Requirements for the identification of 'components' – an EDItEUR discussion paper

Mark Bide and Graham Bell, EDItEUR

Summary

This paper is intended to inform discussion of the issues surrounding the potential requirements for standardised identification of highly-granular content – content such as individual chapters or extracts, diagrams or illustrations, and audio or video clips – small chunks of content that may be disaggregated components of 'larger' content items or may have been created independently. It discusses the nature and need for clear identifiers, and considers whether these chunks or 'components' are a distinct class that requires a dedicated identifier.

The authors contend that these components are 'works', a generalisation of the work class that is currently identified using ISTCs[1], and that these works have peer relationships with other works. Some have direct relationships with intellectual property contracts, some have direct relationships with products, some have both, and some have neither but inherit rights and product links indirectly through their relationships with other works.

EDItEUR recommends further work to ground any conclusions about identifier requirements in up-to-date use cases, and consideration of the potential application of the ISTC and DOI in this field – neither of these is fully 'ready to go' but might form a part of a standards-based solution.

1 Background to this paper

A number of publisher members have approached EDItEUR informally over the last two years about the need to initiate a piece of standards work for the identification of what they variously called "components"[2], "fragments", or "micro-content". The question has always been the same: is anyone doing anything about this? The discussion usually focused on the use of ISBN to identify **products** "smaller than a book" – typically people think in terms of selling a "chapter".

However, on closer examination, it seemed to us that these questions were only rarely focused on *product* identity (the domain of the ISBN) but on identification of the components that might go to make up a product. EDItEUR therefore decided to hold a workshop to explore the underlying requirements, and whether these were issues for standardisation or simply for proprietary implementation. This paper is the output from the workshop, and from subsequent discussion with those who attended.

[1] The International ISTC Agency has indicated it is considering an extension of the scope of the ISTC to encompass the wider class of 'works'.

[2] We have chosen 'component' over 'fragment' for reasons which should become clear later in this paper.

This paper explores the requirements at a theoretical level and puts forward some propositions as to how this issue might be resolved. It grounds these propositions in practical Use Cases created by publishers over the last couple of years, and offers suggestions for further work. For reasons of commercial confidentiality, the Use Cases and all other inputs to this paper have been anonymised.

Our purpose in creating this paper is to drive forward the discussion on a global basis. We look forward to engaging with our colleagues round the world on the issues that it raises.

2 The need for identity

There has always been a need to identify arbitrary pieces of content. Any of us who started our lives as production assistants are likely to remember going through huge piles of artwork, making sure each line drawing and photograph was correctly identified with the author's name and the figure number (always in soft pencil). Even in very large production departments, this mechanism for identification was usually adequate as a book went through the production process. Once a book was published, the artwork would either be returned to the author or filed, in the more or less certain knowledge that it would never be needed again; if it was, a dusty search through the archive might be able to locate it, but it was a distinctly hit or miss process.

This approach to content identification has been an inadequate option for some time. Publishers have been establishing Digital Asset Management systems for storing content of all types, with a view to re-use and repurposing of this content. A primary aim of DAMs is to make content discoverable, and that means giving content unambiguous and unique identity – as well as associating appropriate metadata with each item that allows its discovery in different contexts.

The content might need to be identified in any of several ways:

- As an 'abstract work' ["a photograph by Mark Bide of the Eiffel Tower" [3]];
- As a particular manifestation of that work ["a thumbnail of that photograph in a particular file format"];
- As a specific instance of that manifestation ["a specific instance of a file of that photograph"];
- In relationship to other content ["as part of a collection of content"];
- In relationship to a product ["as part of that collection of content that is available as a product in a particular digital format or in a particular physical form"].

A significant task of this paper is to consider which of these classes of entity requires a specific identifier, and which is best identified by 'clustering' through the use of metadata. Occam's razor warns against the unnecessary multiplication of entities, and this is a principle to which we believe it makes good sense to adhere.

[3] The choice of the Eiffel Tower as the subject of this photograph is not entirely coincidental, since there is a significant claim of image rights in photography of the Eiffel Tower **at night**. The Société d'exploitation de la Tour Eiffel claims author and brand rights over its night-time illuminations, and licences those rights to sponsoring organisations who may require a credit to be associated with any view of the Tower. However, daytime views of or from the Tower fall within the public domain.

3 Understanding the nature of abstract works

There is frequent misunderstanding of the term 'work' in casual conversation in publishing. In order to avoid ambiguity, we will explain what we mean by a 'work' in this paper.

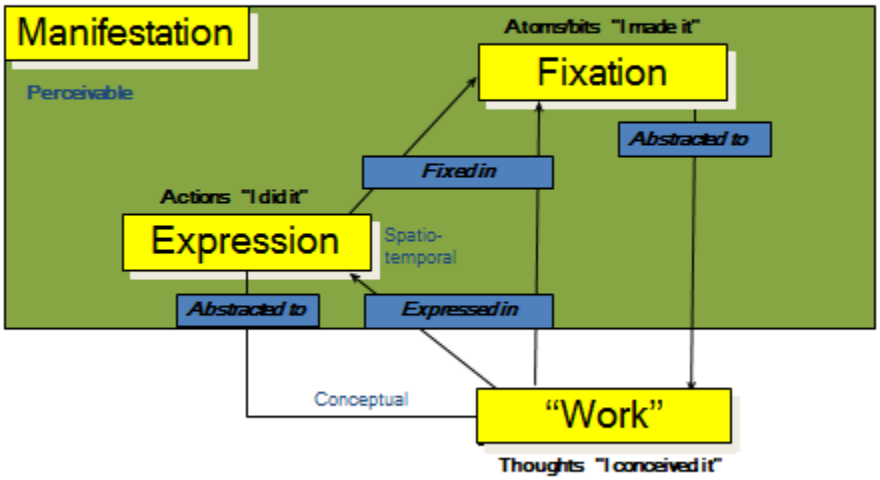


Figure 1 The <indecs> model of making [after Rust G & Bide M 2000 The <indecs> metadata framework: Principles, model and data dictionary http://www.doi.org/topics/indecs/indecs_framework_2000.pdf]

Whenever we refer to a 'work', this diagram shows exactly what we mean[4]. A work is entirely abstract and perceivable only through a manifestation of that work, but it exists entirely separately of any particular manifestation.

Manifestations are concrete embodiments of a work. There may be many different manifestations of a work, and many instances of a particular manifestation. For example, an ISBN identifies a class of instances which may be treated as being 'the same thing' for the purposes of a supply chain) or as individual instances.

4 Some principles of identification

It is useful to remind ourselves of some basic principles of identification. First of all, what is an identifier?

Here is a definition, drawn partly from Wikipedia:

- An identifier is a unique expression in a written format either by a code, by numbers or by the combination of both to distinguish variations from one to another among a class of substances, items, or objects;

[4] 'When I use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I choose it to mean – neither more nor less.' Carroll, L (1871) *Through the Looking-Glass, and What Alice Found There.* However, this is not an arbitrary definition; it is shared by standards like the ISTC and ISWC (and also – at least to some extent – by FRBR; see Section 9 below).

- In computer science, identifiers are lexical tokens (that is, 'nouns') that name entities (or 'things'). The concept is analogous to that of a 'name'. Identifiers are used extensively in virtually all information processing systems;
- Naming entities makes it possible to refer to them, which is essential for any kind of symbolic processing.

In other words, it is a name that needs to be unique in a given (systems) context that can be used to process information about the entity which it identifies. A definition of entity may be helpful here: in this paper, an entity is 'something that is identified'.

