Identifier and Metadata Standards in the Publishing Industry

A REPORT ON CURRENT STATUS FOR **IPA** AND **IFRRO**

October 2021

By Mark Bide

International Federation of Reproduction Rights Organisations

ifrro

IPA international publishers association

'Part of IPA's State of Publishing Collection'

Contents

All links in this document were checked on 24 July 2021 and were confirmed to be functioning.

1 Introduction

1.1. BACKGROUND

This report was commissioned in February 2021 by two global organisations: the International Publishers Association (IPA); and the International Federation of Reproduction Rights Organisations (IFRRO). Its objective is to provide a high-level overview of the deployment of technical standards in the global publishing value and supply chains, with a particular emphasis on where the standards infrastructure may need development if it is to meet the needs of some or all participants in those chains – creators (authors, illustrators, photographers); publishers; intermediaries of many different kinds; collective organisations (particularly Collective Management Organisations, CMOs); and consumers (end users).

This is a broad ambition for a brief project. Even within what might be called "book publishing" there are many different value and supply chains. Extending this remit to journals, magazines and newspapers makes the task daunting.

However, it has never been our remit to write an encyclopaedic account. While our ambition is to look at a high level at all aspects, we focus most of our effort on a specific set of challenges identified by the commissioners of this report:

- Collective rights management
- Out of commerce works
- Orphan works
- eCommerce in the developing world
- Accessibility for people who are print impaired.

Our primary interest is in standards for naming and description – identifiers and metadata (and messaging about these) – although it is important to recognise that the boundaries

between these classifications are fuzzy. We will touch on standards for the formatting of content itself only to the extent that these have a particular impact on a sector of specific interest – notably, accessibility.

We have in mind a number of perspectives:

- The effectiveness of existing standards which are widely deployed
- The reasons why some existing or attempted standards developments have failed to meet their objectives
- Where there are gaps in the infrastructure and how the intervention of relevant stakeholders might prove effective in filling these gaps.

1.2. THE STRUCTURE OF THIS REPORT

Structuring this report proved challenging. Ultimately, we decided to manage our material in the following way:

- **Section 2** looks at questions of definition, and how standards are formulated
- **Section 3** looks at a number of individual standards and specifications, particularly (but not exclusively) those named in the terms of reference
- **Section 4** looks at the issues on a sectoral basis, considering both different sectors of the publishing industry and different areas of standards activity
- **Section 5** looks specifically at issues of rights management, pulling together themes from earlier sections as well as discussing specific rights management issues including Orphan and Out of Commerce Works as they relate to publishing
- **Section 6** brings together some conclusions and makes recommendations for the future.

As well as the main text, the report has many links to other documents. Some of these are simply citations for the source of information provided. But many are links to documents which contain additional information, some of which we regard as essential reading for any real understanding of this report. If we had rewritten all this information to include it in the text of this report, it would have been three or four times the length it is now (and would have taken many weeks).

1.3. ACKNOWLEDGEMENTS

This report would not have been possible without the input of a wide range of participants who allowed us to interview them on Zoom against a very tight timetable. Their input has been invaluable, but all errors, misunderstandings, omissions and all the opinions expressed are entirely our own.

Participants are listed alphabetically *by affiliation.*

Monica Halil Lövblad	ABC; WIPO
Piero Attanasio	AIE
Giulia Marangoni	AIE; TDMREP
Owen Atkinson	ALCS
David Grundy	ALCS
Beat Barblan	Bowker; ISO TC46/SC9
Alan Danskin	British Library
Andrew MacEwan	British Library; ISNI
Michael Healy	CCC; IFRRO
John Balean	CEPIC; TopPhoto
Alex Cope	CLA
Heather Walmsley	CLA
Kevin Gohil	CLA
Paul Jessop	County Analytics; ISO TC46/SC9; DOI Foundation
Richard Orme	DAISY
Graham Bell	EDItEUR
Chris Saynor	EDItEUR
Angela Mills Wade	Europe Analytica; European Publishers Council
Enrico Turin	FEP: EU IPO OOC initiative

Anna Vuopala	Finnish Ministry of Education and Culture
Cristina Mussinelli	Fondazione LIA
Caroline Morgan	IFRRO
Alicia Wise	Information Power
José Borghino	IPA
Michael Steidl	IPTC
Stella Griffiths	ISBN International; ISO TC46/SC9
Tim Devenport	ISNI; EDItEUR
Sven Fund	Knowledge Unlatched
Max Mosterd	Knowledge Unlatched
Sebastian Posth	Licium; ICC
Paola Mazzucchi	mEDRA; AIE
Emma House	Oreham Group
Andrew Hughes	PDLN
Nicolas Roebben	Phidias Consulting
Philippe Rixhon	Philippe Rixhon Associates
Paul Seheult	PICSEL
Vincent van den Eijende	Pictoright
Sarah Faulder	PLS
Tom West	PLS
Clare Hodder	Rights2
Godfrey Rust	Rightscom
Giulia Marangoni	TDMREP; AIE
Victoria Owen	ARL/CARL; University of Toronto; ABC
Anita Huss-Ekerhult	WIPO
Benoît Müller	WIPO
Dimiter Gantchev	WIPO
Michel Allain	WIPO Connect

2 Some definitions

We were asked to write for "the intelligent layman" rather than for standards experts, and some initial orientation on the topics at the centre of this report may be helpful to ensure that we are communicating with our readership unambiguously.

2.1. WHY STANDARDS?

Why do we need technical standards of the type we are discussing in this report? The answer may seem self-evident, but so much hangs on it that we cannot leave it entirely unspoken. As a first approximation, we will take as the answer that identification and metadata standards meet the need for information to cross boundaries between different computer systems while retaining as much semantic integrity as possible. With a well specified standard (or suite of standards), information passes between systems without any need for human intervention to resolve issues of ambiguous meaning.

Within a single organisation, internal semantic standards are sometimes very well specified and documented; but unfortunately, they may equally be based on the principle that "everyone knows" what a specific term means; particularly within large and geographically distributed organisations that means that disparate systems that need to use what is ostensibly "the same information" become unable to communicate unambiguously with one another. So much is commonplace.

The problem grows exponentially as information flows beyond the boundaries of individual organisations, as the number of nodes within a communication network grows. In a complex automated transactional environment, different people and organisations must have total confidence that their individual systems are talking to each other in a completely common language, if they are to have an appropriate level of trust in the transactions that depend on these exchanges of data. This implies a process through which the meaning and structure of shared information can be tightly defined.

Sometimes, this can simply be a more-or-less informal arrangement between trading partners. But as a supply channel becomes more complex, many-to-many communications using poorly defined or poorly implemented methods of communication

become increasingly burdensome to every participant.[1] The only response is the development of standards which:

- "Codify the boring" – allowing organisations to focus on bigger problems
- Radically increase efficiency (through reduction in complexity) particularly in any environment involving many-to-many communications
- Make it easier to keep track over existing market transactions as well as allowing entirely new ones, particularly as business moves to digital
- Reduce entry barriers and reduce risk for all participants
- Make for easier system development and reduced maintenance
- Reduce supplier lock in.

Properly defined and implemented, standards reduce risks and costs while increasing potential revenues.

2.2. FORMAL STANDARDISATION

Standardisation may come about in a number of different contexts – through ISO, through W3C, through formal and/or informal trade associations. Ultimately, perhaps the most critical issue with standardisation is the system of *governance,* not least because no standard can ever be wholly neutral. Every standard has an unspoken "point of view", a set of usually undocumented assumptions that informs its specification. Any organisation implementing the standard had better be clear that it shares that point of view.[2]

It is also critical for an implementer that the standard is stable[3] – because otherwise there is significant and unquantifiable risk in committing to implementation. The specification must represent a consensus "point of view" among those who will implement it – or at least something close to a consensus. ISO has a precise definition of what reaching consensus means in its process – it means bringing the discussion to a point where there is "no sustained opposition". Getting to consensus takes time, which is one of the reasons that standardisation processes can sometimes seem interminable.

Within our frame of reference, there are standards defined through the ISO process (for example, ISBN); through W3C (we will touch on EPUB3, now a W3C "Business Group Report"; and on the growing potential of ODRL, a W3C "Recommendation"); and through international trade bodies (for example, the ONIX family developed by EDItEUR).

All these organisations develop *non-proprietary s*tandards – in other words, they are managed in common by the members of the standards organisation. The standards we discuss here are, with few exceptions, "open" for anyone to implement, and for the most part do not demand licence fees from end-users;[4] all require financial and time commitment for organisations that want to be directly involved in the standards development process. Some standards – notably the identification standards managed by ISO – allow registration agencies to charge in order to "recover costs" from users.

Identification standards can be expensive to administer and maintain; they may require an explicit business model beyond membership payments, although once established the issue (sometimes called "minting") of a single incremental identifier may be so small that it is close to costless.

Some standards develop in less formal groupings; this can particularly be seen in the area of STM publishing (see Section 4.3.1). The point at which a more-or-less informal grouping of interested parties creates a "standard" rather than just a shared specification might be an interesting debate, but it would not be a fruitful use of time in the context of this report.

It can of course be much easier for a dominant player to impose (or attempt to impose) a "proprietary standard" on a market. However, it is hard to see anywhere in our frame of reference where this has happened. Dominant players sometimes have had (indeed continue to have) an extremely important role to play in the success or failure of the development and implementation of standards, but this is different from imposing a set of proprietary rules on the market as a whole.

Consider the position of Amazon. It has an internal identification standard, the ASIN (Amazon Standard Identification Number) which identifies every Amazon SKU.[5] It sometimes incorporates a numerical overlap with an ISBN but does not seek to replace that ISBN in anything other than Amazon's own system. Similarly, Amazon's Kindle ebook readers support a proprietary text format (AZW) and Amazon converts publishers' files

1 Many businesses have been built on the premise that intermediaries have a necessary role in a supply chain in normalising and/or tailoring otherwise disparate flows of data.

2 In this context, it is worth noting the dominance of North America and Europe in standards development and governance in this sector at least; that suggests that there is a proper space for more active intervention by a more diverse range of organisations.

3 It is important to distinguish between stability and stagnation. Good standards are designed so that they are structurally stable but able to adapt as the environment demands. This underlies the ISO practice of regular reviews of all its standards to ensure that they retain relevance.

4 This is not true of some of the library standards, nor of standards issued by BISG. We are also unclear about exactly how some of the newer "standards organisations" that have arisen in the academic space are funded (see Section 4.3.1). The common definition of an "open standard" includes all those released under FRAND (Fair, Reasonable and Non-Discriminatory) licence, as is the case of many of the technology standards underpinning the ICT industry.

5 Stock Keeping Unit – anything that is for sale on Amazon from a book to a garden fork.

to this format rather than use the standard EPUB3 text format – but there has been no attempt to make AZW a more widely implemented "standard" for all ebooks.

It is self-evident why Amazon uses the ASIN rather than ISBN as its primary identifier of the things it sells. No external standard identifier would meet the requirement, and in any event, *it is good practice **never** to use an external standard as the primary key in a database.*[6]

2.3. WHY STANDARDS SOMETIMES FAIL

There are many and various reasons why standards efforts may fail to lead to wide-spread deployment, and these are not infallibly predictable (otherwise the considerable effort and cost involved in their development would never have been invested). The failure of a standard (either broadly or in specific implementation) may have many roots, often occurring in combination. These include:

- The standard as eventually developed or promulgated may not meet a real need or the costs of deployment may exceed its value to some or all players in the chain
- The cost of implementation may fall unevenly between different players in a particular value/supply chain; the financial incentive to implement may not be great enough if the financial benefit is apparently gained entirely in efficiency elsewhere[7]
- Effective implementation may depend on the engagement of individuals or organisations without the necessary skills or understanding; at its most extreme, there will be individuals and organisations that do not know that a standard even exists
- "Legacy systems" (that is, any system that is already in production or even close to being in production) may not allow for the implementation of a standard
- Lack of priority in IT departments with scarce resources; this may be as simple as the undoubted fact that standards implementation is an unexciting task for developers, and they will often find "good reasons" not to implement them
- The commercial interest of an influential player (or group of players) in the chain in obstructing or delaying the deployment of a standard, for their own benefit[8]
- The business model adopted by a Registration Authority (for example high cost of entry; lack of exclusivity) may deter potential Registration Agencies
- Market inertia may so substantially delay the building of implementation value through network effects that the return on investment of creating the standard is lost or at least substantially reduced[9]
- The technical or social environment for the effective deployment of a standard may not exist; it is a rarely spoken truth about standards implementation that building the right social infrastructure (including but not limited to governance) is likely to far outweigh any technical design issues in success or failure

6 The use of the ISBN as the primary key in publishers' domestic systems has often been the cause of considerable problems.

7 We have long argued that efficiency benefits in any supply chain are ultimately shared by all players in the chain, but this argument sometimes fails against the reality of annual corporate budgets and inter-organisational politics. It should also never be forgotten that one person's efficiency gain is often another person's livelihood.

8 As we have said elsewhere, lack of public availability of standardised data has created many substantial businesses which depend on information asymmetry.

2.4. GETTING THE TIMING RIGHT

One of the greatest challenges in standards development is timing. To what extent should organisations develop standards to meet an "obvious" future requirement? Too soon, and perhaps a few early adopters may keep a standard deployed long enough for the market to catch up – otherwise that standard may never be deployed or may fade into disuse. Too late, and markets may have already become dysfunctional or dependent on proprietary solutions promulgated by dominant market actors.

We will develop at least some of these themes as we look at specific issues in this report.

2.5. MODELS OF GOVERNANCE AND WHY THEY MATTER

What creates the circumstances in which an organisation adopts an identification or metadata standard? This is perhaps a misleading question, inasmuch as it is not an individual organisation that usefully adopts a standard but a whole group of partners who share information with each other. We cannot say this often enough, the adoption of standards is essentially more a social than a technical process; if the social process is wrong, even what may be the best technical standards will fail.

We see in this report several different models of standardisation, and we will briefly consider the standardisation processes and how they are governed.

9 Sadly, this is an issue which recurs through this report. It is particularly frustrating when those who have championed the development of a standard then fail to implement it in a timely fashion. However, this is so common as to be completely predictable.

2.5.1. ISO (despite appearances, the initials are not an acronym)

The ISO standards mentioned in this report are: ISBN, ISSN, ISNI, ISTC, DOI. (ISCC may become an ISO candidate standard, but it is not yet – see more below.)

ISO is an international membership organisation, its membership being 165 national standards organisations. Its Council is drawn from among the Members on a rotational basis. It has a sizeable secretariat, but this is more involved in process than in the standards themselves (certainly in our sector).

Its standards development[10] is undertaken under the auspices of specialist committees – in the case of all the standards listed above, this is TC46/SC9 (Technical Committee 46/ Sub Committee 9). Participation in technical committees is through nomination as "an expert" by a national body.

The practical processes of standards development is through the formation by the appropriate committee of subgroups of experts set up to undertake a specific task; they operate by consensus. When developing a standard, there are two rounds of voting in which the national members of ISO are entitled to vote. The process is extended – ISO say that the typical time between starting work on a standard and its final approval is 3 years.

TC46/SC9 plenary meetings are held all over the world (when travel is permitted); subgroups of experts meet at such places and times as they agree, with individual members undertaking work between meetings (drafting etc).

2.5.2. The World Wide Web Consortium (W3C)

The W3C Recommendations and related specifications mentioned in this report are XML, ODRL, RightsML, EPUB3. (There is also a Community Group currently working urgently on a Text and Data Mining profile of ODRL.)

The W3C is a membership organisation with over 400 members; technology companies predominate. It is not an incorporated body and has no physical location of its own, but is "hosted" by a number of academic institutions worldwide.

A few large publishing companies are members, but annual subscriptions are high. However, the real cost of membership is substantially higher than it appears, as members are expected to dedicate significant specialist human resources to the work of W3C. W3C has a number of employed experts, who act as gatekeepers and provide specialist input and guidance at all levels.

There are various ways in which work in W3C[11] work strands may begin, but for the specialist type of specification which are of primary interest to us in this report,

10 See this guide to the ISO process **https://www.iso.org/publication/PUB100269.html**

11 The W3C process (see **https://www.w3.org/2020/Process-20200915/**) is complex and we will not try to explain it in detail here.

Community and Business Groups are the typical entry level. Participation is open to non-members – the group simply has to gather sufficient interest. These groups are designed to "socialise ideas". Work may be completed in a "report" or may proceed to formal standardisation (Recommendation) through a Working Group (members and invited experts). Approval of Recommendations requires a vote of members.

Community reports can be completed quickly. Formal Recommendations are much more tortuous and sometimes remain in "draft" form for many years. W3C went through a major management reorganisation in 2016 at least in part as an attempt to improve the effectiveness of its processes.

2.5.3. ANSI-NISO

The only ANSI-NISO standard mentioned in this report is DAISY3. Mention is also made of two NISO Best Practice Recommendations (one still in draft); these are not formal ANSI standards.

ANSI is the US national standards body; ANSI delegates responsibility to NISO (National Information Standards Organization) for US national standards in its areas of competence. NISO has a small specialist secretariat.[12] It is a membership organisation with around 100 voting members (all from the US). It is based in Baltimore.

Most of NISO's work is focused either on libraries or the interface between libraries and publishers. Its standards and best practice recommendations are developed by committees of members with some expertise provided by the secretariat, and approved by members.[13]

So far as we are aware, NISO is the only official national standards body in this sector which provides standards that are used internationally.

