bid textos universitaris de biblioteconomia i documentació
ISSN 1575-5886
DL B-19.675-1998
número 13
desembre 2004
Facultat de Biblioteconomia i Do
Universitat de Barcelona

L'automatització de tesaurus i la seva utilització en el web semàn

[Versión castellana]

JOSÉ RAMÓN PÉREZ AGÜERA
Departamento de Biblioteconomía y Documentación
Universidad Complutense de Madrid
jose.aguera@ccinf.ucm.es

Resum [Abstract] [Resumen]

Es presenta una proposta bàsica d'automatització i utilització de tesaurus documentals en entorns distribuïts de recuperació d'informació mitjançant serveis web basats en l'arquitectura RDF (*resource description framework* o marc de descripció de recursos). Per aquest motiu, es revisen, en primer lloc, les propostes d'etiquetatge descriptiu aparegudes en els últims quatre anys per a la codificació de tesaurus documentals. A continuació, es mostra una arquitectura bàsica d'un servidor de tesaurus implementat en Java. I, finalment, es repassen els diversos protocols de comunicació i d'intercanvi de dades entre aplicacions que es poden usar per implementar aquest servei. El text s'acompanya de l'aplicació informàtica que s'ha desenvolupat.

1 Introducció

Els tesaurus documentals són un tipus de llenguatge combinatori que consta de llistes de termes, els quals representen un àmbit científic i tècnic determinat i mantenen una sèrie de relacions semàntiques entre si. Aquestes relacions semàntiques són de tres tipus concrets: equivalència, associació i jerarquia. Aquests tipus de llenguatge documental tenen una gran flexibilitat i capacitat d'especialització, cosa que els fa molt útils en entorns de recuperació de la informació (RI), com ara Internet. La definició més acceptada de *tesaurus* és la de 'llenguatge documental d'estructura combinatòria, de caràcter especialitzat, que es basa en expressions conceptuals, anomenades *descriptors*, proveïdes de relacions semàntiques'.

Hi ha diverses normes internacionals que estableixen les directrius per a la construcció de tesaurus, entre les quals destaquen l'ISO 2788:1986 i la posterior evolució Z39.19:1993. Segons aquestes normes, els tesaurus són realment instruments de control terminològic en entorns de RI i, encara que es poden trobar certes analogies amb altres recursos, com ara les ontologies, l'estructura dels tesaurus sol ser més molt més simple i menys definida, a més de tenir una diferenciació lexicosemàntica menor.

La utilització de tesaurus documentals en entorns de RI és una constant

des de fa molts anys. Malgrat això, els processos d'automatització de RI no sempre han inclòs aquestes eines de desambiguació i normalització semàntica dels termes utilitzats, com demostren el gran nombre d'aplicacions de RI que es basen únicament en càlculs de tipus estadístic i anàlisi de freqüències d'aparició de termes, com ara les diverses variants del *term frequency - inverse document frequency* (TF-IDF).

No obstant això, hi ha alguns projectes de RI en què s'ha treballat amb recursos lingüístics informatitzats similars als tesaurus per a tasques de desambiguació. Un exemple d'això és la profusió amb què s'ha utilitzat *Wordnet* durant la dècada de 1990 per al desenvolupament d'aquest tipus de tasques. L'ús de tesaurus documentals, encara que més restringit, també és present en aquest tipus de projectes. Per aquest motiu, considerem d'interès la migració d'aquest tipus de recursos als processos de RI basats en Internet.

L'RI a Internet és un procés que tendeix a dur-se a terme de manera distribuïda. Així doncs, és interessant que la integració de tesaurus documentals en aquest procés s'adapti a aquest tipus d'arquitectures. Aquesta adaptació pot concebre's de diferents maneres. En el cas que aquí ens ocupa, proposem la definició d'un servei d'informació que permeti la utilització distribuïda de tesaurus mitjançant una aplicació específica destinada a servir d'interfície entre el tesaurus i l'aplicació que necessiti fer-ne ús. La idea és permetre l'ús transparent del tesaurus a totes aquelles aplicacions de RI que necessitin utilitzar-lo, de la mateixa manera que treballen els serveis web. Per desenvolupar aquestes tasques hem d'automatitzar el tesaurus i implementar-ne les funcionalitats bàsiques d'accés i de consulta.

Per al desenvolupament d'aquest experiment s'ha utilitzat el tesaurus *Spines*, en la versió espanyola publicada pel CINDOC,[2] encara que el sistema informàtic desenvolupat s'ha pensat perquè pugui treballar amb qualsevol tesaurus que compleixi la norma ISO 2788.

2 Models d'etiquetatge per a tesaurus

En primer lloc, abans de pensar en la implementació del programari que desenvoluparà les funcions de servidor de tesaurus, hem d'estudiar les diverses maneres de codificació que hi ha a Internet. Destaquen les propostes dutes a terme en el marc del web semàntic per a la codificació de tesaurus i sistemes d'organització de coneixement.

Si partim de la base que tots els tesaurus tradicionalment utilitzats en centres de documentació i biblioteques contenen sempre algun tipus d'etiquetatge, és fàcil arribar a la conclusió que l'única cosa que hem de fer per a una primera adaptació d'aquests recursos a l'àmbit del web semàntic és codificar aquest etiquetatge de manera que s'adapti als estàndards establerts. D'aquesta manera, podem aprofitar una gran quantitat de feina ja feta i començar a utilitzar-lo a Internet immediatament.

L'aparició de l'XML i la seva àmplia acceptació ens dota de l'eina necessària per desenvolupar aquesta conversió de l'etiquetatge i solament ens cal decidir quin model és el més adequat per dur a terme aquesta tasca.

2.1 RDF/XML

Si bé l'aparició de l'XML (*extensible markup language*, llenguatge d'etiquetatge extensible) marca el començament d'una tendència, l'aparició de l'RDF (*resource description framework* o marc de descripció de recursos), sens dubte, suposa un punt d'inflexió quant a la creació d'una infraestructura semàntica que doni suport a la informació que hi ha a Internet.

L'RDF és un llenguatge per a la representació d'informació que ha de ser processada per màquines sense que això suposi una pèrdua de significat. L'RDF es basa en la idea que podem identificar els elements a partir dels URI (*uniform resource identification* o identificador uniforme de recursos), descrivint els recursos en termes de propietats simples o parells propietat-valor, cosa que permet representar les declaracions simples sobre recursos com un graf de nodes i d'arcs que representen els recursos, les propietats i els valors.[3]

En el graf RDF següent, s'observa com es pot descriure una persona, incloent-hi informació complementària.