This definition of an identifier also provides the reason why we need to assign these names: collocation and disambiguation – in other words, so that we can talk about two things as being the same or about two things as being different. But whether two things are 'the same' or 'different' is not absolute; it is *always* contextual. It is clear that 'the same book' should have the same ISBN; but clearly two different copies of the same book are quite distinct things. We treat them as being the same thing for the purpose of the supply chain, but this collocation of two different things may not be appropriate in a library where each individual copy of 'the same book' needs to be identified separately.

This brings us to a final principle, which like the 'model of making is drawn from the <indecs> framework (see above). This is the principle of 'functional granularity'. So important is this issue, that we are reproducing the relevant section of the <indecs> framework document only lightly redacted:[5]

> **The Principle of Functional Granularity**
> *It should be possible to identify an entity whenever it needs to be distinguished.*
>
> When should an identifier be issued? In this deceptively simple question lies the most basic question of metadata: for which data is it meta-? Resources – stuff – can be viewed in an infinite number of complex ways. Taking this document as an example, it has an identifier in the <indecs> domain: WP1a-006-2.0. But to what does this refer? Does it refer to the original Word document, or to a PDF version available on the Website? Or does it refer to the underlying "abstract" content irrespective of delivery format?
>
> If it refers to the Web document, is this also adequate as a reference to local copies that have been downloaded onto other computers or servers? The document's parts may require identification at any level. If you wish to make a precise reference to this sentence from another document, you will need a more precise locator, and its nature will depend on whether your reference is intended to allow automated linking. As this document has been through many stages of preparation, how many different versions need to be separately recorded? Each of these requires the exercise of functional granularity: the provision of a way (or ways) of identifying parts and versions whenever the practical need arises.
>
> The application of functional granularity depends on a huge range of factors, including the type of resource, its location in time and place, its precise composition and condition, the uses to which it is or may be put, its volatility, its process of creation, and the identity of the party identifying it. The implication of this is that a resource may have any number of identifiers. The same entity may be subjected to functional granularity across a range of views. The basic "elements" of a resource may be entirely different according to your purpose. Stuff may be analysed, for example, in terms of

[5] From http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

> molecular entities (chemistry), particles such as electrons, quarks or superstrings (physics), spatial co-ordinates (geography), biological functions (biology, medicine), genres of expression (creations), price categories (commerce), and so on.
>
> In the digital environment, stuff can be relatively easily managed at extreme levels of granularity as minute as a single bit. Each of these processes will apply identifiers of different types at different levels of (functional) granularity in different "dimensions"; these may need to be reconciled to one another at a point of higher granularity**. Functional granularity does not propose that every possible part and version is identified: only that the means exists to identify any possible part or version when the occasion arises** [emphasis added].

5 Identification standardisation

We need to consider why we might choose to develop *standard* identifiers – that is, identification schemes that are interoperable beyond corporate boundaries. This would imply (for example) that the identifier syntax is standardised; that identifiers of the same class have mechanisms to ensure that they are globally unique; that identifiers of the same class identify the same class of entity; and that identifiers of the same class have the same 'Reference Descriptive Metadata'[6].

It is clear that at least some of the identifiers we are considering in this document do not, as a matter of course, cross corporate boundaries. However, our requirements workshop surfaced three requirements which imply that it is necessary to consider an appropriate degree of standardisation:

1. There is a need for standardised identifier requirements to create the conditions for sharing development costs. As things stand, different publishers are specifying slightly different proprietary identification models and identifiers. In an industry that is characterised by a small number of developers each with an international installed base of more-or-less standard systems, these different requirements are unhelpful in that they lead to non-standard implementations of otherwise similar systems – with higher development and maintenance costs over time;
2. Any business-to-business content trading – including, for example co-publishing ventures and content aggregation services – will benefit from common standards of identification;
3. This is an industry that is characterised by a great deal of M&A activity; the cost and time taken in assimilation of businesses can be substantially reduced if they have broadly similar data and identification architectures (even if they do not share the same system).

We appreciate the power of these arguments, but the extent of standardisation needed to satisfy these requirements is something that needs to be further explored. Such exploration needs to take into account that the boundaries of 'the corporation' are not fixed, and that the need to communicate unambiguously is increasing all the time.

[6] Reference Descriptive Metadata is the set of metadata associated with an identifier whose primary purpose is disambiguation – in other words different 'referents' (different things that are identified) should always have different Reference Descriptive Metadata. To the extent possible, Reference Descriptive Metadata should be drawn from controlled value lists and avoid free text fields (with the obvious exception of names and titles). This paper will not say a great deal about the metadata associated with identifiers – that is another and rather wider topic than we are approaching here.

6 Fragments or components?

As Appendix 1 makes clear, there is a distinct difference between identifying fragments and identifying components. To make the matter explicit, for the purposes of this document:

- A **fragment** is a piece of content that is defined by its <isPartOf> relationship to a larger aggregation of content;
- A **component** is a piece of content that is identified in its own right as a 'first class object'[7] although it still may have one or many <isPartOf> relationships.

Either is of arbitrary size.

The justification for identifying something as a 'fragment rather than as a component is primarily related to the ability easily to 'resolve' its identifier to the identity of the higher level aggregation – usually because this aggregation identifier is explicit within the fragment identifier, with no need to resort to external metadata. For this reason, fragment identifiers are sometimes termed 'parent:child' identifiers, because the identity of the parent is evident within the identity of the child. It is possible that there could be a justification for doing this in the case of product identifiers; however, doing so leads to the slightly paradoxical situation that the same product fragment may have multiple identities (if it <isPartOf> more than one product aggregation). We think it is highly unlikely that there is any justification for using fragment identifiers in the case of other entity classes (although there is frequently a requirement to maintain in metadata a proper relationship with an entity from which a component has been extracted[8]).

This isn't necessarily an argument against using fragment identifiers in any circumstances, but we would only be able to make the argument for their use in the context of a given use case (or set of use cases).

7 Intelligence and Uniqueness

Fragment or parent:child identifiers have some degree of 'intelligence' (or internal semantic content). So can product identifiers – the prefix parts of the ISBN have well known (if poorly understood) meanings. We hesitate to re-enter the discussion of 'intelligence' in identifiers, which was discussed in detail in a much earlier document[9]. That paper argued that:

> We have not yet entered the 'information society'; we are only just approaching its threshold. Much of the content in which our real customers are interested is not available in digital form – and much of it may never be unless and until real demand is established. In the meantime, users will identify content by reference to physical manifestations of that content – printed sources. To facilitate this without ambiguity, it is essential to develop numbering systems which have a high degree of 'affordance'. This rather curious technical term, which seems to spring from the definition of 'afford'

[7] First class object = "one that has an identity independent of any other item".
[8] The significance of "relationship" in metadata management should never be underestimated. The <indecs> framework (*op cit*) defines an item of metadata as: "a relationship that someone claims to exist between two entities". Insufficient work has been done on the standardisation of relators, although this has been recognised as a potential work item by ISO TC46/SC9 (responsible for all the familiar ISO identifiers including ISBN) for several years.
[9] Green B & Bide M (1999) Unique Identifiers: a brief introduction (revised edition) http://www.bic.org.uk/files/pdfs/uniquid.pdf

meaning ‘to supply from its own resources’, in this instance describes the ability of the end user to construct a unique reference number from the physical product or from a bibliographic record. This must truly be an ‘intelligent’ number.

In essence, we are convinced that there is no single answer to this debate on identification numbers and that attempts to seek a single solution are as likely of success as the search for the unicorn. On the other hand, we cannot support an uncontrolled proliferation of standards and quasi or would-be standards essentially seeking to answer the same problems. What we need to establish is the smallest possible number of universal standards able to answer the challenges of trading in digital content that we can currently identify.

