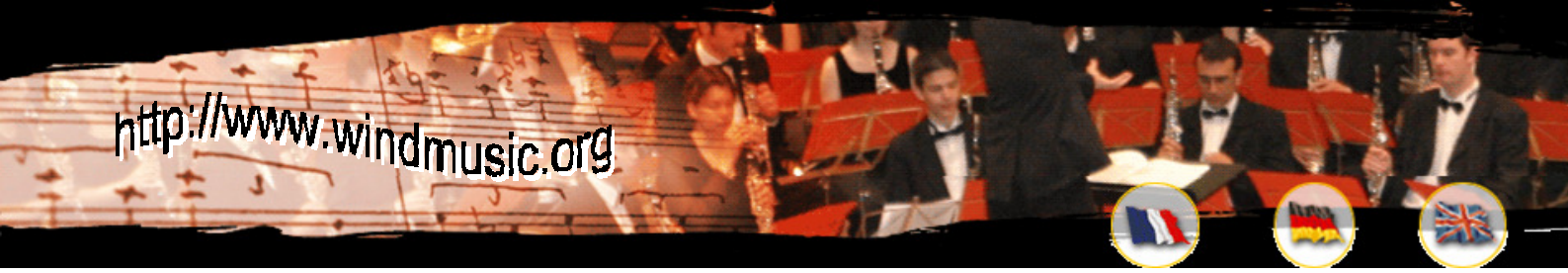


WindMusic, example of the new possibilities for DSpace when adding SKOS thesaurus and authority lists management

DSUG 2009 – Göteborg, SE

Authors:

- Christophe Dupriez, Computer Scientist, DESTIN inc., rue des Palais 44 boîte 82, Bruxelles, Belgium 1030 : dupriez@destin.be
- Julien Schubnel, CDMC, Les Dominicains - B.P. 95, Guebwiller Cedex, France 68502 : julien@cdmc68.com

Introduction:

WindMusic is one of the most complete reference sources for music scores and recordings for Wind Orchestras and Bands. WindMusic is multilingual: French, German and English. This computer based catalogue allows visitors to discover full scores of interesting works and to listen directly their recordings with headphones on multimedia stations.

The CDMC (Conseil Départemental de Musique et de Culture de Haute Alsace = *Upper Alsace Council for Music and Culture*) is an organization supported by public founds since 1969. It promotes the cultural production (music, dance and theatre) of the Upper Alsace Department.

The CIM (CDMC's Music Information Centre) wanted to give the trainees, teachers, musicians and conductors a structured information tool, which would fit their needs. All types of documents are available: from encyclopaedias to CD-ROMs, with periodicals, monographs, but mainly scores (with about 20.000 items) and audio recordings (with about 25.000 items).

Created in 1996, the WindMusic database was encoded in UniMarc with the strict validations and authority lists enforced by the Loris Library Information System (provided by Ever).

Aiming to a larger accessibility, the database was made available on the Internet in 2000 (http://www.windmusic.org). Audio recordings are only available within the CIM to respect publishers' copyright.

WindMusic has been recently re-implemented using DSpace and a whole set of functions to manage and take advantage of authority lists and thesauri.

Working for the CDMC and for the Belgium Poison Centre, DESTIN provided the development services necessary for this realisation.

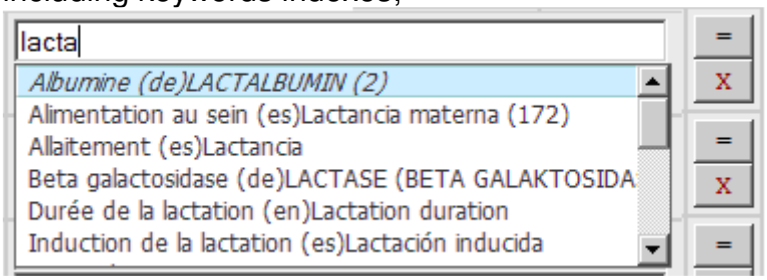
This development was also triggered by the needs of the internal scientific documentation database of the Belgium Poison Centre (Dutch, French and English); Background document: http://convegni.cilea.it/conf/viewpaper.php?id=197&print=1&cf=11

These results are there to share in new projects with the DSpace community.

