Metadata and multi-lingualism; OCLC projects and initiatives

Janifer Gatenby
Research Integration and
Standards
Trento 9th September 2009

OCLC™ The world's libraries.
Connected.

The paper will address metadata organisation and enrichment to support multi-lingual interfaces. It will cover OCLC projects and initiatives including name and place authority data (VIAF and WorldCat identities), subject authority data (terminologies and Dewey) and parallel records in WorldCat and related data files.

WorldCat Current situation

- Parallel bibliographic records
- Name authorities – VIAF
- Dewey classification and subject authorities
- Data architecture changes
- Issue: transliteration

Parallel Bibliographic Records (language of cataloguing = eng, language of work = spa)

OCLC

```
040 __  XXX ‡c XXX ‡d XXX
245 00 Antología de textos, la reforma y el segundo imperio (1853-
       1867) / ‡c introducción y selección de textos, Silvestre Villegas
       Revueltas.
260 __  México, D.F. : ‡b Universidad Nacional Autónoma de México,
       ‡c 2008.
300 __  xxxvii, 424 p. : ‡b maps ; ‡c 18 cm.
490 1_ Biblioteca del estudiante universitario ; ‡v 145
440 _0 Cultura e historia mexicanas
651 _0 Mexico ‡x History ‡y 1821-1861 ‡v Sources.
651 _0 Mexico ‡x History ‡y European intervention, 1861-1867 ‡v
       Sources.
653 __  México ‡a Historia ‡a Siglo XVIII
700 1_ Villegas Revueltas, Silvestre.
830 _0 Biblioteca del estudiante universitario (Universidad Nacional
       Autónoma de México) ; ‡v 145.
936     PR 238572899
```

Up until 2003, there was a one record per resource policy. Allow libraries and users to view records described in a preferred language

142 million records, est. 8 million linguistic variant bibliographic records. 1.46 billion holdings

Language of cataloguing & of work = spa

OCLC Control Number: 238572899

040 __ XXX ‡b spa ‡c XXX

245 00 Antología de textos : ‡b la reforma y el segundo imperio, ‡f 1853-1867 / ‡c introducción y selección de textos Silvestre Villegas Revueltas.

260 __ México : ‡b UNAM, Coordinación de Humanidades, Programa Editorial, Instituto de Investigaciones Históricas, ‡c 2008.

300 __ xxxvii, 424 p. : ‡b [3] h. pleg. de mapas ; ‡c 18 cm.

440 _0 Biblioteca del estudiante universitario. ‡p Cultura e historia mexicana ; ‡v 145

504 __Incluye referencias bibliográficas.

651 _4 México ‡x Historia ‡y Intervención y segundo imperio

651 _0 Mexico ‡x History ‡y European intervention, 1861-1867.

700 1_ Villegas Revueltas, Silvestre, ‡e Colab

Spanish language of cataloguing - especially notes which can include substantial summaries and abstracts. Also Spanish subject headings

Hybrid bibliographic records

040   S3O‡b **swe** ‡c S3O

020   0333993373

020   9780333993378

043   e-uk---

100 1   Oulton, Carolyn. ‡d 1972-

245 10 Literature and religion in mid-Victorian England : ‡b from Dickens to Eliot / ‡c Carolyn Oulton.

260 __ Houndmills, Hampshire ‡a New York : ‡b Palgrave Macmillan, ‡c 2003.

300 __ 221 **s.**

504 __ **Includes bibliographical references and index.**

650   7 Religion och litteratur ‡z England ‡x historia ‡y 1800-talet. ‡2 **sao**

650   7 Engelsk religiös litteratur ‡x historia ‡y 1800-talet. ‡2 **sao**

Slowly correcting these so that they have only one language of cataloguing (subject headings are the exception)

"6" = répertoire de vedettes matière (Canadian French headings)

"7" = Source specified in subfield $2 - specifies source but not language

Records with multi-lingual subjects are not ideal – current pragmatic solution.