Arguments in favour of the introduction of a degree of intelligence in numbering systems are probably now harder to make than they were 12 years ago (and the first version of the Green & Bide paper is considerably older than that). Nevertheless, globally unique identifiers – particularly if they are relatively compact – may require a structure which includes a ‘prefix’ (the equivalent of a namespace) for each issuer of identifiers (whether these are different organisations or different parts of the same organisation). These uniqueness mechanisms should be no more than that – something the International Digital Object Identifier Foundation (IDF) stresses with respect to DOIs, repeatedly making the point that a DOI should always be treated as an indivisible “opaque string” – but there is always some temptation to read intelligence into these identifiers (even though the implicit metadata encoded within them can frequently be incorrect).

There are other ways of achieving uniqueness that should be considered:

1. The use of a centralised ‘minting’ authority, which provides unique identities in the appropriate format on demand: this is a realistic option in a world of ubiquitous connectedness and web services. This is broadly the approach taken by the ISTC;
2. The use of arbitrary random strings of sufficiently large size: if the string is large enough *and is genuinely random* the odds against two systems minting the same identity becomes so small as to be entirely negligible.

Our task here is not to select the solution but simply to lay out some of the options.

8 What classes of media need to be identified by book and journal publishers?

The answer to this question is “any and all”. In our exploration of this topic, we listed:

- Text (without or with markup);
- Structured and unstructured data;
- Graphics (including photographs);
- Audio;
- Audio-visual;
- Executable software.

A component or a fragment might be any one of these. It could also (at a higher level) be any aggregation of any arbitrary combination of these media types. To fulfil the requirement, it must be

possible for publishers to identify components of any of these types at the appropriate level(s) of abstraction.

9 A note on different models for the identification of Works

Shortly before the publication of the <indecs> framework (see above), IFLA (the International Federation of Library Associations) published a similar substantive analysis in a document called *Functional Requirements for Bibliographic Records* (FRBR, usually pronounced ˈfɜrbər) [10]. There was considerable consonance between the two approaches recommended, but one fairly critical distinction exists between the two which needs to be noted here.

FRBR recognises four classes of resources:

- Works;
- Expressions;
- Manifestations;
- Items.

This looks very similar to the <indecs> framework but differs in a crucial respect – the term ‘expression’ is used to represent a very different concept in the two schemes.[11] In FRBR, works and expressions are seen as having a hierarchy; so the work might be *Romeo & Juliet*, with the play, a screenplay and (perhaps) *West Side Story* being seen as ‘expressions’ of *Romeo & Juliet*. <indecs> doesn’t see these as works and expressions existing in a hierarchy, but rather as a network of works which are all at the same level of abstraction (where works are not related in a hierarchy, but in a peer network).

The International Standard Text Code (ISTC), the standard designed to identify textual works, follows the <indecs> model.

10 Existing standards

There are several existing standards which need to be taken into account in any analysis.

10.1 ISBN

The role of the ISBN as a supply chain product identifier does not appear to be in question. The standard is designed to be able to manage products at any level of granularity. It is possible that a case could be made for parent:child identification of products (see above), particularly in solving some aspects of product metadata management. However, we already know about the difficulties that can be caused if ISBN is used to identify anything other than supply chain products.

10.2 ISTC

ISTC is proving slow to gain adoption. While it has (in the text of the standard at least) precisely the same advantage as ISBN in terms of arbitrary granularity, it is largely being promoted only for the

[10] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[11] It is important to understand that this isn’t a disagreement about what the ‘expression’ means. In any formal model, the word (‘lexical token’) used to represent a particular defined concept means what the modeller says that it means.

identification of works which are the precise equivalents of products. This would not allow it to be used for many of the Use Cases identified in this paper (see Sections 11 and 12 below) where the entities are small components rather than whole products. However, a significant number of work-to-work relators are already included in the standard, and these could relatively easily be extended if need be.

But the primary challenge is that ISTC is only applicable to textual works (or primarily textual works). As we have seen, we need to be able identify works of any type. ISTC could therefore be a contributor to this process but as currently constituted could not be a complete solution.

10.3 DOI

The DOI is already, of course, widely used in scientific and academic publishing through the CrossRef registration agency. Now that the DOI has been approved as an ISO standard (although it has yet to be officially published as such), it might be a serious contender for providing wider identity services for publishers' assets in exactly the same way as it is now being used for identification of audiovisual assets in the EIDR initiative.[12]

Essentially the DOI can be used to identify *anything*. Note that it is not even limited to identifying digital objects, as its name might be taken to suggest[13]. However this strength is also a weakness. In order to be useful, the DOI needs a Registration Agency offering domain-appropriate services – not only around registration but also around resolution and use of the identifier. So for example, although CrossRef delivers invaluable services to its members in the provision of citation linking, this is just one application and has no direct relevance to the type of services that might be required around a different domain such as asset management.

10.4 Other identifiers

There are a number of other identifiers of relevance to this discussion. These include other standards from ISO TC46/SC9, including:

- **ISWC:** the international standard (musical) work code, the music industry's equivalent to the ISTC; this is extremely well established because of the requirement for the collective management of primary rights, which is universal in music publishing.
- **ISRC:** the international standard recording code, used by the record labels to manage the identity of recordings. Has suffered from inconsistent implementation for a considerable period; about to be substantially revised.

[12] The Entertainment Identifier Registry www.eidr.org. This is an extract from the EIDR FAQs:

> At the most basic level, an EIDR is registered for a 'work'. EIDRs for new objects can be derived from it or associated with a previously registered object. Each object is referred to by its EIDR and has metadata associated with it. Some example levels are:
>
> - Basic objects from which inheritance flows, such as a movie, series, TV show, or commercial.
> - Derivative and related assets, such as edits, language versions, clips, and trailers.
> - Instances of an asset, or manifestations which include encoding information.
> - Composite works, such as mash-ups or sequences of clips.
> - Adjunct material, such as alternate content.

[13] It is a digital identifier for an object, not an identifier for a digital object.

- **ISAN & vISAN:** the international standard audiovisual number; the vISAN is designed to be used to identify 'related entities' and is a parent:child identifier. The ISAN is used not only for movies but also for games. We are currently unclear about the extent of its implementation (although an ISAN is a required data field in the metadata in the Blu-ray standard).
- **ISNI:** although perhaps not directly relevant to the topic, the ISNI (a name identifier for parties associated with any type of creation – authors, composers, publishers) is potentially going to have a significant impact on identity management in publishing.

Also of note (since it has been suggested that publishers need something similar) is the Global Release Identifier (GRid)[14]; this is an industry (non-ISO) standard, administered by IFPI[15]. Its purpose is to provide an identifier for an aggregate of assets ("a release") disconnected from any particular product. From the GRid Handbook:

> GRid provides a system for the unique identification of 'Releases' of music over electronic networks, so that they can be managed efficiently. A Release is defined precisely in the Standard but can be understood as a collection of recordings or other media that are grouped together for commerce. Products can be made from Releases by, for instance, choosing a technology to encode the recordings (such as MP3 or AAC) or a business model (such as sale or rental).
>
> By assigning a unique GRid to a Release, it can be identified without ambiguity in, for instance, reports of sales of products based on the Release.

Individual products are characterised by a unique combination of Grid plus other metadata, and may also be allocated a supply chain identifier such as an EAN or UPC.

11 What levels of abstraction need to be recognised as discrete entities?

In our workshop, there was complete agreement that publishers need to be able to identify:

- **Works**: these are the fundamental building block of any content identification system. Almost all rights subsist in works; this means, for example, that royalties or fees are normally payable for the use of the work, not for the use of a particular manifestation of the work.
    - Note that when defined like this, a work is considerably more generalised than an ISTC work
- **Products:** things that are available for sale, although the requirement for unique identification of a product may be limited to those instances where a particular product is specifically merchandised: there are certainly cases where products may be customised for a unique transaction where the product itself does not need identification (only the transaction itself and the content components at 'work' level).
- **Instances**: there is a clearly a need to be able to identify uniquely the particular file or files that are individual copies of a particular manifestation; this may be supported by system directory structures and file names, or by DAM system 'accession' numbers (but need not necessarily be).