2.5.4. EDItEUR

Standards managed by EDItEUR and mentioned in this report are the ONIX family of standards, Thema, and a number of transactional messaging standards.

EDItEUR is an international standards organisation, based in London but with "over 110 members drawn from 25 countries around the world". These are drawn from all sectors of the book and journal supply chains and supporting industries. Its Board has 14 members, 13 of whom are representatives of "charter members" of EDItEUR (who pay enhanced membership fees); these are either international trade organisations (including IPA and IFRRO) or national trade organisations. The 14th member is the Executive Director.

12 For information only, the Executive Director of NISO currently provides the secretariat for ISO TC46/SC9

13 Because it is part of the national standards body, NISO processes are quite complex for an organisation of its size. See **https://groups.niso.org/apps/group_public/document.php?document_id=21703&wg_abbrev=staff**

EDItEUR has a number of sectoral standing committees to which any of its members may belong. These committees typically meet face to face twice annually (at international book fairs in Frankfurt and London); during periods of active development, groups meet more frequently, usually remotely. EDItEUR has a small but technically-focused secretariat which undertakes much of the work on standards development, supported by specialist paid consultants. Approval of standards lies with the committees, with decisions ultimately ratified by the Board.

Because of its size and structure, EDItEUR is sometimes able to move faster than larger and more formal organisations, which is attractive for the management of dynamic standards.

2.5.5. Other organisations which manage standards mentioned in this report

A number of other organisations manage standards mentioned below, but few of these show any significant variation from those we have looked at:

- IPTC is an organisation in the international newspaper industry which is much like EDItEUR
- BIC and BISG are organisations similar to EDItEUR but nationally based (in the UK and the US respectively)
- The Library of Congress manages MARC, LCSH and (with OCLC) DDC
- There are a number of organisations managing single standards/specifications in the academic space. Two are specifically mentioned in this report: ORCID and ROR

2.6. DIFFERENT FORMS OF GOVERNANCE – THEIR PROS AND CONS

As we stress above, governance is at the heart of creating the necessary confidence and trust in standards for them to be widely adopted. We were asked to provide a comparative overview of different models.

Governance	Pros	Cons
ISO	• Branding/universal global recognition • Highly respected process and clear governance • Specialist staff process expertise • Participation of international experts	• Lack of speed and urgency in a rapidly changing environment – can generate frustration • Lack of flexibility • Surprising lack of attention to interoperability (no overarching model) • Cost of participation can be high because of face-to-face meetings in international locations • Participation is usually drawn from a small group of highly specialist "standards experts"
W3C	• Branding/universal global recognition • Specialist staff technical and process expertise to assist in the development process • The "only game in town" if you want standards that will be recognised by the World Wide Web community • Interoperability with other aspects of the WWW technology stack • Appropriate use of the Community Group reports can deliver quick results	• Lack of formal corporate structure • Lack of speed in formal standardisation process possibly driven by overfull agenda – as a result, many Recommendations are widely implemented while still in draft • Lingering concern on the part of the media community that W3C is dominated by "big tech" and is hostile to improving the management of IP on the internet • Cost of membership and participation is a significant barrier to entry • Participants require high degree of technical expertise

Governance	Pros	Cons
National standards body (eg NISO)	• Good national sectoral branding and may also be widely recognised overseas in some sectors • Locally respected process, unlikely to be widely understood • Specialist staff expertise in the sector	• The main clue is in the name – national standards bodies do not enjoy international membership • Formal standards organisations may have rather slow consensus processes
EDItEUR or similar	• Branding/global recognition within the sector • Specialist staff expertise developing standards "in house" with cross-media contacts • Interoperability • Can move extremely quickly when necessary	• Small organisation – limited capacity and risk of lack of financial stability/ sustainability (particularly if dependent on a small group of members for a significant element of funding) • Dependent on individual expertise of staff and a limited pool of consultants – another sustainability risk • Lack of recognition beyond the sector
Single standard organisation	• Dedicated to solving a single problem; should be able to act at high speed • Likely to have access to considerable specialist expertise – otherwise would be unlikely to begin the process • Those establishing the organisation may have sufficient brand value to provide confidence in the output	• Splintered governance increases costs and may be opaque • May not involve a wide-enough stakeholder group • May not be widely interoperable if developed for a narrow application • Long term sustainability

There is no single "right answer" as to which of these is "the best" model for developing standards. Every choice involves compromise, each one has some advantages. All of these (except perhaps the last) allow for the engagement of a wide group of stakeholders from across the community, which is ultimately essential to credibility.

A wide set of expertise is required in standards development, skills that are rarely found in a single individual. The tendency to see standards as the province of technologists is profoundly mistaken. Of course, technological input is necessary, but the technology must be placed within a much wider context – commercial, legal, political. This is particularly the case with standards in the rights management domain.

This is a topic to which we will explicitly return in Section 6.1.

2.7. WHAT ARE IDENTIFIERS?

"An identifier is a name which is unique within its type and domain." This definition is the opening sentence of The Linked Content Coalition's (LCC) paper *Principles of Identification*[14]. We know of no better primer on identifiers and identification. We strongly recommend that anyone who reads the present report is also fully familiar with this LCC paper. It is a short paper – only 4pp. It would be difficult to cover the ground more economically.

2.8. WHAT IS METADATA?

Three separate and very different projects asked this question in the mid to late 1990s. The standard definition ("data about data") was clearly an inadequate answer. These projects were:

- The Resource Description Framework (RDF), a W3C Recommendation, a generic approach which was the forerunner of "linked data",
- Functional Requirements for Bibliographic Records (FRBR), an IFLA (library) project,
- indecs (Interoperability of Data for Electronic Commerce Systems), a cross-media multistakeholder project on identifiers and metadata, funded by the European Commission.

14 See **http://www.linkedcontentcoalition.org/phocadownload/principles_of_identification/LCC%20Principles%20of%20Identification%20v1.1.pdf**

All came to much the same conclusion. In a slightly modified version of the indecs definition of metadata:

"An item of metadata is a relationship that someone claims to exist between two referents"[15]

In other words:

- All items of metadata depend on identifiers – underlining the challenge of metadata schemes that depend on undefined "free text"[16]
- All items of metadata depend on typed relators; these may be one-directional or bi-directional
- All items of metadata are claims; it is important to know who is making that claim in order to assess its trustworthiness.

This definition was published in 2000, six years before Tim Berners-Lee publicly named the concept of "Linked Data".[17] However what indecs proposed went beyond linked data; the simple "triple" structures of Linked Data are not enough (a realisation which has subsequently arisen in many other contexts). The identity of the asserter of the relationship may be as important as the assertion itself.

It will be extremely helpful for readers of this paper to be familiar with *The indecs metadata framework*.[18] The thinking that is incorporated in this paper has informed much that has happened since, at least in commercial metadata in the media: ONIX, DOI, DDEX.[19]

FRBR[20] was probably rather more challenging to the library world than indecs was in commercial metadata development, because of the already well established standards infrastructure in libraries. By the turn of the century, MARC (Machine Readable Cataloguing) already had an astonishing 30+ year history,[21] and was the data model that informed more or less every library system in the world. However, this data model was based on the two-dimensional library catalogue card, well suited to computing in the 1960s and 1970s, less well to the age of the World Wide Web.

15 Referents are "things which are identified"; the original definition used the term "entity" which is more ambiguous.

16 For this purpose, "controlled values" act as identifiers

17 See **https://en.wikipedia.org/wiki/Linked_data**

18 See **https://www.doi.org/topics/indecs/indecs_framework_2000.pdf**

19 The music-industry standards organisation, broadly similar in structure and approach to EDItEUR in publishing. See **https://ddex.net/**.

20 See **https://www.ifla.org/publications/functional-requirements-for-bibliographic-records**

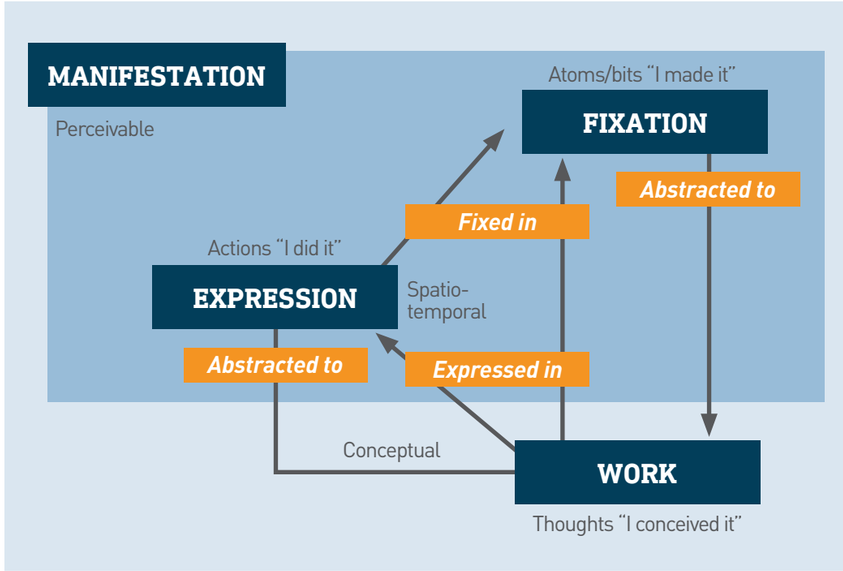
21 See **http://tandfonline.com/doi/abs/10.1080/07317131.2011.574519.** Paywalled except for the abstract.

Nevertheless, MARC and the various tools that surround it have survived FRBR for another 20 years. New tools are now slowly coming into use: Resource Description and Access (RDA),[22] a "linked data" model for creating cataloguing records is already in use in some national libraries; BIBFRAME,[23] a proposed replacement for MARC itself is now in v2.0. This should ultimately be the replacement for MARC but (in the light of library funding challenges, from national libraries downwards) is unlikely to be widely adopted for many years.

2.9. THE CHALLENGE OF THE "EXPRESSION"

There was one important disagreement between FRBR and indecs, which remains unresolved. It may seem theoretical and rather unimportant, but it is part of the challenge of developing a textual work identifier (see Section 3.2). This centres on the concept of an "expression".

To explain this issue, we start with the indecs model of making:



22 See **https://www.loc.gov/aba/rda/**

23 See **https://www.loc.gov/bibframe/**

This model recognises one abstract entity "the work" and two types of "manifestation" – things that can be perceived: an "expression", a spatio-temporal entity (like a performance) and a "fixation" (like a book).[24]

FRBR, on the other hand takes a different view:

> *"A Work, according to FRBR, is a distinct intellectual or artistic creation. It is an abstract entity. The boundaries of a work are sometimes dependent on the cultural or national view, but FRBR suggested some boundaries. I like to think of it as the ideas that a person has. A work is realized through an expression - another abstract entity. An expression is the intellectual or artistic realization of a work in the form of alpha-numeric notation, musical notation, choreographic notation, sound, image, object, movement, etc., or any combination. The person decides how he or she wants to express the ideas – as text, sound, image, etc. and in what language. Using this model, one could even collapse both "work" and "expression" into something called an "abstraction" when that was found to be useful for a particular application. It should be noted however, that FRBR kept them separate."*[25]

It is easy to see the challenge. An "expression" in the FRBR model is another class of abstraction, not a manifestation as in indecs.

At first sight this seems trivial – surely, it's "just" a question of one or the other model changing the word "expression" for something else? But words have huge resonance, and semantics are never entirely neutral.

And there's another particular problem which flows from this difference of view: derivative "works". In the indecs model, all works have the same status as any other work: <work B><is a revision of><work A>. But in FRBR, some works exist in a hierarchical relationship: <expression B><is a revision of><work A>.

Neither of these is "right" or "wrong". They are simply different ways of modelling the same thing, but in considering issues like work identification, it is critical to understanding some of the problems of incomprehension that sometimes get in the way of collaboration. Fortunately, many on both "sides" now understand and sympathise with this difference in understanding (which may indeed be necessary to fulfil the very different requirements, say, of library cataloguing and rights management). But it doesn't seem to bode well for the development of a standard textual work identifier shared between these communities.

Some compromise is clearly going to be necessary; but the reality may be that it is not possible to bridge the differences in requirement within a single framework.

24 Both FRBR and indecs also recognise individual "instances" or "items". In computer science terms, of course, all classes of instances are abstract, but this would lead to a discussion on a rather different level.

25 See *The FRBR Model* a 2004 presentation by Dr Barbara A Tillet (**https://www.loc.gov/catdir/cpso/frbreng.pdf**)

3 Reflections on some specific standards and specifications

In this section of our report, we look at a list of specific standards that were raised with us in the Terms of Reference (and a few that were not). As already discussed, this is not intended as a complete survey. Note that some standards are covered in the next section of the report, which is sectoral in focus.

3.1. THE INTERNATIONAL STANDARD BOOK NUMBER (ISBN)

"ISO 2108 establishes the specifications for the International Standard Book Number (ISBN) as a unique international identification system for each product form or edition of a separately available monographic publication published or produced by a specific publisher that is available to the public. ISBN is applicable to monographic publications (books), not to textual works (content). Monographic publications include individual sections or chapters where these are made separately available[26] *and certain types of related products that are available to the public irrespective of whether those publications are made available for sale or on a gratis basis."*

The ISBN hardly needs introduction to the audience of this report. It is an extremely successful, globally implemented system for the identification of books as products

26 Although it is permitted in the standard, we have not come across instances of its use to identify chapters of books. Such a practice is certainly rare and the circumstances in which it can be applied very constrained.

which has been in place for over 50 years.[27] The specification has been through a number of iterations over that period and has proved adaptable to the changing environment in which it operates, not least the exponential growth in the number of new titles being published each year since the 1980s and, after a slightly faltering start, the recent explosion in ebooks and audiobooks.

An explanation of its success lies in numbers. Bearing in mind the enormous number of books that are available for sale at any one time in developed economies — and the number of new books being published each year — it is hard to imagine an orderly and effective supply chain for books without robust and standardised identification of products.

However, its original adoption is perhaps a valuable case study of the power of dominant players to enforce the implementation of a standard.

The ISBN has from the outset been managed by local registration agencies (of which now it has over 150), whose activities are co-ordinated by the ISBN International Agency, ISBN-IA. Registration agencies, which have a high degree of local autonomy, range from large commercial companies that act as registration agencies alongside the provision of broader commercial services to the book supply chain, to National Libraries and other government agencies.

CASE STUDY – ISBN AND WH SMITH

When the ISBN was first adopted in the UK in, there was a considerable problem with one aspect of the introduction – the ISBN Bar Code. Designers were horrified that their designs might be spoiled by printing ugly barcodes on the back. There were some signs that they might win the argument and that barcodes would be printed on the inside of covers rather than the outside.

WH Smith was then the predominant bookseller in the UK and it issued an edict: No ISBN bar code on the back cover, no sale. Bar codes were quickly adopted universally.

The role of dominant players in rapid standards adoption would be hard to overstress.

ISO rules (and the ISBN specification) allow for charges to be made to cover the cost of ISBN allocation – although in some territories they are provided free to anyone who needs them. ISBN rules provide territorial monopolies for ISBN allocation, to avoid any opportunity for international arbitrage.

Despite, maybe indeed because of, its extraordinary success and effectiveness, ISBN faces some challenges and complaints about its operation. We have been able to identify (and asked to comment) on the following:

- "**ISBN is too expensive"**: this has been the centre of many complaints particularly in the US but also in the UK.[28] While some larger publishers complain about the charges, the reality is that the purchase of ISBNs is hardly a major item in their operating costs (although the monopolistic nature of supply may be galling). The problem is rather more acute for self-publishers or publishers with a very small title output. It costs £89.00 to purchase a single ISBN in the UK;[29] $125.00 in the US;[30] €28 in France;[31] €70 in Germany.[32] As quantities increase, costs per ISBN come down quite dramatically. In contrast, in India, where the ISBN is administered by a department of government, self-published authors can get an ISBN free.[33] Similarly, Amazon offers authors who are self-publishing through its Kindle Direct service free ISBNs when these are needed (i.e. where there is a physical book as well as an ebook – no ISBN is required for a Kindle edition – and of course all receive ASINs).[34] In view of the commercial interests involved, we are doubtful that this is a soluble "problem". With the best will in the world, ISBN-IA is not in a position to force individual agencies (particularly the largest and most influential ones) to change their pricing structures unless perhaps it could somehow be shown that their charging regime goes beyond "cost recovery" – in itself, an extremely slippery concept. It seems to us likely, for example, that the real cost of establishing a new account for a self-publisher seeking a single ISBN could be convincingly demonstrated to be somewhere in the region of the prices being charged.

- "**Ebook-only publishing without ISBN means that titles are escaping legal deposit schemes"**: legal deposit schemes have long depended on ISBN and related services to ensure that their activities are complete. Whatever may be the arguments about current processes for legal deposit, the value of maintaining a complete record of the national literary output for future generations would be difficult to deny. However, in light of the way that the "ebook only" market has developed, it seems clear that ISBN cannot be the only tool to ensure compliance with legal deposit regulations, and other solutions will need to be found unless part of the cultural heritage is potentially to be lost. If the ISBN is not valuable to a self-publisher, or indeed a commercial publisher only of ebooks, it would be utopian to imagine that publishers will adopt it just to suit the needs of national libraries. Unless it were to be compulsory – which brings us neatly to the next challenge.

