

Unique Identifiers: a brief introduction

by Brian Green and Mark Bide (February 1999)

- [Some overall considerations](#)
- [A crucial debate - "intelligent" and "unintelligent" numbers](#)
- [What are we identifying anyway?](#)
- [Granularity](#)
- [What do we need from unique identification systems for content?](#)
- [The Serial Item and Contribution Identifier \(SICI\)](#)
- [The Book Item and Component Identifier \(BICI\)](#)
- [The Publisher Item Identifier \(PII\)](#)
- [The Common Information System \(CIS\) and the International Standard Work Code \(ISWC\)](#)
- [Internet developments: URNs and PURLs](#)
- [The Digital Object Identifier \(DOI\)](#)
- [Conclusions](#)

Some overall considerations

In recent months, there has been an explosion in interest in and comment about methods of uniquely identifying "content" in the digital environment. In parallel, we seem to be faced with growing complexity in the arcane discussions surrounding the solutions to the various challenges which unique identifiers pose within the publishing community as a whole. This in turn has led to growing perplexity among those who are not intimately involved.

In this brief document, we seek both to explain some of the significant issues in the debate and to describe some of the standards, *de jure* and *de facto*, which are currently proposed. First, though, in case it is not immediately obvious, we should explain why BIC and EDItEUR are taking an active interest in the establishment of identifier standards.

Identifiers are critical to the successful implementation of all forms of electronic commerce. It would be impossible to imagine how any form of electronic ordering of books could have been successfully implemented without the universal acceptance of the ISBN. In the digital environment, as the realistic potential arises to trade in fragments of works smaller than the book or the journal issue, a similar universal system will become essential if any sort of trading in content is to survive. From our point of view, the debate on unique identifiers has far-reaching implications on many different areas of activity affecting our entire constituency - EDI transactions (including rights transactions); electronic tables of contents; bibliographic and product information; Copyright Management Systems. There are other issues which may be more limited in scope including, for publishers, their internal administrative and production tracking systems and, for libraries, aspects of bibliographic control.

A crucial debate - "intelligent" and "unintelligent" numbers

There is one theoretical issue which we need to discuss before moving to specifics. This is the crucial debate on whether a numbering system should adopt intelligent or unintelligent numbering. This is probably best described by reference to what an unintelligent numbering scheme is: this is a purely random number which can only be interpreted by reference to a central database; examining the number itself tells you nothing about the object which it identifies.

All other numbers have some degree of intelligence. Just how "intelligent" the number is another question. In our context, the most obvious example of an intelligent number is the ISBN. The first part of the ISBN identifies the country, language or geographic area in which the book was published; the second part identifies the publisher to whom it was issued. It is of course the case that the country in which an ISBN was issued is not a reliable indicator of the language of the book. It is equally true that when book lists are sold from one publisher to another, the original ISBN is often used for many years after the transaction has taken place. This is one of the key reasons why EDI in the book trade has depended to a considerable extent for its success on the "clearing house" exemplified by TeleOrdering, where the purportedly "intelligent" ISBN is compared against a complete database of numbers to test the veracity of the intelligence it carries.

There is another generalised criticism of the ISBN and similar "intelligent" constructs in the digital environment. Other interest groups, particularly authors' groups, see them as being "publisher-centric", failing to identify the underlying rights owners in the content. This may be largely irrelevant in a world where physical goods are being traded, where the necessity is to identify the source of those goods. It clearly becomes more crucial in the digital world where the publisher of the physical product may have no electronic rights or where rights originally owned by a publisher may have been transferred - or reverted to the author.

The trend in information technology in general is away from intelligent to unintelligent identifiers, and the reason is obvious. It is indeed hard to imagine a system where, in the long term, numbers with any real intelligence can be maintained as a way of identifying content, particularly bearing in mind that the types of content to be described are not only those in which publishers have a traditional interest - text and graphics - but also encompass sounds and moving pictures, even computer code.

However, we have a problem which those who would whole-heartedly abandon all intelligent numbering systems immediately overlook at their peril. 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" 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.

