Standard Identifiers:
an overview of the current landscape

A presentation to USPTO Open Meeting: **Facilitating the Development of the Online Licensing Environment for Copyrighted Works**
by Mark Bide April 1, 2015

Why do we always start conversations about improving the management of media online – and particularly about licensing – with a presentation about identifiers?

Let's begin at the beginning. What is an identifier? What is a standard? And why are they important?

The first of these questions is the easiest to answer. An identifier is simply a name to call something by. But, in the context of what we are talking about today, the way we understand "what an identifier is" goes beyond simply saying "it's a name".

By "an identifier" we mean a name that is suitable for machine-to-machine communication. And that means it has to have certain other characteristics. The most important of these is that it should be unique in its own context and in its own domain – in other words, an identifier should uniquely identify only one thing.

Because machines don't manage ambiguity well. For the time being, at least, people are better at inference than are machines. And in inference context is everything. People can guess at the context, machines have to work without guessing.

Identifiers help machines to understand whether two things are "the same thing" or "two different things" – or to use the slightly more technical language we sometimes favour, to collocate and to disambiguate. So, identifiers are powerful tools in the process of machine-to-machine communication. And if we are to enable people to manage rights and licensing online, using automated or semi-automated processes, we have to have certainty about identification – are we all talking about the same thing?

But what do I mean when I say two things are "the same"? That's a question as old as philosophy – think of the "Theseus paradox", a conundrum also sometimes known as "my grandfather's axe". If every part of something has been replaced, how can it possibly be "the same thing" – and yet we call it by the same name.

Or an example closer to my topic today, consider the case of these two books – clearly distinct and different objects and yet, from the point of view of the ISBN – the most venerable of all ISO identification standards – they are indeed the "same thing". But if one of these things is a hardback and one a paperback, they have different ISBNs.

And this is exposing the context in which the ISBN functions as an identifier – it identifies distinct products, items of trade that need to be distinguished from one another in the book supply chain. ISBNs have been doing this for nearly 50 years.

Before that, books had to be ordered from their publishers using proprietary catalogue numbers. Automation was more-or-less impossible. The introduction of the ISBN allowed the development of automated systems for the management of the book supply chain – for ordering, for invoicing, for the myriad communications that need to take place to facilitate an orderly market place. Before they had them, no one knew they needed them. Now ISBNs are used to identify books in book supply chains all over the world.

That doesn't mean you can't use another identifier instead – and some organisations do use proprietary identifiers within their own systems. But to deploy one on the same scale as the ISBN you would have to reinvent the ISBN…an expensive exercise in futility.

However, the challenges of identification in books are as nothing in comparison with the challenges of identification in more complex media like music or audio-visual.

So here's something for you to think about as your attention drifts from my presentation. To what extent are two sets of recordings of Beethoven's Nine Symphonies the same thing and to what extent are they different things? In what contexts?

This key issue – deciding when two things are the same and when they are different – we call "granularity", the extent to which we have to divide the world we are describing into the individual elements that need identification. The answer to the question "the same or different?" in any computer system – and in talking about identifiers we are always really talking about computer systems – is that granularity must follow the functional requirement that the identifier has been designed to fulfil.

Of course, all computer systems have inbuilt identification systems, they couldn't operate without them. Systems have "database keys" automatically assigned to each record created in the system.

So what is the difference between a database key and an identifier?

At its simplest, this is simply a matter of scope. Within any single system, there is no need for a database key to be given any context – the context is the system in which it is used. Even within an organisation, that context can simply be assumed (although this in itself can often cause problems in large and distributed organisations). Move beyond the boundary of the organisation, become a "public identifier", and context is everything.

And this is where standards become critical.

In a complex automated transactional environment, different people and organisations must have confidence that their individual systems are talking about "the same thing", if they are to have an appropriate level of trust in the transactions themselves.

So, we need some sort of process through which the meaning and use of a shared identifier can be defined.