[14] http://www.ifpi.org/content/section_resources/grid.html
[15] The International Federation of the Phonographic Industry, the global trade association for the recording industry.

The question of whether there is a useful entity (and thus something that requires an identifier) at a different level of abstraction between 'work' and 'product' needs further exploration, but the following view arose from the workshop.

Some 'works' are single works of intellectual property with a simple one-to-one relationship with a set of rights, and are manifested in one or more products (see Figure 2, for example).

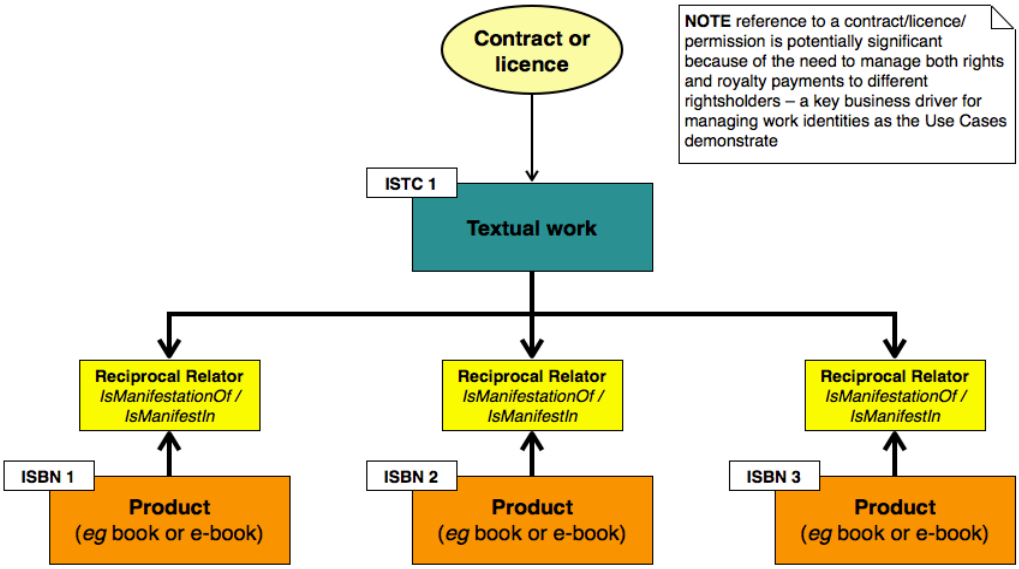


Figure 2 The simplest case – a single textual work manifested in several products

But a work is not necessarily an 'atomic' component – indeed it is perhaps more commonly an aggregate of many components, each of which might themselves be viewed as works (see Figure 3). So a 'work' may be the text delivered by an author to the publisher, over which the publisher acquires a selection of rights. Another 'work' may be a set of diagrams commissioned by the publisher to accompany the text. The photographs licensed by the publisher from a picture agency are without doubt 'works' in their own right. And the combination of the text, the diagrams and the photographs is a 'work' that can be manifested in several products (this last 'work' is a work in the sense used by the ISTC, as is the text delivered by the author – though the latter has no manifestation).

Based on this, there are apparently two sorts of works – those works where use of the work is controlled by a single, homogeneous set of rights such as the text delivered by the author or the set of diagrams or photographs, and those works where the rights over the work are emergent properties based on the rights subsisting in the various component works contained within.

Yet the 'work' that is the text delivered by the author could be viewed as an aggregation of separate chapter-sized works, each manifested in Chapter products sold individually (see Figure 4). And the set of diagrams may consist of several individual illustrations that could be used independently within other aggregate works. So the difference between the two sorts of works is more apparent than real. The fact that one work is 'bigger' than, or is identified before another does not make it of a different class; such a distinction is entirely arbitrary. Whether the publishing rights associated with a work derive directly from a contract with the creator of the work, or are secondarily derived from rights in other works is a largely artificial distinction. On this basis, works may have many

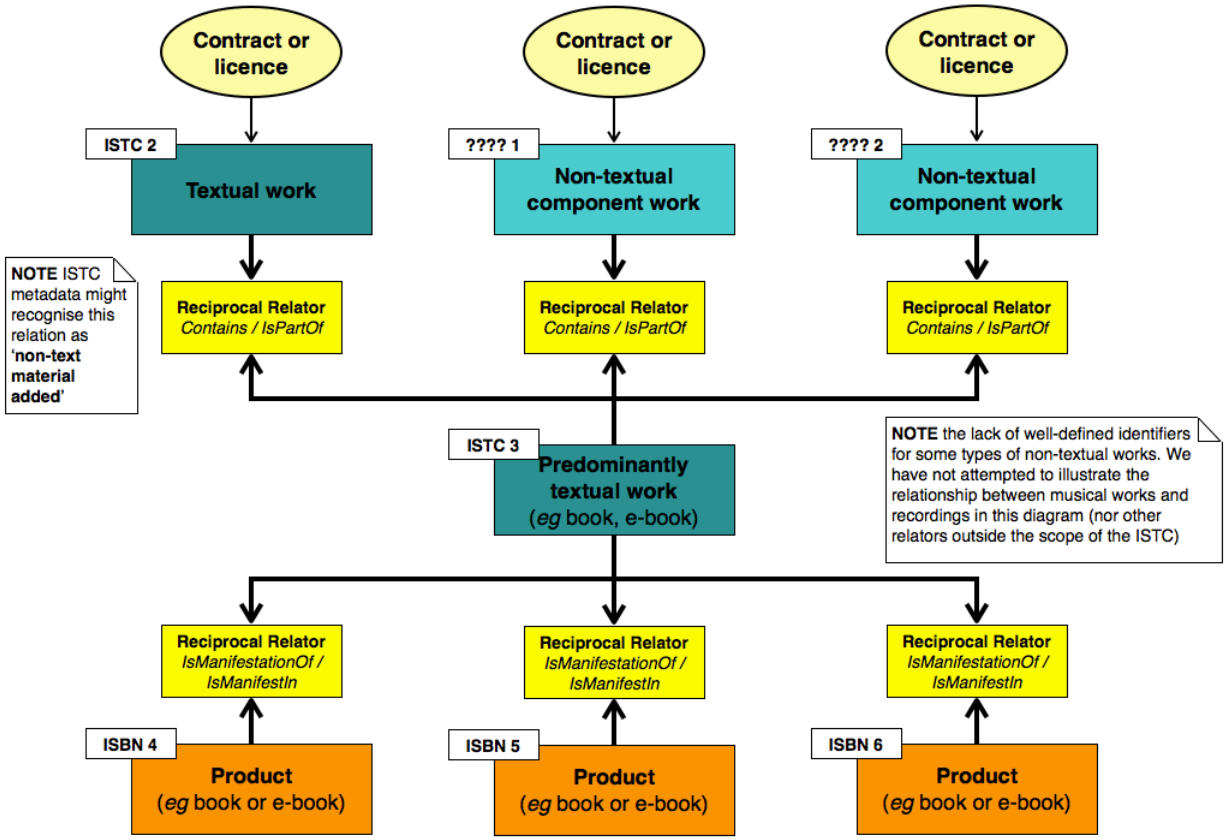


Figure 3 Probably the commonest use case: a ‘predominantly textual’ work that includes non-textual components (for example, illustrations) with separate rights. This structure can be extended to describe any aggregated work

