NETSL

Annual Spring Conference

April 15, 2010

Mapping Bibliographic Metadata

Carol Jean Godby, Research Scientist

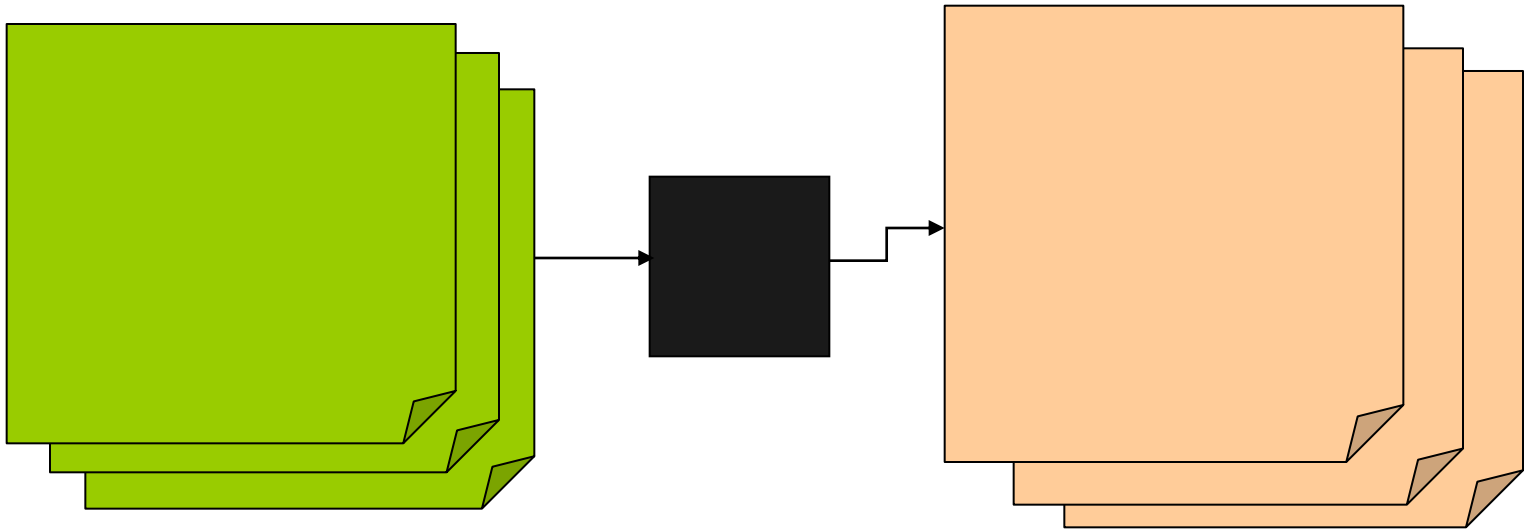
OCLC™ The world's libraries. Connected.

Agenda for today

- 1. Crosswalks and how they are used at OCLC
- 2. A report of successes on some difficult problems
- 3. What is still unresolved
- 4. Some topics for group discussion

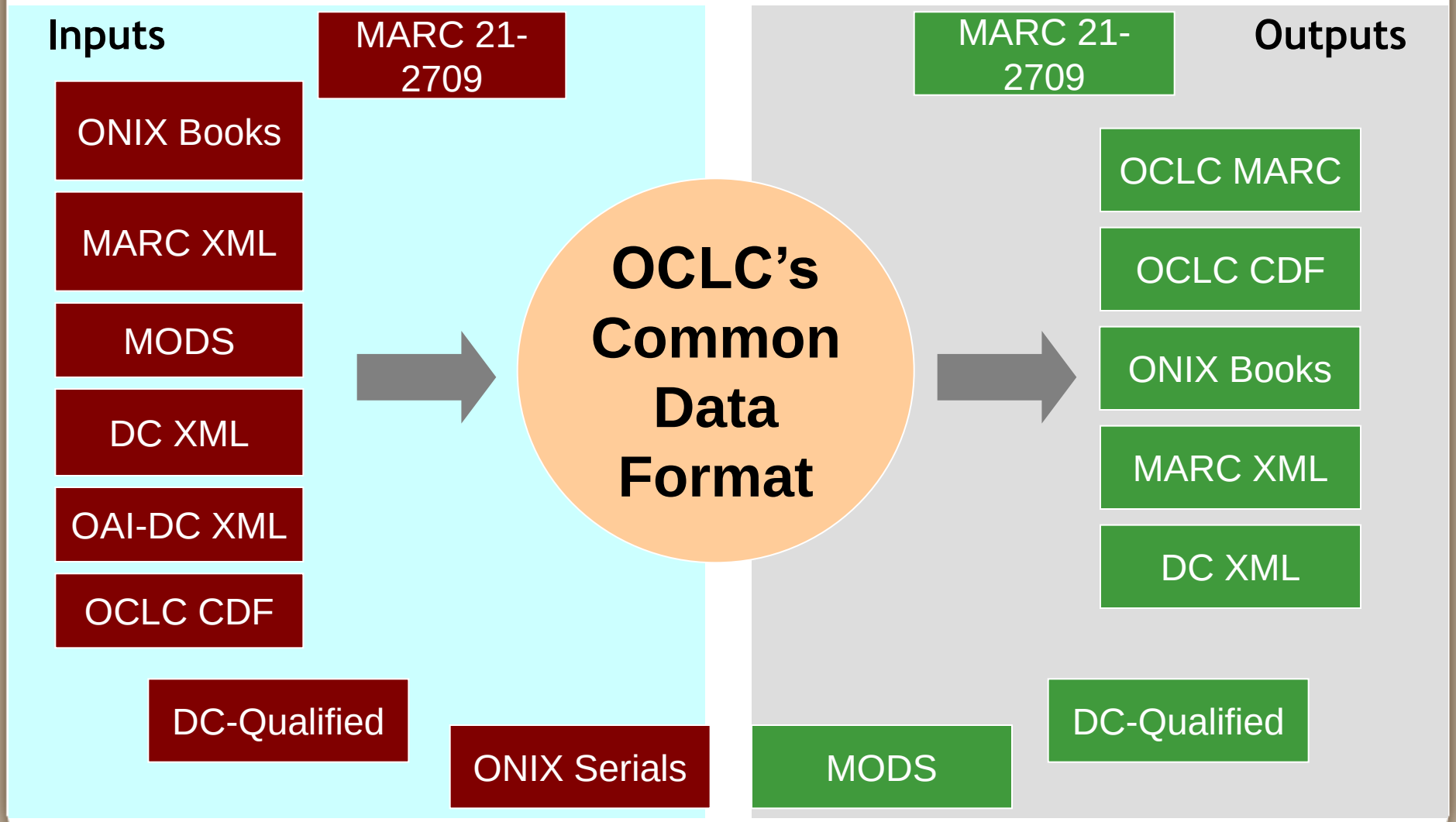
The problem

My records are in this format…



…but they need to be in this other format.

The supported translations



What is a crosswalk?

"Crosswalks are used to 'translate' between different metadata element sets. The elements (or fields) in one metadata set are correlated with the elements of another metadata set that have the same or similar meanings. This is also sometimes called 'semantic mapping.'"