Sometimes, this can simply be defined between trading partners in a more-or-less informal way. But as the supply chain becomes more complex, many-to-many communications using poorly defined or poorly implemented standards become increasingly burdensome to every participant.

Standardisation may come about in many different ways – through ISO, through W3C, through formal and informal trade associations.

The critical issue with standardisation is *governance*. Why is governance so important? Because – while we all talk about "neutrality" in the standards world – there is a hidden guilty secret in our work – no standard can ever be wholly neutral. Every standard has an unspoken "point of view", a set of assumptions that informs its entire specification.

From the point of view of an organisation that implements a standard, it is critical that the standard is stable – because otherwise there is a risk in committing to implementation. It is also important that the specification represents a consensus "point of view" among those who will implement it – or at least something close to a consensus.

"Consensus" is an important word in standards setting, and 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 can take time, which is one of the reasons that standardisation processes often seem interminable.

How much easier to impose a proprietary solution? But how should we feel about this in the context of governance and neutrality?

Standards

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

In other words, standards reduce risks and costs while increasing potential revenues. So what's not to like?

Well, I guess what's not to like is that the rules are laid down by the community not by the individual, and this can create costs and other barriers which organisations can find hard to justify in their own specific circumstances.
It is also very often the case that the costs of implementing any standard in the supply chain, and the benefits to be derived from its implementation, are unequally distributed.

This can prove to be a considerable disincentive.

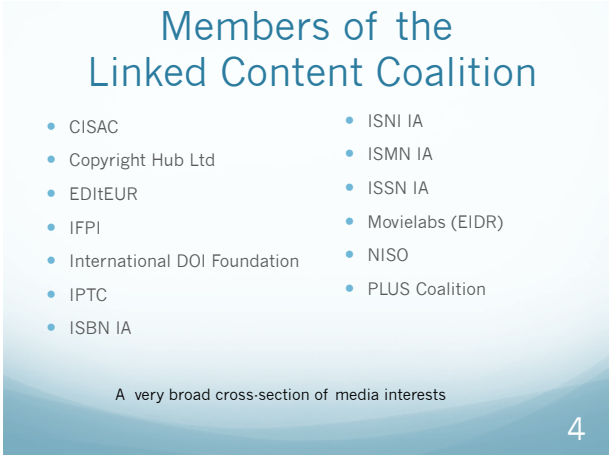
Which takes us to another point about standards that is often overlooked by those who are not part of this world. It is often assumed that standards are about technology. In reality, as those of us who have worked in standards know – another of our guilty

secrets – is that standards implementation is mostly about establishing acceptable social norms to which people willingly adhere.

This is why my interests have turned more to governance in later years – having what you believe to be a better technical solution is no guarantee of acceptance, in standards as elsewhere. Governance goes to the heart of the challenge.

So, having thought a bit about identification and a bit about standardisation, we turn our attention to what a media identifier standard should look like in the context of facilitating online licensing.

One project – now transformed into an organisation – has taken a good shot at answering that question – the Linked Content Coalition, or LCC.



LCC's objective was the development of the technical framework for an orderly online "content and rights supply chain" – of which identifiers form an intrinsic element.

This cross-media-industry initiative (which was funded out of the industry) has developed a set of principles in answer to four questions about identifiers. Version 1.1 of the LCC *Principles of identification* was published a little under a year ago.[1]

These questions were:

1. What should be identified?
2. What form should an identifier take?
3. How should an identifier be assigned?
4. How should an identifier be used?

That they were able to answer these complex and far reaching questions in a document only 4 pages long – although admittedly with extensive appendices of supporting material – is remarkable. The document is packed with detail that is too extensive for me to deal with in its entirety today. But I will pick out some of the key issues we need to bear in mind.

First, what needs to be identified?

LCC says that each entity (another word for "thing") which needs to be recognised distinctly in the digital network should be assigned at least one "persistent public identifier" so that it may be denoted unambiguously wherever that is useful. They define the term "public identifier" as "one that is accessible and recognisable by people or machines within the digital network."