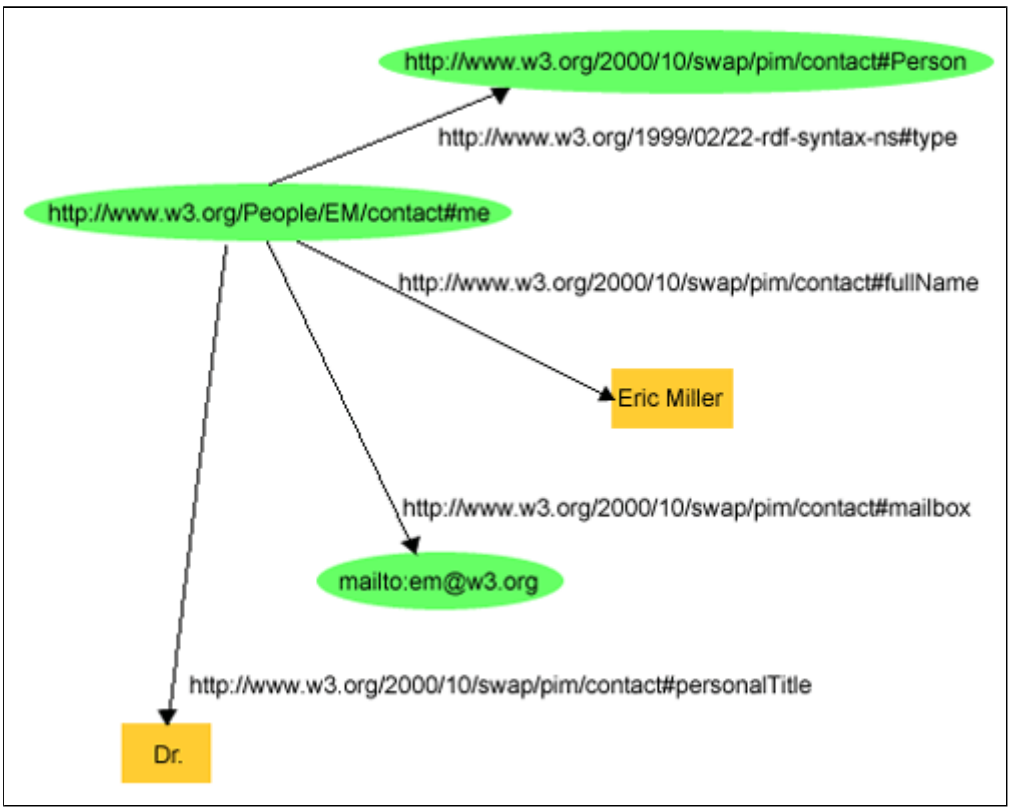

Figura 1. Graf RDF
(font: http://ww.w3.org/TR/rdf-primer/)

En la figura anterior descrivim un individu, Eric Miller, identificat amb l'URI: http://www.w3.org/People/EM/contact#em. Al seu torn, hi ha un altre URI que en defineix el tipus: http://www.w3.org/2000/10/swap/pim/contact#Person. Finalment, entre altres propietats, hi ha l'adreça electrònica, identificada per l'URI: http://www.w3.org/2000/10/swap/pim/contact#mailbox, el valor del qual és mailto:em@w3.org.

Aquest graf es pot representar mitjançant la sintaxi XML de la manera següent:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
    <contact:Person
    rdf:about="http://www.w3.org/People/EM/contact#em">
        <contact:fullName>Eric
        Miller</contact:fullName>
        <contact:mailbox
        rdf:resource="mailto:em@w3.org"/>
        <contact:personalTitle>Dr.</contact:personalTitle>
    </contact:Person>
</rdf:RDF>
```

2.1.1 RDF per a la codificació de tesaurus

La utilitat de l'RDF per a la representació de tesaurus no va trigar gaire temps a ser descoberta. L'any 2000 es proposa el primer esquema RDF per a la codificació de tesaurus. Posteriorment, sorgeixen altres propostes fins a l'arribada de l'SKOS-Core, de què parlarem àmpliament més endavant. En aquest apartat, però, ens referirem amb detall únicament a la proposta del projecte CERES, com a antecedent principal, ja que obre la via de desenvolupament d'aquest tipus d'especificacions.[4]

La proposta CERES va ser impulsada per la California Resources Agency,[5] i es defineix a partir de l'estàndard americà per a l'elaboració de tesaurus monolingües Z39.19:1993, el corresponent internacional del qual és la norma ISO 2788.[6] L'objectiu d'aquesta iniciativa és proporcionar un mètode per a l'intercanvi de tesaurus entre aplicacions, ja sigui en la seva totalitat o parcialment. La raó per la qual s'escull l'RDF/XML rau en el fet que aquest llenguatge permet especificar, de manera natural, les diverses relacions que hi ha entre els conceptes que formen el tesaurus. A això, s'hi ha d'afegir l'auge d'aquest tipus de tecnologies, que ja existia a Internet en el moment d'elaborar aquesta proposta.

La manera en què treballa aquesta proposta és molt senzilla, ja que es basa en la utilització bàsica d'XML perquè els documents puguin ser analitzats per qualsevol analitzador (*parser*). Segons aquesta especificació, els recursos són etiquetats com a *Descriptor*, *Category* o *EntryTerm*. Al seu torn, són subtipus de *Term*. Aquesta implementació

utilitza les formes de producció *typedNode* per a tots aquests recursos.

A la vegada, els *Term* poden tenir les *propertyTypes* següents:

- SN (*scope notes*): notes aclaridores per als descriptors.
- CN (*cataloger notes*): notes del catalogador per als descriptors.
- HN (*historical notes*): notes històriques per als termes.
- *Source*: indicació de la font d'un terme.
- *Status*: indicació de l'estatus d'un terme.

Totes les *propertyTypes* que segueixen a continuació són, al seu torn, recursos de *Label* i de *Term* (*Descriptor*, *EntryTerm* i *Category*), que ja hem vist abans:

- IC: descriptor en una categoria.
- CAT: categoria del descriptor.
- UF: terme d'entrada (*EntryTerm*) per al qual s'ha d'usar un descriptor preferent.
- TT (*topmost term*): cap de jerarquia per a un descriptor.
- BT (*broader term*): terme general per a un descriptor.
- RT(*related term*): terme relacionat per a un descriptor.
- NT (*narrower term*): terme específic per a un descriptor.
- USE: descriptor preferent pel qual s'ha de substituir un terme d'entrada (*EntryTerm*).

Mitjançant la utilització de les *propertyTypes* anteriors, obtenim una representació dels termes que componen el tesaurus codificada mitjançant la sintaxi següent:

```
<?xml version="1.0"?>
<?xml:namespace ns='http://www.w3.org/TR/WD-rdf-
syntax/' prefix='RDF' ?>
<?xml:namespace ns='http://www.w3.org/TR/WD-rdf-
schema/' prefix='RDFS' ?>
<?xml:namespace ns='http://ceres.ca.gov/thesaurus/'
prefix='Z19' ?>
<RDF:RDF>
<Category RDF:id="01">
    <Label>Natural Environment</Label>
    <IC>
        <Label>Biosphere</Label>
        <Descriptor RDF:resource="0101"/>
    </IC>
    <IC>
        <Label>Lithosphere</Label>
        <Descriptor RDF:resource="0102"/>
    </IC>
</Category>
<Descriptor RDF:id="0101">
    <Label>Biosphere</Label>
    <CAT>
        <Label>Natural Environment</Label>
        <Category RDF:resource="01"/>
    </CAT>
```