Source: Canadian Heritage Information Network (http://www.chin.gc.ca/English/)

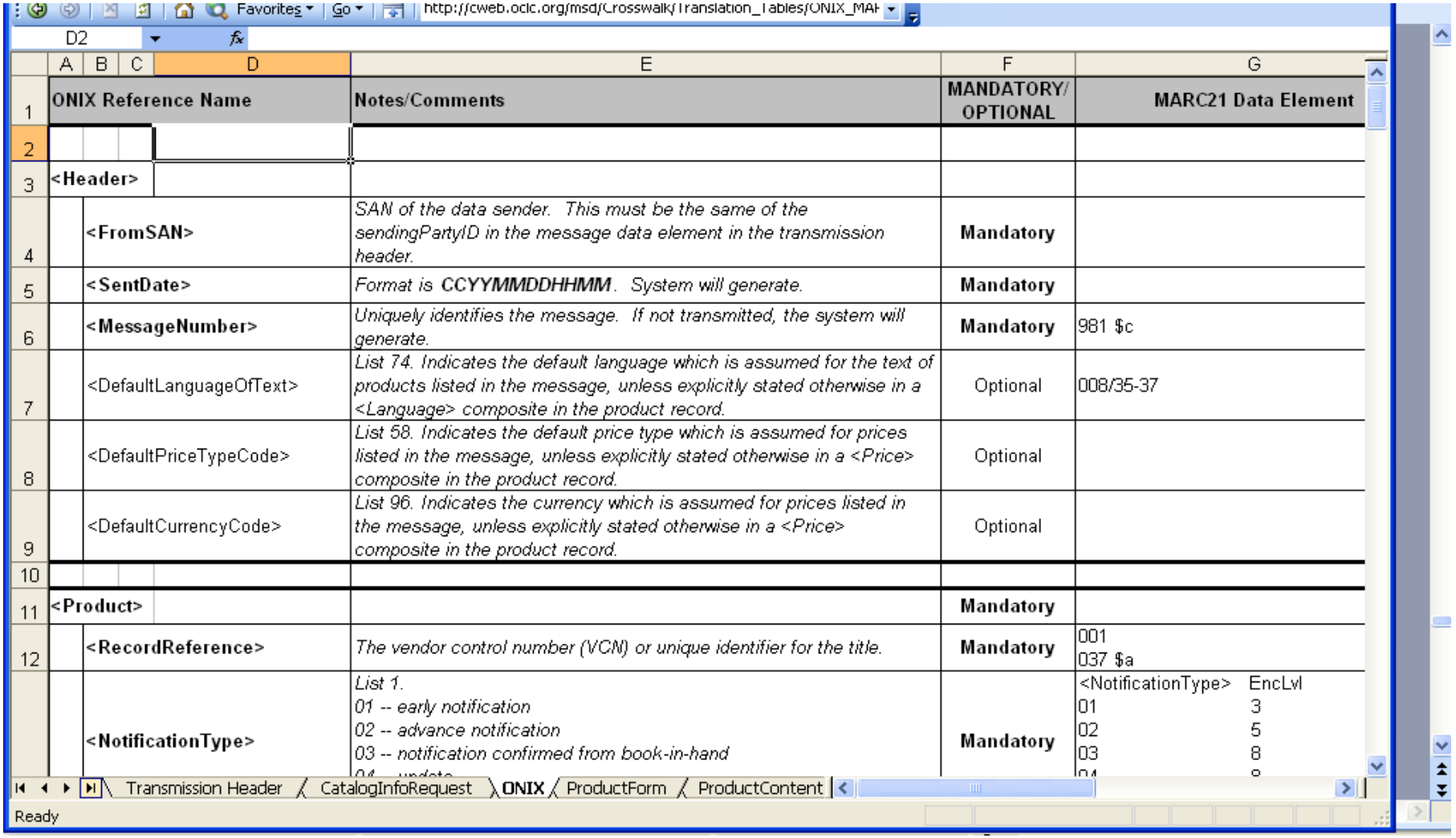
ONIX and MARC records

```
<Product>
    <RecordReference>0892962844
    <ProductIdentifier>
        <ProductIDType>03 </>
        <IDValue>9780892962846</>
     </ProductIdentifier>
     <ProductForm>BB</>
     <Title>
        <TitleType>01</>
        <TitleText>McBain's Ladies
     </Title>
     <Contributor>
         <ContributorRole>A01</>
         <PersonNameInverted>Hunter, Evan</>
     </Contributor>
     <Subject>
        <SubjectSchemeIdentifier>02</>
        <SubjectHeadingText>
                Policewomen--Fiction.
```

```
Leader  00000 jm a22000005  4500
008                g         eng
020      0892962844
024 3#   9780892962846
100      $a Hunter, Evan
245      $a McBain's ladies
260      $b Mysterious Press $d
1988
300      $a 320 p.
650 #2   $a Policewomen -- Fiction
```

ONIX is the international standard for the book industry

OCLC's ONIX to MARC Crosswalk



Favorites Go http://cweb.oclc.org/msd/Crosswalk/Translation_Tables/ONIX_MAI

D2

	ONIX Reference Name	Notes/Comments	MANDATORY/ OPTIONAL	MARC21 Data Element
2				
3	**<Header>**			
4	**<FromSAN>**	*SAN of the data sender. This must be the same of the sendingPartyID in the message data element in the transmission header.*	**Mandatory**	
5	**<SentDate>**	*Format is **CCYYMMDDHHMM**. System will generate.*	**Mandatory**	
6	**<MessageNumber>**	*Uniquely identifies the message. If not transmitted, the system will generate.*	**Mandatory**	981 $c
7	<DefaultLanguageOfText>	*List 74. Indicates the default language which is assumed for the text of products listed in the message, unless explicitly stated otherwise in a <Language> composite in the product record.*	Optional	008/35-37
8	<DefaultPriceTypeCode>	*List 58. Indicates the default price type which is assumed for prices listed in the message, unless explicitly stated otherwise in a <Price> composite in the product record.*	Optional	
9	<DefaultCurrencyCode>	*List 96. Indicates the currency which is assumed for prices listed in the message, unless explicitly stated otherwise in a <Price> composite in the product record.*	Optional	
10				
11	**<Product>**		**Mandatory**	
12	**<RecordReference>**	*The vendor control number (VCN) or unique identifier for the title.*	**Mandatory**	001 037 $a
	<NotificationType>	*List 1.* *01 -- early notification* *02 -- advance notification* *03 -- notification confirmed from book-in-hand*	**Mandatory**	<NotificationType> EncLvl 01 3 02 5 03 8

Transmission Header | CatalogInfoRequest | ONIX | ProductForm | ProductContent

Ready

OCLC Services

Contract Cataloging Services for Publishers

OCLC Metadata Services for Publishers

- Overview
- Ordering
- Support

See the OCLC press release.

