects in considerable depth, as part of the process of establishing standard ways of communicating of rights and permissions information between publishers and libraries within digital library programmes.

It is be very important to establish a continuing relationship between ARROW and the Google programme and the BRR on this topic to ensure that all requirements are properly identified and met in the standards development process.

6. Standards for e-content formats

To enter any digital collection (e.g. Gallica 2, Libreka, eBog or the Google BookSearch programme), a books can either be digitised from a paper edition or provided by the publisher (or other rightholder) in a digital format. Subsequently, digitised books will be provided to third parties in different ways, and according different business models.

There are problems in the technical format of the digitised books:

- format or formats accepted by platforms from publishers. It is clear that ePUB files are going to become the standard for XML-structured e-books, while pdf remains very important as the “digital equivalent” of a printed version.

- Format or formats used by libraries or further intermediaries (e.g. Google digitises a collection of a library) to return the output of a digitisation to the rightholders.

For example, within the Google partners programme, up until now, publishers have not received digitised files from Google, while libraries have received the files for their collections. The argument from Google for refusing the provision of the files to publishers has been the absence of a standard format (although the same problem apparently does not apply to libraries).

7. Resolution mechanisms

Up to now we have spoken about “digitisation”. However, digitisation is just the prerequisite for the use of a work, and in particular for the publication on the web or other “display uses”. For instance:

- in Libreka users can access a preview of the title and than be redirected to services to access the full text (in different forms) or to buy the printed book;

- in Gallica 2, copyrighted works are accessible via intermediaries that can have different business models (e.g. subscription, pay per view, pay per download, free access paid via advertising, etc.);
- in ebog.dk users can either "buy" the digital version or "borrow" it, i.e. access for a limited time;
- in the Norwegian "Bookshelf" users can access copyright books on limited basis;
- in the Google BookSearch programme, according the new businesses envisaged by the Settlement, copyrighted books will be available in preview, or on sale for online access, or via institutional subscription.

A key aspect of all those services are the links provided to users for improving their experience and getting further value, in particular when a preview of the book is displayed. The way these links are created can be improved by using standard "resolution services", like that provided by the DOI system. The combination between the ISBN and the DOI, called ISBN-A, can substantially benefit the user experience in all these initiatives.

8. Interoperability between different e-content collections and with search engine

Any digital library or e-content collection provides own search facilities to discover individual books. These can be based on full text search, or on metadata, or on combinations of the two. However, users usually start their search using "general purpose" search engine, and – in the current European market – Google in large prevalence.

There is here a key aspect in use of standards in order to avoid creation of dominant positions in the digital library sector. As soon as the Google BookSearch project will be integrated within the normal Google search, as announced, there is a problem of interoperability between Google BookSearch and the other initiatives, like Libreka, Gallica 2, ebog.dk, Bookshelf in Norway, the forthcoming similar initiatives in other European countries, and – above all – with the Europeana project. All such digital library collections may be seen as being in competition with the Google BookSearch library programme.

As with any Internet resource, these collections can be made available to be indexed by any search engine. When Google indexes third party of this type, a number of issues emerge, including:

- the fair "ranking" of competitors' digital collections results displayed within the same framework as those from Google BookSearch

- the possibility for competitors' digital collections to communicate access rights to Google through a standard protocol, such as ACAP

The adoption of standards is once again a key element of the future development of the market. The need for a constructive dialogue between Google and the other existing digital collections is entirely self evident and has been promoted by ARROW.

Conclusion

The evolution of next months will determine long term equilibrium in this sector in Europe and at global level. The commitment of the European book world, also through ARROW, is to foster the adoption of standards along the whole value chain for digitised or "digital born" books so to contribute to the evolution of content distribution in a way that ensures for the future the same level of pluralism and cultural diversity that characterised the book sector in the past.