The things – entities – that require identifiers include not only items of content (or "creations") but also each party (person or organization) who is recognised as, or claims to be, a contributor to or rightsholder in content (or someone who asserts metadata about that content). We also need identifiers for rights and licences.

[1] http://bit.ly/1BNLqKY

In terms of structure, LCC has little to say except that identifiers should be expressible as Universal Resource Identifiers – URIs – in order to be resolvable, a key requirement we will discuss a little more under the way in which identifiers should be used.

The other structural requirement proposed by LCC is that identifiers should be as dumb as possible – in other words, should include as little metadata as possible about the thing being identified, leaving all information to be retrieved from metadata repositories rather than inferred from the identifier itself.

People always want to infer meaning, and will often try to teach machines to do the same. The problem is that any apparent meaning in the structure of an identifier is all too often misleading (and may be used, for example to infer rights ownership, an issue on which identifiers should be entirely neutral)…but these are perhaps conversations for another day.

The LCC-recommended approach to the assignment of identifiers is more complex – and potentially more controversial. There are relatively straightforward (if critical) points about the need for uniqueness and persistence, and the requirement for a trusted authority. And a recognition that, in order to be authoritative, identifiers should be assigned as early as practicable in the creation process.

But more controversially, LCC also recommends that identifiers must be associated with metadata repositories that enable the referent – the thing that is identified – to be "discovered and unambiguously recognised".

LCC make clear that this isn't *necessarily* the role of the identifier registration authority itself – third party databases may fulfil part or all of the requirement – but the preference is clear for authoritative metadata registration in an identifier repository managed by the identifier authority.

This is a proposal that has proved unpalatable in some sectors and unimaginable in others – at least without significantly more joined up thinking than has currently been achieved.

The final set of recommendations – those regarding deployment – include some similarly ambitious – and equally controversial – implications.

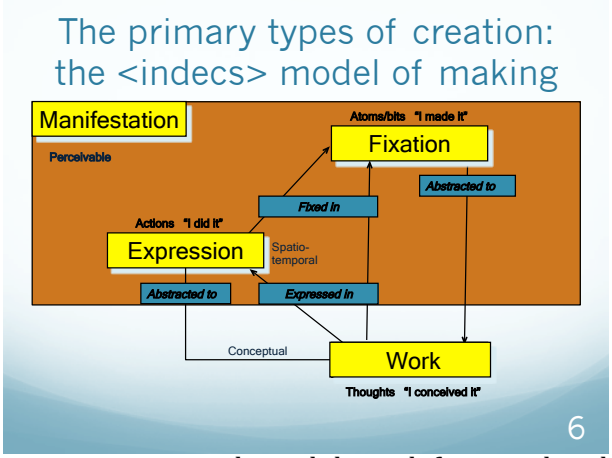
- Appropriate mechanisms need to be developed to associate identifiers directly with the object they identify so that people can find the identifier (by in some way embedding the identifier into the object itself)
- Every public identifier needs to be resolvable, using standard internet protocols, to the thing that it identifies and/or to sources of information about it
- The identifier should offer the potential of "multiple resolution" – in other words, any number of different data sources should be accessible from a single identifier, using standard methods to identify what type of data the user is trying to retrieve

Throughout the document there is a consistent stress on the need for people to be able to place unequivocal trust in the identification network. This speaks to the need for transparent but effective governance.

So, measured by this exacting yardstick, how are we doing today?

Let us think first about "creations" – one of the key types of entity that needs to be identified.

About 15 years ago, the <indecs> project put forward a model to explain the relationships between the various different aspects of creations that need to be identified and managed, particularly to support the management of rights and licensing. This model, which we called the Model of Making, has not been seriously challenged in the interim and, with all apologies due to those who have seen it many times before, I am going to use it again here today.