OCLC Metadata Services for Publishers

Enhanced ONIX metadata for publishers

AT A GLANCE

- **Improves discoverability**—through the addition of subject analysis, classification numbers, reviews and more
- **Improves efficiency**—in creating and distributing metadata to supply chain partners
- **Increases marketability**—through the addition of subject analysis, classification numbers, evaluative content, reviews and more

View complete Overview ››

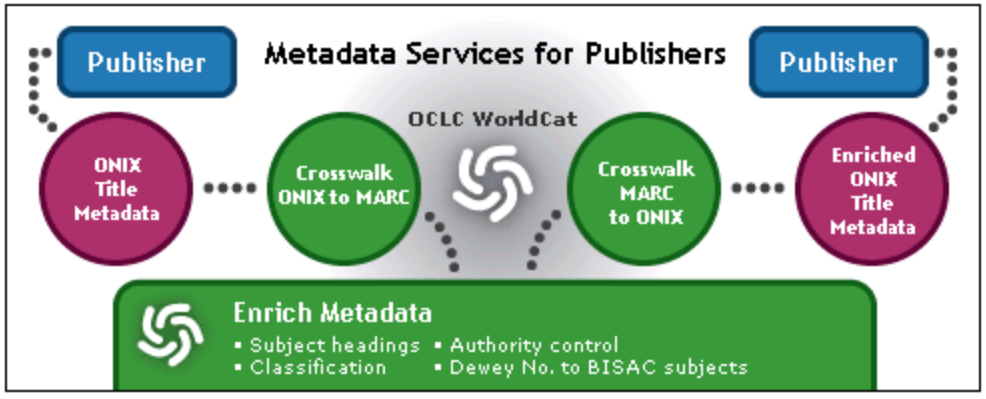
Download the brochure ››

NEWS

- Webinar introduces Metadata Services for Publishers

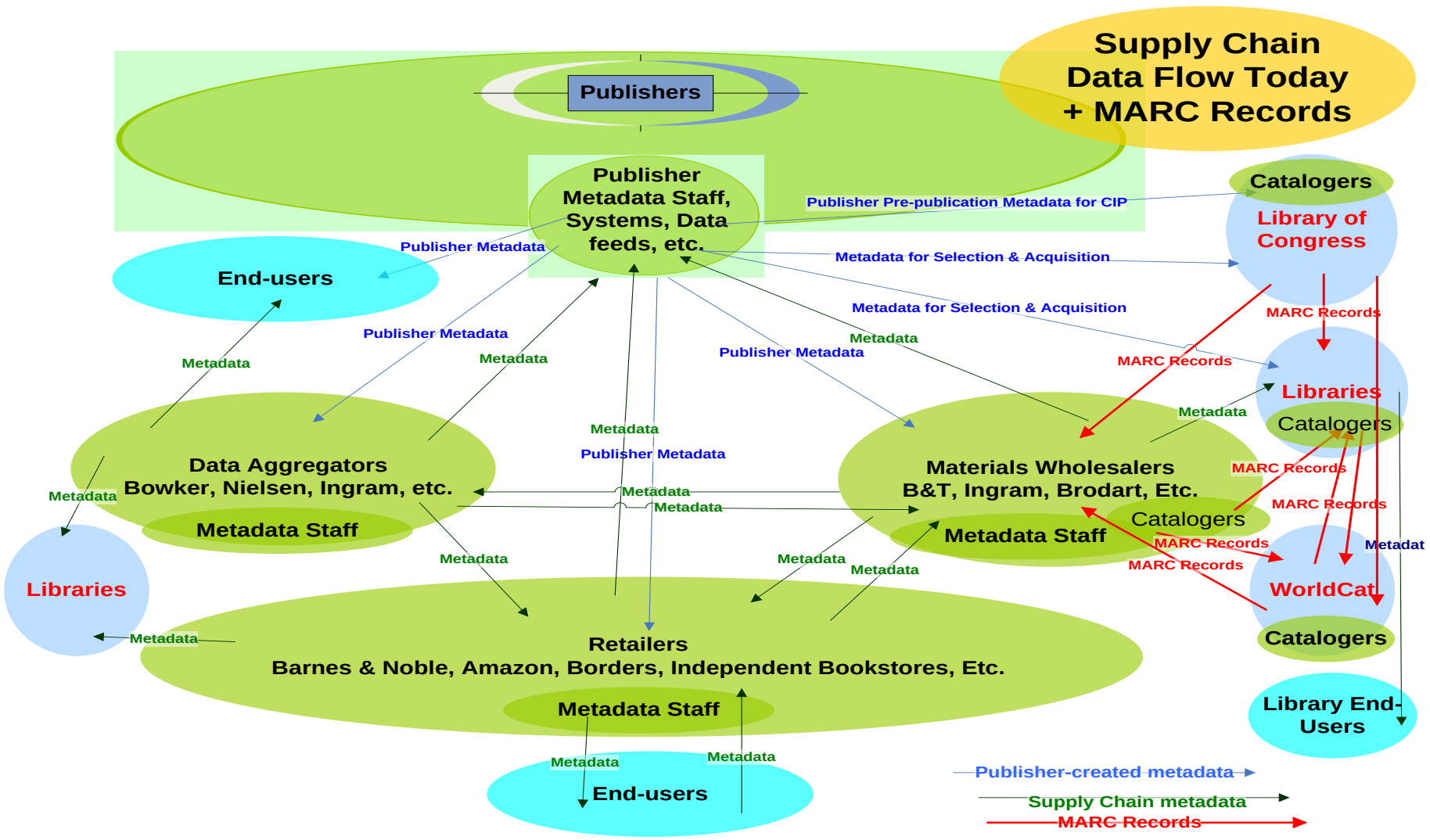
More news ››

OCLC's Metadata Services for Publishers can help you sell more titles by accepting your title metadata in ONIX format, enriching it by mining the WorldCat database, and delivering an enhanced ONIX file ready to use in supply chain systems and marketing communications.



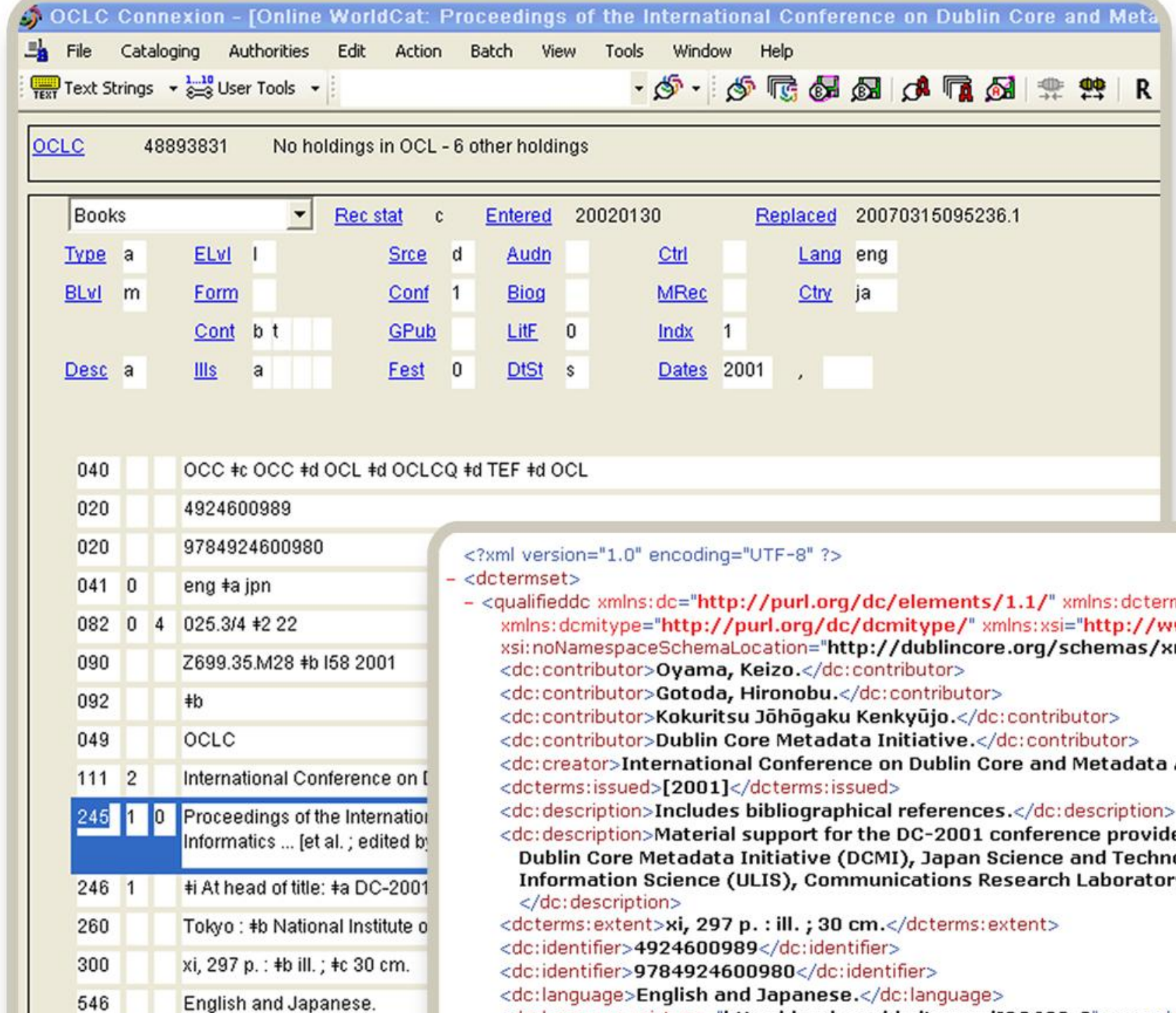
Learn how to order or request a quote for OCLC Metadata Services for Publishers ››