27 Standard book numbering schemes go back to about 1965, but the ISBN first became an ISO standard in 1970.

28 There have also been consistent complaints about government-controlled agencies in other countries overcharging small publishers

29 **https://www.nielsenisbnstore.com/Home/Isbn**

30 **https://www.myidentifiers.com/identify-protect-your-book/isbn/buy-isbn**

31 **https://www.afnil.org/tarification/**

32 **https://www.isbn-shop.de/**

33 **https://www.24by7publishing.com/isbn-for-self-publishers---independent-authors-in-india.html**

34 Note that authors are also given the option of buying their own ISBN through Bowker (the US national agency) at a small discount on the standard price. We do not know how this works for non-US authors although we have been assured that Bowker would not issue an ISBN prefix to a publisher without a US address. **https://kdp.amazon.com/en_US/help/topic/G201834170**

- **"ISBN is used as a tool for censorship"**: there can be little doubt that this is true. Where a regime demands that every book that is sold in a country must have an ISBN; and that every publisher who gets an ISBN prefix must be government approved, and the local ISBN agency is government controlled, the opportunity for censorship is self-evident. Again, this lies outside anything that can easily be tackled by the ISBN-IA,[35] even though it may lead to some real challenges beyond the "freedom to publish" issue itself.[36]

- "**ISBN is badly administered in some territories"**: the ISBN-IA would be first to admit that there are a few agencies which are not well run, with errors in ISBN allocation and other administrative issues. Some of these are in less-developed countries, but this is not invariably the case. Some are in relatively prosperous countries and are run by government agencies – but badly.

CASE STUDY – ISBN 10 BECOMES ISBN 13

When it was launched in 1970, the ISBN was a 10-digit number; but 30 years later it was becoming apparent that it would run out of capacity. There had been an explosion in the number of new books being published which began in the 1980s and continues to this day. The revised 13-digit standard was published in 2005, with phased introduction expected to complete at the beginning of 2007. The switch to 13 digits also enabled compatibility with EAN13, through a unique and extremely ingenious solution agreed with ISBN-IA. EAN13 prefixes are normally representative of countries; the 978 and 979 prefixes used by ISBN represent "Bookland". It is the implementation of the 979 prefix which increases the capacity of the standard. The first 979 ISBNs appeared in 2008 in some national markets, although notably the first 979 prefixes are only just now becoming used in, for example, the US. It cannot be a surprise that this change was strongly opposed by some vested interests. It involved extensive changes to IT systems, and the necessity for change in some major markets was a long time away. (There are complex rules governing the construction of ISBNs which mean that some countries ran out of 978 prefix ISBNs many years ahead of others.) This experience teaches some important lessons, not least about the importance of looking ahead – and about the unexpected consequences that a major change in market conditions may have on mechanisms for ensuring uniqueness in numbering systems. It also reminds us of the importance of a strong governance system that has the determination to push back on powerful lobbies.

35 It is hard to conceive that any government-controlled ISBN agency would cease to issue ISBNs, even if the ISBN-IA were to expel it from the global network. There would be no obvious motivation for it to do so.

36 There is at least anecdotal evidence that, where an ISBN is required on books found for sale in bookstores, it doesn't actually matter what the ISBN is. This has led to a practice by which apparently valid ISBNs are improperly printed on covers of books other than the title to which they were allocated. It is not, of course, possible simply to look at an ISBN and to know whether or not it has been correctly applied to the book in hand.

- **"ISBN is inconsistently applied to ebooks"**: as ebooks became commonplace in the earlier years of the 21st century, different publishers applied ISBN to their ebooks at different levels of granularity, with some following ISBN rules of applying different ISBNs to each different ebook format, others using a single ISBN for all ebook variants. To a large extent, this was an argument about what constitutes "a product" and was also driven by concerns about unmanaged proliferation of ISBNs. We understand that, among larger international publishers at least, a common understanding has been reached about how the ISBN is most appropriately used for ebook identification (in line with ISBN-IA guidelines).[37]

- **"ISBN is not available in every country in the world":** this is true, but only just! Although the vast majority of countries have their own agency, larger agencies sometimes provide cover to countries and territories where there is no local agency. For example, the agency in France assigns ISBNs to a number of Francophone territories which don't have their own agency (including Djibouti, Madagascar, Monaco). There is a small number of countries for which no such arrangements currently exist. These include Somalia, South Sudan, Yemen and Equatorial Guinea. For these countries (where the local book supply chain is also weak and not really using ISBN) ISBN-IA encourages publishers who want to export *beyond their borders* to work with an international distributor who can provide ISBNs as necessary. The ISBN-IA does not formally offer itself as a "registration agency of last resort" although has acted as such in a real emergency.

- "**ISBN doesn't work as an identifier for anything other than products"**: this is simply a truism. ISBN is a product identifier for books, designed and implemented solely for that purpose. That hasn't stopped many organisations – including publishers, intermediaries and CMOs – from using ISBNs for all sorts of other purposes. In most instances, this has been a pragmatic solution that provides a stop-gap solution within a given context. Many publishers and intermediaries use "dummy ISBNs" perhaps to identify items (including products) that are not books[38] but need to be stored and located in their warehouses; this works fine as long as the dummy numbers do not escape "into the wild" (which from time to time inevitably they do, sometimes causing chaos).[39]

Issues around ISBN, the management of rights and CMOs are discussed elsewhere in this report. A further subset of this problem also emerged during our research for this report on the deployment of standards in academic publishing, particularly in the light of the growing trend towards open access books; again, we will discuss this elsewhere.

37 **https://www.isbn-international.org/content/guidelines-assignment-e-books**

38 In the education sector, it is equally the case that ISBNs are given to books that are not strictly products – are not separately available.

39 They also use up part of the finite stock of available ISBNs which annoys ISBN-IA but is largely ignored and is not a major issue, at least in the short to medium term.

We should make one thing clear though: neither of these issues is *prima facie* a problem with the ISBN itself or its formal rules of deployment. Rather they are the problems that inevitably occur when seeking to re-use an identification standard for a purpose for which it was never intended.

3.2. THE INTERNATIONAL STANDARD TEXT CODE (ISTC)

"The purpose of the International Standard Text Code (ISTC) is to enable the efficient identification of textual works. The ISTC provides a means of uniquely and persistently identifying textual works in information systems and of facilitating the exchange of information about those textual works between authors, agents, publishers, retailers, libraries, rights administrators and other interested parties, on an international level." (ISO 21047)

Unfortunately, the ISTC (in its current iteration, launched in 2009) has not succeeded in getting wide deployment. While it appears that there is lingering support for a textual work standard among at least some of the communities of interest listed in the preamble quoted above, that support has not been sufficient to keep the existing standard afloat. The ISTC International Agency (ISTC-IA) failed as a business and was closed. Existing ISTC records are being maintained in dormancy, while a decision is made about the future.

Following the failure of the ISTC-IA, ISO TC46/SC9 established an ad hoc group to make recommendations on the future of the standard. Our understanding is that this group has submitted a report, recommending that the current standard should be withdrawn, but that a project should be undertaken to see whether a new initiative might find a way through some of the problems that caused ISTC to fail.[40]

The fundamental cause of the failure is, as might be expected, the failure of any players in the chain identifying sufficient value to them to justify the cost of paying for ISTC.

40 It is possible that this new initiative will adopt a different name to avoid confusion with the previous (failed) standard. While we would agree that this would be a good idea in principle, we will continue to use the ISTC name for the identifier of textual works for convenience and clarity. It should be obvious from the context whether we are looking forward to a successor standard, backwards to the failed standard, or both.

CASE STUDY – THE FAILURE OF ISTC

Why did ISTC fail? A difficult question to answer, but several contributory factors can be identified. Most critical perhaps was the inability to gain consensus on the "killer app".

There was a strong initial focus by ISTC-IA on its adoption within the retail supply chain. This was not a sufficiently strong value proposition for publishers, who felt it was just an opportunity for the registration agencies to sell them something they didn't need; or for intermediaries or large retailers who were confident in their own proprietary approaches to clustering products.

The significance of ISTC application in rights management may have been grasped at a theoretical level. However, the fact that both publishers and CMOs have managed without it for over a decade since its launch indicates that any necessary work-rounds have not proved to be a major problem in the "real world".

Publishers,[41] who need to be a – perhaps the – primary "point of entry" to any textual work identification system, perceive any internal system requirement for a **standard** textual work identifier to be low. They have survived well enough without one for decades, when necessary, using either a proprietary "project" identifier or the hardback ISBN as a surrogate. There is no obvious justification for deploying a standard identifier which their supply chain is not even requesting, let alone insisting upon.

There may be a deeper challenge. For trade books, there are often different publishers of the same work in the same language in different territories – one work identifier or two? Clearly, by most definitions, there is only one work,[42] but identifying that work may not be in an individual publisher's interests in the supply chain, since it might end up promoting sales of a competitor's product available at a lower cost to a consumer if purchased from another territory.

Retailers might be seen as having a substantial interest in being able to cluster different manifestations of the same work; but major retailers can again achieve much of what they need to do algorithmically within their own systems – or can purchase proprietary linking

41 It is arguable – indeed it has often been argued – that the only proper place to enter a work identification scheme in any medium is with the original creator. Many author groups (and their agents) have been strong in defence of this position. Our view is that this is something of a counsel of perfection; to have any meaning, a work identifier in any medium has to be linked at least one identified manifestation of that work. In the limited case of self-publishing, it is certainly arguable that the author-as-publisher is always the appropriate point of entry to a work identification system. But authors are not (for the most part) likely to deal with the problem of applying work identifiers to the historic record (which includes orphan and out-of-commerce works). Authors should, of course, be able and encouraged to register works but if all textual works are to be given identities, then other groups will also need to be involved – not only publishers but libraries. CMOs are also likely to have a role to play. It is an open question how valuable an identifier system is if it is not comprehensive or moving towards being comprehensive in a particular domain.

42 It is (just?) arguable that changing spelling and cultural references to suit readers on the different sides of the Atlantic would be sufficient to define one of them as a derivative work but even that has immediately obvious problems if the link is publicly accessible (and might be perceived as giving the "original" work a preferential value).

information from third parties (Books in Print organisations, for example). It is also an inescapable truth that retail markets have seen a sharp reduction in competition.

Libraries would find ISTC a valuable additional tool in their cataloguing armoury, seeking under newer cataloguing models to "cluster" records for the same works and "expressions" of works.[43] However, the appropriate model for clustering in libraries would potentially have a substantially different level of granularity from what might be useful in a publisher or retailer environment.

Some substantive projects have already been undertaken clustering library records of manifestations (primarily books) as evidence of the existence of a work which could then be given an ISTC. These include the Europeana ARROW[44] project which completed its initial phase in 2011; and the more recent GLIMIR project undertaken by OCLC (although GLIMIR did not have the explicit intention of identifying "works", as its name makes clear).[45]

The ARROW project points in the direction of the use case which looks the strongest argument for the deployment of a textual works identifier. Is it possible to imagine a future comprehensive infrastructure for rights management without an explicit textual work identifier? ARROW used clustering of library records to support the identification of textual works (books) which were potentially "orphan". Similar challenges exist in other rights-related applications (not least out-of-commerce works).

However, even those CMOs which maintain detailed repertoire databases (primarily those in the "common law territories") do not report any problems in meeting *current* operational requirements without having a textual work identifier (primarily by using manifestation identifiers as a surrogate). There are also some challenging questions here about the granularity of identification, as CMOs would probably expect any work identifier to have a granularity which follows changes in rights ownership (something that might look more like the publisher requirement).

Although it is possible to be entirely convinced *in theory* of the long-term requirement for a textual work identifier, so far at least it has proved difficult for any of the possible constituencies that might find it useful to be sufficiently convinced that there is a compelling reason to pay for the work to create it, particularly as its value will be very limited until a substantial proportion of the current and historical record has been identified and described.

These fundamental challenges of granularity will need to be addressed at an early point in any process intended to revive an ISTC-successor standard; while we can see some potential pragmatic approaches (which might look more like FRBR than indecs in the way that they approach the identification of abstractions), it may be extremely difficult to persuade all interested parties to compromise on a viable solution.

Whatever happens, it is clear that any revival would be a long-term project requiring long-term management and financial commitment. We return to thinking about the future of ISTC in Section 6.3.1.

3.3. INTERNATIONAL STANDARD NAME IDENTIFIER (ISNI)

"ISO 27729 specifies the International Standard Name Identifier (ISNI) for the identification of public identities of parties, ie the identities used publicly by parties involved throughout the media content industries in the creation, production, management and content distribution chains. The ISNI system uniquely identifies public identities across multiple fields of creative activity and provides a tool for disambiguating public identities that might otherwise be confused. The ISNI is not intended to provide direct access to comprehensive information about a public identity but can provide links to other systems where such information is held."

The ISNI was standardised in 2012, in part because the only extant name/party identification system in the media space, the CISAC CMO's Interested Party System (IPI) was – and indeed remains – proprietary to that sector (creators, primarily in music and audio-visual). ISNI was never intended precisely to mimic the IPI, inasmuch as it was focused entirely on public personas; the IPI distinguishes clearly between parties and the names that they use. However, while different IPI name identifiers are given to each different expression of a name, the ISNI clusters different versions of names (or even in some cases different names)[46] for the same person.

However, ISNI itself maintains no metadata about the underlying party, seeking instead to act as something akin to a predicate identifier, a link between systems.[47]

ISNI started identifying personas using algorithmic techniques for matching databases, beginning primarily with resources in the library sector. By matching data in (for example) library name authorities and the databases of author-facing CMOs, combined with limited information about their publications, a reasonably clear picture could be derived of the identities of authors. This was imperfect – even people with relatively unusual names were sometimes given two ISNIs – but it is always easier to reduce two to one than to

43 See Section 2.9 for discussion of the difference between FRBR and indecs modelling of works and expressions

44 "Accessible Registries of Rights Information and Orphan Works towards Europeana" **https://pro.europeana.eu/project/arrow**

45 Global Library Manifestation Identifier. "Items in a GLIMIR cluster may represent manifestations of a work in different languages of cataloguing or in different physical formats, such as original print text and microform versions or ebooks with print versions that are not new editions." **https://www.oclc.org/content/dam/research/presentations/Gatenby/GLIMIR_thepotentialimpact.pdf**

46 The IPI system is specifically designed to preserve pseudonymity for song writers, some of whom may for commercial or legal reasons publish under a considerable number of different names. The ISNI specifically links pseudonyms when knowledge of these is in the public domain. See for example the ISNI for Ruth Rendell which, as well as listing a number of name variants, specifies a relationship with Barbara Vine (a pseudonym).

47 Seeing metadata as a triple <subject><predicate><object>, ISNI effectively defines its role as being the predicate. In this case, of course, the triple can be read in either direction. <name123 in database A><is used by the same public persona as><name456 in database B>. The same predicate can be used to link other databases. In theory, it would have no declared metadata other than the identities in the databases that it links. The practice looks a little different.

separate two different people who have been given a single ISNI. (This is true of any referents in an identification scheme – it is easier to lump than to split after the event.)

ISNI now has over 30 registration agencies worldwide. Although (unlike, for example, the ISBN) ISNI offers no territorial or specialism monopolies, these agencies tend to focus either on specific territories or on specific media sectors (or both).[48] Registration agencies were originally drawn from the library sector, but the balance is now close to parity with other sectors.

ISNI can be used not only to register people but also organisations – one of the registration agencies[49] maintains ISNIs for worldwide academic institutions.

There are now over 12 million ISNIs in issue, around 11 million to individuals and 1 million to organisations (broadly defined). One reason for its success is rigorous quality control, something which appears to be missing from its main competitor in the publishing space, the ORCID (see Section 4.3.1.2)

The substantive development with ISNI has been its adoption in relatively recent times by online music services and other organisations in the music space. Here, it is being used as something much closer to a party identifier.[50] The most obvious evidence for this is that individuals can register themselves and be given an ISNI which they can use in other contexts.

Having observed the success of the implementation of ISNIs in other sectors, book publishers are also becoming more active,[51] and it is our view that use of the ISNI should become commonplace in that sector over coming years. We can reasonably hope that as a result it will also begin to be implemented in the RRO sector.

ISNI is a success story, although it has taken the better part of a decade to come into its own (which may seem slow to some outside observers). But notably it has achieved that success without the intervention of any dominant player forcing implementation – instead fulfilling a need on the back of real-world applications and the dedicated hard work of a small number of organisations and of ISNI-IA itself. The ISNI-IA can be regarded as a model of a strong registration authority with board members drawn from many different stakeholder communities (and with a sound commercial outlook). IFRRO has participated in its governance from the outset.

Like every other standard, it has not survived entirely unscathed from its contact with the real world, and our view is that the time for a formal review of the standard may now be near, to ensure that its specification remains fully in line with the reality of its current and future applications. This should not be seen a criticism nor a suggestion that it should change anything that it is doing.

However, with the introduction of self-registration, we do see ISNI changing from its originally defined role as a simple mechanism for linking public identities in different databases. In these new applications, at least, it has more of the appearance of a party identifier. This has implications.

48 For the current list of registration agencies see **https://isni.org/page/isni-registration-agencies/**

49 Ringgold **https://www.ringgold.com/isni/**

50 See Sound Credit **https://www.sound.credit/about** for example

51 There is for example an active group of publishers in the UK which has been undertaking some initial data matching activities with library data. MVB in Germany, which is that country's provider of ISBN, books in print and other metadata services, has been an ISNI Registration Agency since early 2020, although we are unable to track any reference to this on their website so for the time being this appears to be a "submarine" activity.