```
        <TT>
            <Label>Biosphere</Label>
            <Descriptor RDF:resource="0101"/>
        </TT>
        <NT>
            <Label>Ecosystems</Label>
            <Descriptor RDF:resource="010101"/>
        </NT>
    </Descriptor>
```

Com es pot veure en l'exemple, l'aplicació de l'RDF a l'etiquetatge de tesaurus és força senzill i funcional, cosa que va permetre l'adopció d'aquest model per diverses entitats d'importància.[7] No obstant això, a dia d'avui, cap d'aquestes entitats conserva aquest model, ja que, si bé en el seu moment proporcionava una bona solució, l'avenç de les tecnologies d'etiquetatge i la proliferació de les recomanacions del W3C més adaptades a la utilització sobre sistemes d'organització del coneixement, va provocar un abandó progressiu d'aquest model i l'evolució cap a d'altres més elaborats.

2.2 OWL

Sens dubte, l'aparició de l'OWL (*ontology web language* o llenguatge d'ontologia web) suposa un nou horitzó en l'etiquetatge de sistemes d'organització del coneixement. Té com a punt de partida les experiències prèvies dutes a terme amb DAML-OIL, en les quals es van inspirar, juntament amb la recerca en lògica descriptiva, els creadors de l'OWL per desenvolupar-lo. L'OWL és un llenguatge d'etiquetatge per a la publicació d'ontologies en el web. Té com a objectiu facilitar un model d'etiquetatge, construït sobre RDF i codificat en XML, que permeti representar ontologies a partir d'un vocabulari més ampli i d'una sintaxi més forta que la que permet l'RDF.[8] Per aquest motiu, l'OWL pot ser utilitzat per representar, de manera explícita, el significat de termes pertanyents a un vocabulari i per definir les relacions que hi ha entre si.[8]

L'OWL es divideix en tres subllenguatges: Lite, DL i Full. Cadascun proporciona un conjunt definit sobre el qual es pot treballar. Entre els tres, el més senzill és l'OWL Lite i el més complet, l'OWL Full. Per als objectius que aquí ens ocupen, l'OWL Lite és més que suficient, ja que, com s'indica en l'especificació del W3C, permet establir, entre d'altres, relacions jeràrquiques entre els conceptes que componen l'ontologia, alhora que aporta una menor complexitat formal que els seus germans grans. De fet, l'OWL Lite, en paraules de la recomanació del Consortium, proporciona una manera ràpida de migrar tesaurus i altres taxonomies en l'àmbit del web semàntic.[10]

Igual que en el cas de l'RDF, hi ha propostes concretes per a la utilització de l'OWL en la representació de tesaurus. Entre elles destaca la proposta de D. H. Fischer per al tesaurus del National Cancer Institute, a Alemanya, el qual es pot descarregar gratuïtament per Internet.[11] L'interès d'aquesta proposta resideix en el fet que l'autor va més enllà de la mera codificació del tesaurus mitjançant l'OWL Lite i es

planteja la utilització de l'OWL DL amb la intenció de convertir el tesaurus en una ontologia, no només formalment, sinó també conceptualment, cosa que el duu a replantejar-se la mateixa organització del vocabulari.

No és objectiu d'aquest treball aprofundir en aquest aspecte, però és sens dubte de gran interès reflexionar sobre els canvis que suposa, en la concepció dels llenguatges documentals clàssics, l'adopció d'aquests models de representació, la utilització dels quals, encara que en principi només afecti el caràcter formal del tesaurus, en modifica l'organització interna, com es pot veure en el treball de Fischer ja esmentat.

2.3 SKOS-Core

Obviant la reflexió iniciada en l'apartat anterior, però encara esquitxats per algunes de les conseqüències, s'arriba a la proposta més concreta per a la representació de tesaurus en l'entorn del web semàntic: l'SKOS-Core. L'SKOS-Core és un esquema RDF per a la representació de tesaurus i sistemes similars d'organització de coneixement. Això el situa en la mateixa línia que el projecte CERES, que hem comentat en la secció dedicada a l'RDF. Ara bé, aquesta aproximació al problema, a més de ser una proposta del W3C, proporciona mecanismes molt més elaborats que la mostrada anteriorment.

L'objectiu fonamental de l'SKOS-Core és proporcionar un model per a la migració de sistemes d'organització de coneixement a l'entorn del web semàntic. A més, serveix per construir esquemes de conceptes simples per utilitzar-los en la web. L'SKOS-Core està pensat com un complement de l'OWL, ja que proporciona un marc bàsic per a la construcció d'esquemes de conceptes, però sense la definició semàntica tan estricta que exigeix la utilització d'OWL. Es tracta, en certa mesura, d'una simplificació més gran que la de l'OWL Lite, cosa que permet que un nombre més gran de persones accedeixin a aquest tipus de tecnologies per a la representació del coneixement.

Veient el desenvolupament de l'SKOS-Core i la seva vinculació amb l'OWL, no podem evitar pensar en el desenvolupament i en la vinculació entre l'SGML i l'XML, ja que en ambdós casos es tracta d'una simplificació d'un model amb l'objectiu de fer-lo més atractiu a un públic més ampli. No és la nostra intenció afirmar que l'SKOS-Core ha de substituir l'OWL, ja que la representació d'ontologies requereix d'unes capacitats que aquest primer esquema és incapaç d'oferir. Ara bé, sí que és significatiu que sorgeixi una especificació del Consortium pensada específicament per a la migració directa de tesaurus ja existents amb l'objectiu que els utilitzin un nombre més gran d'usuaris.

Des del nostre punt de vista, es tracta de la mateixa fórmula que va fer popular en el seu moment l'HTML, o que va provocar la substitució en un gran nombre de casos de l'SGML per l'XML, una fórmula basada en la posada a disposició del gran públic d'un model senzill, d'utilització fàcil i d'aprenentatge ràpid, que permeti una expansió del web semàntic

generalitzada i no només en forma d'illes.[12] És prematur afirmar que l'SKOS-Core es convertirà en un esquema d'àmplia utilització, fins i tot per usuaris no especialitzats, però sí que ens atrevim a afirmar que aquest tipus d'iniciatives posen a l'abast dels professionals de la informació unes eines que s'adapten al seu nivell de coneixements tècnics, cosa que serà, sens dubte, una de les claus de l'èxit del web semàntic globalment.

Una vegada situat l'SKOS-Core en el seu context, podem passar a conèixer l'estructura i la composició de la resta de maremàgnum que forma el web semàntic. La idea clau d'aquest esquema RDF resideix en la seva capacitat de permetre la definició de conceptes i esquemes de conceptes. Un *concepte* es defineix com una unitat de pensament que pot ser definida o descrita. Al seu torn, un *esquema de conceptes* no és altra cosa que una col·lecció de conceptes. Un concepte pot tenir una sèrie d'etiquetes associades, on cada etiqueta és una paraula, frase o símbol que sol utilitzar-se per referir-se a aquest concepte.

Cada concepte, i aquí ja entrem en idees familiars per als documentalistes, només pot tenir una etiqueta preferent, anomenada *descriptor* o *terme preferent*, i un nombre il·limitat d'etiquetes alternatives, anomenades *no-descriptors* o *termes no preferents* en la terminologia habitual dels vocabularis controlats.

Les mateixes relacions que trobem en els tesaurus tradicionals (equivalència, jerarquia i associació) poden ser assignades entre conceptes pertanyents a un mateix esquema, el qual es pot assimilar a un tesaurus amb relacions semàntiques. A més, podem establir mapatges o equivalències entre conceptes que pertanyen a diferents esquemes, cosa que es pot entendre com una associació semàntica entre conceptes de tesaurus diferents.

El conjunt dels tipus de relacions bàsiques previstes per l'SKOS-Core poden veure's en la figura següent:

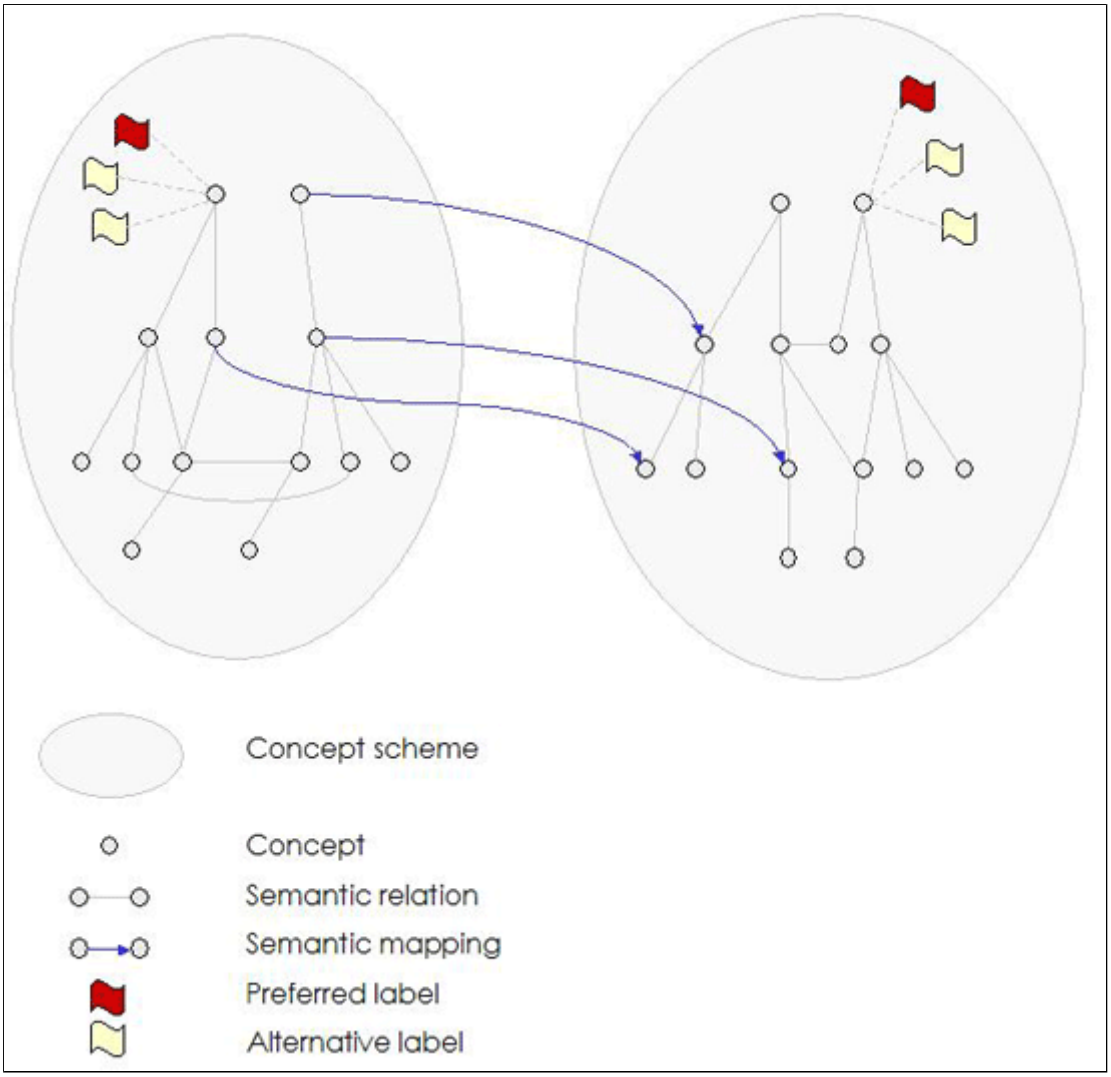

Figura 2. Representació gràfica de dos esquemes de conceptes i els seus components (font: http://ww.w3.org/2001/sw/Europe/reports/thes/1.0/guide/)

Com hem dit, l'esquema de conceptes es pot assimilar a un tesaurus. La creació d'un esquema de conceptes skos:ConceptScheme és bastant senzilla, ja que tan sols es necessita definir-lo de manera unívoca mitjançant un URI. Al seu torn, podem utilitzar metadades —per exemple les Dublin Core—, per descriure'l de manera global, com es pot veure en el codi següent:[13]

