

Aplicacions per a la gestió de taxonomies

Seminari DigiDoc
Miquel Centelles
24.04.2006

SUMARI

- Nou marc normatiu dels vocabularis controlats.
- Quina era la situació del mercat abans de la norma NISO Z39.19-2005.
- Quines noves exigències incorpora la norma NISO Z39.19-2005.

Nou marc normatiu dels vocabularis controlats

- *NISO Z39.19-2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* (<http://www.niso.org/committees/MT-info.html>).
- *BS 8723: Structured Vocabularies for Information Retrieval – Guide* (<http://www.bsi-global.com/News/index.xalter>).
- *IFLA Guidelines for Multilingual Thesauri* (<http://www.ifla.org/VII/s29/wgmt-invitation.htm>).

Canvis incorporats per la norma NISO Z39.19-2005

Versions anteriors a 2005

- Cobertura: documents.
- Tipus de vocabularis: thesaurus.
- Post-coordinació.
- Entorns analògics.
- Vocabularis monolingües.

Versió de 2005

- Cobertura: recursos d'informació (*content objects*).
- Tipus de vocabularis: llistes, anells de sinònims, taxonomies i tesaurus.
- Pre-coordinació.
- Entorn Web.
- Vocabularis multilingües (general).
- Interoperabilitat.
- Anàlisi de facetes.

Àmbits d'incidència de les aplicacions informàtiques

- Apartat 11.1.3.4: Assistència informàtica per a determinades tasques.
- Apartat 11.4: Sistemes de gestió de vocabularis controlats.

Assistència informàtica per a determinades tasques (11.1.3.4)

- La norma assumeix que la construcció del vocabulari controlat es fonamenta en decisions intel·lectuals.
- L'assistència informàtica pot aplicar-se per a tasques d'identificació de termes com:
 - Identificació de termes candidats.
 - Registre de freqüències dels termes.
 - Registre de termes procedents de les cerques dels usuaris.

Sistemes de gestió de vocabularis controlats (11.4)

■ Recomanacions generals dels sistemes/1:

- Ha de dependre i/o ha de ser una part integral d'un sistema d'informació complet.
- Ha de poder gestionar les necessitats especials d'un vocabulari controlat.
- Ha de ser el suficientment flexible per a permetre l'aprofitament de les tecnologies emergents.
- Ha de donar cobertura, com a mínim, a les relacions bàsiques d'U/UP, TG/TE, i TR.
- Ha de donar cobertura a les notes d'abast (NA) i d'història del terme (HT).

Sistemes de gestió de vocabularis controlats (11.4)

■ Recomanacions generals dels sistemes/2:

- Ha de permetre les presentacions jeràrquica i alfabètica.
- Ha de ser preferiblement no propietari.
- Ha de ser multiusuari.
- Ha de ser independent del maquinari
- Ha d'operar en un sistema obert o en un sistema operatiu estàndard.
- Ha d'exigir un entrenament mínim als usuaris.
- Ha de proporcionar documentació detallada als usuaris.

Sistemes de gestió de vocabularis controlats (11.4)

■ Funcionalitats que han d'oferir els sistemes/1:

- Diversitat tipogràfica.
- Ordenació de caràcters alfabètics i numèrics.
- Diferents tipus de presentacions dels vocabularis: alfabètica, jeràrquica, etc.
- Definició de camps descriptius dels termes.
- Registre i presentació dels termes.
- Referències creuades entre termes.

Sistemes de gestió de vocabularis controlats (11.4)

■ Funcionalitats que han d'oferir els sistemes/2:

- Eliminació de termes.
- Incorporació i designació de termes i relacions candidates.
- Detecció d'errors.
- Recuperació (*searching*) i exploració (*browsing*) de termes i relacions entre termes.
- Generació d'estadístiques i informes.
- Sistemes de comprovació i avaluació.

Quina era la situació del mercat abans de la norma NISO Z39.19-2005

■ Basat en Morante i Walker (2003).