3.4. OPEN DIGITAL RIGHTS LANGUAGE (ODRL)

"The Open Digital Rights Language (ODRL) is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. The ODRL Information Model describes the underlying concepts, entities, and relationships that form the foundational basis for the semantics of the ODRL policies. Policies are used to represent permitted and prohibited actions over a certain asset, as well as the obligations required to be met by stakeholders. In addition, policies may be limited by constraints (eg temporal or spatial constraints) and duties (eg payments) may be imposed on permissions."[52]

ODRL started its development as a "standard" around the year 2000, outside any formal standardisation structure. Its original and continuing purpose was to be an open standard for the communication of rights and licensing information, in sharp contrast with proprietary "rights expression" languages (RELs) like ContentGuard's XrML.[53]

ODRL gained significant traction when it was adopted by the Open Mobile Alliance (OMA), to manage access to content on mobile networks. It was very widely implemented on mobile devices. However, it made little further progress until it was taken to the World Wide Web Consortium (W3C) in 2011 as a "Community Group" activity – the entry level towards standardisation. It achieved the status of a W3C Recommendation (standard) in 2018, in version 2.2. The Working Group[54] established to steer it through standardisation was called the *W3C Permissions and Obligations WG* – perhaps a signal of the intense distrust of the concept of intellectual property rights within significant parts of the internet technology community. Nevertheless, ODRL is now a W3C Recommendation and (in line with normal W3C processes) the Permissions and Obligations Working Group has been closed, with continuing work being undertaken by Community Groups.

52 **https://www.w3.org/TR/odrl-model/**

53 XrML struggled to find traction until it was incorporated into the ISO/IEC MPEG-21 standard (at which point, ContentGuard ceased to develop it). See **https://mpeg.chiariglione.org/standards/mpeg-21** Some work has continued within MPEG-21 as recently as 2019 on the development of Rights Management Ontologies. See **https://mpeg.chiariglione.org/sites/default/files/files/standards/docs/w18500_MPEG IPR Ontologies.docx** However, we are not aware any substantive implementations in the publishing sector.

54 See Section 2.5.2 for description of the W3C standardisation process.

There have been some well-informed criticisms of ODRL in terms of its fundamental data model, but in a very real sense these have to be seen as irrelevant – in terms of standards for expressing rights and permissions in machine readable form for use on the World Wide Web, ODRL is "the only game in town". There may be arguments for re-opening the standardisation process at some point in the future, but until the limits of the existing data model are demonstrated to be a challenge in implementation of real use cases, the sensible approach to ODRL is to make best use of it.

3.4.1. RightsML

This is the approach taken by IPTC in the development of RightsML, an IPTC standard developed to communicate rights relating to news content on the Web.[55] In the context of ODRL, RightsML is a "profile", which is the expected direction for any implementation of ODRL. A profile builds on the ODRL model and its core semantics but includes specialist semantics for the management of the particular set of rights and transactions.

There have been some implementations of RightsML, but none of these are generic expressions of rights and permissions on the "open web", intended to be publicly interpreted. Rather, they are constrained to specific business relationships. It is understood, for example, that one publisher of very high value financial data uses RightsML for the expression of usage rights to their customers. We also understand that at least one implementation may have been established in the audiovisual broadcast sector. But both of these are within pre-identified supply chains. While RightsML is a profile of ODRL, individual applications might themselves be seen as representing specific profiles of RightsML.

3.4.2. Text and Data Mining

A new ODRL Community Group[56] has very recently been established by a group of European publishers seeking to devise a mechanism to opt-out from the exception for commercial text and data mining, in compliance with Article 4 of the European Copyright Directive. Because of the way that the legislation is drawn, this project has to complete its work in a very short time frame – with completion by June 2021. Conceptually, the ODRL profile is intended to make it clear that rights for commercial TDM are reserved, and to point in the direction of a licensing mechanism. Why this project is not being undertaken within the framework of RightsML, we are far from clear.

55 RightsML was originally a development of work initially undertaken in the ACAP (Automated Content Access Protocol) project. **https://en.wikipedia.org/wiki/Automated_Content_Access_Protocol**. This project, which ultimately became more political than technical, was closed in 2011 and its intellectual property passed to IPTC.

56 More **https://www.w3.org/community/tdmrep/**

While we might be reasonably confident that the necessary technical work can be completed within this time frame, we have some concerns about the socialisation of the project (which seems to have some things in common with the ACAP[57] project, although it seems to us that its focus may not be on content published on the "open web").

3.5. THE "INTERNATIONAL STANDARD CONTENT CODE" (ISCC)

We need first of all to get some things out of the way.

Most importantly, the ISCC is emphatically not an international standard. Secondly, it is not (by any definition that we would understand) an identifier. This might seem to be damning the entire enterprise from the outset. But in reality, if it works as advertised (and that's still an "if"), it could form a very useful tool as part of a broader digital rights management infrastructure.

Essentially, it creates a code derived from a set of intrinsic properties of a digital file. Rather than attempting to summarise the technology being used to do this, we draw the reader's attention to its published specification.[58] In essence, the code is a concatenation of four "hash"[59] codes derived in different ways from a digital file instance. If two digital file instances are identical, they will return identical ISCCs when "hashed" in accordance with the specification. More controversially, if two digital files include "the same content" but it has in some way been altered, the proponents of the ISCC claim they can make reliable estimations of the likelihood that the content is indeed "the same thing".

Some of the technologies that would make this possible are familiar enough; image recognition technology is known to work with (for example) massively cropped and otherwise reformatted images. The claims of the ISCC go beyond this; by extracting semantic data from text, and hashing that, ISCC's proponents believe they should be able to recognise (for example) translations.

While there are demonstrations of its efficacy "in the laboratory", we are not alone in believing that some of the claims made about ISCC can only be really tested in proper trials "in the wild". We understand that this had yet to be seen to be done at the time of writing.

57 Op. cit. It is clearly the case that all standards have dimensions well beyond the technical, but those standards that seek to work at the interface between the news media and "big tech" have the problems much more obviously than others. Any move towards their implementation risks becoming a tussle between giant vested interests, even when the legal framework is clear.

58 Available here **https://iscc.codes/**

59 Hashing is the algorithmic transformation of a string into a shorter fixed-length value or key that represents the original string. ISCC uses different hashing algorithms for different parts of the process and for different media types.

An initial working group was established by ISO TC46/SC9 which we have been told was established to define the *requirements* for standardising a technology that does broadly what the ISCC claims – uncovering exactly and approximately "the same" content in different file instances. This is clearly not an identifier, since it fails to answer the question "the same as what?" It would not even be a very helpful identifier, since it would identify two files with identical content but some other differences with non-identical ISCCs – a potentially useful service but not an identifier. It is a service that might have a particular application in deduplicating registries.

Our view (one that we know is shared by others) is that ISCC may not yet be ready for standardisation but would benefit from a period of real-life testing and deployment. It may be more appropriate for any standard of this type ultimately to be more like the ISO DOI standard, which is the standardisation of an abstract method rather than of a specific technical implementation.[60]

We will return to the potential role of the ISCC in a comprehensive rights management infrastructure in Section 5.4.3.

3.6. ONIX

3.6.1. Origins

We will finish this section with a brief account of ONIX, a set of initials[61] which appears regularly throughout this report. It can trace its origins back to an initiative from the Association of American Publishers (AAP) which developed a format in the late 1990s for communicating book metadata to online booksellers. What became ONIX for Books 1.0, a "flat" file format, was never implemented. But responsibility was given to EDItEUR to develop the standard for international implementation as an XML format. ONIX for Books 2.1 became ubiquitous in the first decade of the century for sharing "rich product metadata".

3.6.2. The ONIX family

EDItEUR subsequently adopted ONIX branding for a family of standards in different sectors including, for example, ONIX for RROs (see Section 5.2.2.1); there are also ONIX standards for registering standard identifiers.[62] These standards share a common approach to structure, and where appropriate share value code lists.

60 DOI was already a well-established and proven technology before it became an ISO standard.

61 Online Information eXchange

62 **https://www.editeur.org/118/ONIX-ISBN-Registration-format/**
https://www.editeur.org/106/ONIX-ISTC-Registration-Format/

EDItEUR also supports a range of formats for the exchange of transaction information; some (the most heavily used) are "flat" EDI formats (from which EDItEUR got its name. Others are XML formats (branded as EDItX). These are outside the scope of this report.[63]

3.6.3. ONIX for Books 3.0

ONIX for Books 3.0, published in 2009/10, was a significant step forward, designed to improve the capability of publishers to communicate information about ebooks, and also simplifying the process of updating information. The challenge with ONIX 3.0 is that adoption in two key markets – the US and the UK – has been frustratingly slow. However, the recent insistence by Amazon that it will only accept ONIX 3.0 appears to have had the impact that might be expected, proving again the influence that market dominance can have on standards adoption.

EDItEUR's suite of ONIX standards demonstrates one of the challenges of deployment of standards, even where there are many stakeholders involved and consensus is reached on the requirement and the technical specification. If deployment is desirable rather than essential, it can be difficult to persuade a sufficient proportion of the market to adopt a new solution when the status quo is "good enough for the moment" – and the value of communication standards for all participants is driven by network effects. That "moment" can last 10 years or more.

CASE STUDY – SLOW TAKE UP OF ONIX 3.0

ONIX 3.0 was approved in 2009 and ONIX 2.1 (its predecessor) was "sunsetted" at the end of 2014. More than 6 years have passed, and yet many major publishers in major markets have not updated their practices, preferring to continue to follow an outdated, less flexible and no-longer-maintained specification (to the enormous frustration of those responsible for maintaining the standard). Some whole countries successfully adopted a "big bang" approach, but many publishers in the US and the UK lagged behind, only moving to catch up when more-or-less compelled to do so by a dominant supply chain partner.

It is not that these publishers were not represented in the development of the new standard. It is not that industry software suppliers were not fully capable of supplying systems that supported it.

We cannot answer this question with absolute certainty, and indeed there are surely myriad answers for each publisher. But the lack of any leadership in the market was probably a major contributor, leading to an absence of "network effects" that are critical to standards adoption.

63 For EDItEUR standards in journals, see Section 4.3.1.

4 Reflections on some specific sectors

In this section of our report, rather than looking at the issue from the viewpoint of particular standards, we look at sectors of publishing and/or of publishing activity. Because of its significance to the commissioners of this report, we have given rights management a section of its own (see Section 5).

4.1. STANDARDS AND PROVISION FOR PEOPLE WITH VISUAL IMPAIRMENT

This section of our report will focus on books, since books are of key interest to the commissioning parties. It will also focus on identifiers and metadata, although for completeness it may be helpful to begin by filling in some background on standards for content formatting, if only in outline.

Until the advent of the ebook, the creation of specialist formats of published works for people with visual impairments was (as to a significant extent they still are) the exclusive province of specialist libraries and support organisations. There is some tendency when talking about people with visual impairment to imagine that we are talking about Braille editions. While of course these are important to those who use them, in truth readers of Braille are a small minority of people with visual impairment, typically those who have been blind from birth. Since the 1920s, much greater use has been made of "talking books" and "talking newspapers"- audio recordings made by volunteers. Originally distributed as analogue recordings on disc, and subsequently on audiotape, the turn of

the current century saw the development of a standard digital format for CD ("DAISY")[64] with specialist playback equipment providing navigation tools designed specifically to help people with visual impairment.

DAISY 3, the current version of the standard, was introduced in 2005, although DAISY 2.02 is the version mainly used by libraries.[65] Here is an abridged description of the capabilities of the format drawn from the DAISY website:

> *"Digital Talking Books (DTB) go far beyond the limits imposed on analog audio books because they can include not just the audio rendition of the work, but also the full textual content and images. Because the textual content file is synchronized with the audio file, a DTB offers multiple sensory inputs to readers, a great benefit to, for example, learning-disabled readers. Some visually impaired readers may choose to listen to most of the book but find that inspecting the images provides information not available in the narrative flow. Others may opt to skip the audio presentation altogether and instead view the text file via screen-enlarging software. Braille readers may prefer to read some or all of the document via a refreshable Braille display device connected to their DTB player and accessing the textual content file. DTBs containing a textual content file but no audio material might be accessed via synthetic speech, screen-enlarging software, or a Braille device."*

We include this description because it clearly illustrates the extent to which "accessibility" in the digital context is a three-dimensional challenge, involving the capabilities of an underlying data file, the capabilities of the user's interface technology, and the precise accessibility requirements of the individual user as well as the use of accessibility features by producers and those in the supply chain.

From early in the development of mainstream ebooks it has been recognised that, in some circumstances, they can substitute for specialist editions – sometimes at marginal additional cost. The most obvious example is "large print" – all ebook readers of which we are aware allow the user to select type size as a universal feature. They may also provide the capability for the user to choose the font, and sometimes foreground and background colours; these can be an aid to people with dyslexia.

64 Digital Accessible Information System. The DAISY Consortium is a standards organisation formed by the global network of talking book libraries. Membership continues to be primarily drawn from this constituency, but there are also "inclusive publishing partners" drawn from both technology and publishing industries: more information **https://daisy.org/about-us/membership/**. In parallel with managing a suite of standards, and actively contributing to the development of others, DAISY also offers a number of software tools for the creation and management of accessible content, as well as advocating and training actively on behalf of its community.

65 DAISY3 is a formal standard governed within the US standards infrastructure: ANSI/NISO Z39.86-2005

4.1.1. The accessibility challenge for the publishing industry

With increasingly widespread adoption of the EPUB3 format, now in v3.2,[66] the foundation is being laid for a complete change in the landscape of accessibility. Although an ebook in EPUB3 is not necessarily accessible, EPUB has supplementary guidelines which define the specifications that must be followed if an EPUB3 file is to be defined as "accessible". DAISY provides tools which can check whether a file is compliant with these specifications.

As legislation (of which the European Accessibility Act is the most immediate example) takes full effect over the coming few years, all ebooks (which will approximate to "ebook versions of almost all published books") ought to be "born accessible" – the same underlying file ought to be equally capable of supporting both fully sighted readers and those with reading impairment.[67] The latter group may need specialist interface technology (for example, Jaws or NVDA software installed on their computers or refreshable Braille displays) but the objective can be achieved of providing simultaneous access at the same price (and without the need for costly intervention by intermediary specialist libraries or others).

In terms of identification and metadata standards:

- Accessible ebooks can be identified (like any other ebook) with an ISBN
- Their accessibility status can be defined in ONIX[68]

ONIX recognises only a very limited set of accessible format values covering compliance with EPUB accessibility specifications and with PDF/A (the accessibility specification for the PDF file format). Publishers do not deliver ebooks in DAISY format. Should there be a requirement for additional formats (as is bound to be the case over time) there will be no difficulty with adding new values to ONIX as they are required.

However, this is not to suggest that there is not a long way to go for publishers to achieve what is (quite properly) expected of them. There are myriad technical and procedural challenges, but we would be reasonably confident that (like most essentially technical challenges) these are soluble.

Of greater concern, though, is the continuing knowledge and expertise gap between publishers. Some publishers – including many of the largest multinationals – have

66 More information here: **https://www.w3.org/publishing/epub3/epub-spec.html**. It is worth noting that EPUB3 is not a formal W3C Recommendation (standard) and is "not on the W3C standardization track". However, as a "Community Group Report" it appears to be able to embody the authority of a formal standard from the viewpoint of the user community – publishers and device manufacturers. The EPUB3 specification is itself entirely dependent on a number of formal W3C Recommendations in the XML family; this was one of the reasons why it was decided that W3C was the most appropriate home for EPUB, an extremely contentious and contested decision of the members of the (previously independent) International Digital Publishing Forum (IDPF).

67 It has often been argued – in our view entirely legitimately – that accessible ebooks provide advantages to all readers of ebooks, not only those with visual impairment.

68 More information **https://ns.editeur.org/onix/en?utf8=%E2%9C%93&search=accessibility&commit=Search**; it is an unfortunate reality that the accessibility fields in ONIX are not often populated.

developed all the necessary capabilities, but in our research we heard continuing reports of a lack of training, in Europe and more widely. This includes not only lack of capability in developing the workflows necessary for the creation of accessible content but also in how to describe accessibility features in ONIX (metadata management is often well separated from production workflows in a publishing house).

There are also problems with how that metadata is "surfaced" in retail, although these may be easier to solve when the metadata is more consistently created and delivered (and as accessible ebooks become the norm rather than the exception).

4.1.2. The metadata challenge facing the Accessible Books Consortium (ABC)

ABC is a global initiative of the World Intellectual Property Organisation (WIPO) to improve access to books for people with print impairment, in support of the implementation of the Marrakesh Treaty.

One of the challenges set to ABC is to develop an information resource which presents, in a common interface, records of books which have been converted into accessible formats and which are available from specialist libraries for cross-border sharing under the terms of the Treaty. This was always going to be challenging to achieve. It is made more difficult by the intention that it should be consumer facing.

There are now over 670,000 titles in the ABC catalogue[69] – a catalogue made up of records submitted by over 100 contributing libraries from around the world. The main standard used by libraries for cataloguing is the MARC standard. There are several reasons why cataloguing is challenging for a global platform such as the ABC Global Book Service:

- While MARC is implemented by most libraries in the world, it is not always implemented in the same version;
- The MARC accessibility fields were only recently added in 2018 (532 and 341) and there is no controlled vocabulary "or standardised description" for the accessibility status of a resource; Cataloguing in MARC is a skilled task and records from different libraries are not likely to be of uniform quality;
- Many of the smaller organizations serving people who are print disabled, particularly in developing or least developed countries, do not use MARC;
- There are issues of translation.