```
<rdf:RDF>
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
    <skos:ConceptScheme
    rdf:about="http:/spines.org/thesaurus">
        <dc:title>SPINES</dc:title>
        <dc:description>Tesauro de política
```

```
            científica</dc:description>
            <dc:creator>UNESCO</dc:creator>
        </skos:ConceptScheme>
    </rdf:RDF>
```

Per definir cada concepte skos:Concept és necessari assignar-li també un identificador unívoc, que sol ser un URI, de manera que es defineixen de la manera següent:

```
<rdf:RDF>
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
    rdf:about="http:/spines.org/concept/0001">
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
    </skos:Concept>
</rdf:RDF>
```

Com es pot observar, el concepte és assignat al tesaurus creat en l'exemple anterior per mitjà de la propietat skos:inScheme. L'especificació d'aquesta propietat és opcional, ja que podem crear un concepte sense necessitat que pertanyi a un esquema de conceptes o tesaurus concret. De la mateixa manera, podem assignar un concepte a diversos tesaurus repetint aquesta propietat les vegades que siguin necessàries, cosa que ens permet estalviar espai en especificacions multitesaurus.

Una vegada identificat un concepte mitjançant un URI, és necessari que li assignem les etiquetes corresponents als termes amb els quals aquest concepte està relacionat, ja que aquesta és la manera de relacionar les paraules o els termes amb els conceptes. Aquesta separació tan clara entre informació lèxica i informació semàntica és típica de l'elaboració d'ontologies. Així doncs, abandonem la línia definida durant anys per *Wordnet* i diferenciem, de manera clara, ambdós tipus d'informació. Aquesta és, per tant, una de les mostres a les quals fèiem esment quan comentàvem que la concepció tradicional dels tesaurus es veu modificada, en major o menor mesura, per l'arribada de noves perspectives que, en aquest cas, vénen de la mà del web semàntic.

La codificació de les etiquetes corresponents als termes preferents i no preferents que pertanyen a un concepte, es duu a terme mitjançant les propietats skos:prefLabel, per als termes preferents o descriptors, i skos:altLabel, per als termes no preferents o no-descriptors. Aquesta segona etiqueta dóna cobertura a la relació d'equivalència o de sinonímia entre termes, tan usual en la totalitat dels tesaurus utilitzats avui en dia. El codi té l'aspecte següent:

```
<rdf:RDF>
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
```

```
    rdf:about="http:/spines.org/concept/0001">
        <skos:externalID>A.01.0001</skos:externalID>
        <skos:prefLabel>Capital</skos:prefLabel>
        <skos:altLabel>Activo</skos:altLabel>
        <skos:altLabel>Riqueza</skos:altLabel>
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
        </skos:Concept>
</rdf:RDF>
```

Com es pot veure en l'exemple anterior, hi ha la possibilitat d'utilitzar, de manera addicional, identificadors per als conceptes que no siguin URI i d'aprofitar aquells tesaurus que els utilitzin. A aquest efecte, s'utilitza la propietat skos:externalID, la qual fita l'identificador que s'estigui utilitzant.

La utilització de notes aclaridores és una altra de les necessitats que es troben cobertes en l'SKOS-Core. De fet, podem anar més enllà i assignar definicions als conceptes, i alhora exemples contextuals i fins i tot imatges que estiguin relacionades amb el concepte o que hi facin referència. Les propietats corresponents són skos:scopeNote per a les notes aclaridores, skos:definition per a les definicions de conceptes, skos:example per als exemples contextuals i foaf:depiction per a les imatges. La raó per la qual el tractament de les imatges és diferent rau en el fet que les imatges són, al seu torn, un recurs en si mateixes i, per tant, han de tenir un URI propi que les identifiqui de manera unívoca.

En el cas de les notes aclaridores, el codi corresponent tindria l'aspecte següent:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
    rdf:about="http:/spines.org/concept/0011">
        <skos:prefLabel>Ergonomía</skos:prefLabel>
        <skos:scopeNote>Adaptación del trabajo a los
        requisitos fisiológicos y psicológicos del ser
        humano</skos:scopeNote>
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
    </skos:Concept>
</rdf:RDF>
```

Per al cas de les imatges, el codi es representaria de la manera següent:

```
<rdf:RDF>
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <foaf:Image
    rdf:about="http:/spines.org/img/red.jpg"/>
    <skos:Concept
    rdf:about="http:/spines.org/concept/0001">
```

```
            <skos:prefLabel>Eritrocitos</skos:prefLabel>
            <skos:altLabel>Glóbulos rojos</skos:altLabel>
            <skos:altLabel>Hematíes</skos:altLabel>
            <skos:inScheme
            rdf:resource="http:/spines.org/thesaurus"/>
            <foaf:depiction
            rdf:resource="http:/spines.org/img/red.jpg"/>
        </skos:Concept>
    </rdf:RDF>
```

Com es pot veure en l'exemple, s'utilitza *foaf*[14] com a propietat per referir-se a una imatge. De la mateixa manera, es poden utilitzar imatges per referir-se als conceptes a partir de símbols gràfics.

2.3.1 Definició de relacions semàntiques

Una vegada s'han definit els elements que poden formar part d'un concepte és important centrar-se en les capacitats d'expressió de les relacions semàntiques que ens ofereix l'SKOS-Core. Aquestes relacions cobreixen de sobres les necessitats de la major part de tesaurus utilitzats avui en dia en centres de documentació. Així doncs, s'estableixen relacions jeràrquiques i associatives, totes agrupades entorn d'una família de propietats destinades a representar relacions simples entre conceptes. El nivell més alt d'aquesta família és la propietat skos:semanticRelation. Després, hi ha les relacions jeràrquiques definides per les propietats skos:narrower i skos:broader, per a termes específics i generals respectivament, així com la propietat skos:related, per a termes relacionats. Recordem que la relació d'equivalència ja s'ha definit prèviament per mitjà de la propietat skos:altLabel.