What are we identifying anyway?

This question also goes to the heart of the debate from the publishers' point of view, and it is essential to recognise the complexities which this issue creates. Publishers may be inclined to see the "work" to be identified as analogous to a book: it can therefore be identified in a similar way to that which the ISBN represents. However, that this does not hold good even for books can best be demonstrated by reference to literary classics: the ISBN of a particular edition of *Pride and Prejudice* does not uniquely identify the content, simply a particular manifestation of that content. That manifestation may, of course, have unique features - in terms both of physical layout and of interpretation.

A similar problem exists in the world of music; there it has been vigorously tackled because of the multiplicity of rights issues which arise in performance, recording and broadcasting.

This further underlines another fact about identification schemes in a digital environment which the publishing industry will do well to recognise. The comfortable dividing lines between different types of content are fast disappearing which emphasises the need for close co-operation between different sectors of the broader "content industry". This may be difficult to achieve but it is essential. There is no room for rivalries or the "not invented here" syndrome. For example, there are many things which the publishing industry can profitably learn from the unique identification schemes which the international music industry is adopting and much to be gained from working to develop at the least a compatibility of approach to the same or similar problems.

Granularity

Among these may be ways of tackling the problem of uncertain granularity. To what level of detail does content have to be identified?

The ISBN identifies the whole book; the SICI identifies the journal issue and, appropriately extended, the individual article within the issue. This may be enough for some uses but is clearly inadequate for others. If we are to be able to identify all rights owners in a particular piece of content, that may require a far finer degree of granularity

of identification, to the level of the individual illustration or quotation from another source. Similarly, if information is to be traded with customers at a level of granularity finer than the "chapter" or the "article", then publishers may have compelling marketing reasons for being able properly to identify and to keep track of what is being traded.

The level of granularity which *may* need to be identified becomes effectively arbitrary in a digital environment. This is another essential pointer in the direction of what will eventually be required from unique identification standards. It might suggest a requirement for relational identification where (like the SICI) smaller fragments are identified by reference to the larger "whole" from which they come, although this would have some drawbacks, not least in terms of the size and structure of the codes.

What do we need from unique identification systems for content?

Before we look at what is on offer, let us first then draw together what we see as the essential threads in our arguments as to what is required. We must first, though, re-iterate that we do not believe that any single standard is capable of meeting all of our requirements; we will need to live with several "layers" of identification standards, certainly for the foreseeable future, possibly in perpetuity.

1. We need to be able uniquely to identify "content" for a number of different reasons. We can identify as a minimum the following:
 1. to facilitate the trading, between publishers and customers, of "fragments" of larger works
 2. to facilitate the trading, between publishers and other rights owners, of the rights in fragments of larger works
 3. to facilitate the development of appropriate Electronic Copyright Management Systems, to control the use of, as well as the rewards for, content distributed on networks
 4. within publishing houses, and within consortia of publishing houses, to track pieces of "content" through the production process up to the point where they are combined into products
2. In the longer run, we recognise that the argument about intelligent and unintelligent numbers is likely to come down in favour of the essentially random, unintelligent number - the only intelligence being incidental to mechanisms devised to ensure uniqueness. In the short run, though, we will need numbering systems with high degrees of "affordance" for certain applications.
3. Identification systems should make it possible to identify both the underlying content and particular manifestations of that content.
4. The granularity required for content identification will be different for different applications; for certain immediate requirements, identification to the level of the individual journal article or book chapter is adequate.
5. Publishers should not take a publisher-centric view of the development of standards, if they are to achieve universal acceptance. Not only do we have to take into account developments in other sectors of the content industry (music, film, TV), we cannot ignore the legitimate interests of others - both the creators and the

consumers of the content which publishers supply as well as intermediaries in the information supply chain (such as libraries).

Current standards and initiatives

The Serial Item and Contribution Identifier (SICI)

SICI, the Serial Item and Contribution Identifier, is ANSI/NISO standard Z39.56. Work on the standard began in SISAC, the US Serials Industry Systems Advisory Committee, in 1983. NISO took over the work at SISAC's request and the original standard was published by them in 1991.