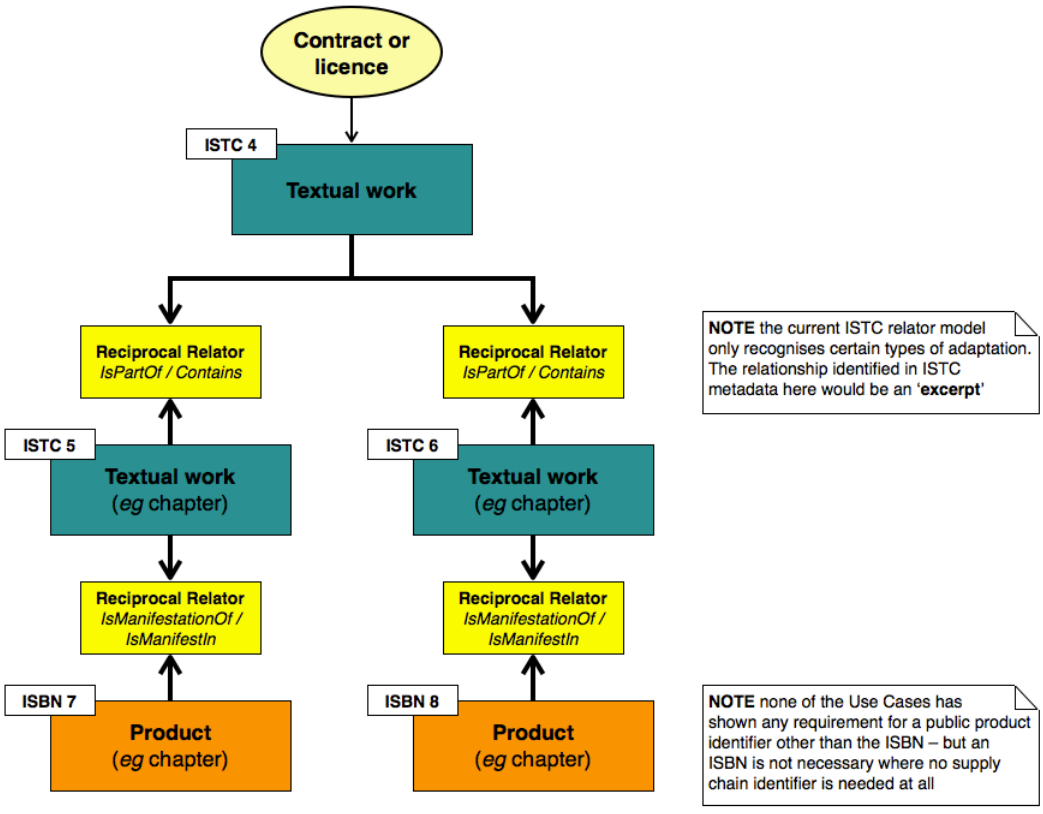


Figure 4 Another simple use case – parts of a single textual work, each identified as a work in its own right and each manifest in a separate product

relationships to other works – they can be disassembled and recombined to create other works – and these relationships are relations among peers rather than relationships within a hierarchy. Figure 5 illustrates some possible complex relationships between works, other works and products.

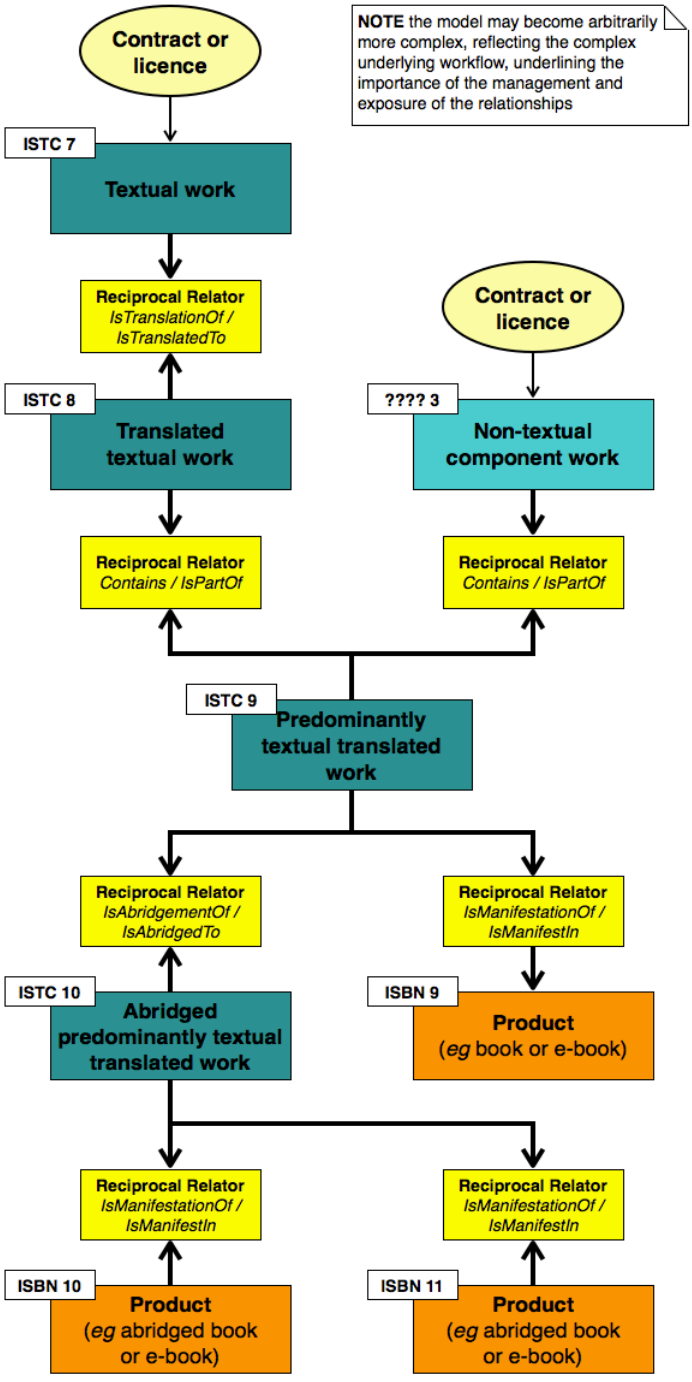


Figure 5 The model may become very complex, underlining the importance of exposing and managing the relationships.

Note that in all these diagrams, these relationships are themselves represented as first-class entities, which might have particular attributes of their own (including for example dates when the relationship was asserted, and by whom).

Note that some works have direct relationships with contracts, some have direct relationships with manifestations, some have both, and some have neither. And while there is a clear distinction

between works (which are abstract) and products (which are concrete), the case for an intermediate class (and identifier) between work and product is not at all clear – those intermediate entities are themselves works.

This line of reasoning perhaps illustrates two reasons for publishers' reluctance to adopt the ISTC. First, although it was originally promoted as an identifier suitable for use in the contracts, rights and royalties arena, the link between the work as identified by the ISTC and the intellectual property rights is not at all simple. A few ISTC works (mostly novels) might have a simple one-to-one relationship with a set of rights in a contract and a simple set of one-to-many relationships with products (as illustrated in Figure 2). But for most works, the rights associated with the work are a complex amalgamation of rights associated with other component works. Second, if you accept the view that works have complex relationships with many other works, it is self-evident that not all works are primarily textual. Publishers may be more receptive to a work identifier that encompasses *all* types of work, rather than the 'primarily textual' subset of those works that lies within the (currently) defined scope of ISTC.

We note that the recorded music industry is struggling with a similar challenge and may be reaching slightly different conclusions for pragmatic reasons (see Appendix 2). However, there is an additional level of abstraction in music – the recording (indecs:expression) of a song (indecs:work) – that complicates the model in a way that does not seem at first sight to be essential for publishers of predominantly text-based assets. Nevertheless, a similar issue exists with respect to AV material, for example with audiobooks; this will need to be taken into account in an identity model for publishers which (as we have seen) must also embrace audio and AV[16].