Main Motivations

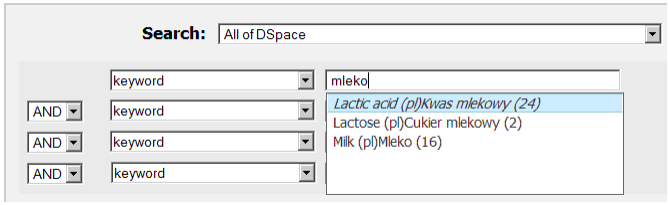
Authority lists and thesauri are there for better:

- Cataloguing:
  - To help cataloguers choose the right terms to fill-in metadata fields, including keywords indexes;

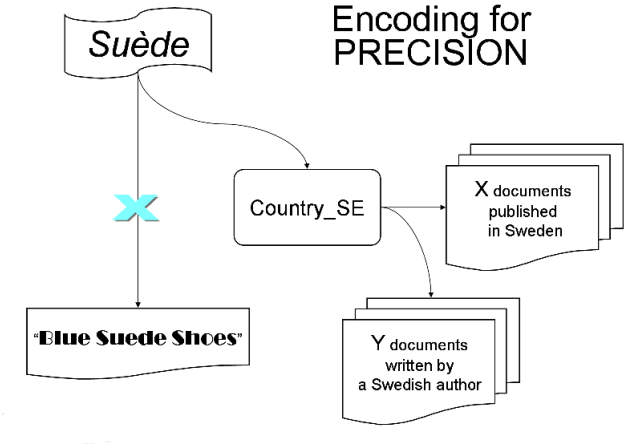


  - Control of Data quality (no incorrect spelling of terms)
- Display:
  - To store “concept codes” which are translated in the user language whenever they are displayed: multilingualism is something free text fields cannot provide without manual translations;

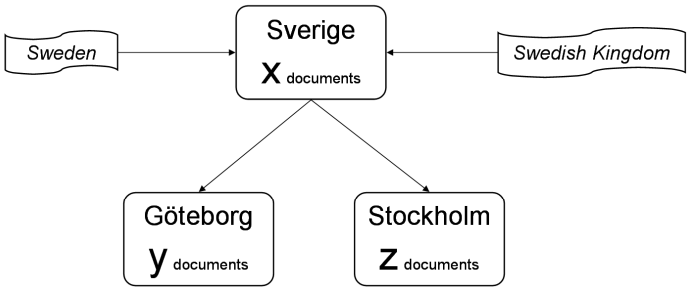
- Searching:
  - Auto-complete to help specify search criteria



- To retrieve a precise concept (with or without its specific concepts) using a list, a tree or an auto-completed field;

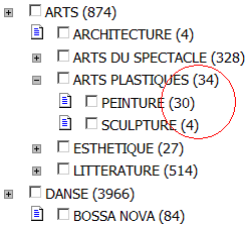


- To allow multilingual word searches on the preferred terms but also synonyms (exhaustivity);
- To use concepts hierarchies to encompass all specific topics (exhaustivity);



$$\text{Exhaustivity} = x \cup y \cup z$$

- Navigation:
  - To provide an overview of the database content (and size) by classification or hierarchies of subjects;



  - To provide clues about the size of the horizontal searches results a.k. provide clues about the coverage of the database for any list of concepts;
  - To propose facetted search: which concepts are frequently used in any search result? (to be implemented)

Authority lists and thesauri

Within a metadata field, it is now possible to refer to a concept in a Thesaurus, an entry in an Authority list, a code in a table and enable users:

- to get the name of the concept / entry / code in the users' language;
- to see the number of records in various application and/or various indexing roles linked to a given concept / entry / code ;
- to display complete information about a concept / entry / code (translations, synonyms, definitions, usage statistics, thesaural relations, etc.);
- to search using any words of any translation or synonyms of concepts / entries
- to search a precise concept / entry by first identifying the concept / entry (auto-complete, index browsing) and then receiving the linked records.