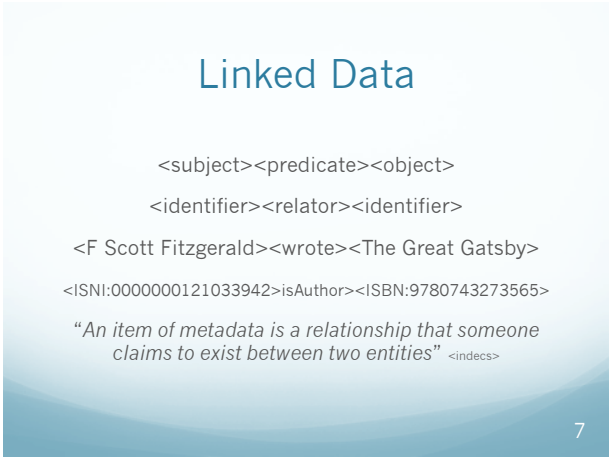


All acts of creation start with someone doing something – writing something down…singing a song…filming an event.

However, in order for that act – what we call an Expression in the <indecs> model – for that act of "doing something" to have any continuing identifiable existence, it must be fixed in some sort of medium. This "fixation" is of course critical to any understanding of how intellectual property is created – and through fixation that the underlying abstraction or "work" can be recognised and protected.

The same abstraction can subsequently be re-expressed in a spatio-temporal performance – or re-fixed in one or many fixations (for example, many printed copies of the same book, or many digital copies of the same piece of sheet music).

Both expressions and fixations are things we can perceive directly – in other words, they are manifestations. The existence of the work on the other hand can only ever be inferred. This means that identification of works is more difficult – because there is no perceivable object with which the identity can be associated. It exists only in its metadata – its description and its relationship with other identities.



And this perhaps takes us to one of the most critical issues of all – a reality recognised originally in <indecs> but given more prominence in the naming of the Linked Content Coalition – the importance of linking, of relationships.

The understanding in <indecs> is the same as that of the basis building block of "linked data" – that all metadata can be expressed in terms of relationships between identifiers.

This last quote from the <indecs> project is critical to our understanding of metadata: first of all, metadata is about claims – and there may be many different claims about the same thing; secondly, it is about relationships. This is the world of linked data into which we now moving online, in which identifiers are linked to other identifiers, and the internet itself promises to begin to function as a single database, in which related data is linked to other data in new and exciting ways.

But this is built on dust if the identifiers themselves are not reliable, and the nature of the relationship is not clear. In this context the precise nature of the relationship is as important as the identifiers themselves.

To say that any two things are related "*in some way*" is a truism of startlingly little interest – but to be able say how they are related in ways that are properly standardised and controlled begins to create capabilities that will bring to life Charles Clark's famous maxim on the future of copyright – "the answer to the machine lies in the machine". Technology can now work for the benefit of copyright not against it. But I am at risk of straying off topic.

In terms of where we are today, which parts of the media industries are doing well and which not so well in terms of universal identification systems? The challenges are mixed – some of them because of the embedded nature of existing systems, some because standard identification systems have not reached the sector at all.

Music can perhaps be regarded as being in the best shape; because automated and semi-automated rights management has been critical to music for a very long time, the infrastructure for managing identity has been well developed with internationally-well-established ISO-governed identifiers for musical works – the ISWC – and for recordings – the ISRC. The music collecting societies were also early in adopting their own standard way of identifying people and organisations – the Interested Parties Information system or IPI.

But there are gaps. There is no definitive database of ISRCs available, and no single authoritative source of the relationship between ISRCs and ISWCs (although there are commercially available sources of data which are becoming more comprehensive).

While very comprehensive in terms of the coverage of the sector, the IPI itself is a proprietary system and closed to those outside the collecting societies.

The book-publishing sector has, as we have already discussed, had a product identifier for getting on for 50 years. This need to identify products dictated its most critical set of requirements and the ISBN has been astoundingly successful in meeting these requirements (although not always so good at other ones).