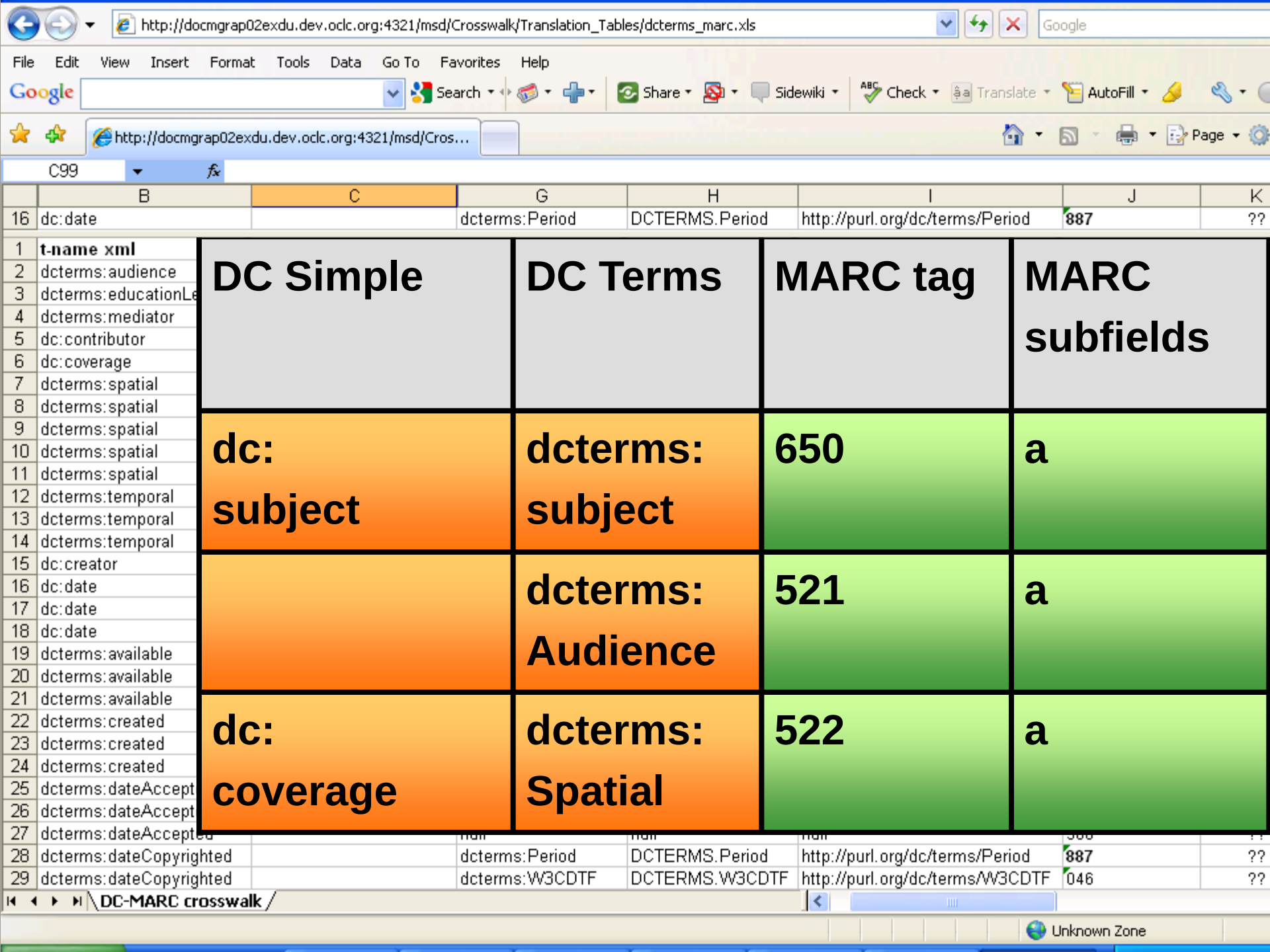
Publisher Supply Chain Data Flow Today + MARC Records



Metadata Services for Publishers: Highlights

- ONIX records are obtained from publishers and mapped to MARC.
- If a match is found in WorldCat, fields from the corresponding MARC record are applied to it.
- If no match is found, the record is populated from fields in the closest FRBR cluster.
- The result is mapped back to ONIX to be delivered to publishers. It is also made available to the library community as an enhanced MARC record.