This original version of the SICI, ANSI/NISO Z39.56-1991, established two levels of coding, a unique code for the identification of a serial title called the Serial Item Identifier and a unique code for individual contributions within a serial, the Serial Contribution Identifier.

SICI is currently widely used, mainly still at the item (i.e. issue) level, by subscription agents and libraries. It is an important element in EDI message transactions and is used in most library systems. It is represented in bar code form (the SISAC barcode symbol) using the EAN128 symbology.

It has recently been significantly revised to make it more suitable for use at the contribution level in EDI, EToCS and as a component of Internet developments (URN).

The main changes are the introduction of a Code Structure Identifier for different uses, a Derivative Part Identifier (DPI) to identify fragments other than articles (e.g. tables of contents, index, abstract) and a Media Format Identifier (MFI) to indicate physical format. The DPI and MFI may be used in all SICI types (CSIs).

The new Code Structure Identifiers (CSI) are:

- 1) CSI 1, which includes chronology and enumeration but does not include the contribution data segment. It is used mainly as the basis of the SISAC barcode and in EDI message transactions (e.g. price lists, orders, claims).
- 2) CSI 2, which allows for the inclusion of contribution identifier including location (page number), contribution title code (the first letter of the first six words of the title), as well as derivative part identifier and media format. It can be used to identify both paper and electronic articles as the location identifier is not required if there is no page information to record.
- 3) CSI 3, which is intended to accommodate locally-assigned identifiers such as PII or Adonis numbers or a publisher's own internal number. It can be used for pre-publication items where the ISSN is known but not the volume or issue. Once published, it is expected that a document will be assigned a CSI 2 SICI.

One shortcoming of the SICI has been its confinement to serials. The imminent arrival of the 'BICI' (see below) will solve that. SICI is, however, unsuitable for use before publication if the contribution has not yet been assigned to a specific journal.

The Book Item and Component Identifier (BICI)

A book version of the SICI, currently nicknamed BICI, has been drafted by Book Industry Communication with support from The British Library's BNB Research Fund. It is very closely based on the SICI, with the ISBN replacing the ISSN and with a number of other changes, either needed because of the different characteristics of books versus serials, or designed to make the code distinctive.

The code can be used to identify a part, a chapter or a section within a chapter, or any other text component, such as an introduction, foreword, afterword, bibliography, index etc.. It can also identify an entry or article in a directory, encyclopedia or similar work which is not structured into chapters; an illustration, map, figure, table or other piece of non-textual content which is physically part of the item, or an enclosure which is supplied with but is not physically part of the item.

The Publisher Item Identifier (PII)

The Publisher Item Identifier (PII) was agreed in 1995 by an informal group of Scientific and Technical Information publishers calling themselves the STI group and consisting of the American Chemical Society, American Institute of Physics, American Physical Society, Elsevier Science and IEEE. It was developed as an identifier for internal use and exchange between consortia partners. It was closely modelled on the Elsevier Standard Serial Document Identifier and the ADONIS number, both of which it has now replaced.

The STI group defined their requirements for an identifier as follows:

- 1) Format (presentation) independent: relates to semantic content.
- 2) Capable of extension to describe differing manifestations or expressions of the same document.
- 3) Identifier is unique to a document; a document has a unique identifier.
- 4) Easy to generate and use.
- 5) Determined by, and generated by, the originator of the published item (i.e. the publisher)
- 6) Not restrictive. Able to accommodate many publication item types.
- 7) Serves only one purpose. Does not carry any "compulsory" explicit meaning other than that of unique identification.

8) Compatible with (not in conflict with) existing related standards.

It was also considered vital that the identifier could be generated at a very early stage, possibly before the article to be identified had been allocated to specific journal.

In addition to the STI group of publishers, the PII has been adopted by Springer and some other primary publishers as well as by their secondary databases including Chemical Abstracts, EMBase, INSPEC and ADONIS.