Com ja hem dit, skos:broader i skos:narrower ens permeten crear jerarquies de conceptes. A més, és interessant destacar que l'SKOS-Core permet assignar diversos skos:broader a un mateix concepte, i així preveu la representació de la polijerarquia. El codi següent mostra, de manera més completa, la utilització d'aquestes propietats jeràrquiques:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
    rdf:about="http://spines.org/concept/0001">
        <skos:prefLabel>Eritrocitos</skos:prefLabel>
        <skos:altLabel>Glóbulos rojos</skos:altLabel>
        <skos:altLabel>Hematíes</skos:altLabel>
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
        <skos:broader
        rdf:resource="http:/spines.org/concept/0002"/>
    </skos:Concept>
    <skos:Concept rdf:about="http://spines.org/concept/0002">
        <skos:prefLabel>Sangre</skos:prefLabel>
        <skos:altLabel>Plasma</skos:altLabel>
        <skos:altLabel>Suero sanguíneo</skos:altLabel>
```

```
            <skos:inScheme
            rdf:resource="http:/spines.org/thesaurus"/>
            <skos:narrower
            rdf:resource="http:/spines.org/concept/0001"/>
        </skos:Concept>
    </rdf:RDF>
```

Com ja hem esmentat abans, la relació associativa entre termes relacionats també es troba representada mitjançant la propietat skos:related, que es codifica d'aquesta manera:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
    rdf:about="http://spines.org/concept/0001">
        <skos:prefLabel>Eritrocitos</skos:prefLabel>
        <skos:altLabel>Glóbulos rojos</skos:altLabel>
        <skos:altLabel>Hematíes</skos:altLabel>
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
        <skos:related
        rdf:resource="http:/spines.org/concept/0002"/>
    </skos:Concept>
    <skos:Concept rdf:about="http://spines.org/concept/0002">
        <skos:prefLabel>Sangre</skos:prefLabel>
        <skos:altLabel>Plasma</skos:altLabel>
        <skos:altLabel>Suero sanguíneo</skos:altLabel>
        <skos:inScheme
        rdf:resource="http:/spines.org/thesaurus"/>
        <skos:related
        rdf:resource="http:/spines.org/concept/0001"/>
    </skos:Concept>
</rdf:RDF>
```

L'SKOS-Core també proporciona mecanismes per a la representació de termes en altres idiomes a través d'un etiquetatge multilingüe. Aquest etiquetatge es basa en la utilització de les propietats skos:prefLabel i skos:altLabel, a les quals s'afegeix l'atribut d'idioma RDF. Per a cada idioma es pot fixar un terme preferent, mitjançant skos:prefLabel, i tots els termes no preferents que siguin necessaris, mitjançant skos:altLabel. La codificació tindria l'aspecte següent:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
    syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept
    rdf:about="http:/spines.org/concept/0001">
        <skos:prefLabel xml:lang="en">English
        cuisine</skos:prefLabel>
        <skos:altLabel xml:lang="en">English
        dishes</skos:altLabel>
        <skos:altLabel xml:lang="en">English
```