OCLC Connexion - [Online WorldCat: Proceedings of the International Conference on Dublin Core and Meta
File Cataloging Authorities Edit Action Batch View Tools Window Help
Text Strings User Tools
OCLC 48893831 No holdings in OCL - 6 other holdings
Books Rec stat c Entered 20020130 Replaced 20070315095236.1
Type a ELvl I Srce d Audn Ctrl Lang eng
BLvl m Form Conf 1 Biog MRec Ctry ja
Cont b t GPub LitF 0 Indx 1
Desc a Ills a Fest 0 DtSt s Dates 2001 ,
040 OCC ǂc OCC ǂd OCL ǂd OCLCQ ǂd TEF ǂd OCL
020 4924600989
020 9784924600980
041 0 eng ǂa jpn
082 0 4 025.3/4 ǂ2 22
090 Z699.35.M28 ǂb I58 2001
092 ǂb
049 OCLC
111 2 International Conference on D
245 1 0 Proceedings of the Internatio
Informatics ... [et al. ; edited by
246 1 ǂi At head of title: ǂa DC-2001
260 Tokyo : ǂb National Institute o
300 xi, 297 p. : ǂb ill. ; ǂc 30 cm.
546 English and Japanese.
<?xml version="1.0" encoding="UTF-8" ?>
<dctermset>
<qualifieddc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.o
xmlns:dcmitype="http://purl.org/dc/dcmitype/" xmlns:xsi="http://www.w3.org/2001
xsi:noNamespaceSchemaLocation="http://dublincore.org/schemas/xmls/qdc/2006/0
<dc:contributor>Oyama, Keizo.</dc:contributor>
<dc:contributor>Gotoda, Hironobu.</dc:contributor>
<dc:contributor>Kokuritsu Jōhōgaku Kenkyūjo.</dc:contributor>
<dc:contributor>Dublin Core Metadata Initiative.</dc:contributor>
<dc:creator>International Conference on Dublin Core and Metadata Applications (200
<dcterms:issued>[2001]</dcterms:issued>
<dc:description>Includes bibliographical references.</dc:description>
<dc:description>Material support for the DC-2001 conference provided by: National In
Dublin Core Metadata Initiative (DCMI), Japan Science and Technology Corporation
Information Science (ULIS), Communications Research Laboratory (CRL), National
</dc:description>
<dcterms:extent>xi, 297 p. : ill. ; 30 cm.</dcterms:extent>
<dc:identifier>4924600989</dc:identifier>
<dc:identifier>9784924600980</dc:identifier>
<dc:language>English and Japanese.</dc:language>



http://docmgrap02exdu.dev.oclc.org:4321/msd/Crosswalk/Translation_Tables/dcterms_marc.xls

File Edit View Insert Format Tools Data Go To Favorites Help

Google Search Share Sidewiki Check Translate AutoFill

http://docmgrap02exdu.dev.oclc.org:4321/msd/Cros...

Page

C99

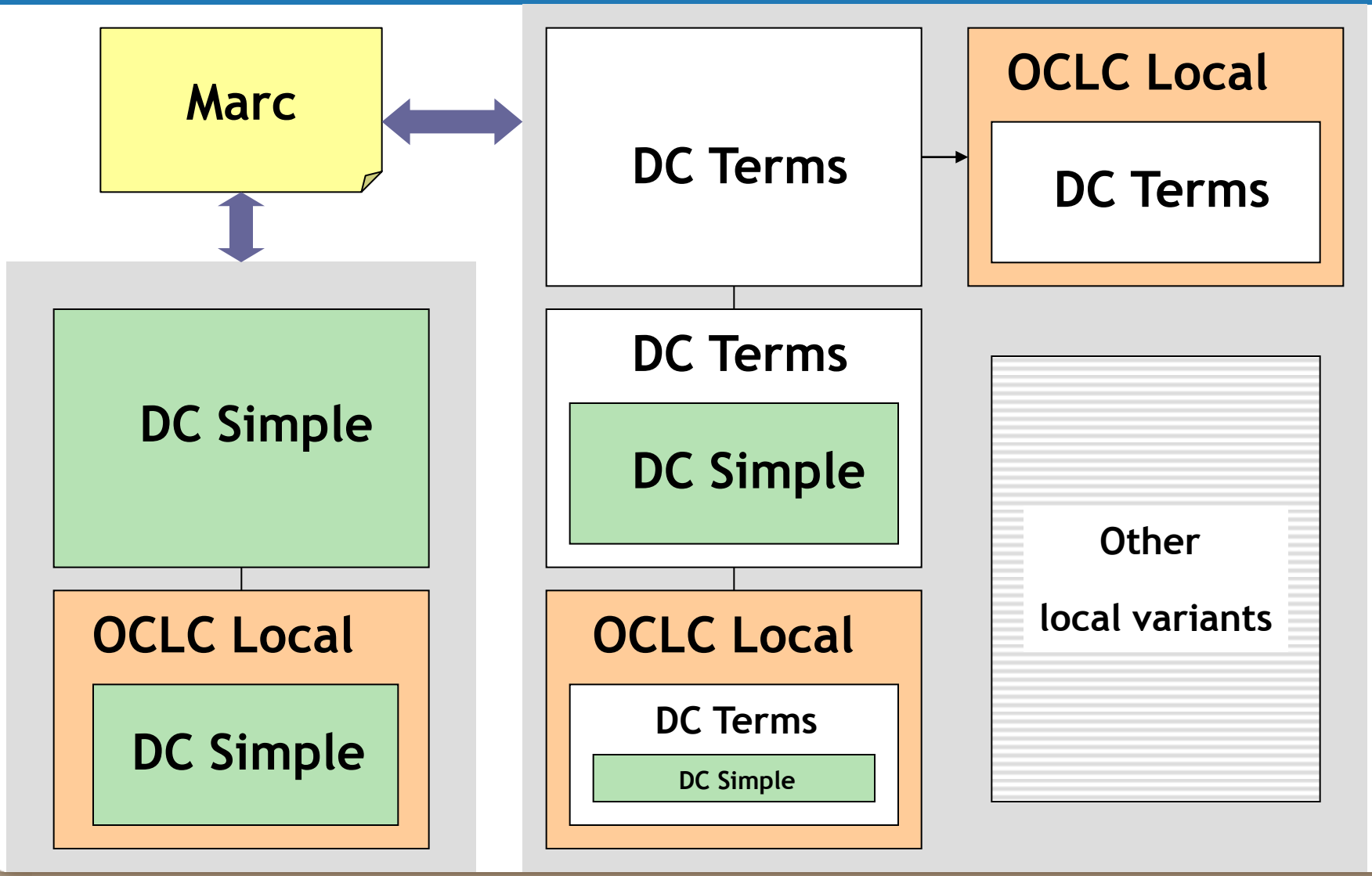
	B	C	G	H	I	J	K
16	dc:date		dcterms:Period	DCTERMS.Period	http://purl.org/dc/terms/Period	887	??
1	t-name xml						
2	dcterms:audience						
3	dcterms:educationLe						
4	dcterms:mediator						
5	dc:contributor						
6	dc:coverage						
7	dcterms:spatial						
8	dcterms:spatial						
9	dcterms:spatial						
10	dcterms:spatial						
11	dcterms:spatial						
12	dcterms:temporal						
13	dcterms:temporal						
14	dcterms:temporal						
15	dc:creator						
16	dc:date						
17	dc:date						
18	dc:date						
19	dcterms:available						
20	dcterms:available						
21	dcterms:available						
22	dcterms:created						
23	dcterms:created						
24	dcterms:created						
25	dcterms:dateAccept						
26	dcterms:dateAccept						
27	dcterms:dateAccepted		null	null	null	506	??
28	dcterms:dateCopyrighted		dcterms:Period	DCTERMS.Period	http://purl.org/dc/terms/Period	887	??
29	dcterms:dateCopyrighted		dcterms:W3CDTF	DCTERMS.W3CDTF	http://purl.org/dc/terms/W3CDTF	046	??