The PII is a string of 17 alphanumeric characters comprising one character to indicate source publication type, the identification code (ISSN or ISBN) of the publication type (serial or book) to which the publication item is primarily assigned; (in the case of serials only) the calendar year (final two digits) of the date of assignment (this is not necessarily identical to the cover date); a number unique to the publication item within the publication type and a check digit.

The PII is a 'dumb' number with no intrinsic meaning. The ISBN and ISSN are used as a part of the number but simply to ensure uniqueness. The PII can only be assigned by the publisher and has no affordance (i.e. it cannot be 'reconstructed' from a published article). The STI group have made it clear that they will not assign PII's retrospectively. There is to be no central registry of numbers.

It thus seems clear that although the PII is well designed to track publication items throughout their life cycle, it is unsuitable for use in ordering or claims transactions, requests for permissions or as an aid to finding published articles. It is therefore essential that any publisher using the PII is also able to accept and deal with SICIs.

The Common Information System (CIS) and the International Standard Work Code (ISWC)

The music industry and community of authors' societies have some particularly complex copyright management requirements and have devised a *Common Information System* (CIS) designed to integrate and standardise a number of key databases outlined below. Initiatives are directed by the International Confederation of Authors and Composers' Societies (CISAC) and the International Federation of Phonograph Industries (IFPI).

The *Compositeur, Auteur, Editeur* (CAE) number currently identifies the creators and publishers of music and literary texts. It is an entirely 'dumb' number which is to be extended to encompass all CISAC repertoires including visual, audiovisual and plastic arts, and renamed as an Interested Party (IP) number.

The *International Standard Music Number* (ISMN) identifies the published edition of printed music.

The *International Standard Recording Code* (ISRC) identifies individual sound recordings (such as make up the music tracks on CDs).

The *International Standard AudioVisual Number* (ISAN) is a new joint development of CISAC and the film producers group AGICOA, which identifies individual audiovisual works such as film or television programmes in a similar fashion to the ISRC.

The cross-industry *EAN/UPC article number*, which can also be expressed as a barcode, is used to identify the carrier of the recorded music (e.g. the CD, tape cassette etc.).

The *International Standard Work Code* (ISWC) identifies the musical composition itself, rather than the printed or recorded expression of the work

The ISWC is a recent development, successfully piloted for music in the first half of 1996, and it has been suggested that it could be extended to cover literature and visual arts as well as music. The ISWC, originally ten characters, has now been extended to eleven. The first character is, for the music world, the letter 'T' (for tune) followed by a unique nine digit number and a check digit. One or more single letter prefixes could be allocated to the literary and visual arts community. The International Federation of Reproduction Rights Organisations (IFRRO) has expressed its intention to adopt the ISWC 'L' for literature as the way in which they will identify literary works.

ISWC is a 'dumb' number which cannot be reconstructed from the actual work. Its level of granularity is arbitrary and it can therefore be assigned to any fragment that needs to be uniquely identified (e.g. separate ISWCs for a whole opera and an aria within that opera).

Doubts have been expressed about the capacity of the system, but the extension of the number has gone some way towards allaying fears concerning its adequacy. ISWC appears to be a complementary and not a competitive initiative to those of the book and serials publishing industry.

Internet developments. Uniform Resource Names (URNs) , Persistent Uniform Resource Locators (PURLs) etc.

The success of the World Wide Web owes much to the standardisation of *Uniform Resource Locators* (URLs), which, however, identify specific locations rather than documents and, as users will be aware, are subject to change.

The Internet Engineering Task Force (IETF) has been working for some time on the development of a system for *Uniform Resource Names* (URNs), designed to persistently identify actual information resources rather than their Internet locations. A number of proposals have been developed but none has yet found widespread support.

An intermediate proposal developed by OCLC, who have devoted considerable time and resources to this issue which they regard as of prime importance, is the *Persistent Uniform Resource Locator* (PURL). Instead of pointing directly to an Internet location, a PURL points to an intermediate resolution service which maintains a database linking the PURL to its current URL and returning that URL to the user client, similar to the use of

email aliases. In this way, references expressed as PURLs should remain viable as long as the resolution service continues to operate.