12 Use cases

During late 2009 and early 2010, a Working Group formed from large UK-based publishers created a set of use cases[17] under the title *Book-centric to Content-centric Publishing*. The primary purpose of these use cases was to explore the extent to which existing and future publishing models where the 'product' is highly granular were supported by existing identifier schemes, and to identify the gaps in support that would require new models of identification. Some cases were based on real-world projects, others purely theoretical. A short introduction and a summary discussion attached to the set of use cases by the Working Group is reproduced here as Appendix 3.

Broadly, the use cases described:

1. Products that were small chunks of content
    a. chunks had to be aggregated into fixed product bundles to ensure that each product was of a sensible value, or that royalties could be calculated properly;

[16] A further need for an expression entity might arise where a particular typesetting of a work becomes important. Publishers who are subrights licensees have long paid 'offset fees' for use of typesetting or page layouts created by their licensors. And this may become an issue if the e-book publishing rights are not held by the same publisher as the conventional (hardback and paperback) volume rights – one publisher would provide edited, marked-up text to the other. In both these case, what is traded between publishers is clearly an expression in the sense defined by <indecs>.

[17] *Book-centric to Content-centric Publishing, Version 1.0*, unpublished report

2. Products that were, in effect, 'access' to a corpus of small chunks of content
    a. this was only deemed possible where the content was effectively 'owned outright' by the publisher and no royalties were required, or where the rights to all chunks in the corpus were similar;
3. Products that were created through recombination of small chunks of content (with varying values and rights)
    a. this worked where the number of chunks was small enough that each could be given an identifier and a price (or at least a 'weighting') to be used in royalty calculations.

The use cases presented by the Working Group illustrated the flexibility of the current ISBN system, but served to clarify the fact that the ultimate granularity of commercially viable products is limited at least as much by relatively high transaction costs as by the costs[18] of maintaining an identifier such as the ISBN. As content is sold or licensed in ever-smaller chunks – from book chapters or articles down to individual learning objects, for example – the cost of each transaction tends to become greater than the value of the product. Very small, low value chunks of content need to be formed into bundles to ensure the per-transaction cost is relatively low compared with the overall transaction revenue. The Working Group did not uncover any case where *economically-sized* content chunks require new types of product identifier.

And the Working Group recognised that where a product is created in a way that is likely to make each instance unique – for example where the purchaser selects content from a large selection of content chunks to create something bespoke – then the 'product' need not have an identifier at all that is persistent beyond that particular transaction. There is no need to identify all possible products that might be created (a requirement which would overwhelm the capacity of the entire ISBN system with only thirteen chunks in total[19]).

The business requirement in this last case is to track usage of the various chunks from which the purchaser might select. *For the use cases presented*, the total number of chunks was not so large as to make the use of individual internal identifiers impractical, and for practical purposes the use cases envisaged using ISBNs (since they were concerned with whether such a use case would be practicable with various legacy IT systems). On the one hand, these chunks could be considered 'works', which would clearly make the use of ISBNs inappropriate, but on the other, the 'product' purchased (a particular selection of chunks) could be viewed as a bundle or pack of multiple products.[20]

However, the study did conclude that, "while ISBNs identify individual traded products, and ISTCs identify works (bundles of content) which may be manifested in products, there is no common identifier for the contracted units of content (chunks) which comprise those bundles." In the Working

[18] Note these 'costs' are mostly the costs of maintaining the necessary metadata, not the cost of the number itself.
[19] There are nearly 17 billion possible non-empty ordered subsets of a set with 13 members.
[20] It is a clear shortcoming of legacy rights and royalties systems that the link between product sale and the royalty that accrues is based solely around the ISBN. There is no choice, and the use of ISBNs as chunk identifiers is forced, even when individual chunks are not available as products in their own right. It means ultimately that ISBNs are being used in two quite separate roles, both as a product identifier, and as an internal work identifier (a similar issue arises with terms such as 'head ISBN' or 'title ISBN', where the identity of the first published product is used to identify the work). And if these internal work identifiers are not carefully kept internal, they appear to the remainder of the supply chain as if they were products.

Group's use cases these chunks ranged from individual learning objects (small items of text, individual diagrams, animations, video files, and so on) through database entries in a corpus of data about birds, to online language learning tests.

The study identified several requirements for a chunk identifier:

- It should uniquely identify an 'atomic' chunk of content, below which the content cannot be sliced up and recombined;
- It would be equally applicable to text, still and moving images, sound, executable code, any form of 'content';
- It needs to be stored in the contracts, rights and royalties system and associated via a contract a set of rights and royalty obligations, to promote visibility of those rights and royalties;
- It needs to be stored in any digital asset or content management system, to permit simple retrieval
- It needs to be handled in any sales system, so that every sale can be decomposed into a set of chunk sales when necessary;
- Ideally, the identifier would be persistently embedded in each chunk of content itself, so it can be discovered from examination of the chunk (however this may not be possible with most current digital file types);
- It would not need to have the typical bibliographic metadata associated with an ISBN (title, pub date and so on) since the chunk would potentially be associated with many ISBNs, but would need enough associated [Reference Descriptive] metadata to permit discovery.

And yet, as argued in Section 11 above, there is no strong conceptual difference between works as bundles of content which may be manifested in products – the view expressed in the ISTC – and works as content over which there is a set of homogeneous intellectual property rights – the 'atomic' chunks or contracted units of content highlighted by the Working Group. A novel might contain only a single 'work', and a single learning object may comprise several 'component works'. Thus the use cases do highlight the need for a work identifier that can be applied to any type of content, at scales from individual chunk to 'whole book'.

13 Recommendations for further work

Quite deliberately, this report has no conclusions and no specific recommendations with respect to solutions. That is not its purpose. Rather, it is intended to provide a background of what we hope will be common understanding of a set of high-level requirements against which future work can be undertaken.

We have some specific suggestions as to this future work:

1. The Use Cases that are described in Section 12 and Appendix 3 were compiled between 12 and 18 months ago. In terms of publishers' thinking about digital products this is quite a long time ago and we recommend that the Use Cases should be updated and extended. In the UK, this work might best be undertaken by the original working group, by the BIC *Future of Metadata* committee or another BIC-convened group; however we also recommend that it should also be extended internationally if at all possible (for example through the work of the BISAC's *Identification Committee* as well as similar groups in other countries). At the same time, we recommend that the Use Cases should be extended from trade and educational publishing (the original focus) to a wider consideration of other sectors – particularly academic and scholarly

publishing (in the latter case, for example, to bring in journals).

2. At the same time, we recommend that the benefits of standardisation of identifiers in this context (see Section 5) be more widely considered, and that the arguments for and against the implementation of formal or informal standards in the management of assets fully explored by publishers in consultation with their system vendors.

3. It is clear that application of the ISTC in this context would require some rethinking of current ISTC strategy and technical scope. The current value of the ISTC is as a collocator for multiple products, but this value is constrained by the lack of emphasis on and exposure of the work-to-work links that can broaden the scope of that collocation (eg 'select all manifestations of this work, its parent work and its sibling works' would collocate *The girl with the dragon tattoo* with *Män som hatar kvinnor* (the original in Swedish) and *Verblendung* (the German translation)[21].

   Any change to strategy and scope is, of course, a decision for the Board of the International ISTC Agency and our recommendation is simply that the implications of this report should be considered by them and due consideration given to the requirements that this report identifies. The International ISTC Agency has indicated that it is currently considering proposing an extension to the scope of the ISTC, when the standard is formally reviewed, with a view to including non-textual works that 'lie within the domain of book publishing' (admittedly, this is a difficult concept to define). Such an extension has the potential to enhance the applicability of the ISTC to the management of components.