DC Simple	DC Terms	MARC tag	MARC subfields
dc: subject	dcterms: subject	650	a
	dcterms: Audience	521	a
dc: coverage	dcterms: Spatial	522	a

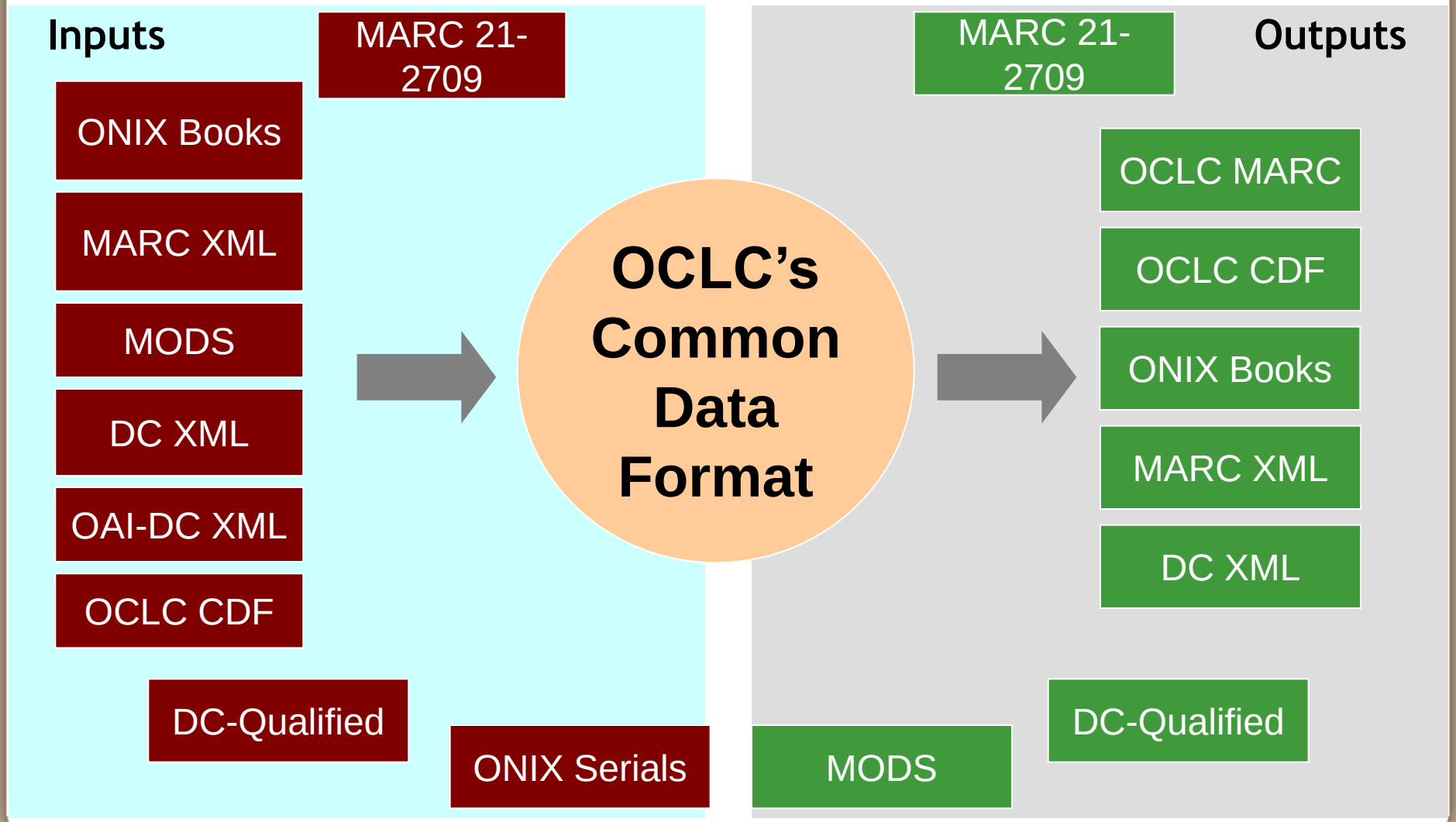
DC-MARC crosswalk

Unknown Zone

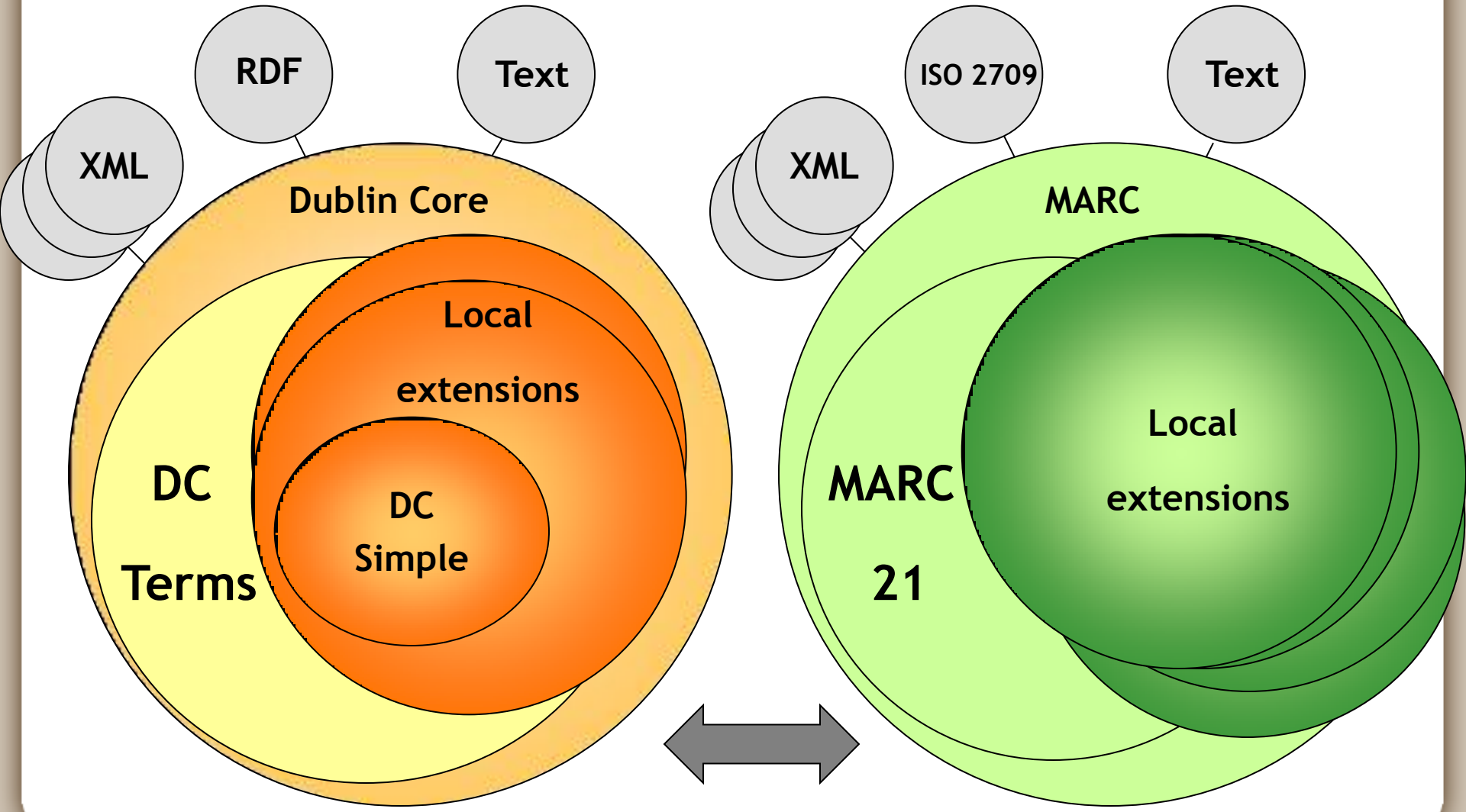
The MARC-Dublin Core relationship



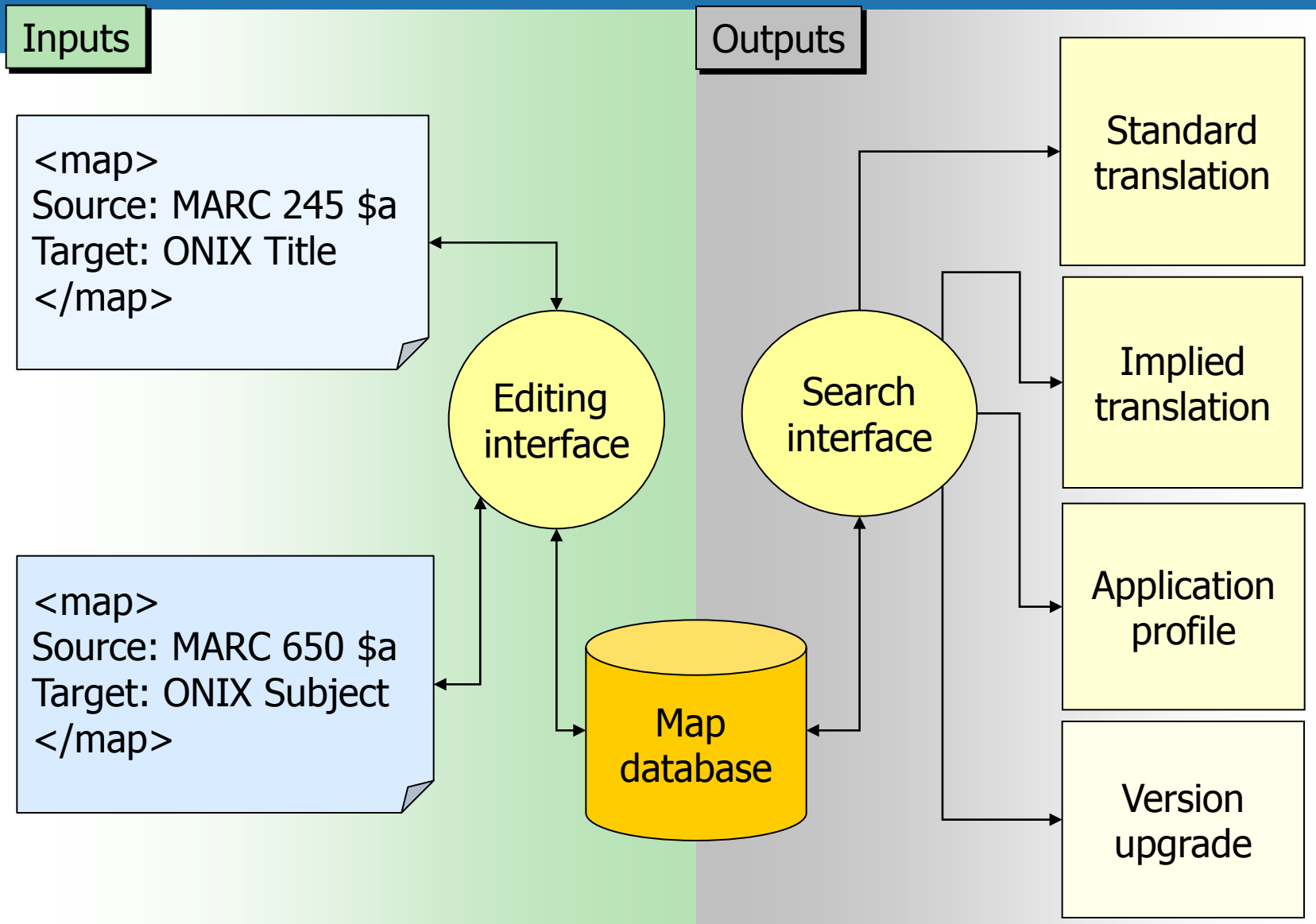
CDF as an application profile



Translations and conversions, expanded