■ Quatre tipus d'eines:

- Eines de gestió de taxonomies o gestors de taxonomies.
- Vocabularis pre-elaborats.
- Funcions integrades en aplicacions més àmplies:
 - Editors de taxonomies.
 - Extracció de conceptes.
 - Sistemes de *clustering*.
- Solucions integrades.

Quines noves exigències incorpora la norma NISO Z39.19-2005

■ Atenció a les diferents aproximacions per a la construcció de vocabularis controlats:

- L'enfocament del comitè (*committee approach*).
- L'enfocament empíric (*empirical approach*).
- Combinació de mètodes.

■ Atenció a les diferents fonts dels termes i de les relacions entre els termes:

- El llenguatge natural dels recursos d'informació o *content objects* (justificació bibliogràfica).
- El llenguatge dels usuaris (justificació dels usuaris).
- Les necessitats i prioritats de l'organització (justificació organitzativa).

■ Atenció a la interoperabilitat: l'habilitat de dos o més sistemes o dels components dels sistemes per a intercanviar informació i per a utilitzar la informació intercanviada sense que exigeixi un esforç especial per part d'algun dels sistemes.

El llenguatge natural dels recursos d'informació...

- **Extracció de termes de documents representatius de l'organització.**
- **Què volem trobar?/1**
 - Relació de termes i variants de termes (lèxic).
 - Informació sobre la freqüència amb què apareixen a les fonts usades.
 - Informació sobre la quantitat de fonts a les quals apareixen.
 - Informació sobre co-ocurrència: termes que apareixen sempre junts.

El llenguatge natural dels recursos d'informació...

- **Què volem trobar?/2**
 - Relació de frases freqüents.
 - Relació de paraules buides, termes que no es consideren útils per representar conceptes.
- **Programes d'extracció de termes/coceptes:**
 - TerminologyExtractor / Chamblon
<http://www.chamblon.com>
 - WordSmith Tools 4.0
<http://www.lexically.net/wordsmith/index.html>

El llenguatge dels usuaris

- **Fonts dels termes: registres de transaccions de visites i de cerques.**
 - Aplicacions d'anàlisi del web. Exemples:
 - Web Log Storming <http://www.dataandsoftware.com/weblog/index.html>
 - WebSTAT <http://www.webstat.com>
 - WebTrends <http://www.webtrends.com>
 - Extracció de registres de consultes dels cercadors. Exemples:
 - En general, tots els cercadors generen registres de transaccions de cerca. SearchTools.com <http://www.searchtools.com/tools/tools.html> relaciona aproximadament 200 productes que es poden implementar com a enginy de cerca intern en un lloc web, i possiblement cada un d'ells té el seu format propietari d'elaboració de registres de consultes.
 - Específicament, HBX On-Demand / WebSideStory <http://www.hbxondemand.com> ha desenvolupat una aplicació de anàlisi de logs para los usuarios de Atomz.
 - Aplicaciones de análisis de la recuperación. Example: BehaviorTracking / Mondosoft <http://www.behaviortracking.com/reports.asp>

El llenguatge dels usuaris

- **Fonts de relacions entre termes:**
 - Anàlisi de coocurrència de termes mitjançant els programes d'extracció de termes.
 - Anàlisi de les rutes de navegació del lloc web:
 - Mostra d'una aplicació d'anàlisi de visites al web: WebSTAT <http://www.webstat.com/demo.php>
 - Representació gràfica de les rutes de navegació: Pthalizer <http://pathalizer.sourceforge.net>
 - Test d'ordenació de targes. Example: CardZort i CardCluster <http://condor.depaul.edu/~itoro/cardzort/cardzort.htm>

Interoperabilitat: estàndards per a la representació i intercanvi de vocabularis controlats

- **Diversitat d'estàndards aplicables:**
 - XML.
 - RDF.
 - Conjunts d'elements i esquemes de metadades per a vocabularis controlats.

