Book metadata and identification: Bridging the divide from print to digital

Mark Bide
Executive Director EDItEUR
About EDItEUR

- Not-for-profit membership organisation
  - Our role is to develop, maintain and promote the use of standards in the book and journal supply chains round the world

- Based in London, global membership
  - Publishers, distributors, wholesalers, subscription agents, booksellers, libraries, system vendors, rights management organisations and trade associations
  - 100 members in over 20 countries
    - US, Japan, China, UK and throughout Europe

- Governing board of national, regional and international trade organisations to provide strategic direction

- Provide management services for ISO standards
  - ISBN, ISTC, ISNI
Book industry standards and the physical supply chain

Metadata standards are not new to the industry

- Managing a huge catalog of products: the ISBN
  - Unambiguous identification of “things that are for sale”

- Managing a huge volume of transactions: EDI
  - X12, EDIFACT, Tradacoms
  - Exchange of commercial transactions

- Managing a huge volume of metadata: ONIX
  - Exchange of rich descriptions
  - An essential tool as commerce started to move from the physical bookstore to online
What is the International Standard Book Number?

- ISO 2108 (1970; most recent revision 2005)

- 13 digit numeric string

- Includes some – but often misleading – affordance


- A class of books – a **product**
Book industry standards and the physical supply chain

Metadata standards are not new

- Managing a huge catalog of products: the ISBN
  - Unambiguous identification of “things that are for sale”
- Managing a huge volume of transactions: EDI
  - X12, EDIFACT, Tradacoms
  - Exchange of commercial transactions
- Managing a huge volume of metadata: ONIX
  - Exchange of rich descriptions
  - An essential tool as commerce started to move from the physical bookstore to online
What is ONIX for Books?

- XML communication format for sharing book industry product information
- Originated 1999 by the American Association of Publishers
- Current status: v2.1 widely implemented, v3.0 growing
- Implemented in many countries throughout the world – most recently Japan, China, Egypt, Turkey
- Allows the communication of information about publishers’ products throughout the supply chain – to distributors, wholesalers, retailers and other partners
- In many markets, data is collected from many sources and redistributed in consolidated form to supply chain partners
- Used by small and large organisations, included in many off the shelf IT systems
What do these standards have in common?

- Unashamedly, they are all about **commerce**

- Metadata and messaging standards are not simply about discovery – they are required for all aspects of commerce
  - Helping people to find and buy things is a key driver of ONIX distribution…but there is lots more in an ONIX message

- Commerce is not constrained by borders or language
  - Standards reflect that reality
ONIX and language

- Language of the standard and of the supporting documentation is English – although many national groups have their own translations.
- No constraints on the use of character sets or reading direction.
  - Active implementations in Japan, China, Korea, Russia, Egypt, Turkey, Bulgaria.
- The codes are a language-independent notation – identifiers for concepts.
- When an ONIX message crosses borders, the tokens continue to convey the same meaning.
What metadata does ONIX for books communicate?

<table>
<thead>
<tr>
<th>Identity and authority</th>
<th>Publishing, including</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record details</td>
<td>Imprint and publisher</td>
</tr>
<tr>
<td>Product identifiers</td>
<td>Publication date</td>
</tr>
<tr>
<td></td>
<td>Territorial Rights</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Descriptive, including</th>
<th>Related material, including</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product form</td>
<td>Related works</td>
</tr>
<tr>
<td>Classifications</td>
<td>Related products</td>
</tr>
<tr>
<td>Titles</td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td></td>
</tr>
<tr>
<td>Edition</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td></td>
</tr>
<tr>
<td>Audience</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collateral, including</th>
<th>Supply, including</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketing resources</td>
<td>Availability</td>
</tr>
<tr>
<td>Supporting text</td>
<td>Suppliers</td>
</tr>
<tr>
<td></td>
<td>Prices</td>
</tr>
<tr>
<td></td>
<td>Discounts</td>
</tr>
</tbody>
</table>
From physical to digital – a mixed economy

- Metadata and identity are the “lifeblood of ecommerce”

- The core challenge is the increased complexity….

- ….of identification, of description, of transaction

- Metadata is as complex as the world it seeks to describe…

- …“simplification” of metadata = loss of information
Industry systems are not designed to deal with this complexity

- ISBN is a **product** identifier – but has been used as the primary key of many systems that have nothing to do with products
- Definition of “a product” has become more difficult
- Hardback, paperback…ebook?

- The potential number of products has become an order of magnitude greater
- How do you collocate all these different products?
  - A work identifier (ISTC)?
  - A “release” identifier?

- How far do we have to manage **instance** identification?
  - The equivalent of RFID – already required for management of DRM
Managing the metadata explosion

- All metadata is essentially about identity
  - Particularly if it is to be unambiguously machine-processable
  - Essential for a commercial environment

- Public identification systems are not primarily technical but social – agreed upon norms and processes
  - Unambiguous rules for what is identified
  - Unambiguous rules of granularity – when are two things treated as being “the same thing” and when as different

- To be useful, public identification systems require publically accessible registries – so that others can know what is being identified
  - Books in Print – registries are not always “freely available”
The creation and management of authoritative metadata is never costless

- Common, authoritative metadata databases, if they are well run and maintained, will save costs for everyone…
- …but inaccurate, inconsistent and out-of-date metadata may be worse than no metadata at all
- Traditional systems for managing metadata and identity in publishing are no longer viable
  - We don’t deal simply in products
- Metadata itself is a service not a good
  - It needs to be managed on an ongoing basis, not just manufactured once
- “Metadata should be free” is too simplistic
  - There are costs associated with metadata creation and management that someone has to pay
Some questions I would like to hear answered

- **eBook identification**
  - What are the classes of referents we need to identify?

- **eBook metadata: in-band or out-of band**
  - What should be embedded and what associated by external reference?

- **Convergence between commercial and library practice**
  - Can we share metadata more effectively?
Book metadata and identification:
Bridging the divide from print to digital
mark@editeur.org