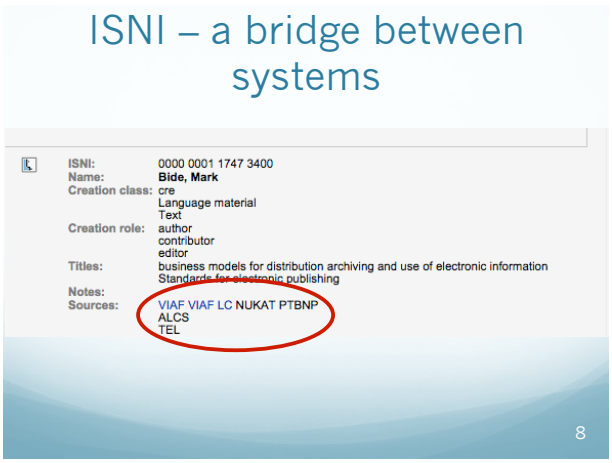
More recently there have been some significant problems with the granularity of application of the ISBN to ebooks, which has threatened the global integrity of the standard and has meant that it is more difficult to understand exactly what an ISBN for an ebook represents. This remains a work in progress.

However, attempts by the publishing industry to introduce a work identifier – the ISTC or International Standard Text Code, equivalent to the music industry's ISWC – have not been marked by signal success. There proved to be insufficient value from identifying works in the existing physical and digital supply chain, and insufficient scale in rights management.

Organisations managing rights in books have to infer the identity of works from the identity of products.

However, book publishing is working on the significant potential for the widespread adoption of a cross-sectoral party identifier, the International Standard Name Identifier, or ISNI. The ISNI started making assignments in the literary sector, linking records held in the giant Virtual International Authority File (domain of the national libraries) with records held in other party identification systems. It now has over eight and a half

million confirmed identifiers issued to creators from all walks of life (writers, musicians, filmmakers…).



I am showing you my ISNI entry only to demonstrate the way in which it is bridging identities between different worlds – in this instance, libraries and authors' societies – for example VIAF and the Authors Licensing and Copyright Society of which I am a member. These worlds, long separate and with little apparent common interest, are coming together very effectively to share a common identity standard. This now extends into other types of requirement like Books in Print catalogues.

Not replacing existing identification systems, but enhancing them.

I believe we can be optimistic about the long-term adoption of the ISNI, although it takes a long time for such systems to become universal, often for relatively banal reasons to do with the cost of reconfiguring legacy systems.

The position of other sectors of publishing is a bit mixed.

Journals publishers have been extremely effective in identifying individual articles using the CrossRef implementation of the Digital Object Identifier, the DOI.

The DOI is an ISO standard that both deals with resolvability and can help with cross-sectoral interoperability.

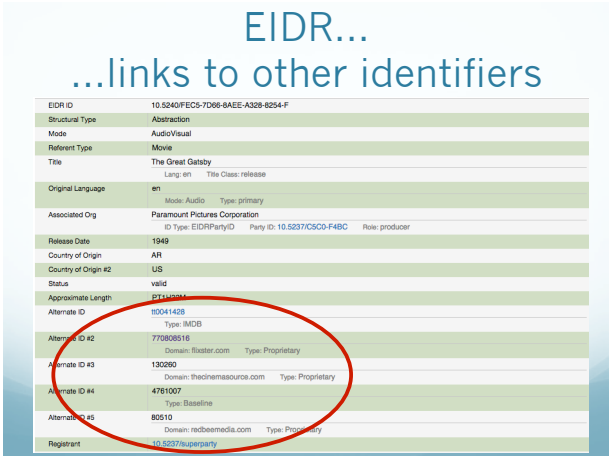
CrossRef has around 70 million items identified – not only journal articles but also scholarly books and book chapters, technical standards, theses and dissertations. CrossRef has also been heavily involved in the development of an identifier for researchers, the ORCID (which is separate from but coordinated with ISNI).