XML

- **ADL Thesaurus Protocol** <http://alexandria.sdc.ucsb.edu/~qjane/thesaurus>
 - "The ADL Thesaurus Protocol is a lightweight, stateless, XML- and HTTP-based protocol for accessing *thesauri*: structured, controlled vocabularies of words and phrases that represent conceptual categories. The protocol is based on the Z39.19 thesaurus model and supports downloading, querying, and navigating thesauri."
- **Zthes: a Profile for Thesaurus Navigation in Z39.50 and SRW** <http://zthes.z3950.org>
 - "The Zthes profile describes an abstract model for representing and searching thesauri - semantic hierarchies of terms as described in ISO 2788 - and specifies how this model may be implemented using the Z39.50 and SRW protocols. It also suggests how the model may be implemented using other protocols and formats."
- **Altres "tradicions":**
 - Tradició de la terminologia. Exemples: TC 37; TBX, implementació de TMF; etc.
 - Tradició dels Topic maps. Example: XTM.

RDF

- W3C Recommendation: *RDF Vocabulary Description Language 1.0: RDF Schema* <http://www.w3.org/TR/2004/REC-rdf-schema-20040210>
- W3C Working Draft: *Quick Guide to Publishing a Thesaurus on the Semantic Web* <http://www.w3.org/TR/2005/WD-swbp-thesaurus-pubguide-20050517>
- SKOS Core i SKOS Core Vocabulary:
 - "SKOS Core provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies.
 - The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF), that can be used to express a concept scheme as an RDF graph. Using RDF allows data to be linked to and/or merged with other data, enabling data sources to be distributed across the web, but still be meaningfully composed and integrated."
 - SKOS Core RDF Vocabulary: <http://www.w3.org/2004/02/skos/core/>
- RDFS/OWL:
 - "The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDFS) by providing additional vocabulary along with a formal semantics. OWL has three increasingly-expressive sublanguages: OWL Lite, OWL DL, and OWL Full."
 - W3C Recommendation: *OWL Web Ontology Language* <http://www.w3.org/TR/2004/REC-owl-features-20040210>

Conjunts d'elements i esquemes de metadades...

- Apartat 5.5 de *NISO Z39.19-2005: Metadades i esquemes de metadades de vocabularis controlats*
 - "Les metadades es poden fer servir amb vocabularis controlats de diferents formes:
 1. Utilització d'un vocabulari controlat com a font per als termes que es poden aplicar a un element particular de les metadades.
 2. Utilització de metadades per a la descripció d'un vocabulari controlat en el seu conjunt per al descobriment de recursos.
 3. Utilització de metadades i d'un esquema de metadades per a representació del contingut complet del vocabulari controlat."

Conjunts d'elements i esquemes de metadades...

- Aplicacions de metadades als vocabularis controlats
 - *eXchangeable Faceted Metadata Language (XFML)* <http://www.xfml.org/>
 - "XFML Core is an open XML format for publishing and sharing hierarchical faceted metadata and indexing efforts. XFML Core is lightweight and easy to implement, yet uniquely powerful."
 - *Language Independent Metadata Browsing of European Resources (LIMBER)* <http://www.limber.rl.ac.uk/>
 - "LIMBER is developing tools to support multilingual access to data distributed across the world wide web by using metadata and a multilingual thesaurus of terms in a restricted vocabulary."
 - *Vocabulary Definition Exchange (VDEX)* <http://www.imsglobal.org/vdex/>
 - "The IMS Vocabulary Definition Exchange (VDEX) specification defines a grammar for the exchange of value lists of various classes: collections often denoted "vocabulary". Specifically, VDEX defines a grammar for the exchange of simple machine-readable lists of values, or terms, together with information that may aid a human being in understanding the meaning or applicability of the various terms. VDEX may be used to express valid data for use in instances of IEEE LOM, IMS Metadata, IMS Learner Information Package and ADL SCORM, etc. for example. In these cases, the terms are often not human language words or phrases but more abstract tokens. VDEX can also express strictly hierarchical schemes in a compact manner while allowing for more loose networks of relationship to be expressed if required."
 - *Vocabulary Markup Language (VocML)* <http://nkos.slis.kent.edu/VOCML-1.DOC>
 - "The Vocabulary Markup Language (VocML) supports the structured representation of a wide range of KOS resources, "including authority files, hierarchical thesauri (including those with polyhierarchies), classification schemes, digital gazetteers, and subject heading lists."