CLASSIFICATION IN LIBRARY LINKED DATA ENVIRONMENT: opportunities and challenges

Aida Slavić
UDC Editor-in-Chief

UDC CONSORTIUM
Knowledge Organization Systems (KOS) thesauri, classifications, subject heading systems

Library Linked data (LLD) bibliographic data, authority data
- 2011 W3C Library Linked Data Incubator Group
- 2012 BIBFRAME
- 2014-2016 LD4L
Resource Description Framework (RDF)

Data presentation model: statements (subject, object, predicate) & resource description by (meta)data “nesting” and pointing to remote data sources.

RDF depends on the following principles

- each semantically significant unit or data set on the Web is described as a resource (a “thing”)
- each resource has a unique resource identifier (URI) which can be a Web address

everything that can be identified can be linked and pointed to
INFORMATION & KNOWLEDGE
in different domains
for different audience
by different agencies
in different form
of different kind
for human use
for machines

DATA SILOS
different
schema
data formats
standards
SCENARIOS
DISCLOSING DATA ON AN OPEN WEB OR NOT

LOD - OPEN LINKED DATA

LINKED DATA BEHIND DATA BARRIER

https://lod-cloud.net/
Conversion of non-RDF data

- the majority of linked data published on the open web is conversion data (this is the case for both knowledge organization systems as well as bibliographic data)

Native linked data

- creating data records using an interface that is integrated with an RDF data model - assumes adopting linked data on a conceptual level (native linked data cataloguing, BIBFRAME)
LINKED DATA
BIBLIOGRAPHIC CONTROL

Publishing bibliographic resources as linked data enables

- identification (as Web resources) any of the following: an entire library collection, a library catalogue, each bibliographic record, each element of the bibliographic record;

- linking individual data within a bibliographic description, such as author’s name, title, publisher, place of publishing, content description to additional information on a remote server;

- free access (by computer programs) to each data element of a bibliographic description and unlimited linking with semantically related resources on the Web.
Classmarks in bibliographic data
- contain no meaning in themselves
- they have to be interpreted externally

Classification schemes when published as linked data:
- can improve and extend semantics and increase number of access points in the process of resource discovery
- improve subject retrieval in a large number of collections in different languages / scripts in which a classification is used in content description
WHY CLASSIFICATION?
UDC IN SUPPORTING SEMANTIC INTEGRATION

Overcoming language barriers
- translations in over 50 languages; control of the meaning/verbal expressions using notation

Legacy data and international presence
- large amount of legacy data in collections world-wide

Mappings to other KOS
- subject heading systems, thesauri and other classifications

Contains additional semantic data
- UDC contains semantic data that is not present in library catalogues
SEMANTIC DATA
THE CASE OF UDC

HIERARCHICAL RELATIONSHIPS

LINKS CONCEPTS IN DIFFERENT AREAS OF KNOWLEDGE
</skos:ConceptScheme>
<skos:Concept rdf:about="http://udcdata.info/rdf/061284">
    <skos:inScheme rdf:resource="http://udcdata.info/udc-schema"/>
    <skos:broader rdf:resource="http://udcdata.info/rdf/059352"/>
    <skos:notation rdf:datatype="http://udcdata.info/UDCnotation">678</skos:notation>
    <skos:prefLabel xml:lang="en">Industries based on macromolecular materials</skos:prefLabel>
    <skos:related rdf:resource="http://udcdata.info/rdf/031559"/>
    <skos:related rdf:resource="http://udcdata.info/rdf/054896"/>
    <skos:narrower rdf:resource="http://udcdata.info/rdf/061480"/>
    <skos:narrower rdf:resource="http://udcdata.info/rdf/061498"/>
    <skos:narrower rdf:resource="http://udcdata.info/rdf/061521"/>
    <skos:narrower rdf:resource="http://udcdata.info/rdf/061555"/>
</skos:Concept>
Experimental phase

- all important KOS were published as open linked data (LOD)
- data access: data-dumps, SPARQL endpoint or a Web API
- problems with simplicity of SKOS

UDC namespace

- 2011 UDC Summary published as LOD
- 2013 UDC RDF STORE (licence protected data)
- 2018-2019 working towards combined approach
National Szechenyi Library (Hungary)

Author: Székely János (1929-1992)
Title (and responsibility): A nyugati hadtest / Székely János
Publication: Bp.: Magvető, 1982 (Győr: Széchényi)
Physical description: 243 p.; 17 cm
Classification no.: 5
894.511-32 (498)
Note: Regény
Holding institution: B1
Call number: A 20.149
MA 20.149