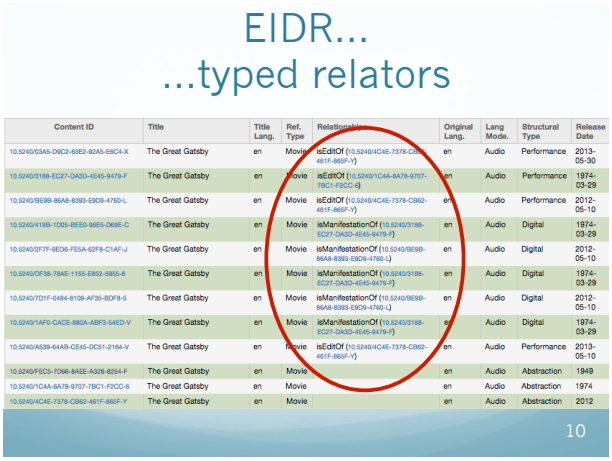
Magazines and newspapers, on the other hand, have found little justification for any public identifier for their articles; although they do have International Standard Serial Numbers (ISSNs) allocated – as indeed do journals, alongside their CrossRef DOIs.

The audio-visual world has two identifiers for works – the ISAN and the EIDR, the latter, like CrossRef, an implementation of the Digital Object Identifier.

EIDR is something of a newcomer, but already has over 700,000 individual registrations and is growing at a very considerable rate. Although it originated in the US, it is rapidly seeing deployment in other countries too.

ISAN and EIDR came from somewhat different sets of requirements, and have pursued very different approaches to identifying AV works.





EIDR is not unlike the ISNI, in that the EIDR database links together the many different proprietary identifiers that are widely used in the distribution of television programmes and movies. It does not seek to supplant these existing identification systems – rather, it seeks to enhance their continuing value by linking them all together.

EIDR data also allows the publication of exactly the type of linkage between different entities that I talked about earlier.

As you can see from this example, EIDR identifies not only AV works but also manifestations, and codifies the relationships between them in a structured way.

From this point of view, it can be seen as a model of identification good practice.

The last individual sector I will look at – briefly – is the visual arts, and particularly photography. But that will also take me on to the whole question of user generated content.

The professional photographic market has never developed a requirement for a standard identifier, presumably because of the very simplicity of the supply chain – from the individual photographer through an intermediary (the photo agency) to the user (typically a professional buyer). Projects like the PLUS Coalition have made a strong argument for the implementation of standard identifiers, but have so far enjoyed only limited traction.

The challenge has been that there is typically no need for aggregation of data by third parties, no need to track complex transactions through a supply chain – and therefore no obvious set of requirements has ever emerged for an industry-wide identifier.

However, the work of projects like the Copyright Hub, about which you will hear more this afternoon, is beginning to make the case for universal application of identification to photographs and indeed to all creations – every type of creation that is being made and posted on the internet by all of us, every day.

This is a considerable challenge; identifier standards have typically been the preserve of large organisations rather than individual creators.

And there is another one. While (as we have seen) creations and parties have identifier standards of varying levels of maturity and breadth of implementation, there is currently no standard for identifying statements of rights or licences. LCC and the

Copyright Hub are working together on prototyping such an ID, which should enable the network of rights data to be distributed and navigated a great deal more easily around the network.

During the course of my working lifetime working in the media, I have seen a steady process of democratisation in the tools available for creation of content, and then for its publication. Now we can all create content of all types – publications, music, photographs, movies – using tools that less than a generation ago would have seemed unimaginably powerful. And we can make them available to a potential market of billions of people at the click of a button. Democratisation indeed.

We now have the opportunity to democratise the tools of rights management and licensing in exactly the same way. As creators, we all have the right – but not always the ability – to decide how our content should be used.

Secure unambiguous identification is the essential first step. We will need to find ways of making the identification network work for everyone.

Thank you