OCLC operates its own PURL service but is distributing the PURL source code in order to promote the widespread use of the system. PURL is universally regarded as a major step towards controlling the current proliferation of identifiers on the Internet

The Digital Object Identifier (DOI)

The Digital Object Identifier (DOI), being developed for the Association of American Publishers (AAP) by RR Bowker and the Corporation for National Research Initiatives (CNRI) is both an identifier and a routing system. It is a URN-compatible system similar in concept to the PURL (see above), designed to provide a persistent way of identifying and linking to electronic documents and their constituent parts.

The originally assigned DOI never changes, even on change in ownership of the document or object. The new publisher simply advises the maintenance agency for the DOI of the change in ownership and the original DOI (once authentication is received from the original publisher) would 'point' to the new owner.

At its core is the 'handle' system developed by the CNRI which uses a directory to link the permanently assigned DOI to the URL containing the object in question. The CNRI routing service is based on powerful servers in Reston Virginia which are mirrored in California and Spain.

The DOI is a two part identifier whose first part indicates numbering agency and publisher. The latter could be the Interested Party (IP) number used in the music industry's Common Information System (see above). The second part of the number, following a backslash, is a publisher assigned 'item ID'. This could be assigned at any level of granularity the copyright holder or assignee may deem appropriate.

Technically, the system allows the second part of the identifier to be *any* alphanumeric chain unique to an individual publisher. Since, however, paper and electronic editions will often coexist in parallel, it would seem sensible for publishers to use similar numbering systems for both. This would seem to indicate adoption of ISBN- and ISSN-based identifiers such as BICI and SICI.

Originally proposed as a copyright management system, with the DOI, possibly embedded in a digital object, routing a query to the copyright owner who returns a standard response screen of bibliographic and copyright information, the DOI clearly has much wider applications and, indeed, early demonstrations have concentrated on use of the DOI to connect the user directly to a document on the Internet.

A high level joint committee of the International Publishers Association and the STM group of publishers is monitoring the DOI and intends to issue recommendations for its global adoption in April.

Conclusions

More than 6 months have passed since we published the first version of this review which achieved gratifyingly wide circulation. Much has happened in the intervening period. The BICI has moved from being purely an acronym to a draft specification. The DOI has been demonstrated, conceptually at least, in Washington in February.

In our first edition of this paper last September we were unable to recommend a best buy; we are still unable to do so unequivocally. However, we are optimistic that the DOI may turn out to have the necessary attributes to provide many of the key requirements of a universal solution for the identification of digital expressions of content. Much will depend on the detail of developments during the rest of this year and the results of live studies with real applications. We are particularly concerned to see how current discussions on the syntax rules of the identifier progress.

However, we remain doubtful that even the DOI will be a solution which comprehensively meets *all* requirements for unique digital content identification. For the foreseeable future at least, publishers' systems will need to be able to handle multiple (and overlapping) identification of the same content. This is not perhaps as troubling as it may appear at first sight - publishers and their trade customers have for many years been familiar with identifying their products by both a number (the ISBN) and a brief description (author and title).

The concern which we had when we wrote the first edition of this paper was the apparent risk of unnecessary proliferation of identification standards. This problem appears to be receding as different groups meet to seek consensus and discover that the issues we have in common are much more significant than the issues which divide us. It remains extremely important that the issues involved in identification of digital content are discussed as widely as possible.

The authors would like to thank the following colleagues for their encouragement, assistance and contributions both to this text and to the ongoing work on unique identifiers:

Julia Blixrud, Association of Research Libraries

John Dowd, OCLC Europe

Norman Paskin, Elsevier Science Ltd

Carol Risher, Association of American Publishers

Godfrey Rust, Mechanical Copyright Protection Society

Book Industry Communication / EDItEUR

39 - 41 North Road

London

N7 9DP

UK

tel: +44 (0)171-607 0021

fax: +44 (0)171-607 0415
email: brian@bic.org.uk