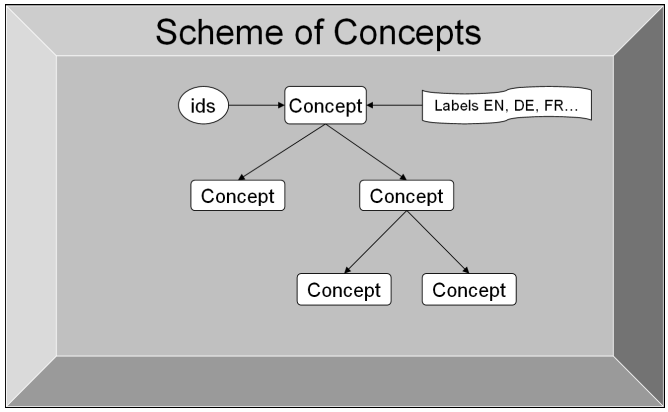
This development is based on the terminology and the recommendations of the SKOS (Simple Knowledge Organisation System) standard established by the W3C: http://www.w3.org/TR/skos-primer/

Some more features have been added foreseeing the size and the complexity of vocabularies like the MeSH (Medical Subject Headings) with its Supplementary Concept Records (http://www.nlm.nih.gov/pubs/factsheets/mesh.html).

Each metadata field can be controlled by one or more "concept schemes" (authority list, thesaurus).

A "concept scheme", as defined by the SKOS specification published by the W3C (http://www.w3.org/2004/02/skos/), is uniquely identified with an URL.
It is a collection of "concepts", each concept having:

- an identification code (RDF "about" URL),
- aliases to this code,
- preferred labels,
- synonyms,
- hierarchical relations,
- "see also" relations,
- scope notes,
- editorial note,
- codes in other notation systems,
- etc.



SKOS data is encoded in any format supporting RDFS (XML, N-triples, N3, etc.).

We propose a simplified way to encode, within DSpace metadata fields, a reference to a SKOS concept:

1. a Scheme identification code (a single word made of letters or digits, not a full URL) e.g. *Country*,
2. an underline character ("_"),
3. the concept identification code itself (any string of letters, digits or underscores) e.g. *SE*.

Country_SE is therefore the simplified way to refer to an URI like "http://www.iso.org/resources/3166-2/version-x/SE": *Country* can also be made equivalent to a different URI whenever the ISO standard changes.

This simplified encoding also ensures that one concept reference is seen by Lucene Search Engine (the search engine used in DSpace) as ONE (and only one) specific, non ambiguous word and that precise search is possible (Precision is THE major request from our users).

- Some prefixes before a reference can specify a more specific role than what the index allows. For instance: `composer_ persons_12345` specifies that the person code 12345 is the composer of current record. Prefixes are codes terminated by an underline character.
- Some suffixes after a reference can specify a quantity or any attribute of the relation. For instance: `instrument_piano _2` specifies that 2 pianos are requested for this music score. Suffixes are codes beginning by an underline character.

Lucene proximity search then allows to find the use of a concept with a given role and/or quantity (for instance, “flutes 3” find music scores for three flutes).

Main characteristics of the implemented system:

Many if not most Java information management tools are combining a storage mechanism and a Lucene search engine (good ones like DSpace easily recreate their Lucene search indexes from their main storage).

We therefore did not modify the DSpace database structure: only the concept codes, prefixed by their scheme code, are stored in the fields. The database remains orthogonal, no information is duplicated: the preferred terms, the synonyms, the translations can change without having to update the DSpace records.

DC Field	Value	Language
dc.contributor.author	Compositeur_ loraut_16196	-
dc.contributor.arranger	Arrangeur_ loraut_13614	-
dc.date.accessioned	2009-04-20T14:35:50Z	-
dc.date.available	2009-01-05T04:25:18Z	-
dc.date.issued	1983	-
dc.date.submitted	2008-01-30	-
dc.identifier	036293	-
dc.identifier.other	210.1 TEM	-
dc.identifier.uri	http:/dspace/handle/68502/56282	-
dc.identifier.loris	70231	-
dc.identifier.publisher	R4519M	-
dc.description.abstract	Cet album mythique des années 80, produit par Quincy Jones, est l'album le plus vendu de tous les temps avec près de 60 millions de copies. Dédié à la mère de Michael Jackson, Thriller est une collection de tubes dont la chanson du même nom. Cette orchestration met en valeur son fameux Groove. Le thème est distribué à tous les pupitres ponctués de riffs de cuivres. Simple et facile à jouer, cet arrangement surprend par son efficacité.	-
dc.audience	niveau_Orchestre_Difficulte_2	-
dc.format.extent	000200	-
dc.format.medium	15 p.	-
dc.publisher	lorpub_15	-
dc.publisher.country	Charnay-lès-Mâcon, France	-
dc.subject	lorthes_457	-
dc.subject	lorthes_381	-
dc.subject	lorthes_499	-
dc.subject	lorthes_393	-
dc.subject.orchestra	typeOrchestre_orchestre_d_harmonie _1	-
dc.title	Thriller	-
dc.type.score	typePartition_Orchestre_a_Vent_Orchestration_Complete	-
dc.type.record	notice_breve	-
Appears in Collections:	1.1. Concert Band Scores	
Submitted by:	gerard	