69 We understand that ABC, as well as holding metadata, is also becoming established as a repository of copies of accessible files themselves. We are not sure that this changes anything about the identification and metadata issues that are in scope for this report, except that with access to the files themselves, it may be possible to improve the metadata through the equivalent of "book in hand" cataloguing.

The implication is that it is not always possible to tell from an ABC catalogue record the format of an accessible version because data is missing or held in a different field in the record or is undiscoverable because of the use of free text. Beyond that, MARC records are specifically designed for use by librarians and the lack of a controlled vocabulary for accessible formats does not make it easy for users who are print disabled to identify accessible format copies of works.

Identifiers, to the extent that they exist within the database, are likely to be either the ISBN of the book from which the derivative accessible edition has been created (a related manifestation for sure, but definitely not "the same thing" by any reasoned definition) or perhaps a locally defined accession number (or perhaps an OCLC WorldCat number).

4.1.3. Resolving the MARC challenge

A project is under way to resolve the challenge of including accessibility data in MARC21[70] records in a standardised way. This work is being given impetus by the Association of Research Libraries and the Canadian Association of Research Libraries (ARL/CARL), and critically involves both the Library of Congress (as members of ARL), specialist libraries for the support of people with print impairment, and OCLC (the extremely influential global library cooperative).[71] This should bring together a well-resourced and skilled group of librarians and technologists, motivated and with the necessary tools to find a solution to the challenge. We were unable to get a sense of how long this work (which will also be co-ordinated with IFLA) will take.

It might be possible, particularly if the libraries that contribute records to ABC also contribute their records to WorldCat, that OCLC might be able to assist ABC with enriching and normalising its MARC records. However, if data is completely missing (for example, there is inadequate information about the format of the accessible file), this can only be resolved through discussion between ABC and the library that submitted the original record.

4.1.4. The intersection between the trade metadata supply chain and ABC[72]

Although there is a mapping convention between ONIX and MARC, it is hard to imagine a process by which metadata from publishers – which in ONIX format might ease the problem of providing a consumer-oriented interface for ABC – could pass into the ABC catalogue. Apart from any commercial considerations, unless there was consistent

70 As has been discussed elsewhere, the library community acknowledges the weakness of MARC for the long-term. However, in the short and even the medium terms, there is no realistic expectation that the majority of libraries will be updating their systems.

71 *"OCLC is a global library cooperative that provides shared technology services, original research and community programs for its membership and the library community at large. We are librarians, technologists, researchers, pioneers, leaders and learners. With thousands of library members in more than 100 countries, we come together as OCLC to make information more accessible and more useful."* OCLC has been involved in a number of the projects discussed in this document. It provides the technical underpinnings of the ISNI.

72 In this context, it may be helpful to have some familiarity with the draft NISO Best Practice *"E-book Bibliographic Metadata Requirements in the Sale, Publication, Discovery, Delivery, and Preservation Supply Chain"*. This paper seeks to address the "MARC is from Mars, ONIX is from Venus" problem.

availability of the ISBNs of the works that have been converted and then had MARC records submitted to ABC, there would be no mechanism for reliably associating any incoming metadata with the records in the database. Furthermore, such data would need to be collected from all the different countries where the records themselves are sourced.

And, of course, none of this includes a controlled vocabulary or "standardised description" for all accessible formats.

There is another problem with the lack of a satisfactory intersection between the worlds of commercial publishing and specialist libraries for people with print impairment that should not go unrecorded. One option in the Marrakesh Treaty that has been implemented in some jurisdictions is exclusion from the exception titles where an accessible ebook version is already commercially available. It is exceedingly difficult to find out whether or not an accessible version is available anywhere in the world – there is no global books in print.

This is similar to one of the problems presented by out-of-commerce and orphan works (see Sections 5.5 and 5.6).

4.2. SUBJECT CLASSIFICATION

Issues of subject classification were specifically raised by the commissioning parties. Here we provide a brief overview of classification standards.

There is an enormous range of standards for subject classification of books[73] and other textual content, ranging from the library Dewey Decimal Classification system (DDC, dating back to 1876) and Library of Congress Subject Headings (LCSH, 1898)[74] to Thema, which was launched as a project in 2012.[75]

ONIX Code List 27, which lists the range of classification systems that can be communicated in ONIX, has over 100 entries,[76] including DDC, LCSH and Thema, but also library schemes in other languages (DDC and LCSH are essentially English language, although there is an international version of DDC, UDC) and, for example, schemes used in individual countries for educational books.[77] Individual bookstore chains and other participants in the supply chain may have their own proprietary classification schemes, and there is a common requirement for the use of automated or semi-automated mapping between schemes. Such mappings are almost always "lossy" – variable amounts of precision will be lost in translation between one scheme and another. Mappings tend only to be really effective in one direction (from a richer to a less rich scheme).

Publishers will often be expected by their customers to provide more than one subject classification for the same product.[78] As in other aspects of metadata, there is a fundamental difference of purpose between subject classifications used by libraries and those used in the publishing supply chain. This is the difference between approaches used primarily for classification and shelving and library patrons, and those used for selling books to consumers.

Schemes may be simple hierarchies, or allow more complex "faceted" classifications, allowing combinations of values drawn from different hierarchies to express more subtle subject distinctions. A detailed discussion of the differences between schemes is beyond the scope of this report. However, since BISAC and Thema schemes are specifically noted, it is worth a short exposition on the dominant schemes in use in the commercial book supply chain.[79]

4.2.1. Dominant book supply-chain schemes

There are two long-established English-language trade standard schemes that remain in widespread use. The BISAC Subject Category Scheme is licensable from BISG[80] in the USA and continues in more-or-less universal use in the North American trade. In the UK, and English-language markets outside North America, the BIC[81] Subject Category Scheme was dominant; it was also translated into a number of other languages.

Thema is an international scheme, which was built directly on top of the BIC scheme, but is different from it. It was designed from the outset to be implemented in different languages, so that the same codes could cross territorial and linguistic borders seamlessly. In its current version (1.4 released in 2020) it is available in over 20 languages (including for example Arabic and Russian).

Thema has gained substantial traction worldwide, but somewhat ironically has not yet been widely adopted in the USA, English-speaking Canada or the UK, where BISAC and BIC schemes respectively remain predominant. The BISAC scheme is still actively managed but the BIC scheme has not been maintained since 2010, and BIC itself is clear

73 For a subject scheme used in the news media, see Section 4.5.

74 LCSH is probably less well known than DDC to those outside the library community but is probably more widely implemented globally. It is substantially more detailed than DDC. For an introduction to LCSH, now in its 42nd iteration, see **https://www.loc.gov/aba/publications/FreeLCSH/LCSH42%20Main%20intro.pdf**

75 For more background on Thema, see **https://bisg.org/news/459676/Thema-Developing-A-Subject-Category-Scheme-for-A-Global-Book-Trade.htm**

76 For a complete list, see **https://ns.editeur.org/onix36/en/27**

77 The best exposition we know of the impossibility of developing a universally applicable taxonomy of knowledge is found in an essay by Jorge Luis Borges called in English **The Analytical Language of John Wilkins**; it includes, in a section on a (presumed fictional) Chinese encyclopaedia called "The Celestial Emporium of Benevolent Knowledge" that rarest thing – a metadata joke. We will always be grateful to the late Norman Paskin (the first Managing Agent of the International DOI Foundation) for having brought this to our attention. It should be known and understood by anyone who has to think about metadata semantics, and particularly about subject classification. An English translation can be found **https://www.entish.org/essays/Wilkins.html**

78 A single ONIX message can carry multiple classifications in different schemes if these are available.

79 Library schemes are designed to be applied to any resource; book trade schemes are limited to books (including ebooks and audiobooks) and are not used (for example) for classifying academic journals.

80 The Book Industry Study Group is a cross-industry book supply chain organisation based in New York. It has a number of functions on behalf of its membership, including management of some domestic standards and participation in international trade standards development, including through EDItEUR.

81 Book Industry Communications is the UK's closest equivalent to BISG, although the organisations have somewhat different remits and structures.

that the BIC scheme is now deprecated. Even after 10 years, legacy systems as ever form a significant part of the challenge, although the US book trade has always been slow to adopt new or revised standards, even where these are clearly an improvement on what has gone before.[82]

Historically, mapping tools for converting records between different schemes (BIC to BISAC, for example) have been proprietary and closely guarded for commercial reasons. However, EDItEUR (the organisation that manages Thema on behalf of the international book trade) offers a number of these tools for download from its website. In common with other EDItEUR standards, these are available for use under a permissive open licence[83] without registration.[84]

Thema is a good example of "the right standard arriving at the right time". While its adoption in the USA as the domestic standard may be a long way away, it seems increasingly likely to be implemented in a growing part of the international book marketplace (including, for example, in English-speaking Canada – it is already in use in French-speaking Canada). It underpins the subject classification tree used by Amazon; the long-term influence that Amazon will have should not be underestimated.

Anyone looking to the future who requires a **consumer-facing** subject classification scheme for books, ebooks and/or audiobooks would need an extremely good reason for adoption of any single scheme other than Thema (unless trading exclusively in North America).

4.2.2. Thema or Dewey Decimal Classification (DDC)?

This is an entirely false dichotomy; the schemes fulfil very different purposes and have very different characteristics. Despite (indeed, because of) its deep history, DDC is extremely widely used and has succeeded in broadly maintaining its cultural currency despite the real challenges brought about by attitudinal change in over a century. Thema is still relatively early in its life cycle but is showing every sign of long-term success.

This is not simply a question of "library or commercial application?" There are library applications in which Thema's comprehensive classification of fiction would show some advantages to library users over library schemes (perhaps in local public libraries, for example). The same might be true of a project like ABC, which is intended to provide a consumer-facing service (although adding Thema classifications to a database of over 600,000 titles is impractical). On the other hand, DDC (and LCSH) have a more fine-grained and extensive classification of non-fiction and academic books, which proves desirable in some commercial applications.

82 It is worth noting that subject classification also may have strong cultural biases which militate against international adoption; nevertheless, Thema appears to have walked that particular tightrope without too much difficulty, although it is possible that some "partial" translations reflect precisely these cultural sensitivities.

83 See **https://www.editeur.org/files/about/EDItEUR%20IPR%20licence%2020200317.pdf**

84 While there is much to be said in favour of this approach, it does have one significant downside – EDItEUR does not have a full picture of which organisations are actually using its standards.

To choose whether to use Thema or a library scheme, it is first essential fully to define requirements, and then to test which classification scheme best fits those requirements. In the real world, the answer may not be "either/or" but "both/and".

4.3. ACADEMIC BOOKS AND JOURNALS

We are taking a separate look at books and journals in the academic sector, as this sector has always been very different from the trade sector. And journals and books are also very different from each other. The sector has undergone a period of enormous change as it has moved online to a significant extent; and now, open access is taking academic book and journal publishing in dramatically new directions.

A comprehensive consideration of this sector and its new standards infrastructure is beyond the scope of this report, but this could easily be the subject of a separate report, given the number of new initiatives and new formal and informal standards being developed in the sector. It is notable that no specific issues were raised in the ToRs and few issues were raised while we were doing our research. Here, we will simply provide an overview, primarily to cover topics that may have broader implications.

4.3.1. Academic Journals

The primary identifier in this space has historically been the ISSN.[85] The ISSN is unusual in having been established under the terms of a UNESCO treaty, as well as being an ISO standard. Like ISBN, the ISSN is managed by an international network of agencies ("centres") but, unlike ISBN, these are all in the not-for-profit and library sectors. An ISSN can be (and often is) assigned to a publication by the relevant national centre rather than at the request of the publisher.[86] There is no charge for issuing an ISSN.

As an identifier, it was originally very straightforward in that it identified a periodical name; a new ISSN was issued when the periodical changed its name. It could perhaps have best been defined as the identifier of a periodical brand.[87] This became more complex when the ISSN standard was developed so that the digital version of a periodical has a different ISSN from the print version. Where two separate ISSNs are issued to a periodical for different publication media, one of these (normally the original print ISSN) is designated the "linking ISSN", collocating the different versions. This is another apparent "predicate" identifier – it is highly suggestive that its only metadata are the ISSNs it links.[88]

85 International Standard Serials Number ISO 3297 **https://www.issn.org/**

86 In magazines, it was certainly not unusual for publishers to be unaware of the ISSN for their publications and not to use them; it is unlikely this was ever the case for journal publishers.

87 We should stress that this is our personal view. It was always disputed by ISSN-IA who said that the ISSN identified "a periodical" without defining what level of abstraction that represented.

88 We believe this rather "clunky" mechanism may have come about because of legacy technology constraints.

In the print era of journal publishing, the ISSN coupled with a publication year was used as the "product identifier" in the supply chain (which typically flowed from the publisher via subscription agent to library). Supporting that supply chain was a standard for the communication of subscription data between subscription agents and publishers, initially on magnetic tape carried by motorcycle couriers, and subsequently delivered by FTP online (in an unchanged "fixed format").[89]

CASE STUDY – DOI STANDARDISATION

The Digital Object Identifier is an ISO standard, ISO 26324. But among ISO identifiers it is (so far as we know) unique, in that what was standardised was not a specific implementation but a method. DOI is a technology coupled with an identification and metadata model and governance process.

The application of DOIs is managed by registration agencies, like CrossRef. But the International DOI Foundation manages the core technology and over-arching governance.

DOI was already "up and running" before it sought standardisation by ISO, which might well be a model for other technologies like ISCC.

However, DOI also has another story to tell – the challenge of initial governance. Both DOI and its first major implementation were financed by the major international STM publishers, who made very substantial loans to finance the start-up phase of the DOI. It took a number of years for these loans to be repaid, and understandably during that period the founders kept a tight hold on governance.

As a result, other industries were highly reluctant to put their trust in "a bunch of scientific publishers". It took a very long time for a second major implementation from outside the sector – EIDR, which developed with extraordinary vision by MovieLabs for the major US film studios (not least because of dissatisfaction with an ISO standard, the ISAN).

If it were not for the overwhelming success of CrossRef, DOI could have become another interesting footnote in the history of identification standards. With the addition of EIDR, it now seems secure.

89 These standards were managed by a committee called ICEDIS (International Committee for Electronic Data Interchange for Serials), which ultimately came to be managed by EDItEUR. EDItEUR also developed a set of XML messages under the general "ONIX for Serials" brand. These were never heavily used.

4.3.1.1. From print to online

The moves from print to online, from individual title subscription to site licensing and "big deals" with library consortia, have completely changed the journals landscape. Today, there are almost no subscription agents left in business, those remaining undertaking a rather different role from their historic ones. Communication about commercial issues apparently no longer needs the support of machine-to-machine messaging standards. However, there is still information that publishers need to share with libraries; where this used to be managed by the subscription agents, other mechanisms have had to come into place to cover the gap.

"Best practice recommendations" from NISO[90] (which oversees a number of strands of work designed to improve library—publisher communication) have come into place to solve some of the challenges – an important example being KBART, a mechanism for the supply of information from publishers to libraries about their "holdings" – the content that any particular library or group of libraries is making available to its patrons, covering both subscription and open access content.[91]

However, there has been something of a proliferation of organisations in this sector, each managing an individual standard or protocol.[92] Sometimes it is possible to find out quite easily who is behind these initiatives; others may be a little opaque.

4.3.1.2. Party Identification

Perhaps the most significant issue in terms of (an apparently entirely unnecessary) splintering of standards may be the ORCID.[93] This is essentially the ISNI for researchers,[94] and their use by authors has been made mandatory by some publishers. The good news is that ORCID and ISNI agreed from the outset that ISNI would set aside a specific subset of a common numerical format for ORCID to use – making it theoretically possible to merge the two at some point in the future. However, it is worth noting that significant quality issues have arisen for ORCID, largely because of its unmanaged self-registration processes.[95]

As well as ORCID, yet another organisation has come into being to manage organisational identities. ROR[96] was established by a group led by the California Digital Library. The structure of the ROR is entirely different from ORCID or ISNI.[97]

90 See list of recommendations, white papers and other publications here **https://www.niso.org/publications**. Although a lot of major publishers are members of NISO, it tends to be perceived as a library organisation. There continues to be considerable tension and distrust between academic publishers and academic libraries.

91 KBART (Knowledge Bases and Related Tools **https://www.niso.org/standards-committees/kbart**) is, like ICEDIS was, a "flat" fixed-format message. A proposal to develop "ONIX for KBART" which would have developed XML formats for KBART never got off the ground, presumably because libraries would not be able to process them.

92 See, for example, **https://www.getfulltextresearch.com/why-use-getftr/, https://casrai.org/about/**

93 **Open Research and Contributor ID**

94 Many authors have both an ISNI and an ORCID.

95 The inappropriate use of the database ORCID is apparently well known; see **https://thegeyser.substack.com/p/orcid-needs-to-clean-its-room** (paywalled)

96 Research Organisation Registry **https://ror.org/governance/**

97 ISNI has a Registration Agency, Ringgold, (op cit) which registers only academic institutions.

There may be perfectly good reasons for each individual and separate initiative; but if I were still operating as a publisher in the sector, it would worry me considerably. It is expensive to manage multiple governance structures, and there must be concerns about interoperability if each organisation pursues its own view of "the best" technical solution to its particular challenge.