```
            food</skos:altLabel>
            <skos:prefLabel xml:lang="fr">Cuisine
            anglaise</skos:prefLabel>
            <skos:altLabel xml:lang="fr">Plats
            anglais</skos:altLabel>
            <skos:prefLabel xml:lang="és">Cocina
            inglesa</skos:altLabel>
            <skos:prefLabel xml:lang="it">Cucina
            inglese</skos:prefLabel>
            <skos:inScheme
            rdf:resource="http:/spines.org/thesaurus"/>
        </skos:Concept>
    </rdf:RDF>
```

2.3.2 Ampliació de les relacions semàntiques

Fins aquí hem vist com l'SKOS-Core proporciona un esquema per a l'etiquetatge de tesaurus prou extens i flexible com per codificar la major part de tesaurus amb els quals es treballa avui en dia. Ara bé, les possibilitats d'aquest esquema RDF no acaben aquí, ja que té propietats que permeten definir amb més precisió les relacions semàntiques entre conceptes. Aquestes propietats estenen l'esquema més enllà del que normalment s'hagués considerat necessari per a un tesaurus tradicional, i es construeixen a partir d'especificacions més generals de les relacions que ja hem vist. D'aquesta manera, com a tipus de propietats concretes per a les relacions jeràrquiques, tenim les propietats següents:

- skos:broaderGeneric i skos:narrowerGeneric
- skos:broaderInstantive i skos:narrowerInstantive
- skos:broaderPartitive i skos:narrowerPartitive

El subconjunt Generic s'ha d'utilitzar únicament per especificar relacions de subsumpció entre dos conceptes. De fet, aquesta propietat hereta la semàntica de la propietat RDF rdfs:subClassOf. D'altra banda, el subconjunt Instantive expressa que un concepte és una mostra d'un altre. En aquest cas, la propietat heretada de l'RDF és rdf:type. Finalment, el subconjunt Partitive expressa la idea que un concepte forma part d'uns altres, cosa que no és més que una manera de concretar encara més la idea de terme específic.

El tipus de relació d'associació expressada pels termes relacionats també disposa d'una sèrie de refinaments definits per les propietats següents: skos:relatedHasPart i skos:relatedPartOf.

Les dues propietats serveixen per establir relacions associatives parcials. Si les observem bé, ens adonarem que la semblança que tenen amb els esquemes jeràrquics skos:broaderPartitive i skos:narrowerPartitive les fa, fins i tot, equivalents. Els autors de l'especificació reconeixen aquesta equivalència, que ja hem pogut veure en altres casos, i la justifiquen amb vista a afavorir la interoperabilitat entre diversos esquemes de conceptes, cosa que en principi sembla bastant raonable.

Per finalitzar aquest repàs de la proposta actual del Consortium en

matèria de tesaurus, cal dir que l'SKOS-Core també admet la definició de caps de jerarquia mitjançant la propietat skos:TopConcept, la qual es pot usar també per a la codificació de tesaurus facetats.

2.4 Altres models d'etiquetatge per a la representació de tesaurus a Internet

Malgrat que l'objectiu d'aquest text és situar l'automatització de tesaurus en l'àmbit del web semàntic, no podem deixar d'esmentar, encara que sigui de manera molt succinta, altres esforços d'automatització de tesaurus i sistemes d'organització del coneixement rellevants a Internet.

2.4.1 Zthes

El llenguatge Zthes descriu un model abstracte per a la representació i recerca de tesaurus seguint la norma ISO 2788 abans esmentada. La idea fonamental d'aquesta proposta és un model que permeti la implementació de tesaurus perquè s'hi pugui accedir mitjançant el protocol Z39.50 i SRW. Aquesta proposta, a causa de l'auge dels llenguatges d'etiquetatge, no pot deixar d'oferir una DTD per a la representació del tesaurus, si bé es tracta fonamentalment d'una iniciativa centrada en el protocol Z39.50 amb els condicionaments que això suposa.[15]

2.4.2 Topic Maps

En segon lloc, hi ha la proposta de Topic Maps, un estàndard per a la navegació conceptual que possibilita la representació de tesaurus. Si bé es tracta d'una iniciativa amb certa tradició i desenvolupada inicialment sobre SGML i HyTime, últimament s'ha renovat mitjançant la creació d'una DTD per a la representació pròpia.[16]

2.4.3 SKOS-Core versus Zthes i Topic Maps

Tant Topic Maps com Zthes vénen de línies de treball més antigues i s'han vist en la necessitat d'adaptar-se al ritme i a les modes imposades per l'aparició de l'XML. Això no vol dir que aquests projectes siguin millors ni pitjors, sinó simplement que tenen un camp d'aplicació més reduït i esdevenen solucions concretes a problemes concrets, de manera que estan lluny de convertir-se en un estàndard de facto, com sol ocórrer amb les tecnologies proposades pel Consortium. Precisament per això en aquest treball hem optat per un desenvolupament basat en RDF, OWL i SKOS-Core.

3 Arquitectura de l'agent de gestió de tesaurus

Una vegada escollit l'SKOS-Core com el format per a la codificació de

tesaurus, podem passar a descriure el disseny i la implementació del nucli del servidor de tesaurus. Aquest nucli del sistema se serveix de dues classes[17] destinades a gestionar les funcionalitats bàsiques d'accés al tesaurus, *Thesaurus* i *Concept*, i d'unes altres dues destinades a proporcionar funcionalitats bàsiques de normalització conceptual, *NormalizarTermino* i *Searcher*.

La classe *Thesaurus* abstreu la idea del tesaurus, amb l'objectiu de modelitzar-lo i permetre'n, d'aquesta manera, el maneig mitjançant exemples per a totes aquelles classes que necessitin desenvolupar-hi accions. La classe *Thesaurus* conté un objecte de tipus *TreeMap*, com a atribut de classe, destinat a l'emmagatzematge del tesaurus, en què el parell clau/valor ve definit pels descriptors i no-descriptors, que serveixen com a punt d'entrada al tesaurus, i pels objectes *Concept*, que modelitzen la idea de concepte, respectivament. Cada clau descriptor o no-descriptor conté com a valor l'objecte *Concept* a què pertany.

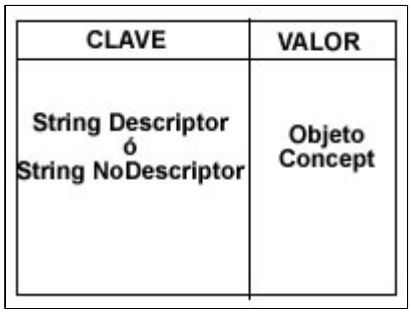

Figura 3. Esquema de l'estructura de dades utilitzada

Com ja hem esmentat, la classe *Thesaurus* utilitza objectes de tipus *Concept* per a l'abstracció del concepte de descriptor. Aquest objecte *Concept* inclou atributs relacionats amb els components que formen part d'un descriptor en els tesaurus tradicionals. D'aquesta manera, entre els atributs o les variables de classe, trobarem un *String* [18] relatiu al terme preferent que representa el descriptor —i, per tant, el concepte— i dos *Strings* relatius a la traducció del terme preferent a l'anglès i al francès. Al seu torn, hi haurà també una sèrie de llistes, implementades mitjançant *ArrayList*,[19] referents a termes generals de primer, segon i tercer nivells; termes específics de primer, segon i tercer nivells; categories a les quals pertany el concepte; termes relacionats, i termes equivalents o no-descriptors. Finalment, també hi ha un *String* corresponent a notes aclaridores sobre el significat i la utilització dels conceptes en el moment de la indexació.

A través d'aquesta estructura, definim un concepte de manera molt similar a com s'ha fet en molts dels recursos lingüístics informatitzats que s'han utilitzat en els últims deu anys en el camp de la recuperació de la informació i el processament del llenguatge natural, és a dir, basant-nos en les relacions semàntiques entre conceptes. Això no obstant, ens ajustem a la concepció pura de tesaurus documental definida mitjançant l'estàndard ISO 2788.

En la figura següent es mostren les classes que formen part d'aquest nucli:

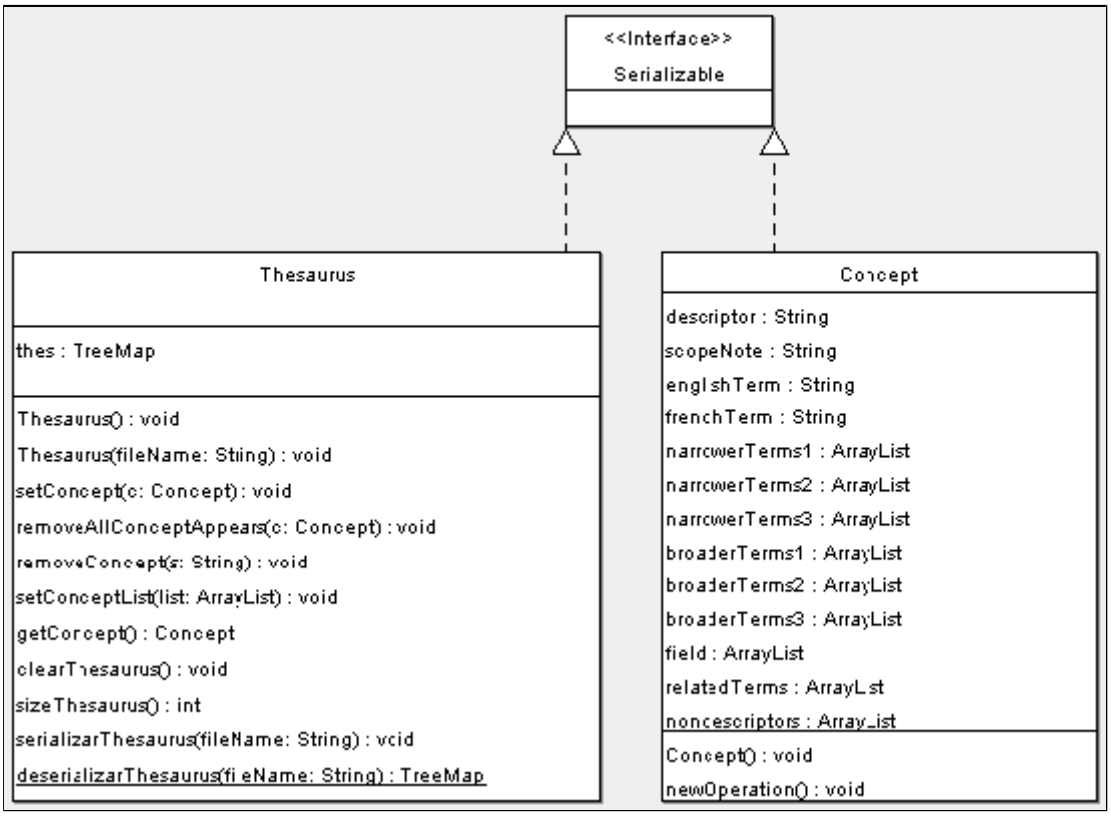

Figura 4. Esquema UML (unified modelling language o llenguatge unificat de modelat)
de les classes que componen el nucli de l'agent servidor de tesaurus

Com es pot veure, les dues classes implementen la interfície *Serializable*. Això es deu al fet que l'emmagatzematge del tesaurus i dels descriptors es duu a terme en un mateix fitxer, on es desen els objectes seriats perquè romanguin de manera persistent més enllà de la utilització del programa que els ha creat. Quan s'inicia l'execució de l'aplicació que utilitza aquestes classes, es carreguen en la memòria primària per ser usats des d'aquí durant l'execució. Això també permet que un mateix agent carregui diversos tesaurus mitjançant mecanismes de seriació, cosa que facilita el mapatge de conceptes entre diversos tesaurus.

Aquesta manera d'emmagatzematge difereix de la utilitzada normalment sobre bases de dades, i té com a objectiu mantenir la consistència dels objectes més enllà de l'execució del programa a fi de no haver de generar, des d'una base de dades, els objectes *Thesaurus* i *Concept* cada vegada que s'usa el programa.[20] Aquesta implementació no perjudica en excés l'eficiència del programa i permet mantenir l'enfocament orientat a objectes en tot moment, amb vista tenir més simplicitat en el disseny i en la implementació del sistema.

A partir d'aquest nucli, hem dissenyat a tall d'exemple un conjunt de

classes encarregades de permetre operacions avançades sobre el tesaurus partint dels mètodes bàsics que conté la classe *Thesaurus*. Així doncs, tenim la classe *NormalizarTermino* i la classe *Searcher*. La classe *NormalizarTermino* permet normalitzar conceptes de manera automàtica en el tesaurus. La funcionalitat és molt senzilla, ja que a partir d'una entrada composta per una paraula o cadena de paraules,[21] aquesta classe és capaç de retornar els termes que normalitzen els conceptes representats pel terme o pels termes introduïts. Aquest procés de normalització es duu a terme sobre la totalitat del tesaurus, utilitzant els termes preferents i no preferents com a punts d'entrada. D'aquesta manera, garantim la coherència del procés de normalització en virtut del tesaurus que estem utilitzant. Un exemple d'utilització d'aquesta classe seria el següent:

```
"jose@leviathan:/eclipse/workspace/ThesaurusManager$ java
NormalizarTermino spines cancer"
```

La consulta és "cancer". El descriptor corresponent és "neoplasmas malignos". El nombre de termes és de 10.772. Encara que en l'exemple fem una normalització simple, per un terme, l'aplicació és capaç de processar cadenes senceres de termes i retornar els descriptors corresponents per separat. De la mateixa manera, permet recuperar els termes relacionats semànticament amb cada descriptor. Fins i tot es pot passar un text complet per dur a terme la normalització sobre els conceptes que hi apareixen.

D'altra banda, la classe *Searcher* funciona de la mateixa manera que la classe anterior, si bé la diferència principal rau en el fet que *Searcher* retorna el resultat de la consulta en format RDF i SKOS-Core. La resposta d'aquesta classe a la consulta anterior té l'aspecte següent:

```
jose@leviathan:/eclipse/workspace/ThesaurusAgent$ java
thes.Searcher cáncer
<rdf:RDF>
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <skos:Concept rdf:about="http://spines/neoplasmas%
    20malignos">
        <skos:broader
        rdf:resource="http://spines/enfermedades"/>
        <skos:related
        rdf:resource="http://spines/transformación%
        20neoplásica%20celular"/>
        <skos:prefLabel>neoplasmas malignos</skos:prefLabel>
        <skos:prefLabel xml:lang="en">malignant
        neoplasms</skos:prefLabel>
        <skos:prefLabel xml:lang="fr">neoplasmes
        malins</skos:prefLabel>
        <skos:related rdf:resource="http://spines/neoplasmas%
        20benignos"/>
        <skos:related rdf:resource="http://spines/hábito%20de%
        20fumar"/>
        <skos:related rdf:resource="http://spines/enfermedades%
```