We rely on Lucene (the standard search engine used in DSpace) to receive the expansion of the indexation to insure:

- Precise searches based on the concept codes (*Country_SE*);
- Searches on the concept labels in any language (*Sweden, Sverige*);
- Searches on the concept synonyms in any language (*Swedish Kingdom*);
- Searches including the narrower concepts in the hierarchy (*Stockholm, Göteborg*);
- Searches including the narrower concepts in the "narrowerMatch" hierarchy (to retrieve, for instance, all the bibliographic references with an author with a given nationality).

The general "à la Google" search remains and is even enriched with all synonyms and translations for every concept references.

Whenever index expansion rules are changed, the existing DSpace Indexation function is enough to rebuild coherent Lucene indexes.

The Lucene tokenizers are modified to ensure they do not lemmatize or otherwise transform words containing an underline character.

Concepts data sources can be:

- Any SQL database (**including the current or any *other* DSpace application**: in http://www.windmusic.org, keywords of the Index are coming from one of the DSpace collections; authors are another one; publishers idem)
- Statically (authority list loaded at start up time) or dynamically (entries are retrieved and cached as needed: the concepts list may be enormous)
- Any CSV file
- XML files: Whole FAO Agrovoc is loaded in 7 seconds from an XML file (nearly 30 thousand concepts in more than 20 languages). Belgium Poison Centre users need similar performance with the whole MESH in 4 languages.
- SKOS/RDF files will be implemented using RIO and/or integration with Standford Protégé

The Java classes implementing the SKOS data structure is mapped to XML (using JaxP and an XSD file) which ensures very efficient marshalling / unmarshalling to XML files.

The uses of each concept in different indexes (even different applications) are counted: those counts are useful to give clues to users about how much data is linked to each concept displayed. Horizontal searches are also provided, whenever possible, to retrieve all records for a given concept mentioned in some display.

Ajax Auto-complete is used for Metadata Fields updates and for Searches: the user types a few letters and proposals are made with terms (preferred or synonyms) in any language. A servlet has been added for this: it receives four parameters:

1. letters typed by the user
2. user preferred language
3. concept scheme to explore
4. Lucene index to use for usage counts

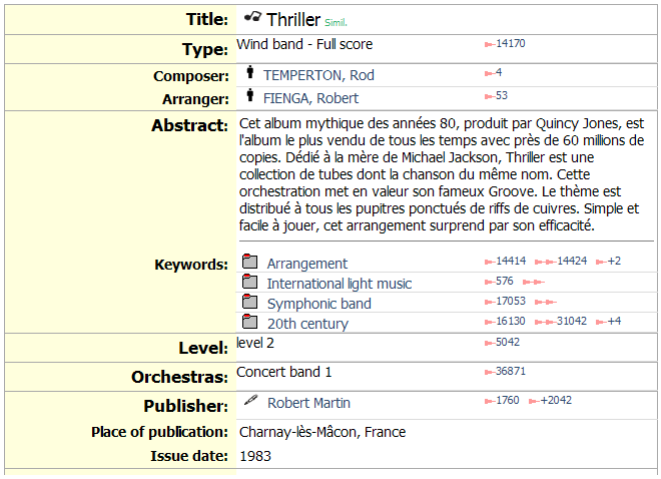
Inter-relations within an application or with other applications:

Dynamic SQL sources allow managing relations between DSpace records. For instance, in WindMusic, there are different collections for CDs and tracks on CDs: the CDs collection is one SKOS source controlling the field “ispartof”. Example: http://www.windmusic.org/dspace/handle/68502/41328

Our general aim is to ensure that any accessible XML/CSV file or SQL database (and other future format to be developed) can be used to link different applications through their common concepts. **Multiple applications (DSpace or others) can be linked**: we will be working to integrate this SKOS tool in JSPWiki to control page names, relations between pages and external applications like DSpace.