4. Similarly, there may be opportunities for the deployment of DOI in the management of publishers' assets. The nascent EIDR may provide a model here. However, this would be dependent on at least one DOI Registration Agency creating services appropriate to the domain that this report describes. In the same way as for the ISTC, we recommend that this is given due consideration by the International DOI Foundation and its membership.

5. At least some of the Use Cases already described involve products which have variable content (and – in at least one case – prices which are dependent on the precise content selected by the user). We recommend that the International ISBN Agency consider the implications of the application of ISBN to such products (particularly in the event that they are made available through a supply chain rather than direct from the publisher) and develop its policy appropriately.

Ultimately, the direction to be taken beyond this point depends on the reception afforded to these recommendations. It is conceivable that all that might be required is the creation of some Best Practice guidelines for minting proprietary identifiers (which, if international in scope, would be the domain of EDItEUR). It is however equally conceivable that much more extensive community standardisation work may be required, focused on either existing identifiers such as the DOI or ISTC, or perhaps something new.

[21] The International ISTC Agency has indicated that such links will be exposed in a near-future version of the ISTC website (http://www.istc-international.org)

14 Appendices

Appendix 1: mechanisms for the identification of 'fragments'

[EDItEUR is grateful to Norman Paskin of the International DOI Foundation for permission to include the following section of this document, which is a redacted version of a document written by him in 2010 but never published.]

A fragment identifier is a string that refers to a resource that is subordinate to another, primary resource. The fragment is not a first class object but instead its identity is defined as a sub-set of the primary resource. A problem raised by fragment identifiers is the existence of an infinite set of possible ad hoc identifiers from one base primary resource (e.g., time ranges in a video). And of course for most people today 'fragments' is used in one specific sense (http) – the piece of a URL that the server doesn't really know about and that the client hangs on to and then processes the html returned to get there or do the right thing (this is a function of the hypertext model that was initially selected for http/html – it's at the file level so to get to some specific point required a second mechanism). In the internet, fragment identifiers are well understood in principle, but not uniformly dealt with: the article http://en.wikipedia.org/wiki/Fragment_identifier gives a good overview and lists some specific proposals; among these I find of particular interest is RFC 5147:

- IETF RFC 5147 *URI Fragment Identifiers for the text/plain Media Type*. http://www.rfcarchive.org/getrfc.php?rfc=5147 "This memo defines URI fragment identifiers for text/plain MIME entities. These fragment identifiers make it possible to refer to parts of a text/plain MIME entity, either identified by character position or range, or by line position or range. Fragment identifiers may also contain information for integrity checks to make them more robust". RFC 5147 proposes a fragment identifier for text/plain documents based on character and line positions and ranges within the document using the keywords 'char' and 'line': e.g. http://example.com/document.txt#line=10,20 identifies lines 11 through 20 of a text document. Hence it has more affordance[22] than the ISMC proposal, but is more limited as it deals only with [precomposed] text. RFC 5147[23] is therefore not identical in scope, but somewhat similar in concept to the idea of the ISMC.
- W3C has a draft specification for Media Fragments: http://www.w3.org/TR/media-frags/ – this is restricted in two senses: (1) it specifies only use of http; and (2) the specified addressing schemes apply mainly to audio and video resources – the spatial fragment addressing may also be used on images. The Media Fragments 1.0 specification, still a working draft, specifies the syntax for constructing media fragment URIs and how to handle them when used over the HTTP protocol. The syntax is based on the specification of

[22] Affordance = "the ability to generate a syntactically correct identifier from content-in-hand".

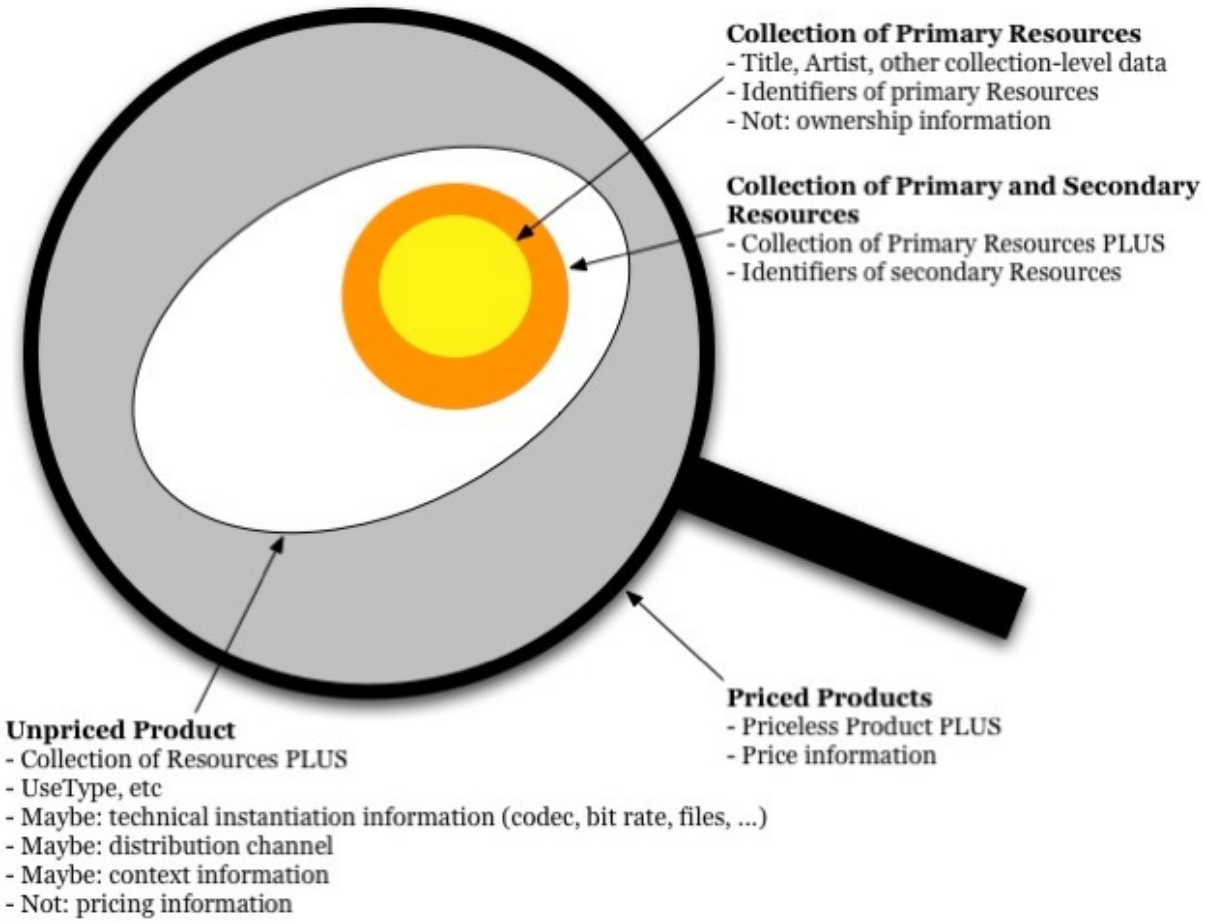
[23] RFC 5147 is a "Standards track" RFC from April 2008, but as far as [NP] can tell it's actually no more developed than an "informational" RFC and so has no particular special standing. Unlike ISO, the RFC process has many "standard track submissions" that are never taken further. I cannot find any evidence of RFC 5147 being adopted or supported. The RFC Standards track is not a particularly rational process: TCP/IP, for example, never was a standard and it is used trillions of times every day. RFC 5147 purports to update 2046, which is the MIME standard from 1996 and is still listed in Proposed Standards despite the fact that it is used in every http header every day.

particular field-value pairs that can be used in URI fragment and URI query requests to restrict a media resource to a certain fragment.