<dcterms:subject>
    <rdf:Description>
        <dcam:memberOf rdf:resource="http://purl.org/dc/terms/UDC"/>
        <rdf:value>894.511-32</rdf:value>
    </rdf:Description>
</dcterms:subject>
Trondheim - Library of Norwegian University of Science And Technology (NTNU)

```
<dc:subject rdf:about="#NTUB00002">
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
<skos:prefLabel xml:lang="no">Abelske varianter</skos:prefLabel>
<dcterms:udc>512.742</dcterms:udc>
</dc:subject>

<dc:subject rdf:about="#NTUB17121">
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
<skos:prefLabel xml:lang="no">Marine sopper</skos:prefLabel>
<dcterms:udc>582.28(26)</dcterms:udc>
</dc:subject>

<dc:subject rdf:about="#NTUB00005">
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
<skos:prefLabel xml:lang="no">Abrasiv slitasje</skos:prefLabel>
<dcterms:udc>620.178.162.44</dcterms:udc>
</dc:subject>
```
UDC in bibliographic records

- bibliographic records contain deprecated UDC notations cancelled 30-50 years ago;

- UDC is a synthetic classification and bibliographic records may contain complex UDC classmarks not present in a standard UDC edition;

- UDC data in bibliographic records are not accompanied by information about UDC source (edition/provenance);

- libraries lack resources or expertise to maintain subject authority files which would provide additional semantics and access points to UDC classmarks.
UDC number may be cancelled but its record and its ID stays permanently

- **DATE** (date of cancellation)
- **SOURCE** (issue of Extensions & Corrections in which this is published)
- **REPLACED BY**: ID of the class to which a cancelled UDC number is redirected
- **REPLACEMENT TYPE**: new class, colon combination; combination with common auxiliary; combination with special auxiliary; other
- **REPLACEMENT SEMANTICS**: exact match, to broader, to narrower, approximation
Reuse of notation for different meaning

cancelled notation

\[
\times 005
\]

previously represented

Organization

now at

001.82

notation reused for

+ 005 Management
Whenever UDC notation is re-used we record this change

! 582.62  Fagales
Notation History: Notation previously used for Hamamelididae, now at 582.62/.63.
Concept History: Fagales were previously at 582.632
Semantics: The scope of class narrowed

Notation history data

• **notation used for**: term describing concept for which the notation was previously used
• **old concept moved to**: ID of the class to which the concept was moved
• **date** of concept move
• **source** of concept move
UDC data assets

- UDC MRF database contains over 12,000 cancelled numbers (historical data) and their redirections to new numbers;
- UDC number lifecycle is recorded in the database and carefully administered
- We have parsing and validation algorithms that can be used to break UDC strings into components

Issues we are trying to resolve:

- Extending UDC SKOS schema (DC terms, LODE) – even better consider using OWL
- building a proper look-up service with parsing/validation option
- enabling service that would handle LOD and licence protected data
OWL-based UDC data model would offer many possibilities through formalization of data presentation and associated rules

• presenting and controlling structural, semantic and syntactic rules
• better control over UDC number life cycle
• better handling of publishing models/rules

UDC ontology would better support automation of

• classification expansion and maintenance
• publishing and translating
• classification use
Linked data enables library problems to be solved outside libraries

- Libraries should not need to worry about resolving the semantics or parsing the components of UDC codes.

- UDC linked data should be supported by a front-end “resolution service” – which would enable parsing, validating and resolving URI for UDC codes.

- UDC namespace, i.e., UDC RDF triple store should contain all data necessary to resolve and interpret strings coming from library catalogues (including historical UDC data).
LINKED DATA POTENTIALS

- benefits for both libraries and KOS communities
- aligns library subject data with the latest knowledge organization tools without libraries having to do re-classifications / re-indexing or re-organizations of collections
- serves as a vehicle for interpreting, validating and enriching and serving back enhanced subject data to libraries
- helps in the discovery of knowledge in heritage collections via subject data (data mining)
FINAL THOUGHTS
EXPERIMENTAL PHASE IS OVER

UDC is used in over 180,000 libraries and is translated in over 50 languages

- UDC namespace and terminological service should be based on real life scenarios;
  - the intended use of data
  - functionalities that needs to be supported
  - governance and curation of UDC data kept in bibliographic domain

- We need to know more about what libraries plan to do with linked data
THANK YOU