Concepts usage information allows listing all the applications effectively using a concept. Therefore any application can make hyperlinks to retrieve the concept in many others (only where an answer is available).

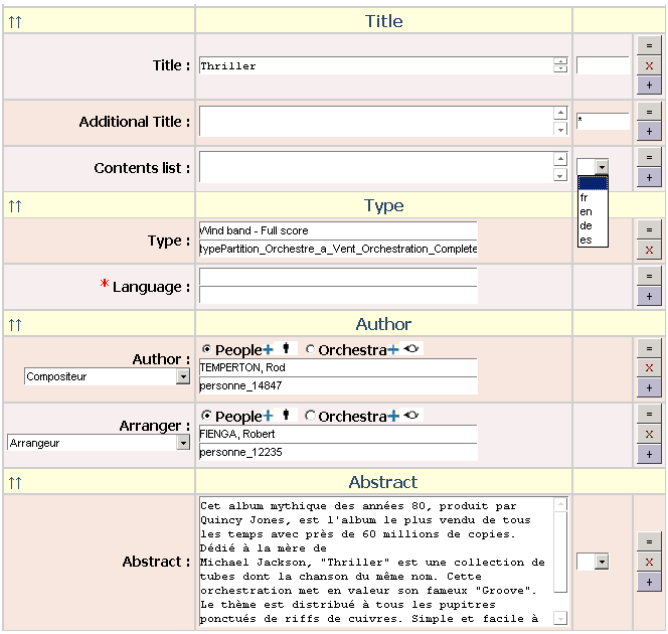
Metadata Display:



- **Title:** Fields presentation is prepared using little HTML templates (JSP UI). Here a collection icon is added and also a search for similar titles (Simil.)
- **Type, Level, Orchestras:** These fields are controlled by small and stable SKOS authority lists. An “horizontal” search is proposed with figures indicating the number of records potentially retrieved.
- **Publisher:** There is a SKOS Scheme (stored in a DSpace collection) for allowed publishers. The two different figures are respectively:
    - the number of CDs or Music Scores published by this publisher
    - the number of Recordings (tracks) on those CDs
- **Author, Keywords:** The role (Composer) of the Author is indicated in the left margin. Theses fields are controlled by Dynamic SQL SKOS Authority lists. In this example, the permitted entries (and their translations) are stored in DSpace collections.
  The number of related records is shown on the right. If there are multiple results possible (for instance, search on a precise concept + search when using it with all its specific concepts), multiple numbers are shown.

Metadata Update Form:

The update form has been completely overhauled:



1. Fields are now in the same order than for the record display; Supplementary fields are allowed at the end.
2. A table of content for groups of fields is provided (direct access to a “section”)
3. Numbers are validated
4. Short lists are shown as menus (Honorific Title for instance)
5. Buttons are “=” to save all updates, “X” to remove current metadata, “+” to add an empty occurrence.
6. Empty metadata fields are mentioned and receive only a “+” button (add).
7. Multiple concept schemes are supported (for instance, persons and orchestras in the field for “performers”): the user can choose which one is used for auto-complete. The user can also press “+” to add a new author or orchestra.
8. A menu of allowed languages can be associated to any field and constrained differently for each.

Our customers have measured that with this form, they achieve the same efficiency to update a record with DSpace than with specialized library management software**.**

A Search Result:

This is the result if one chooses “Arts Plastiques” in the Subject Search Index. It is sorted and contains exactly the records using the selected subjects and their specifics (if so configured).

- Circled in red:
  - the author names coming from the Persons (or Orchestras) authority list;
  - their roles (composer, arranger, etc.) are also coded and translated using an SKOS Concept Scheme of the « Authors’ roles »;
  - the type of the record coded and translated in the user language, using a SKOS Concept Scheme of the records’ types.
- Circled in green, the print basket function allows choosing records to be printed together. An export function could be based on this mechanism.
- Circled in blue, the possible sort keys (configured for this application). The second column displays the sort key for each records (here the availability date of each document). The user can choose anytime a different sort key.