It all suggests that the spirit of "not invented here", and an unwillingness to engage with a wider community (perhaps because of fear of delay), continue to be strong drivers. ORCID and ISNI, for example, continue to exist in parallel but separate universes serving two distinct communities and with no touching points between governance structures. It is clear that this is far from ideal, but it is also difficult to see how change might be brought about.

4.3.2. It is by no means all bad news

The most significant standard to have come into near-universal deployment in this sector (and also in academic books) is the CrossRef implementation of the DOI. The DOI is an ISO standard;[98] CrossRef is one of several implementations,[99] but is not itself a formal standard.

It is hard to measure the application of the CrossRef DOI, in its totality. But it is impressive: at the time of writing, 123m records in the database, the majority (nearly 90m) for individual journal articles from nearly 90,000 journals; also over 1.5m book titles and conference proceedings, and around 25m individual chapters/conference papers. And all this in only just over 20 years.[100] It has become ubiquitous in the sector.

CrossRef has been an extraordinary story of success as a result of being an outstanding example of network effects. Designed with the simple purpose of allowing readers of journal articles to "click through" direct from a reference list to the article or chapter being cited, CrossRef only really became effective as it became ubiquitous. And because it had the backing of (and had indeed been financed by) the major STM journal publishers, it had a substantial head start in creating network effects. But it has also been very responsive to its member publishers and to changing market conditions.

In creating the mechanism for linking references, CrossRef has also developed an extremely effective identifier at "article work" level – a potential substitute, in this sector at least, for an ISTC (although CrossRef have often in the past declined to claim that this is a work identifier, insisting rather that it simply identifies a citation). We will consider this further when writing about standards for rights management (see Section 5.2.1).

98 ISO26324 **https://www.iso.org/standard/43506.html**

99 Other than CrossRef, these are either relatively small or otherwise outside the scope of this report. The most significant are DataCite (**https://datacite.org/**) who provide identifiers for research data resources; and EIDR (**https://www.eidr.org/**) a highly successful application in the audiovisual sector.

100 For current statistics, see **https://www.crossref.org/06members/53status.html**

4.3.3. Some specific issues with online book provision to academic libraries

Provision of books to academic institutions are increasingly (but not exclusively) in digital form. Many of these are managed by intermediary services,[101] who provide access to packages of content from multiple publishers. This applies both to textbooks for students and to academic monographs.

Publishers are expected to provide MARC21 records for these; publishers tend to outsource the creation of MARC records to specialist metadata creation services.[102] We were told that while these may be of high quality, aggregating intermediaries may downgrade the MARC records that are delivered, for reasons that are a little hard to discern. Whatever the reason, it appears that individual libraries may then have to either improve the records, or source improved records from elsewhere. This seems peculiarly wasteful.

Another problem – which understandably seems to create some real difficulties – relates to the application of ISBNs to books in these packages.[103] Since each individual book in each package may be "a separate product" – and *prima facie* there would appear to be an arguable case for this – then it makes third-party usage analysis difficult, unless the book also has a CrossRef DOI (which many but not all do) to act as a (proxy?) work identifier.

The specific case that was brought to our attention related to open access books, where funders are keen to have usage analytics to understand whether the funding of a particular book has been justified by its subsequent use. Some consistency in the approach to identification would clearly be helpful.

4.3.4. Open Access[104]

One final point to note in this section relates to the identification of Open Access resources (both journals and books). It should be simple enough to find information at the title level on which resources are fully OA and which are not.[105] The big challenge is identifying individual OA articles in hybrid journals.[106]

101 We should add a note here about accessibility. As we have noted elsewhere, accessibility is a function of the platform as much as the underlying file; direct and indirect legislative and commercial pressure means that the various textbook platforms now compete with each other in the provision of accessibility.

102 As we have noted elsewhere, librarians are unlikely to make use of ONIX records – we were told that they see them as "commercial" and therefore inappropriate for use in libraries. See also this NISO draft Recommendation **https://www.niso.org/standards-committees/ebmd** currently waiting final publication.

103 ISBNs are still the invariable identifier of "books as products". We were told that some ISBNs were being issued by intermediaries rather than publishers; this is surprising but not entirely unbelievable.

104 We have not looked at any of the requirements for standardisation around the commercial management of Open Access. These appeared to us to fall outside our ToRs.

105 See, for example, the Directory of Open Access Journals. **https://doaj.org/about/**

106 This is significant for RROs inasmuch as they should not be licensing (or distributing revenues relating to) OA Articles; this is difficult to manage if they cannot identify OA articles.

The only source of comprehensive article-level metadata is from CrossRef; however, CrossRef says of its own metadata *"Metadata is supplied by our members and, as such, not all records have the same completeness (or quality) of metadata. Bibliographic metadata is generally required. All other metadata,* ***eg license and funding*** *information, ORCIDs, etc. is optional (though very much encouraged)."* (our emphasis). Clearly, the problem is recognised by CrossRef.

4.4. EDUCATIONAL BOOKS

For completeness, we will include a paragraph on educational (school) books. Although they follow very different supply chains in different countries, they are none of them characterised by complexity or by much requirement for identification or metadata standards - although the ISBN continues to be ubiquitous.

Even as elements of the provision of educational content move online, educational publishers in most countries are showing a strong preference for delivery using their own proprietary platforms, leaving little space for any necessary deployment of standards.

4.5. NEWS MEDIA

Implementation of standards in the news media can best be described as patchy.

Individual publications are almost certain to have ISSNs issued (see Section 4.3.1) but these are largely unused outside libraries. No standard scheme for the identification of individual articles has ever gained traction, even between the organisations that comprise the Press Database and Licensing Network (PDLN)[107], where proprietary article identification schemes must be commonplace. The ability to share content identifiers between organisations (whether other publishers or RROs and publishers) simply hasn't been necessary.[108] Newspaper workflows do not encourage anything that might delay publication processes in any way and any attempt at article identification is likely always to be a post-publication process, the preserve of aggregators (like the members of PDLN) or of particularly assiduous newspaper archivists.

107 See **https://www.pdln.info/**

108 Perhaps because rights issues in this sector are normally relatively straightforward.

CASE STUDY – ACAP: AN ARM WRESTLE BETWEEN GOOGLE AND THE NEWS MEDIA

ACAP – Automated Content Access Protocol – can trace its origins to a meeting of very senior European newspaper executives, seeking a solution to what they perceived as egregious and persistent copyright infringement. Rather than go to law, which was seen as uncertain of outcome (although a case was subsequently brought – and won – by Copiepresse in Belgium), it was decided that a technical solution should be sought. This became the ACAP project.

Google was at best ambivalent about the project. It provided some technical input but insisted on the use of an obsolescent rather clumsy technology as the "solution". When it appeared that even this inadequate technology might be able to achieve its objective, as demonstrated by a small European search engine that had become an active participant in ACAP, Google's involvement waned.

Ultimately, the project closed, bequeathing some technical deliverables to IPTC, where it became RightsML. The project achieved some political momentum and certainly demonstrated that technological solutions could be made to work. But it overall it couldn't be adjudged to have been a success.

Dominant players can make standards work, but they can also sometimes have a more doleful impact.

4.5.1. IPTC (The International Press Telecommunications Council) Standards

Conversely, in one sector – the news agencies – the need to share content in a standard format is essential. The International Press Telecommunications Council (IPTC)[109] has a mature set of content-sharing standards, under the broad NewsML-G2 brand, for the sharing of syndicated data. These are very widely used by organisations including (for example only) Agence France Presse (France), Associated Press (AP), and Reuters (UK).

As well as an XML serialisation, these standards (which include those which are designed for specialist subsets of news such as SportsML) are also available in JSON serialisation (ninjs). The formats are supported by an extensive set of controlled vocabularies,[110] including a hierarchical subject scheme called IPTC Media Topics.[111]

109 See **https://iptc.org/**

110 See **https://iptc.org/standards/newscodes/groups/**; this comprehensive set of values includes some party identifiers (for news providers) and comprehensive controlled vocabularies; the latter could be of use in the context of Orphan Works

111 See **https://cv.iptc.org/newscodes/mediatopic/**

IPTC also has a standard for sharing of video news. These standards not only define the way in which the content is formatted, they also define the metadata that is required alongside the content itself.

But in none of these standards has there proved to be a need for the widespread deployment of standard content identifiers.

Importantly, from the standpoint of this report, IPTC also has a standard for Photo Metadata (often referred to simply as "IPTC Metadata") which is a widely implemented standard for recording information about photographic images. We will deal with this in a little more detail later in this report (see Section 4.7). There is perhaps something slightly ironic about this, as in general the news media seem to persist in stripping metadata from images they publish.[112]

Finally, IPTC manages a standard called RightsML, which is a "profile" of ODRL. ODRL 2.2 is a Recommendation (the equivalent of a standard) of the W3C Permissions & Obligations Expression Working Group. ODRL and RightsML are discussed in more detail above (see Section 3.4.1).

4.5.2. The Content Authenticity Initiative

The Content Authenticity Initiative[113] is a recently formed standards organisation, initiated by Adobe, the New York Times and Twitter, but now growing into a larger coalition of news and technology businesses. They have an ambitious objective; this is an edited version of their mission statement:

> *With the increasing velocity of digital content and the democratization of powerful creation and editing techniques, robust content attribution is critical to ensure transparency, understanding, and ultimately, trust. We are witnessing extraordinary challenges to trust in media. As social platforms amplify the reach and influence of certain content via ever more complex and opaque algorithms, mis-attributed and mis-contextualized content spreads quickly…**inauthentic content** is on the rise. Currently, creators who wish to include metadata about their work (for example authorship) cannot do so in a secure, tamper-evident and standard way across platforms. Without this attribution information, publishers and consumers lack critical context for determining the authenticity of media…Ultimately, the solution to the problem of inauthentic content and the erosion of trust it causes will rely on efforts in three distinct areas. First is **detection** of deliberately deceptive media……As malicious purveyors of content become faster and better, detection techniques will struggle to keep pace. Second, **education** is essential. Well-intentioned creators and consumers will need to understand the danger of disinformation and the use of techniques to eradicate it…Finally, we must consider **content attribution**….Often referred to as provenance, attribution empowers content creators and editors, regardless of their geographic location or degree of access to technology, to disclose information about who created or changed an asset, what was changed and how it was changed. At the same time, it is critically important that those same content creators be able to protect their privacy when necessary…*

This is an extremely ambitious set of objectives and one which has metadata (and, although unspoken, identifiers) at its core. We cannot see how it would be possible to speak about authenticity and attribution without clarity of identity. It also raises questions about image metadata (see Section 4.7) – perhaps authenticity will prove to be the trigger for the wider retention of image metadata.

This initiative should also be seen alongside projects like the Journalism Trust Initiative (JTI).[114] Based on a CEN (European) standard, this initiative is far from being universally popular with media businesses, some of whom believe that projects like this may pose a threat to press freedom and would prefer their trustworthiness to be judged on their brands.

Whichever way this falls out, secure processes to prove identity – of content and of parties – will be crucial.

4.6. MAGAZINES

So far as we are aware and have been able to ascertain as part of this research, magazines make little use of any identification and metadata standards. Like newspapers, they usually have ISSNs but that appears to be the extent of their engagement in identifier and metadata standards.

4.7. IMAGES

The management of rights in images in the online environment has for decades been a source of great concern to photographers, photographic agencies and those managing primary and secondary rights in other visual works of art.

112 See **https://blog.imatag.com/state-of-image-metadata-in-news-sites-2019-update**
113 See **https://contentauthenticity.org/**
114 See **https://www.journalismtrustinitiative.org/**

Various initiatives have looked for solutions to the particular problems faced by photographers and the sharing of photos online.[115] None has been a signal success; examples include:

- The PLUS Coalition[116] was established over 15 years ago as a highly ambitious attempt to develop a standards infrastructure combining a registry (identification) with the communication of licensing information. This project never managed to get enormous traction although it remains extant.[117]

- The UK's Copyright Hub,[118] founded in 2012, while it was intended to provide copyright services across all media, it adopted a major focus on photography from the outset. It has not made the transition to long term sustainability.

We can suggest several contributory reasons for why initiatives like these have failed to gain traction in the management of rights in photographic images:

- The overwhelming majority of people whose photographs appear and are copied on the internet don't care – and would be hugely reluctant to put any effort at all into asserting their rights as they would see no value in it.

- This is not the case for professional photographers or for the agencies which manage their commercial interests, who care a great deal about infringement of rights. But infringement is only one part of the story as far the loss of revenue from photography is concerned. Although the amount of "publishing" using the World Wide Web as a platform has increased demand for photographs, the supply of photographs from amateur photographers as well as professionals has increased at a staggering rate.[119] Although the vast majority of these photos will never be seen by anyone other than the person who took them, supply and demand sometimes tell a difficult story for those who make their living from photography.

115 See for example the ARROW PLUS project documents on image identification **https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ARROW_PLUS/Deliverables/D6.2_Annex_II_images_identifiers_0.pdf** and the feasibility study image rights **https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ARROW_PLUS/Deliverables/D6.2_Feasibility_study_images_0.pdf**

116 See **https://www.useplus.com/aboutplus/system.asp**

117 To be fair, there is one August 2020-dated "newsflash" on the site so some basic maintenance was continuing 9 months ago. The newsflash relates to the Google "Licensable" badge – see more below.

118 See **https://en.wikipedia.org/wiki/Copyright_Hub**; the Hub's own web site has now been closed.

119 According to one source of data we found (**https://blog.mylio.com/how-many-photos-will-be-taken-in-2020/**), around 1.5 trillion photos will be taken in 2021 (note this was a forecast made over a year ago). We cannot verify this data, but even if wrong by an order of magnitude it points to the reality of the problem. Most of these photos will of course be taken on mobile phones.

- While the number of photos taken each year may be rising, the number of photos already "published" on the internet is enormous. The vast majority of these have no metadata associated with them, either because they never had any or because metadata has been stripped from them.

The stripping of metadata in photographs by publishers, particularly the news media, continues to be seen as a potent problem. Although the most recent survey we were able to identify is from 2019,[120] nothing we have heard suggests that anything has changed in this respect in the last two years. To some extent this is because content management systems used by the news media strip photo metadata by default; one explanation offered for this is that the metadata slows the loading of webpages on readers' devices.

The only significant change in attitudes towards rights in images that we have seen recently occurred in 2020, when Google agreed that its image search will in future display a "licensable image" badge (or shield) when displaying thumbnails of photographs with appropriate metadata – and will then make that metadata available for the user to view by clicking on the badge. While giving preference to ISO XMP metadata,[121] Google also recognises the IPTC[122] Photo-Metadata standard. This metadata standard, dating back to the 1990s but much updated, is very widely used by professional photographers and photographic agencies, is being incorporated into the desktop software packages that are predominantly used in the sector – where default values (such as the name of the photographer) can be added without any manual intervention (although some values, like titles, do have to be added manually).

Subject to compliance with the terms of the protocol,[123] the shield will appear automatically; any user can click on the shield and from the metadata presented follow a link either to a licensing service or to a Creative Commons licence. (The approach is reminiscent of the Copyright Hub photo licensing model, except that the linkage to the licensing service is slightly more arbitrary and of course at least for the time being this is limited to Google Images.) It has a slight feeling of being more symbolic than commercial in its impact, but on the other hand it is important to start somewhere.

The other activity in the image sector that is loosely related to the topic of this report involves tracking infringement through the use of image recognition technology.[124] However, this is not standardised but depends on a range of proprietary technologies.

120 See **https://blog.imatag.com/state-of-image-metadata-in-news-sites-2019-update.**

121 See **https://www.iso.org/standard/75163.html**

122 See **https://iptc.org/standards/photo-metadata/iptc-standard/**

123 For more detail, see **https://iptc.org/standards/photo-metadata/quick-guide-to-iptc-photo-metadata-and-google-images/**

124 For a fuller account of the use and potential for image recognition in rights management, please see the report from the French Ministry of Culture *Towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward*, perhaps better known as the Mouchon report. The point we have made elsewhere is that content recognition and identification are not the same thing, although they can be linked.

The AIR project established by ADAGP in France[125] is seeking to establish a global service to track infringement of fine art, potentially establishing a standardised service. We were told that the key issue is unique identification of the artist rather than of the individual works of art.

The ISNI might have a number of different applications in the image sector, but so far as we can tell it has not yet been deployed, nor does it seem to be widely known.

4.8. THE LESS-DEVELOPED WORLD

We were asked to comment specifically on the deployment of standards in the less developed world. We find ourselves with little to say.

The deployment of the ISBN is close to universal (see Section 3.1 for a limited list of exceptions). There may be some questions about the quality of ISBN registration services, and there may be some lack of understanding about the role and application of ISBN among publishers.

It is undoubtedly the case that important services that are built on ISBN in developed economies – like Books in Print and consistent metadata availability in the supply chain – are frequently entirely absent (presumably for lack of commercial opportunity). This of course has major implications for any attempt to establish RROs in these countries, because of the lack of any coherent source of repertoire data (even though WIPO Connect or IFRRO WISE might provide the necessary technical infrastructure at an affordable cost).

But our understanding – such as it is – is that there are much bigger infrastructure challenges to the more efficient distribution of books domestically, of which the most significant may be the lack of bookshops.

In terms of participation in the international book market, we might point to organisations like the African Books Collective.[126] It appears that all the titles represented by this co-operative are available on Amazon (in the UK at least) supported by apparently well-constructed and comprehensive metadata (including reviews and author biographies). We do not know whether this material is reaching Amazon in the form of ONIX messages – but there is no sign that there is a serious problem caused by lack of local availability of standards.