Records with multi-lingual subjects

- No relationship between the heading in one language and in another

- Language is not explicitly coded – just the name of the subject heading scheme

- Multi-lingual subject authorities and classifications are a better solution
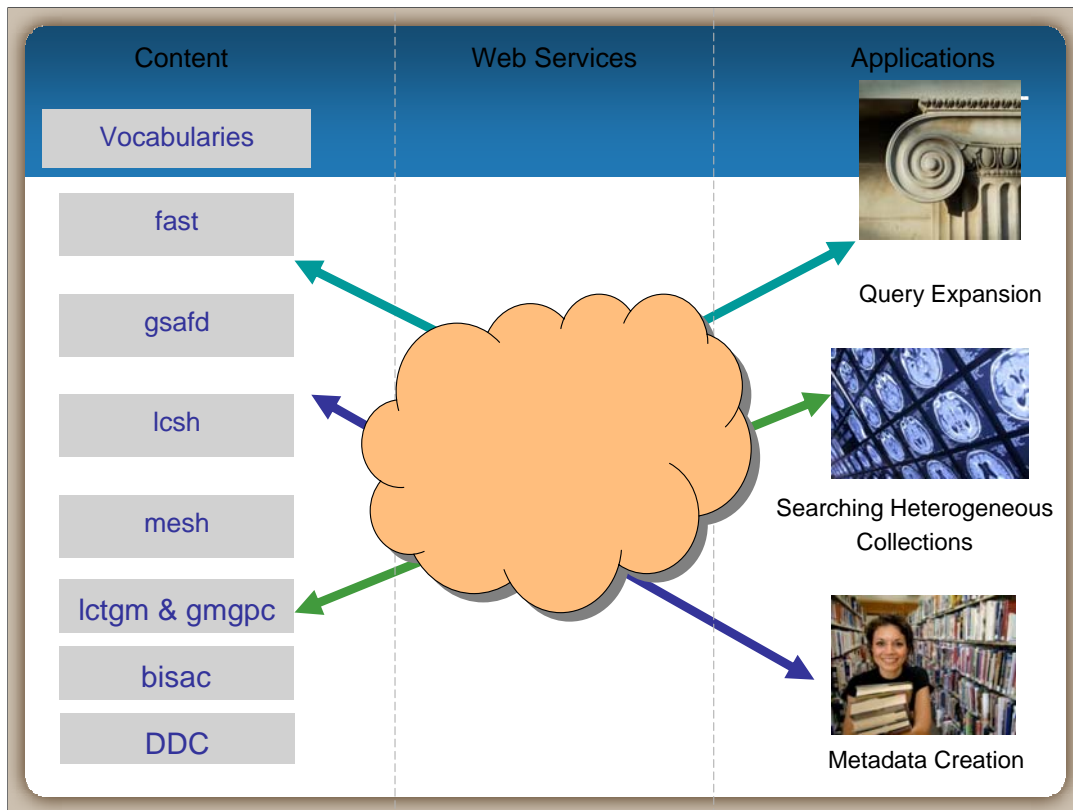
Linking translations
Jack Curtis / David Harsent

A bird's idea of flight	ALCS				OCLC	
After dark		LC			OCLC	
Another round at the pillars	ALCS	LC			OCLC	
Confessor	ALCS				OCLC	
Conjure me	ALCS				OCLC	
Crows' parliament	ALCS	LC	DNB		OCLC	Sam Lawrence
Der blick des magiers			DNB		OCLC	
Der schrei der schwalbe	ALCS		DNB		OCLC	
Die spur der krahe	ALCS		DNB		OCLC	
Dreams of the dead	ALCS	LC			OCLC	
From an inland sea	ALCS				OCLC	
Gawain libretto	ALCS	LC		BNF	OCLC	
Glory	ALCS	LC	DNB		OCLC	
Le parlement des corbeaux	ALCS			BNF	OCLC	
Legion	ALCS	LC		BNF	OCLC	
Les enfants du matin	ALCS			BNF	OCLC	
Livewire chillers - down came a spider	ALCS				OCLC	no author attribution

We are experimenting with combining data from the rights management societies with library data. It can help show us where we can link translations – e.g. here "Le parlement des corbeaux"

Content	Web Services	Applications
Vocabularies		

fast

gsafd

lcsh

mesh

lctgm & gmgpc

bisac

DDC

Query Expansion

Searching Heterogeneous Collections

Metadata Creation

There are three main components to moving vocabularies to the network level:

•Content (controlled vocabularies)

•Web services

•Applications (created by software developers internal & external to OCLC)

Dewey in a multilingual environment

- Language-independent representation
- Interoperable translations (130 languages)
- Mappings and crosswalks
- Large amount of categorized content
- Organizational support
- Worldwide user community

We'll say a few words here about the value of one of the resources listed on the previous slide, the Dewey Decimal Classification, in a multilingual environment. The DDC offers a language-independent representation that can be used to drive discovery and access in a variety of languages and forms. Behind the language-independent representation are a rich set of terminology derived from interoperable translations, and mappings and crosswalks between the DDC and subject heading systems and other classification schemes. There is a large (and growing) amount of categorized content associated with the system and available in WorldCat and other bibliographic utilities and repositories. OCLC provides strong organizational support for the system, and the worldwide Dewey user community takes an active role in its ongoing development and use.