- Other: Handle 7.04 deals with potentially infinite fragments by introducing a delimiter, with the base as a registered handle [an identifier of the primary resource], and defining a transformation on any possible tail. The template handle construction makes use of <template> tags in XML-structured handle values. When a server receives a resolution request for a handle which is not in its database, it determines if there is template for constructing the handle values; if so the server looks up the base handle (i.e. the part before the delimiter) and adds the part after the delimiter from the template XML <value> tags defining the handle values of the result. Hence infinite fragments can be managed as they are created, through templates built on the primary resource.[24]

[24] A single template handle can be created as a base that will allow any number of extensions to that base to be resolved as full handles, according to a pattern, without each such handle being individually registered. This would allow, for example, the use of handles to reference an unlimited number of ranges within a video without each potential range having to be registered with a separate handle. If the pattern needs to be changed, e.g., the video moves or a different kind of server is used to deliver the video clips, only the single base handle needs to be changed to allow an unlimited number of previously constructed extensions to continue to resolve properly. When a server receives a resolution request for a handle which is not in its database, it tries to determine if there is template for constructing the handle values. It is planned to use this in the DOI application to videos in the EIDR. [www.eidr.org]

Appendix 2: possible directions in the music industry



DDEX, the nearest equivalent to EDItEUR in the music industry, is currently undertaking research into the identification of products and of the aggregations of components which go to make up products. This graphic shows the different entities which are under consideration.

- The GRid identifies the resources (although makes no distinction between primary and secondary resources
- An ‘unpriced product’ appears to be an entity at the same level of granularity as the ISBN (in the sense that there is no change in the ISBN related simply to a change in price).

Appendix 3: Book-centric to Content-centric Publishing: extracts from Working Group documentation

From introductory text supplied to the Working Group, Sept 2009

For the most part, publishers allocate new ISBNs to digital products in the same way as to conventional physical products […]. And at least in the UK, most publishers follow industry best practice in allocating different ISBNs to – for example – each different format of e-book made available at retail (ePub, Mobipocket and so on).

In the past, there was a worry that the ISBN system would not be able to cope with the 'explosive' growth in the number of new product identifiers required every year. [But] recent modification to the governance of ISBN has increased the flexibility of the existing system: ISBNs can legitimately be applied to chunks of text at chapter level, and can be allocated by entities other than the publisher (eg an e-book conversion company may allocate ISBNs to specific formats of e-book), if required. Yet the 'cost' of each ISBN remains. This is not the cost of each individual number – though for very small publishers or very low-value products, this too may be a concern. The 'cost' of the ISBN is the cost of managing and maintaining the metadata associated with that ISBN.

However, future publishing activity will be concerned both with selling products like traditional books, and with selling much smaller packages or chunks of content and recombinations of those chunks. Is there a requirement for a lighter weight identifier for such small chunks of IP? And how should recombinant products be handled?

The purpose of the work is to draw out a number of use cases where identifiers beyond those we already in use might be required. There may be experience from other industries or sectors of publishing that can help guide our developments.

From summary text supplied with the final use cases from the Working Group, May 2010

For any content-centric organisation, identification of content at the 'atomic' level is an important missing piece of the identifier jigsaw. Individual chunks of content may be acquired separately, may have different associated rights and royalties associated, and may be combined and recombined with other chunks to form various products. Yet tracking rights and royalties from the contract to the product is often impossible. Typical chunks of content might include manuscripts, but might also include photographs and illustrations, additional text content, and in future might also include various digital-only objects including as video, audio, and other interactive and non-interactive digital objects.

Traditionally, publisher's contract, rights and royalties systems store contracts, and associate products (ISBNs) with those contracts. Each ISBN therefore can be linked with a set of rights (which may be territorial, time-limited, constrained by product format *etc*) and a set of royalties payable when those rights are exploited. A single contract might cover several manuscripts – or more generally, the contracted unit of content – each of which might be manifested in several products (*ie*

hardback, paperback, e-book *etc*) – but at heart, there is a clear link between an ISBN and the contract it is derived from. The contract specifies the rights that the publisher is exploiting, and the royalties due.

Note that in principle, the contracted unit of content – manuscripts or works – might be identified by an ISTC, but in only the simplest cases (*eg* novels) would there be a clear one-to-one relationship between the contracted unit of content and an ISTC ‘work’.

Outside of simple fiction, it is much more likely that any ISTC work (and any ISBN products that are manifestations of that work) would comprise more than one contracted unit of content – a manuscript, perhaps some illustrations or photography, perhaps a separate unit of text content for a foreword. A digital product might contain all of these plus other digital-only content and objects. Each of these contracted units of content might be associated with a separate contract, separate contributor, separate royalty arrangements and so on.

Typically at present, publisher’s contract, rights and royalties systems deal only with the primary textual or illustrative content of the works. Other chunks of content used to supplement the primary contributions might be handled separately: photos licensed from a picture agency, a contributed foreword *etc* are usually not handled in the rights and royalties system, and are likely to be licensed via one-off payments. There is no strong linkage between these separate deals and the products.

A typical problem might be that the publisher holds World rights in the primary contributions to a work, but only licenses the images for Commonwealth use because the publisher only plans to distribute the product in its traditional Commonwealth markets. Later, the publisher decides to do an e-book version that can easily be made available globally, and this leads to an inadvertent breach of the image licence. The alternative is to acquire much broader licences for the images – wasteful if done up front (before clear plans for the global product have emerged) and potentially expensive if done reactively.

The current system works well enough for publication of works with a single or small and fixed list of primary contributors, and with a small amount of additional content that can be handled manually. However, it is not well suited to a publishing model where the publisher may combine and recombine several much more granular contracted units of content in multiple ways to create new products.

In this model, each contracted unit of content (chunk) might be of relatively low value, but each product created and sold comprises many chunks. While each chunk could in principle have a different set of rights and royalties associated with it, the overall rights the publisher can exploit in a particular product depends on the exact combination of chunks in that product – the product rights would be the ‘inner join’ (the Venn diagram overlap) of all the chunk rights.

So the identifier problem is this: while ISBNs identify individual traded products, and ISTCs identify works (bundles of content) which may be manifested in products, there is no common identifier for the contracted units of content (chunks) which comprise those bundles.

If publishers produce new products, including bespoke products created on-the-fly for individual purchasers, by recombining chunks, there needs to be a chunk identifier that can be used to link

what is offered for sale or sold (a potentially unique combination of chunks) back to the contract(s) covering those chunks.

Note that this series of case studies has not yet surfaced a specific case where chunks cannot be handled within current systems – either by identifying each chunk with an ISBN, or by greatly simplifying the rights and royalties – but there is little doubt that such a case will arise at some point in the future.

What would an identifier need to do?

- It should uniquely identify an ‘atomic’ chunk of content, below which the content cannot be sliced up and recombined
- It would be equally applicable to text, still and moving images, sound, executable code, any form of ‘content’
- It needs to be stored in the contracts, rights and royalties system and associated via a contract a set of rights and royalty obligations, to promote visibility of those rights and royalties
- It needs to be stored in any digital asset or content management system, to permit simple retrieval
- It needs to be handled in any sales system, so that every sale can be decomposed into a set of chunk sales when necessary
- Ideally, the identifier would be persistently embedded in each chunk of content itself, so it can be discovered from examination of the chunk (however this may not be possible with most digital file types)
- It would not need to have the typical bibliographic metadata associated with an ISBN (title, pub date and so on) since the chunk would potentially be associated with many ISBNs, but would need enough associated metadata to permit discovery in a publisher’s digital asset management system

No such identifier exists – though proprietary schemes are simple to devise. But ideally, the chunk identifier would be standardised rather than proprietary, to facilitate exchanges of chunks between publishers and to allow systems vendors to offer off-the-shelf systems to manage the chunks.