However, product availability on the Amazon website does not equate to sales. It is simply another book to compete with many hundreds of thousands.

125 A short video about AIR can be found here **https://www.youtube.com/watch?v=VVujmkpvYC8**
126 See **https://www.africanbookscollective.com/.**

We are also only too well aware that successful authors from less developed countries are likely to find publishers in more developed countries to sell and market their books.

It may be possible to find ways to build capacity through training and/or through the development of technical tools to deliver "Books in Print in a box", through collaboration with ISBN-IA. However, there may equally be real difficulty balancing the commercial nature of Books in Print services with the non-commercial nature of many ISBN-RAs in the less-developed world. Perhaps models of non-commercial cooperative ventures might be a way forward?

5 Rights Management

If we were to "write the book" about rights management, it would take far longer than we have available, and this chapter would completely dwarf the rest of the report. That is not a realistic option. Rather, beyond a brief discussion of available standards in publishing as they relate to rights management, we will skim the surface of the topic and point the way to some "other reading". In due course, it may be that relevant organisations might seek to look at these issues in more detail.

5.1. RIGHTS MANAGEMENT IN PUBLISHING – WHAT IS CURRENTLY AT STAKE?

What is striking as we look across the landscape at standards for identification and metadata in publishing is how few standards have been developed specifically to support rights management. Publishing is often compared unfavourably with music[127] in this respect. However, we believe that there is a simple enough reason for this, which is the relative significance and scale of collective rights management in these two sectors.

We will try to illustrate this with some numbers. It is very difficult to develop truly comparable data, and what follows makes no claim to be sound economic analysis. It omits large elements of the music industry which have no direct equivalent in the written word (like music streaming). Publishing has no substantive equivalent to "performing rights" which are the bulk of revenues earned by CMOs. But it is important to have some grasp on the differences in scale.

127 Although this report about the UK music industry has a story to tell **https://www.gov.uk/government/publications/music-2025-the-music-data-dilemma.**

First, publishing as an industry dwarfs recorded music. The most recent numbers I have been able find (for 2019) show this unambiguously:

Publishing annual revenues (2019) **$258.7bn**[128]

Recorded music annual revenues (2019) **$20.2bn**[129]

In other words, publishing in all its forms (books, journals, magazines, newspapers) is more than a full order of magnitude larger than the recorded music business. Considering the breadth of "publishing", this probably ought to surprise no one, although it may be more surprising to learn that book publishing contributes around half of the total. The numbers may not be wholly comparable, but the scale of the difference is unlikely to be far wrong.

And annual revenues from collective rights management?

RROs **$1.0bn**[130]

CISAC Societies **$10.0bn**[131]

These numbers are **definitely not directly comparable** (see footnotes) but the scale of the difference is salutary. Not only is collective management an order of magnitude smaller in publishing than in music, it is perhaps three orders of magnitude less significant to the scale of the industry. These figures may be far from precise, but we contend that they tell a story.

This is not, however, anything like the whole story when it comes to rights management in publishing. We have not been able to trace an estimate of total global rights and licensing revenue, but in 2017 in the UK they represented around 7% of total industry revenue.[132] The UK is likely to be highly unrepresentative of the global position, (if a cliché may be forgiven) punching far above its weight in book publishing and in rights sales. Nevertheless, that figure suggests that total rights and licensing revenues made by book publishers globally are several times the total revenues of the RROs (and they of course collect revenues across other parts of the publishing industry).

Despite this, publishing as a whole remains slow to make the most effective use of technology in managing their rights, even when the payback on investment looks completely convincing, as a recent report published by BISG in the US makes clear.[133]

5.2. STANDARDS FOR RIGHTS MANAGEMENT IN PUBLISHING

5.2.1. Identification

We have already discussed ISTC (see Section 3.2), which was arguably developed as a rights management tool, but has failed in that role as in others. Other identification standards (including ISBN and ISSN) have typically been adapted to rights management roles by RROs as surrogate work identifiers (where repertoires are managed at title level).[134] They work "well enough" for current purposes, but the challenges of using surrogates are sure to grow in the context of any comprehensive infrastructural approach to copyright management.

As the management of rights becomes more granular, academic book and journal publishers already have a more granular identifier in the CrossRef DOI (see 4.3.2) which to all intents and purposes is a universally available work identifier at journal article level; it is becoming more consistently used at chapter level (although it is not uniformly deployed, which is already creating some challenges around open access books – see Section 4.3.3).

However, it is not in use outside this sector; it also doesn't provide any immediately self-evident mechanism for the identification of arbitrary selections of content. Even were an RRO to apply a standard identifier (an ISTC, a DOI) to (for example) a two-page spread in a school textbook, it is hard to see how exactly this could be deployed usefully in "the real world".

128 Publishing Global Industry Almanac 2015-2024
https://www.researchandmarkets.com/reports/5215015/publishing-global-industry-almanac-2015-2024

129 IFPI (**https://www.ifpi.org/ifpi-issues-annual-global-music-report/)**

130 This is an informal but well-informed industry estimate of revenues; about half this revenue is generated by the largest three or four RROs.

131 This is an estimate based on a CISAC press release about the loss of distribution to creatives due to the pandemic. CISAC estimate that creatives have lost 35% of their income, which they equate to $3.5bn; simple calculation suggests *distributions* pre-pandemic from CISAC societies are roughly $10.0bn. Their revenues will therefore likely be at least $1bn more than that. But there is also some double counting here, because some RROs are also members of CISAC and their distributions to authors will be included in the CISAC total. The CISAC total includes payments for some AV rights (which we understand are becoming increasingly significant over time presumably because of the growth of channels to market). Conversely, there are rights distributions in music which are not included in CISAC numbers.

132 The UK Publishers Association
https://www.publishers.org.uk/value-of-uk-publishing-industry-increases-5-to-5-7bn/

133 See *Positioned for Growth* (**https://bisg.org/store/viewproduct.aspx?id=17421021**); this report is paywalled. There are significant differences in the take up of systems (bespoke or off-the-shelf) between different sectors and different sizes of publishers, but if this report is to be accepted, further investment could have a significant RoI.

134 By no means all RROs manage repertoire in the way that might be expected. While we believe it is highly desirable that distributions of collective licensing revenues to rightholders should, to the extent that it is reasonable possible, follow usage and rights, this is not an invariable point of view within the RRO community. It is out of scope for this report to rehearse the pros and cons of statutory, voluntary and extended collective licensing, but different approaches to licensing and distribution of licensing revenues are extremely significant in the requirement for communication of data and therefore of the requirement for standards. Our approach to collective management in this report is unapologetically coloured by our own experience and viewpoint. We find it hard to believe that future rights management initiatives will not drive consistently in the direction of an increasingly granular transactional approach both to licensing and to distribution of revenue.

5.2.2. Metadata

5.2.2.1. Collective Management

To the extent that repertoire metadata is shared within the RRO value chain, so far as we know the great majority of standardised communications are those between publishers and RROs providing information about their books – sometimes using ONIX for Books.[135] An ONIX for Books message from a publisher to an RRO contains a lot of information that an RRO has no need of but crucially it includes a (wholly implicit) claim of rights – "if we give you information about this book of which we are the publisher, we own the publishing rights". This works well enough for current purposes (although when titles change hands between publishers, there may of course be conflicting claims for the same title, dispute resolution being an everyday challenge for all CMOs). It may also ultimately appear to be a rather weak mechanism for the rather critical role that it plays in the "chain of rights". But for the time being, it serves its purpose.

The great majority of repertoire and transaction/distribution information passing between RROs is shared either on paper, or on spreadsheets or by direct data entry. Globally, the same thing is true of communications between rightholders and RROs.

However, two standards[136] were published by EDItEUR in 2008 and were developed in collaboration with IFRRO for the communication of information between RROs. One of these, ONIX for Distribution (ONIX-DS), is in limited use – there appear to be five or six[137] users of the format globally. The fact that it hasn't required any significant attention in over 10 years suggests that it is robust for the purpose for which it was designed. It allows RROs to share arbitrarily detailed distribution data.

The other standard developed at the same time, ONIX for Repertoire (ONIX-RP) has, so far as we are aware, only ever been implemented between two elements of the UK RRO infrastructure.[138]

Here, it was used in its original form for many years to communicate repertoire information (covering books, journals and magazines) from PLS to CLA in a continual exchange of data. However, its XML structure and the declarative nature of the messaging proved cumbersome and resource intensive, and recent system changes have allowed a more streamlined approach to data sharing.[139] We believe that this has retained the original logical structure, and that data dictionaries have also been maintained.

We suggest that RROs might usefully reflect on the desirability of a common methodology for the identification of arbitrary fragments of books which are being copied, in order to be able to communicate usage more accurately. This might have some relationship with the development of a functional ISTC, but also depends on finding common answers relating to mid-to-long-term trends in global RRO data and distribution polices.

5.2.2.2. Rights and licensing in book publishing

BISG in the US has been trying for a number of years to develop standards for the communication of rights and licensing information between literary agents and publishers and between publishers and other publishers. It has proved a hard task,[140] foundering repeatedly on the problem of agreeing semantics and the failure to find any real "killer application".

The most recent iteration is a "rights and licensing taxonomy", which has been released for comment by members. We have not seen this but understand that it could form the basis for a useful communication toolset, although serious hurdles remain in developing a well-structured data model.[141]

5.2.3. Other

We have discussed elsewhere in this report a number of initiatives that have a bearing on rights management including:

- ISNI (see Section 3.3)
- ISCC (see Section 3.5)
- ODRL (see Section 3.4)
- RightsML (seen Section 3.4.1)
- IPTC Photo Metadata (see Section 4.7)

135 This use of ONIX will only be found among a handful of larger RROs.

136 ONIX for RROs **https://www.editeur.org/23/ONIX-for-RROs/**

137 The majority have implemented the IFRRO WISE software suite, which we believe was designed around the data model implicit in the ONIX messaging standards. Why have other RROs, even those who participated in the development of the standards, not implemented them? You would need to ask them, but some of the many reasons why standards do not get implemented are set out in Section 2.3. Anecdotally, the reason we have heard most frequently is that RROs believe that the greatest benefit from increased automation would be enjoyed by someone other than themselves. What lessons can be learned? At least these two: check from the outset the circumstances in which there will be a clear and necessary cost/benefit to participants in the development; and continually check that there is a real commitment to deploy (while remembering that those representing an organisation in the standards process may have little or no executive authority).

138 To understand the unique model adopted for RRO licensing in the UK, see **https://www.cla.co.uk/who-we-represent**

139 Serialised in JSON, and with a more implicit structure.

140 The first iteration of this work created what was in our view a completely valueless "word list" with no structure or definitions.

141 We have been told that the first likely application would be in automating the creation and ingestion of royalty statements between publishers and authors' agents. Whether this has sufficient value to justify the necessary investment we can only speculate.

We do not intend to repeat here everything we have already said.

CASE STUDY – WHAT HAS HAPPENED TO THE UK COPYRIGHT HUB?

It would be facile to try to assign the slow progress made by the UK Copyright Hub to a single cause. With 20/20 hindsight, we can identify a few factors which we posit caused it to lose momentum during its development phase and carry lessons to learn. These include:

- Lack of a clear line of accountability for technical development – the organisation undertaking the technology build had its own agenda, objectives and management structure
- An unwillingness (on the part of the technical team) to use adequate enough existing technology – substantial investment went into building an elegant technical identification solution, for example
- An initial focus on building "hidden" infrastructure rather than more easily demonstrable value
- A focus on relatively low-value use cases – unthreatening but also unexciting
- Conversely, some reluctance to introduce services which might be seen as competing with the vested interests of some of its supporters

The project was established in order to demonstrate that widening exceptions and limitations to copyright was not the only possible response to the challenges faced then and now by copyright. It had until recently an extremely enthusiastic and expert Management Board, and early in the process enjoyed considerable UK Government and industry support. But it cost a lot to establish and has subsequently struggled to find a sustainable business model.

This will likely be the fate of all such attempts at resolving the rights management challenge – until one day, suddenly it isn't. As always, timing will be everything.

5.3. FINDING THE WAY FORWARD FOR A COPYRIGHT MANAGEMENT INFRASTRUCTURE

What would it mean to develop a universal rights management infrastructure? Such an infrastructure would presumably go far beyond simply a standard set of structures for identification and metadata (although that would appear to be a *sine qua non*). In many respects, consideration of the totality of such an infrastructure lies outside the realm of this report. Our approach to the discussion here will be far from complete – it deserves a report of its own.

The idea of a comprehensive approach to copyright management in the internet age, built on a firm foundation of identification and metadata standards, has fascinated many of us for at least the last two decades, not least the European Commission and some individual governments. We have also been aware of WIPO interest since the end of the last century, starting with discussions around the then developing indecs project (see Section 2.8).

Some relatively ambitious projects have been undertaken, including the UK's Copyright Hub. These have helped to understand the challenges that implementing a full scale "rights management infrastructure" would represent, but none of them has come really close to resolving them. This is a challenge which it seems is discovered anew in every generation, sometimes without a full understanding of what has gone before.

It is clear, however, that starting from the Finnish presidency[142] in the second half of 2019, the European Commission has shown renewed interest in the development of a "European Copyright Infrastructure". The timing of this renewed interest may be appropriate, because elements of the technology that might be necessary to implement a convincing infrastructure are just now becoming sufficiently mature and well understood; this includes, for example, distributed ledger technology (sometimes known as "Blockchain", although it is important to recognise that Blockchain is a specific branded implementation of a more generic technology concept); this a subject to which we will return shortly (see Section 5.4.1).

Not only is the Commission itself pressing ahead with exploratory work on deploying DLT, we are aware that several smaller European countries (including Finland) are at the same time pressing ahead with national projects to do the same thing.

The US Copyright Office is also engaged in a major project to update processes and systems, although (unlike Europe) it seems *prima facie* unlikely that they will intervene in the market to create technical infrastructure.[143]

This points to one of the more intractable questions. Is it possible to think about a rights management infrastructure that operates at national level or even at European level, when what is needed in a networked world is a metadata and technology infrastructure that crosses all types of boundary – to quote from indecs, metadata that interoperates:

142 See **https://data.consilium.europa.eu/doc/document/ST-15016-2019-INIT/en/pdf** for their "stocktaking" report. To quote briefly "...it is not necessarily new metadata or other standards that are needed, but rather a facilitation of the data exchange architecture for copyright information for all sectors. The creative sectors are using data differently and have different levels of take-up of data standards." This is not a view with which we wholly concur – but neither is it one with which we would wholly disagree!

143 Other than, we might speculate, to create the technical infrastructure necessary for its own operational purposes and thence to recommend messaging standards for any party wishing to interact digitally with the USCO (eg for copyright recordation purposes).

- Across media (such as books, serials, audio, audiovisual, software, abstract works, and visual material).
- Across functions (such as cataloguing, discovery, workflow and rights management).
- Across levels of metadata (from simple to complex).
- Across linguistic and semantic barriers.
- Across territorial barriers.
- Across technology platforms.

There are many questions to answer, from the legal to technical to commercial to social; here is at least one critical question in each domain:

- Can copyright be managed across borders in the absence of international title?
- Does interoperability of metadata imply a single metadata infrastructure common to all media types, or can we work outwards from what we've got?
- What is the future commercial structure, who will see themselves as being disintermediated[144], and can therefore be expected to oppose it?
- How is this infrastructure to be governed?[145]

Our purpose in asking these questions is not to seek to answer them, but to indicate the spread of the questions that remain to be answered about any large-scale development of a copyright management infrastructure.

144 "One person's efficiency is another person's job." This is not a reason for not doing anything but has a big influence on how any process of stakeholder engagement can be managed effectively.

145 Developing an appropriate model of international and intersectoral governance will require seeking carefully considered answers to a number of questions, including: what must be centralised and what is more appropriately distributed; what is the appropriate model for ownership and operation of the technology; how is the market to be regulated; how is the entire infrastructure to be governed? A UK Copyright Hub governance document provides a template for considering some of these questions and involved input from a wide range of stakeholders; however, it is no longer easily accessible.

5.4. "NEW" TECHNOLOGIES

In the terms of reference, we were asked to comment on the potential role of two technologies in any future rights management infrastructure – BlockChain and Artificial Intelligence.

The first thing to point out is that neither of them is exactly "new"; the origins of distributed ledger technology (DLT) go back to the 1980s and of AI to the 1950s. Both have made substantial advances in more recent years, not least as computing power has caught up with their conceptualisation.

5.4.1. Distributed Ledger Technology

Although it may be BlockChain and Bitcoin that have made DLT headline news,[146] the reality of the technology is both more complex and perhaps more interesting. There are many different types of DLT, some of which may have a useful role to play in a copyright management infrastructure.[147] But BlockChain is not "the answer" as has sometimes been somewhat naively proposed. It seems unlikely that that specific technology is scalable to the extent that would be necessary to manage any reasonable volume of fragmentary rights transactions, for example.[148]

But DLT as a class of technologies is "another set of tools in the computer engineer's toolbox" as was suggested by an engineer during our research for this paper. Properly managed, "permissioned" distributed ledger technology (that is, with identified users given permission to post) certainly should have a part to play. The extent to which "permissionless" systems have a role is perhaps more open to question but is not one we feel completely qualified to answer.

We know that the European Commission has an active interest in deploying DLT in a number of applications which correlate with aspects of a rights management infrastructure.[149]

Regardless of whether DLT can be applied successfully to manage rights, the need for standardised semantics will remain. DLT is a technology, not a magic box.