Thesaurus management directly in DSpace:

Subjects can be catalogued in a DSpace collection like any other bibliographic record: http://www.windmusic.org/dspace/handle/68502/22360

- A title may be specified for each language (translation in progress for WindMusic)

- Next to the title, the two numbers are for “horizontal searches”:
  1. the number of records specifically linked to the concept (here: 1)
  2. the number of records linked to the concept or any of its specifics (here: 34)
- dc.title.alternative can be used for synonyms.
- “in:” (dc.relation.ispartof) indicates that this concept is a subdivision of ARTS. The number of linked records is also indicated for ARTS, the parent concept (the more generic subject).
- “lorthes_565” is a SKOS notation, equivalent to keyword “sujet_22360” but in another application.

Other Examples:

- DSpace index dynamically created using keywords records stored within a DSpace collection: http://www.windmusic.org/dspace/subject-search
- Advanced search modified to provide concepts lists on many search criteria: http://www.windmusic.org/dspace/advanced-search
- Custom Search forms can use the concepts lists either for auto-complete either for menus: http://www.windmusic.org/dspace/scores-search
- Search results with sort possibilities
- Records display with "horizontal searches": http://www.windmusic.org/dspace/handle/68502/24405
- Corresponding "detailed record" gives a clue about concepts lists encoding (conceptScheme_conceptAbout): http://www.windmusic.org/dspace/handle/68502/24405?mode=full
- Concept records can reside within DSpace collections: http://www.windmusic.org/dspace/handle/68502/627
- A search on this concept record: http://www.windmusic.org/dspace/simple-search?query=author:personne_627
- Horizontal search on a specific keyword: http://www.windmusic.org/dspace/simple-search?query=keyword:sujet_22302
- Horizontal search on a keyword and all its narrower terms: http://www.windmusic.org/dspace/simple-search?query=broadkeyword:sujet_22302
- I18n issues are addressed, non latin alphabets included

Future developments:

Future developments, that we would be very happy to share in terms of results but also in terms of workload, could be:

- Support of SKOS RDF representation (input and output). Stanford Protege could then be used to edit a thesaurus or a authority list;
- Support of ISO 25964 XML Schema and semantic;

- SKOS module is independent of DSpace but not Lucene indexer: we would like to "plug-out" DSIndexer and replace it by SolR which would provide better flexibility (and facetted browsing) and possible integration to other search interfaces like BlackLight.
- DCAP (Dublin Core Application Profile) is a standard to define the necessary parameters for metadata field validation and management: aligning our work on this standard would enable the reuse of other tools like the SHAME metadata editing framework (http://shame.sf.net).
- SKOS management applications: this module is in reality a "link manager" between applications; the cataloguer/indexer certainly appreciate to have a suite of validation tools which examine the authority lists/thesaurus (search for ambiguous terms, missing translations, relation loops, etc.) but also their usage by the different applications
- Federated search: tools like Carrot2 allows parallel search in many applications.

Integration in DSpace "trunk":

This development touches many parts of DSpace (input forms, controlled vocabularies, update form, search forms and results, item display, configuration, metadata registry, etc.). It has been made on the basis of DSpace 1.4.1/2 JSP-UI.

With the support of new projects from the DSpace Community, more specific contributions could be defined for easier integration and maintenance by all interested members of the Community.

For instance:

- Uniform use of "*vocabularies*" for every terms appearing in the user interface (metadata included): those vocabularies could come from local or remote sources: SQL, SKOS, ISO 25964, CSV files, property files, Web Services, etc.
- "plugged" indexing and retrieval engine: Apache SolR (this would also open DSpace to direct interfacing with UVA BlackLight);
- Dublin Core Application Profile (DCAP) for metadata validation + Annotation Profile for an RDF based parameterisation of metadata maintenance (SHAME metadata editor);
- Flexible, broad and standardized classes for DSpace object accesses (display, interlinking, search and retrieval, updates, etc.) from User Interface scripts (simplification of JSP UI using EL, enrichment of XML UI flexibility, richer AJAX interfacing, possible integration with other frameworks like Tapestry…)

Conclusion:

This work paves the way for a much more precise encoding of metadata in DSpace. It adds many new functions to assist visitors to find what fits most their needs and cataloguers to choose the right key elements (keyword, performer, composer, arranger, instrument, etc.) to index the documents in the database.

cdmc
Conseil Départemental pour la Musique et la Culture de Haute-Alsace

Destin