Dewey Web Services
http://dewey.info

Experimental space for linked DDC data

Initial linked data set: DDC Summaries in 9 languages
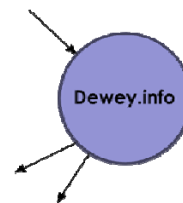
Features:

- Classification semantics encoded in SKOS
- Actionable URIs for every DDC class
- Representations for machines (RDF) and for humans (HTML)
- The linked data is made available under a Creative Commons BY-NC-ND license (more permissions are available here)

We've just launched dewey.info, an experimental space for linked data.

The initial data set available is a linked data version of the DDC Summaries (the top three levels of the DDC) in nine languages.

The classification data is encoded in SKOS (Simple Knowledge Organization System) with actionable URIs (Uniform Resource Identifiers) for each DDC class.

The service responds to regular HTTP requests (in the form of Dewey URIs) with either a machine - (RDF) or human-readable (HTML) presentation.

Publishing parts of Dewey as linked data means that any application or resource description can establish live links into the DDC, instead of just using the Dewey number.

The linked data is available under a Creative Commons BY-NC-ND license—this license allows linking, reuse, and experimentation with the data for noncommercial purposes. We're looking forward to hearing about creative applications of the linked data.

```xml
- <rdf:RDF>
  - <rdf:Description rdf:about="http://dewey.info/class/641/2009/08/about.en">
      <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
      <xhtml:license rdf:resource="http://creativecommons.org/licenses/by-nc-nd/3.0/"/>
      <cc:attributionName>OCLC Online Computer Library Center, Inc.</cc:attributionName>
      <cc:attributionURL rdf:resource="http://www.oclc.org/dewey/"/>
      <dct:isVersionOf rdf:resource="http://dewey.info/class/641/"/>
      <dct:language rdf:datatype="http://purl.org/dc/terms/RFC4646">en</dct:language>
      <skos:notation rdf:datatype="http://dewey.info/schema-terms/Notation">641</skos:notation>
      <skos:inScheme rdf:resource="http://dewey.info/scheme/2009/08/about.en"/>
      <skos:prefLabel xml:lang="en">Food & drink</skos:prefLabel>
      <skos:broader rdf:resource="http://dewey.info/class/64/2009/08/about.en"/>
      <cc:morePermissions rdf:resource="http://www.oclc.org/dewey/about/licensing/"/>
  </rdf:Description>
```

<skos:prefLabel xml:lang="de">**Essen und Trinken**</skos:prefLabel>

<skos:prefLabel xml:lang="fr">**Nourriture et boisson**</skos:prefLabel>

<skos:prefLabel xml:lang="es">**Alimentos y bebidas**</skos:prefLabel>

<skos:prefLabel xml:lang="sv">**Mat & dryck**</skos:prefLabel>

<skos:prefLabel xml:lang="pt">**Alimentos e bebidas**</skos:prefLabel>

<skos:prefLabel xml:lang="ru">**Продукты питания и напитки**</skos:prefLabel>

<skos:prefLabel xml:lang="zh">**食 & 品**</skos:prefLabel>

<skos:prefLabel xml:lang="ar">**الأغذية والمشروعات**./></skos:prefLabel>

VIAF: 10.7 million name authority records

Selected Covers

RICORDI

ISBN: 8875926913

Preferred Forms
- 100 ‡a Donizetti, Gaetano ‡d 1797-1848
- 100 ‡a Donizetti, Gaetano, ‡d 1797-1848
- 100 ‡a Donizetti, Gaetano, ‡d 1797-1848
- 200 ‡a Donizetti ‡b Gaetano ‡f 1797-1848
- 100 ‡a Donizetti, Gaetano, ‡d 1797-1848
- 100 ‡a Donizetti, Gaetano, ‡d 1797-1848