A graphical user interface to crosswalk maps



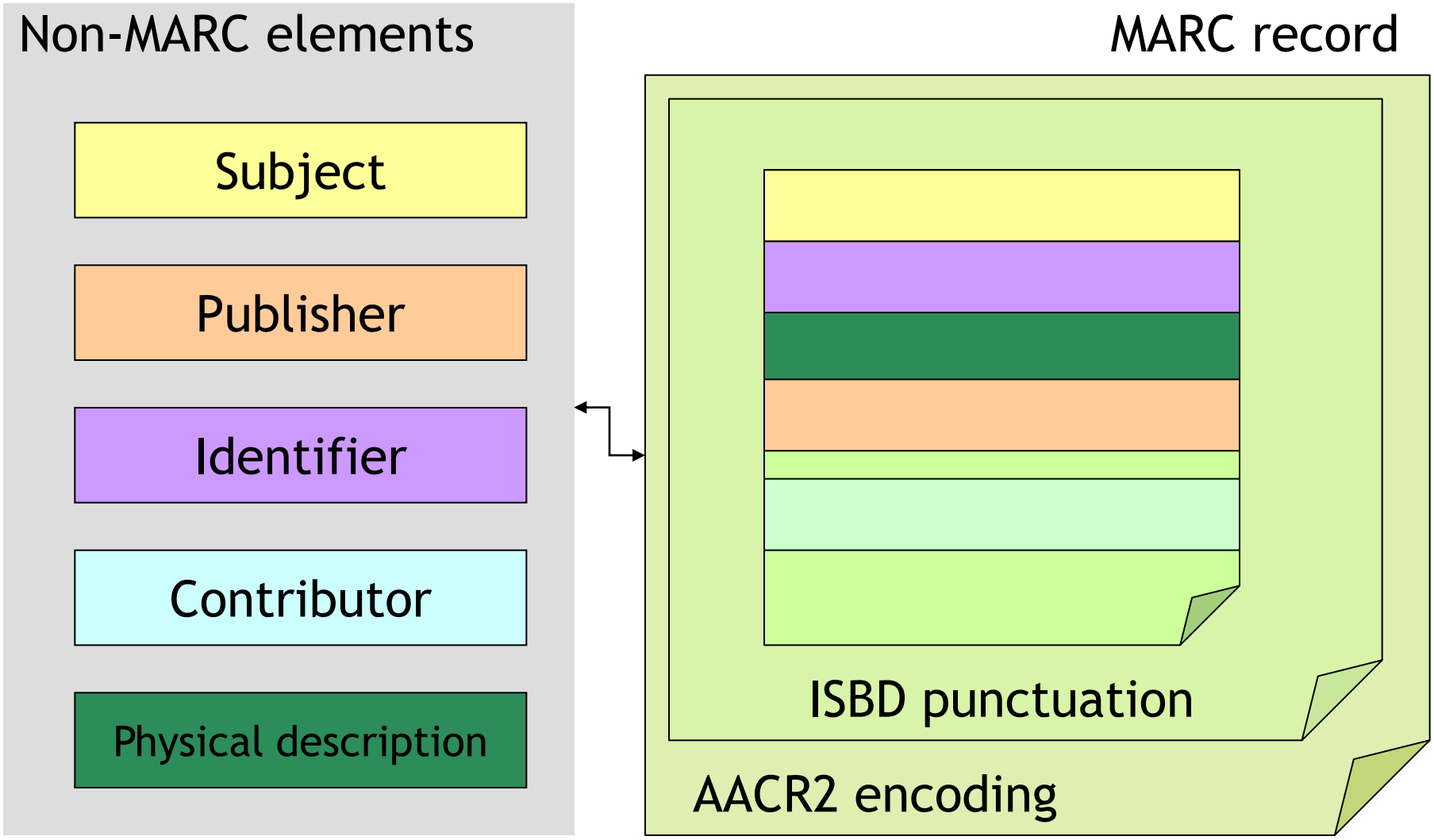
In sum:
What we have learned to do pretty well

- Manipulate MARC and MARC-like formats.
- Manage variation in structure and character encoding.
- Manage different versions.
- Reduce and (almost) eliminate data processing silos.
- Automate what is systematic and regular.
- Reuse software and metadata subject matter expertise.