146 And more recently "non-fungible tokens" which are perhaps the tulipomania of pandemic times. **https://www.economist.com/finance-and-economics/2021/03/18/non-fungible-tokens-are-useful-innovative-and-frothy** (paywalled)

147 There is considerable interest in some RROs about the potential application of BlockChain technology to specific use cases – see this example at Access Copyright in Canada **https://www.accesscopyright.ca/media/announcements/canadian-blockchain-innovation-will-benefit-visual-artists/**

148 Questions of scalability are complex, but BlockChain may not ultimately be scalable to managing Bitcoin in a sustainable way – see **https://www.bbc.co.uk/news/science-environment-56215787**.

149 See **https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/EBSI.**

5.4.2. Artificial Intelligence

The term "artificial intelligence", like DLT, raises any number of questions. We are certainly not competent to answer those in this report. Maybe humankind is ultimately threatened by machines smarter than we are; and AI is progressing at terrific (to some terrifying) speed.[150]

But at its heart, AI – at least to the extent that it is likely to find active application in our sector today – is primarily about machine learning, about pattern matching and the application of rules-based algorithms. We are not talking here about autonomous, human-analogous "thought".[151] But we certainly have another set of tools in the engineer's toolbox when it comes to rights management.

Does AI have a role to play? Like DLT, of course the answer is yes – for example, in cleaning up metadata resources, and in the sort of "clustering" that we have already spoken about with respect to ISTC and ISNI. Indeed, these would not have been possible without the application of some sort of AI, however primitive.

We remain convinced that we are a long way from AI taking the place of standardised identification and semantics for communication between systems. That may be where the long-term future lies, but that is well beyond the boundaries of useful speculation as far as this report is concerned.

5.4.3. A technological view of a future copyright management infrastructure

An account of the role that DLT, AI, and the ISCC (or equivalent technologies) might play in the development of a comprehensive global digital rights management infrastructure, can be found in an interesting paper: *DiSTri, a distributed trusted rights framework for digital content.*[152]

This is a technical paper but is for the most part very readable. In our view, the paper rather underplays the requirement for effective governance, and perhaps overplays the likely engagement of individual creators. Nevertheless it is a timely updating of a generalised technical model for a comprehensive digital rights management infrastructure.

An objective report, elaborating on the issues raised above and their implications for the immediate future and beyond, would make interesting reading. The challenge will be to find author(s) with the right mix of technical, commercial and rights management expertise to provide the necessary insights.

150 We will certainly make no attempt to discuss the vexed question of copyright in machine-created works.

151 For a discussion of the nuances of AI, this article may be useful **https://towardsdatascience.com/clarity-around-ai-language-2dc16fdb6e82**. We also recommend a 2021 New Scientist interview on "The hidden costs of AI" (paywalled) which makes for thought-provoking reading **https://www.newscientist.com/article/mg24933271-000-kate-crawford-interview-how-ai-is-exploiting-people-and-the-planet/** .

152 Personal communication from one of the authors.

5.5. METADATA AND ORPHAN WORKS

We have tried hard to think about how we might best provide useful input to the question of communication standards for literary orphan works. The challenge is considerable. The very absence of an identifier for literary works points to the heart of the difficulty. Works, of course, always require at least one manifestation in order to be able to prove their existence – so a manifestation identifier can stand as surrogate (so long as it is clear that that is what it is doing). So, a book with an ISBN can be given an identity. But a book from before the ISBN age (and that means any book much more than 50 years old, so many books still in copyright) has the same problem as any other literary work. If found in a library, it may have a local accession number; and it may have an OCLC WorldCat number. If it has a library record, it will likely be a MARC record, which is useful for all sorts of things, but not rights management.

Any record for a book[153] is likely to require some sort of "book in hand" cataloguing from scratch, so an easy-to-understand structure and the use of extremely tightly defined code values will be critical (the more so as presumably any interface will need to be multi-lingual).

The next question is whether it is realistically possible to develop a manageable structure which completely encompasses all different media, and perhaps allows the use of different namespaces drawn from other standards. It would certainly be a mistake to attempt to draw up a completely different set of values from those which are already being widely used in an industry setting. The single biggest question (from our perspective at least) is "How can you make this as simple as possible, at least to start?" Judging by 5 years' usage of the UK IPO's Orphan Works licensing register,[154] usage of a suite of orphan works messages may not be very high.

5.6. OUT OF COMMERCE WORKS (OOC)

What we have to say about orphan works applies in equal measure to out-of-commerce works, except that the search needs to include a search of commercial metadata sources (in the case of books, Books-in-Print). The challenge of finding out whether a book identified only with an ISBN is still available (perhaps from another publisher) is self-evident. IFRRO has published an extensive guide,[155] including the entirely reasonable expectation that the search for "in commerce" versions should include audiobooks and ebooks. In the absence of a functioning textual work identifier, this will be very difficult to

153 We have not given much thought to other types of literary works – periodical articles, for example. Images look like an even bigger challenge.

154 See **https://www.orphanworkslicensing.service.gov.uk/view-register**; similarly, the EUIPO Orphan Works portal has records for only about 4500 literary works at the time of writing.

155 See **https://www.ifrro.org/node/3501**.

achieve in reality.

The German national model for managing out of commerce literary works, operated by VG Wort, the German RRO, appears to combine data from the German National Library with data from MVB (who operate the German Books in Print service); this process is managed by the Library. There is considerable pressure in Germany for the mandatory use of ISNI, not least in the context of OOC works.

The EUIPO has recently launched the European Out of Commerce Works Portal. From our perspective, next steps are dependent on gaining a wider understanding of the technical approach being taken by EUIPO, specifically the data model.

However, challenges with cross-border search of nationally based Books in Print resources will remain, and we wonder about the extent to which this problem is technically and commercially soluble. However, possible solutions to this issue might usefully be explored in a future project.

6 Conclusions and Recommendations

6.1. CONCLUSIONS

This has been a wide ranging (although we know incomplete) account of the present state of identification and metadata standards for "publishing". In the limited time available, we certainly never expected to deliver a comprehensive treatise on standards in general or even about the particular standards we have written about. There is plenty in the links we have given to external documents to provide many weeks of reading for any who wish to understand all or parts of the field of study more completely. And much more besides.

Our major task, as we have understood it, has been to identify significant limitations and gaps in the current provision or implementation of standards which currently look unlikely to be filled without thoughtful intervention; and where there are current activities of which relevant stakeholders ought to be properly aware and in which they might usefully be more directly engaged. We hope that in this objective we have succeeded.

We have primarily looked to the present and to the relatively near-term future – perhaps a 3-to-5-year time horizon (bearing in mind the ISO forecast that it takes 3 years to create a new ISO standard). We have deliberately avoided "blue sky" thinking.

The findings of this report should be entirely unsurprising to anyone with even a passing knowledge of the publishing industry and how it really works. For the most part, good standards are in place where they are needed to support "business as usual" (which sadly sometimes means "business as it was ten years ago, but we can still paper over any cracks for now"). Elsewhere, standards can be slow to be adopted.

The tendency remains for dominant players (commercial but perhaps increasingly political or legal) to continue to define those moments when everyone in a sector must finally accept change and sign up to adhere to some part of the standards infrastructure. Peer pressure – through network effects – can sometimes achieve the same end.

With a few notable exceptions, Chief Executives of publishing[156] businesses know little of the technical standards on which their businesses depend. Senior industry champions of standardisation are few, outside those who run or actively participate in national or international industry standards bodies.[157]

To be fair, where a sector of the industry faces overwhelming change in its business model – in STM journals publishing for example – standards have been developed rapidly to meet needs. But (at least as it looks from the outside) they have recently followed an uncoordinated and inefficient development and governance process, which seems to have paid little attention to prior art or existing solutions. Nevertheless, their business goes on, apparently without undue stress, so their main objective is being achieved.

Interestingly, we see many in the library sector as all-too-well aware of the lack of currency of the standards which they have no option but to continue to deploy; the sector does not have the resources to bring about the comprehensive change that many practitioners would welcome. Change will come – but it may take several more decades for it to have full effect.

Work must be encouraged to bridge the continuing challenges in the relationship and incomprehension between the publishing and library sectors.

When it comes to managing rights, the key issues remain. For commercial reasons, collective rights management does not appear high on the radar of many senior executives in every part of the publishing industry (although attention has been growing, not least in STM).

In our direct experience, RROs manage their businesses pretty well on the whole, despite the lack of apparently "essential" data standards. Introducing these standards would be a "nice to have" in many cases but not worth a great deal of cost or effort in terms of payback. At a time of innovation, certainly in some markets, will changes to business models change requirements to share data more effectively with each other or with rightholders? We see that as likely, based on our own experience; but others will disagree.

Quite a large number of people to whom we spoke in the preparation of this report (for the most part, people involved in defining and/or implementing standards) inevitably have had their expectations substantially raised. They expect that the international community will do something.

One general recommendation we would make to relevant organisations is to get more involved in standards on a day-to-day basis, watch closely what is going on, making sure they understand it and how the different pieces of the standards jigsaw fit together (or sadly sometimes don't). Get directly involved where it is most appropriate and where their involvement can most make a difference in achieving their (or their members') strategic objectives.

No one should be deterred by the idea that standards are difficult and "technical". They are difficult but that is not (for the most part) because they are technical. As Norman Paskin often used to say: "The technical problems are the easy ones, they have technical solutions." It's the socialisation challenges that pose the big (and interesting) opportunities for engagement.

One question we have not yet addressed is whether current mechanisms for the development of technical standards are "fit for purpose" in an age of fast-moving technology. They are frequently criticised (by technologists at least, but also by many of those involved) for being too slow and cumbersome.

On the evidence of history, the speed of standards development is rarely the governing factor in their adoption and deployment. The fact that W3C is often heavily criticised for the delays in its process suggests that the problem of gaining wide consensus is no easier in a technologically-led organisation than in any other. But without that wide consensus, could we establish viable standards at all?

No doubt, there could be process improvements, but (for the time being at least) these are probably better achieved from the inside rather than the outside. However, the problem of managing participants who have a vested interest in defeating a standards development project needs acknowledgement, particularly in rights management (see Section 6.3.3).

6.2. AN OVERVIEW OF OUR RECOMMENDATIONS

In this section, we provide an overview of our recommendations for the international community to consider. Where we have specific recommendations for action, these will be found in Section 6.3.

Our first, and most general recommendation is to engage and participate. To engage with the agencies responsible for standards management, such as ISBN-IA; to participate in standards development forums such as ISO TC46/SC9 and to maintain a watching brief on other areas of interest, such as the work of the TDM Community Interest Group and the Content Authenticity Initiative.

At various points throughout this report, we have suggested areas which might merit further study. These areas include standards deployment in academic publishing, the application of new technologies in rights management and content identification, and the issue of cross border search for Out of Commerce Works.

156 This isn't uniquely true of publishing CEOs. And it is also far from true of **all** Publishing CEOs. During the 1990s, a group of far-sighted STM publishers at very senior level were very closely involved in – indeed paid for – the development of the DOI and its CrossRef implementation (which grew out of a project undertaken by the Association of American Publishers). Equally, a number of CEOs in the news media were very closely involved in the ACAP project. No doubt there are many other such examples.

157 At this point, we should acknowledge the part that both IPA and IFRRO play in the financing and governance of EDItEUR. IFRRO has also been an active participant in a number of significant rights-related standards projects in Europe, from indecs onwards.

We do not suggest that the commissioning organisations are equally interested in each of these issues. It may be that other stakeholders could also be interested in engaging in these further reports. However, we suggest that these areas for further analysis and review are important and the right report in each case should yield rich and compelling results.

And our third general recommendation is education – to engage and educate the community and constituency. Would ISNI be useful for the image sector? Will DLT be a silver bullet for managing rights in images? The answers to these questions are best found by dialogue, supplemented by targeted research and study.

And finally, development. For standards to be truly global they need to be inclusive, reaching all communities and also to be widely used and accessible all parts of the world. We suggest that development projects such as training, capacity building and the development of technical tools are necessary, in both specific communities such as in support of the visually impaired and also in less developed parts of the world to ensure equity of access. We understand that development is part of the global mission of the commissioning parties, and recommend future collaboration in this area.

6.3. MORE DETAILED RECOMMENDATIONS

6.3.1. ISTC

There is a very real question about the wisdom of "having another go" at the ISTC, but as we have seen at various points in this report there are challenges that will be difficult to resolve absent a standard mechanism for the common identification of textual works; this is not solely the case in rights management, although the most convincing arguments in the long term are probably about the management of rights (including, for example, in orphan and out-of-commerce works).

As we have seen, the problem with ISTC is that, while deployment of an ISTC might be desirable in many different applications, none of them is sufficiently compelling for any of the sectors involved to be willing to carry the costs. Authors, publishers, supply chain intermediaries, libraries, CMOs – all see some value, but all see that value in terms of the potential for benefitting from a standard where the costs are carried by someone else. The real value can only begin to be realised when a very substantial part of the repertoire has been identified. This will take a long time and involve the willing participation of a lot of different organisations.[158]

158 It is worth noting that it has taken the better part of a decade for the ISNI to get into a position where it is possible to be confident of its long-term success. That required courageous persistence from a number of organisations, incurring significant costs without any certainty of return. There may be questions about the willingness of some organisations to expend effort on ISTC in the same way because of potential conflicts of interest.

This risks a "tragedy of the commons". There is certainly no dominant player to force others into compliance.

Our recommendation is that all relevant stakeholders (collectively and/or severally) make sure that their interests are properly (directly or indirectly) represented in the ISO TC46/SC9 technical sub-committee that we believe will be established to examine the future of the ISTC. This will give them the opportunity to assess whether there is any realistic likelihood of establishing a functioning standard.

Much depends on whether the key issues of granularity discussed in this report can be resolved to everyone's satisfaction – or perhaps shown to be irrelevant.

In the event that there appears (on the basis of the discussions in this committee) to be a realistic opportunity for the successful development of a textual work identifier, the early phases of this work will without doubt require financial support. That support will be a necessary pre-condition, but it will not in itself be sufficient. It will also be essential that – if the identification of works is to be achieved by clustering of data in existing databases – libraries (and books in print organisations and/or RROs) prove willing to participate by submitting comprehensive data records. Without this, no critical mass of work identifiers can be created which would be the necessary first step in demonstrating sufficient value to encourage widespread implementation – bear in mind that this would only be creating historic work identifiers for books. Note that aspects of this work could commence even before the work of confirming the ISO standard itself has been completed (3 years is a long time to wait).

6.3.2. ISNI

Whatever new thinking may be needed about the future alignment of the published standard and aspects of its current implementation; this is a standard that is achieving its objective; ISNI is succeeding. At a minimum, stakeholders should be endorsing its wider implementation, talking about it publicly. This might help the standard to get established in some applications we have identified (like images) where it might be able to make a significant difference.

6.3.3. The development of an international copyright infrastructure

This is the elephant in the room, the strategic question that faces not just publishing but the whole of the "value network" of creativity. How will creativity be supported and rewarded in the future? How will society as a whole get to enjoy the fruits of that creative effort? It brings into play the whole future of societal attitudes to creators and creativity, of the legal and political framework around intellectual property, of our creative industries. Does a new intellectual property architecture pose existential threats to current participants? Can copyright survive at all in the networked world?

These are questions well beyond the scope of this report to discuss, let alone to answer.

The underlying assumption that we can see in many of the tentative answers being put forward is that in the long term "technology will provide the answer". It certainly seems that if a solution is to be found that looks anything like today's model, it will be dependent on a technological solution.

If that is to be so, then a network of data and data standards will lie at the heart of any solution (absent some sort of *deus ex machina* intervention from AI). What might that data network look like? We have already referred to one paper from the Linked Content Coalition in Section 2.7 (the definition of an identifier). We here refer the reader to the outputs of this project as a whole.[159]

It is worth an extensive quote from the LCC Web Site:

> *The Linked Content Coalition is a not-for-profit global consortium of standards bodies and registries. LCC members are organizations who create and manage data standards associated with content of one or more types, particularly for identifiers, metadata and messaging. The purpose of the LCC is to facilitate and expand the legitimate use of content in the digital network through the effective use of interoperable identifiers and metadata. The LCC supports interoperability between the computer systems of any and all legitimate participants in the digital network, including creators, rightsholders, publishers, aggregators, rights and content exchanges, retailers, consumers, cultural institutions (including libraries, museums and archives) and their agents and associations. Participation may be on any scale, from that of private individuals to multi-national organizations.*

If we have demonstrated anything in this report, we hope we have underlined how far the global publishing industry is from having developed and deployed the sort of infrastructure envisaged as necessary components of a global digital rights infrastructure. Much work has been done to define what a comprehensive solution might look like. Rather too little of it is in place.

It is here that we stray into long-term strategic questions for all readers of this report that are clearly not within our remit. If there is to be "a global digital rights infrastructure", what part(s) could or should relevant organisations[160] play in its development and governance? We have already suggested a more active engagement in aspects of the standards making process that might make a difference.

Steps beyond that are for others to determine.

MB
Wells, Somerset
31 May 2021

159 See **http://www.linkedcontentcoalition.org/** We understand that LCC is no longer active but their objectives and technical output remain highly relevant.

160 In this context, see **https://iprinfo.fi/artikkeli/copyright-infrastructure-a-recipe-for-recovery-and-resilience-of-the-creative-sectors/**