Donizetti, Gaetano, 1797-1848

Participating libraries are:
- Bibliothèque nationale de France
- Bibliotheca Alexandrina (Egypt)
- National Library of the Czech Republic
- Deutsche Nationalbibliothek
- National Library of Israel
- Library of Congress/NACO
- National Library of Sweden
- Vatican Library

We have also loaded several files into VIAF for testing:
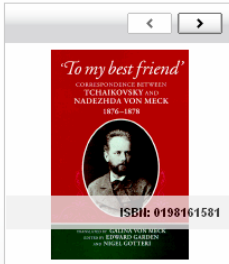- National Library of Australia -test
- Istituto Centrale per il Catalogo Unico (Italy) -test
- Biblioteca Nacional de Portugal -test
- Biblioteca Nacional de España -test
- Swiss National Library -test

www.viaf.org - making links among national authority records – not making a fused record – enables the best form to be selected depending on the user.

Different name forms fro different sources – French and Swedish form different from Australia, US, Israeli form.

In Progress: One page per work
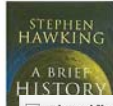
**A brief history of time : from the big bang to black holes**

162 editions in 21 languages, held by 11402 libraries

Table of Contents
Acknowledgments --Introduction /by Carl Sagan --[ch.] 1. Our picture of the universe --[ch.] 2. Space and time --[ch.] 3. The expanding universe --[ch.] 4. The uncertainty principle --[ch.] 5. Elementary particles and the forces of nature --[ch.] 6. Black holes --[ch.] 7. Black holes ain't so black --[ch.] 8. The origin and fate of the universe --[ch.] 9.
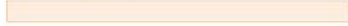
Classification

Dewey Decimal Classification

530.11 (6.98%)
532.1 (1.86%)
523 (0.27%)
521 (0.07%)
529.1 (0.01%)
Unclassified (1.29%)

523.1 (94.87%)
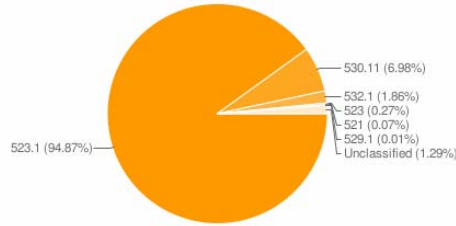
[View DDC Chart] [View LLC Chart]

Classify prototype

- Aggregates bibliographic data above the record level

- Contains all records for a work if any record has a DDC, LCC, or NLM classification number, including all language versions with or without class numbers

- Provides class number and FAST heading recommendations based on holdings

- Accessible through a user interface and as a Web service

Classify
An experimental classification web service

⚙ Classify is powered by the Classify Web Service.

Basic Search | Title/Author

New Search: [_____] [Title ▾] [Search]

Summary

Title: A brief history of time : from the big bang to black holes /
Author: Hawking, S. W.
Format: Books & Audio Books   **Editions:** 164   **Total Holdings:** 11402

Classification Summary

DDC:	Class Number	Holdings
Most Frequent	523.1	10507
Most Recent	523.1	10507
Latest Edition: 22	523.1	919
Latest Edition: 22	532.1	166
Latest Edition: 22	530.11	41

LCC:	Class Number	Holdings
Most Frequent	QB981	9805
Most Recent	TB6717	1

All | DDC | LCC

DDC

530.11 (9.05%)
532.1 (1.60%)
523 (0.23%)
521 (0.06%)
523.12 (0.02%)
Unclassified (2.60%)

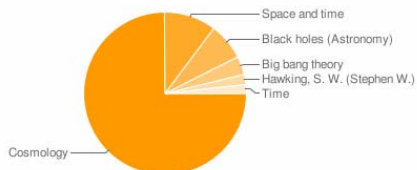523.1 (92.15%)

FAST Subject Headings

🖥 🌐 Internet   🔍 100% ▾

http://deweyresearch.oclc.org/classify2/classify?swid=041431706

17

FAST Subject Headings

1. Cosmology (10705)
2. Space and time (1481)
3. Black holes (Astronomy) (1050)
4. Big bang theory (523)
5. Hawking, S. W. (Stephen W.) (287)
6. Time (251)