But…

- We must move toward new paradigms for metadata creation and maintenance that permit:
  - Greater interoperability and shared metadata.
  - Mechanisms for allowing metadata to “grow up” over time.

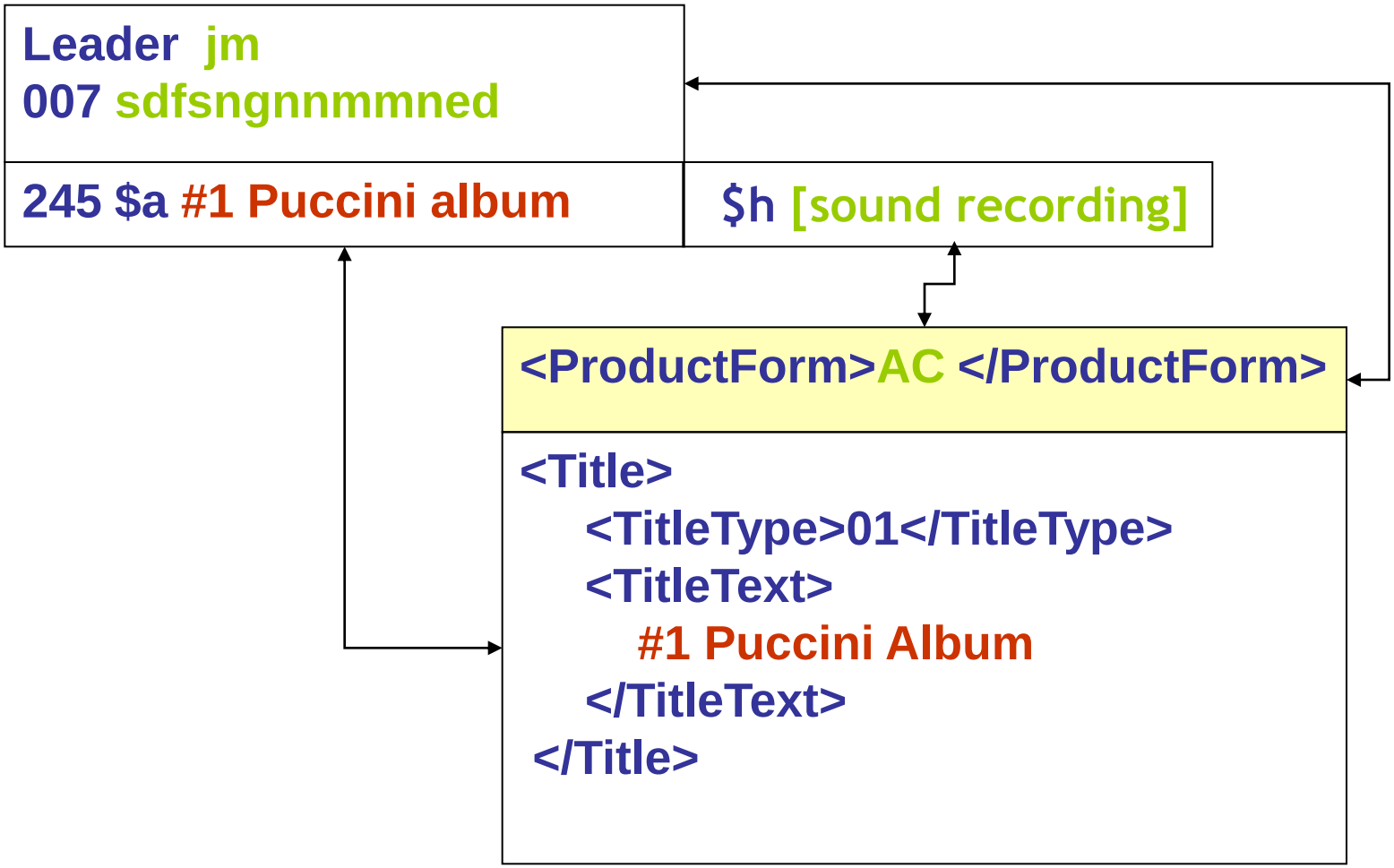
The old and new paradigms



Some problems

- Extra effort is required to add, validate, and dismantle ISBD and AACR2 rules, adding complexity to the metadata translation/conversion problem space.
- AACR2 introduces concepts that the other standards don't have.
- The ISBD and AACR2 layers are not a worldwide standard.
- Vocabulary and semantic concepts are different.
- Differences in punctuation and formatting require the crosswalk to peek at the data. As a result:
  - The mappings are brittle.
  - Duplicate detection is difficult.

An example:
Physical descriptions in ONIX and MARC



The two paradigms, redux

The MARC standard	Modern non-MARC standards
• Record-oriented	• Element or field-oriented
• Tailored to applications in the library community	• Agnostic about how the data will be used.
• Designed for storage	• Designed for transmission
• Static	• Dynamic

Next steps 1:
Participate in the transition to RDA

- The ONIX/RDA Framework solves some of the problems with physical descriptions by proposing a registered common vocabulary that both standards share.
- RDA also has far-reaching recommendations for linking data (instead of copying it).
- But the RDA draft standard still has many formatted fields that will be difficult to process algorithmically.

Next steps 2:
Apply lessons from studies of MARC field usage

- Only 21 to 30 tags occur in 10% or more records in OCLC's WorldCat database.
- Fixed-length data elements, identifier fields, main entry and title fields are the most common.
- Only a subset of fields is indexed by library search systems
- Notes fields are common, but machines aren't good at interpreting them.

So…?

"MARC data cannot continue to exist in its own discrete environment, separate from the rest of the information universe. It will need to be leveraged and used in other domains to reach users in their own networked environments. The 200 or so MARC 21 fields in use must be mapped to simpler schema."

Source: Implications of MARC Tag Usage on Library Metadata Practices

For more information

- Carol Jean Godby. 2010. “Mapping ONIX to MARC.”

  http://www.oclc.org/research/publications/library/2010/2010-14.pdf.

- Gordon Dunsire. 2007. “Distinguishing Content from Carrier: The RDA/ONIX framework for Resource Categorization.”

  http://www.dlib.org/dlib/january07/dunsire/01dunsire.html.

- EDItEUR. http://www.editeur.org/.

- Karen Smith-Yoshimura, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlison de Ortiz, and Hugh Taylor. 2010. “Implications of MARC tag usage on library metadata practices”

  http://www.oclc.org/research/publications/library/2010/2010-06.pdf