Edition Details

Editions: 26 to 50 of 164

< First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last >

	Title / Author	Lang.	Holdings ▼	Tag	Class #	Library	Src.
26	Une breve histoire du temps : du big bang aux trous noirs Hawking, S. W.	fre	42	050	QB981		
				082	530.11		
				082	523.1		
27	A brief history of time Hawking, S. W.	eng	41	082	530.11		22
28	A brief history of time : from the big bang to black holes Hawking, Stephen W.	eng	40	082	523.1		19
29	A brief history of time from the big bang to black holes Hawking, S. W.	eng	36	050	QB981		
				082	523.1		
30	A brief history of time : from the big bang to black holes. Hawking, S. W.	eng	35	050	QC173.59.S65		
				082	530.11		20
31	Die illustrierte kurze Geschichte der Zeit Hawking, Stephen W.	ger	32				
32	A brief history of time Hawking, S. W.	eng	26	050	QB981		
				082	523		
33	A brief history of time : from the big bang to	eng	25	050	QC173.59		

Done | Internet | 100%

Transliteration

- Transliteration comprehension is local. Need many different schemes
  - Cilla in the UK did not accept the LC transliteration of Indic languages
- Could do transliteration "on the fly" for display BUT also require search capability which implies indexing
- Lack of MARC21 coding for the transliteration scheme is a problem
- Dutch use ISO tables with variations; tailored IBM ICU tables
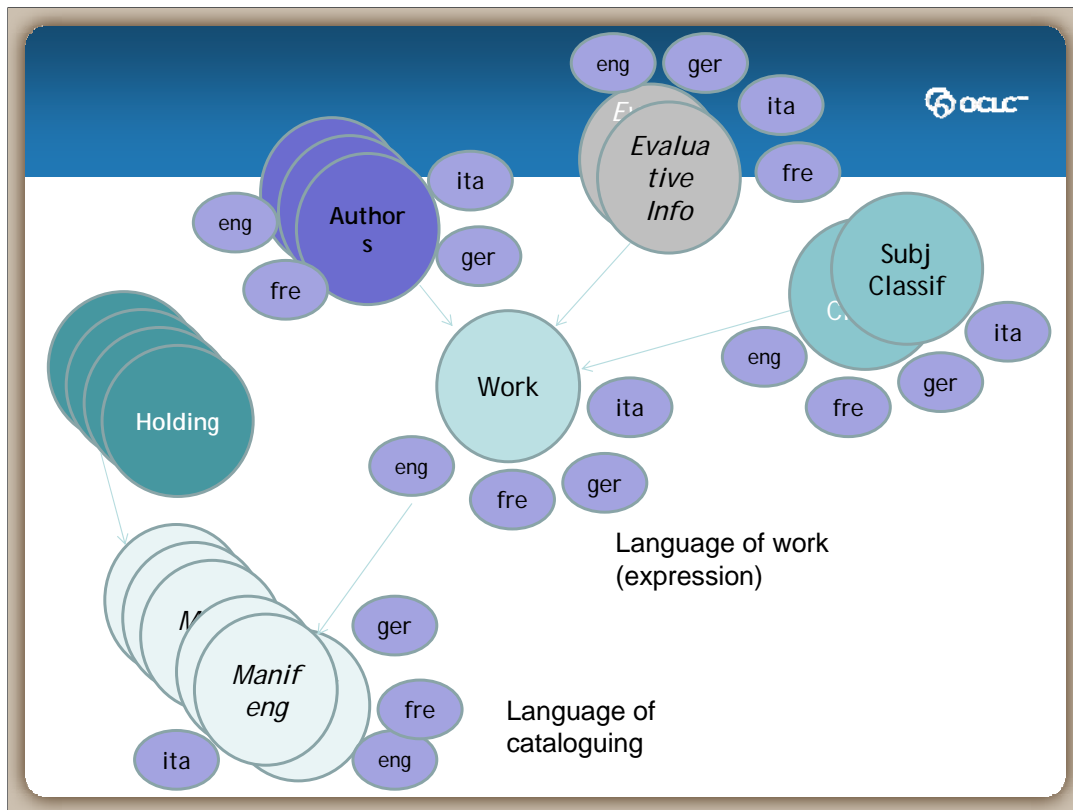
Where we are heading

Holdings are delivered at manif or work level regardless of the record to which they are attached. The "appropriate record" is displayed as an assemblage

Conclusion

- Data needs work on many fronts
- OCLC data better at some things **& all** is evolving
  - Language of cataloguing, Dewey, VIAF
- Need to evolve in tandem, especially vocabularies and to move together on transliteration
- APIs and Linked data

Grazie