```
                20incurables"/>
                <skos:altLabel>cáncer</skos:altLabel>
                <skos:altLabel>carcinoma</skos:altLabel>
                <skos:related rdf:resource="http://spines/i+d%
                20médica"/>
                <skos:related rdf:resource="http://spines/neoplasmas%
                20experimentales"/>
                <skos:related rdf:resource="http://spines/pechos"/>
                <skos:narrower rdf:resource="http://spines/neoplasmas%
                20inducidos%20por%20radiación"/>
                <skos:related
                rdf:resource="http://spines/antineoplásicos"/>
                <skos:related rdf:resource="http://spines/enfermedades%
                20de%20la%20mama"/>
                <skos:related rdf:resource="http://spines/enfermedades%
                20gastrointestinales"/>
                <skos:broader rdf:resource="http://spines/neoplasmas"/>
                <skos:related rdf:resource="http://spines/condiciones%
                20precancerosas"/>
                <skos:related rdf:resource="http://spines/enfermedades%
                20ginecológicas"/>
                <skos:related rdf:resource="http://spines/cancerígenos%
                20ambientales"/>
                <skos:related rdf:resource="http://spines/enfermedades%
                20del%20aparato%20genital"/>
                <skos:related rdf:resource="http://spines/amianto"/>
                <skos:narrower rdf:resource="http://spines/leucemias"/>
                <skos:narrower rdf:resource="http://spines/sarcoma"/>
            </skos:Concept>
        </rdf:RDF>
```

Com es pot veure en aquest cas, la consulta retorna un document RDF amb l'etiquetatge propi de l'esquema SKOS-Core, amb tota la informació referent al concepte sol·licitat.[22]

L'objectiu de les dues classes és mostrar alguns exemples de les funcionalitats que ofereix l'estructura *Thesaurus* i *Concept* anteriorment descrita, tant pel que fa a la consulta per part d'usuaris humans —és el cas de normalitzar termes—, com pel que fa a l'ús per part de sistemes d'indexació automàtica —és el cas de *Searcher*.

A partir d'aquestes descripcions de classes podem definir un nucli bàsic del servidor de tesaurus que implementa tant el tesaurus com les operacions que s'hi desenvolupen, i també els conceptes. D'aquesta manera, el disseny de l'aplicació es correspondria amb la figura següent:

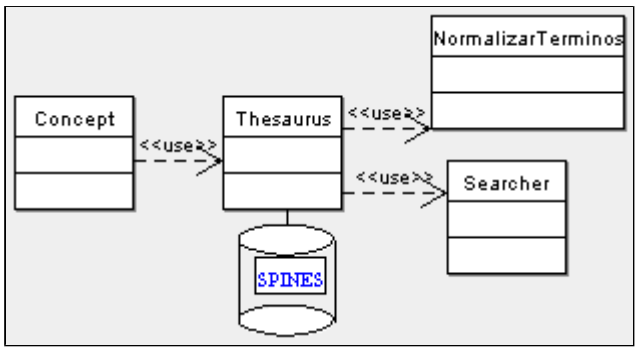

Figura 5. Disseny UML (unified modelling language o llenguatge unificat de modelat) del sistema presentat en aquest article per a la normalització conceptual basada en tesaurus

3.1 Funcionalitats internes del servidor de gestió de tesaurus

Com hem esmentat anteriorment, el nostre concepte de servidor de tesaurus va més enllà de permetre la mera consulta del llenguatge, i passa pel desenvolupament de tasques de manteniment intern i d'actualització del tesaurus. La necessitat de l'actualització de tesaurus és un problema que es tracta des de fa temps en la literatura de l'àrea de documentació. Aquesta afirmació es pot comprovar en el *Manual de lenguajes documentales*, de la doctora Blanca Gil Urdiciain, en què es dedica una secció del capítol referent a tesaurus precisament a aquest tema.[23] En aquest text, la doctora Gil destaca el fet que l'extensibilitat que facilita l'estructura dels tesaurus documentals permet ampliar i modificar el vocabulari en funció de les necessitats que sorgeixin al llarg d'usar-lo continuadament. Les necessitats de modificació es fan més paleses sobretot en els tesaurus de nova creació, ja que es requereix un temps d'adaptació del llenguatge documental a l'entorn en el qual s'utilitza.

A més, l'actualització del tesaurus s'ha de fer tant per incorporar la terminologia derivada del desenvolupament de la ciència o matèria a la qual es dedica com per cobrir llacunes o errades detectades durant la seva utilització, així com per adaptar-lo a les necessitats de recuperació manifestades pels usuaris a través de les seves recerques. Aquest segon procés de correcció i d'adaptació pot ser automatitzat a partir de la detecció, per part del servidor de tesaurus, d'aquestes errades o llacunes. Un dels exemples més clars de l'automatització d'aquest procés s'executa a partir de l'estudi de les consultes fetes pels usuaris en recuperar documents del sistema.

Aquestes consultes poden ser processades de manera estadística per establir la capacitat de recuperació dels descriptors utilitzats en el tesaurus. A partir de la mesura de la capacitat de recuperació de cada descriptor i no-descriptor continguts en el tesaurus, podem establir un índex de rellevància que mantingui sempre com a descriptor el terme amb més capacitat de recuperació. Podríem dir que l'agent de gestió de tesaurus pot ser capaç d'aprendre, a partir de la pràctica dels usuaris, quins termes són els més utilitzats per a la recuperació de determinats

documents, de manera que assigni sempre a aquests documents els termes amb més capacitat de recuperació. A aquesta tasca d'aprenentatge, podem afegir-hi distintes tècniques automàtiques d'avaluació, de manteniment i de generació de tesaurus que en permetin mantenir la consistència en tot moment.

4 La comunicació amb altres aplicacions

L'automatització del tesaurus, tal com l'hem descrita en la secció anterior, possibilita l'ús de tesaurus documentals tant en entorns d'indexació manual com en sistemes d'indexació automàtica i, en general, per a qualsevol sistema de RI. Ara bé, la utilització distribuïda del tesaurus requereix adoptar estàndards que permetin la comunicació, de manera que el tesaurus pugui ser consultat per altres aplicacions i usat en sistemes de recuperació d'informació distribuïts.

En la secció següent, descriurem succintament algunes de les fórmules d'accés remot que pot proporcionar un servidor de tesaurus. Farem un breu repàs sobre les opcions que ha d'implementar un servidor perquè sigui flexible i perquè tingui una capacitat de comunicació àmplia. Des de la perspectiva dels agents de programari, considerarem el FIPA-ACL com un llenguatge de comunicació entre agents, mentre que, amb vista als serveis web, descriurem breument el protocol SOAP.

4.1 Comunicació mitjançant FIPA-RDF

El paradigma d'agents proporciona una teoria sòlida sobre la qual es basa la comunicació entre aplicacions informàtiques. Per aquesta raó, la primera consideració que farem del nostre sistema de gestió de tesaurus serà la funcionalitat que té com a agent. La idea fonamental d'aquest enfocament rau en el fet de concebre aquest sistema com un agent que ofereix com a servei la normalització conceptual de termes extrets de textos per altres agents destinats a executar altres tasques del procés de RI —per exemple, la indexació de paraules clau o l'extensió de consultes en cercadors.

Si hi incloem, per exemple, un agent d'indexació automàtica, la visió del flux de treball es clarifica, ja que tindrem, d'una banda, un agent d'indexació de documents HTML i, de l'altra, un agent de gestió de tesaurus com el descrit en la secció anterior. Quan l'agent d'indexació extreu els termes literals que apareixen en un document ha de normalitzar-los abans de dur a terme les operacions de comptabilització de freqüències i d'assignació de rellevància, ja que es necessita una normalització semàntica abans de dur a terme les operacions quantitatives que permetran que l'agent d'indexació acabi la seva tasca correctament. D'aquesta manera, s'estableix una comunicació entre tots dos agents amb vista a resoldre el problema de la normalització conceptual dels termes d'indexació.

L'estàndard FIPA divideix la comunicació entre agents en actes de

comunicació, protocols d'interacció i llenguatges de contingut. Els *actes de comunicació* es componen de blocs constituents del diàleg entre agents, on es defineix el significat dels missatges independentment del context. Els *protocols d'interacció* defineixen una seqüència de missatges que representen un diàleg complet entre dos agents. Finalment, els *llenguatges de contingut* estableixen el llenguatge per al contingut del missatge.

El nostre servidor de tesaurus, que aquí tractem com un agent de gestió de tesaurus, es pot adaptar a aquesta manera de comunicació. Per això, pot comunicar a un altre agent que ho sol·liciti els resultats de la normalització de termes.

La sol·licitud de normalització d'un terme es basa més concretament en la utilització del protocol de comunicació FIPA-Request, que s'usa quan un agent demana a un altre que desenvolupi una acció. En el nostre cas, l'agent de gestió de tesaurus que actua com a destinatari pot acceptar o rebutjar la petició i, en cas d'acceptar-la, haurà de desenvolupar la normalització i indicar-li a l'altre agent quan finalitzi. En el diagrama següent (vegeu la figura 6), s'observa el flux dels missatges. Els que estan en blanc són els que envia l'"Initiator" (Agent1), que es correspon amb l'agent d'indexació; mentre que els de "Respondre" (Agent2), que es corresponen amb l'agent de gestió de tesaurus, estan en gris.

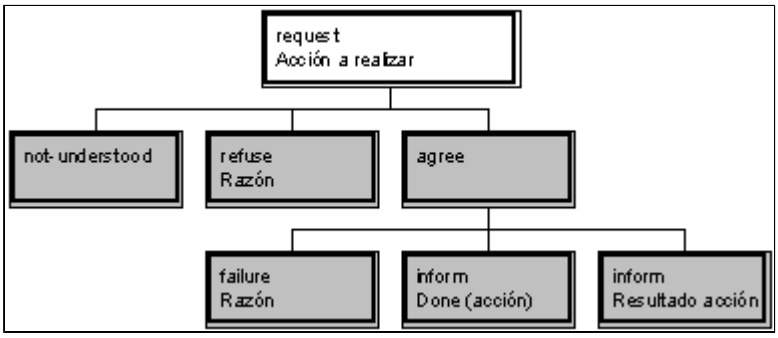

Figura 6. Esquema de comunicació FIPA-Request
(font: http://grasia.fdi.ucm.es/SP/index.html)

Gràcies a aquesta seqüència de missatges, el diàleg coordinat entre els agents és possible i, a més, es permet la utilització dels serveis de l'agent de gestió de tesaurus a tots aquells agents que ho necessitin. En un sistema multiagent, el nombre d'agents de gestió de tesaurus pot ser tan alt com el nombre de tesaurus que hi hagi en el sistema.

D'altra banda, l'atribut *Content* inclòs en la resposta de l'agent de gestió de tesaurus pot contenir documents RDF com els que hem mostrat abans,[24] de manera que es manté el contingut del missatge i s'adapta el servidor de tesaurus a la manera de treballar que tindria com a agent de gestió de tesaurus.

4.2 Serveis web de normalització conceptual amb SOAP

L'enfocament d'agents no és l'únic que ens permet afrontar el problema de la comunicació entre aplicacions de RI. En els últims anys, han proliferat els serveis web lligats al desenvolupament dels llenguatges d'etiquetatge. Aquests serveis web, basats en XML, permeten que les aplicacions comparteixin informació i que, a més, invoquin funcions d'altres aplicacions independentment de com s'hagin creat, quin sigui el sistema operatiu o la plataforma sobre la qual s'executin i quins siguin els dispositius utilitzats per obtenir-hi accés. Encara que els serveis web XML són independents entre si, poden vincular-se i formar un grup de col·laboració per desenvolupar una tasca determinada.

Els serveis web no pretenen eliminar del mapa les biblioteques o els mòduls de programació, ja que no en són una versió millorada, sinó una eina amb diverses aplicacions en determinats casos. Així, per exemple, si necessitem una rutina que descodifiqui un fitxer de vídeo, no és aconsellable utilitzar un servei web, ja que fer-ho suposaria enviar el fitxer de vídeo al servidor del servei web, perquè el descodifiqui i l'enviï en format pla, sense compressió de cap tipus. Això suposaria un consum d'amplada de banda tan gran que fa que processar el còdec de vídeo en local sigui molt més eficient que processar-lo remotament.

No obstant això, hi ha altres ocasions en què és interessant utilitzar un servei web, en comptes d'una rutina d'una biblioteca. Per exemple, si volem que una aplicació sàpiga el preu d'un determinat llibre donat l'ISBN[25] corresponent, podem utilitzar un servei web amb vista a implementar una aplicació que utilitzi i processi aquestes dades per oferir a l'usuari un servei de valor afegit sobre la informació presa inicialment. En el cas de la recuperació d'informació, destaca el servei web basat en SOAP que ofereix *Google* per a la utilització de les funcionalitats del seu cercador en diverses aplicacions.[26]

En el nostre cas, utilitzarem els serveis web per permetre que diverses aplicacions utilitzin les funcions de normalització i consulta del tesaurus sense necessitat d'implementar aplicacions *ad hoc*. La idea és que es pugui disposar d'aquestes funcionalitats sobre el tesaurus a través d'Internet com si es tractés d'una caixa negra.[27]

Per a la implementació d'aquest servei es pot utilitzar SOAP (*simple object access protocol* o protocol d'accés a objectes simples), ja que és un protocol proposat pel W3C i basat en XML per a la comunicació d'informació estructurada entre aplicacions, cosa que s'adapta molt bé a l'ús que li volem donar. A més, SOAP permet, igual que FIPA-RDF, embeure el codi RDF generat pel nostre servidor de tesaurus en el cos del missatge,[28] cosa que ens proporciona una altra manera de comunicació sense necessitat de modificar el format de la resposta del nostre servidor de tesaurus.

5 Conclusions

Com es pot veure per la multiplicitat d'opcions comentades, és possible proporcionar una gran capacitat de comunicació a l'aplicació

encarregada de la gestió del tesaurus. L'objectiu de les diverses interfícies de consulta mostrades aquí és precisament aquest, ja que la utilització del tesaurus com a base de coneixement és un procés prou estandarditzat com perquè sigui d'utilitat, per als desenvolupadors, tenir aplicacions que permetin aquesta funcionalitat de manera transparent.

L'omnipresència de l'RI a Internet i l'alt nombre d'aplicacions que treballen per facilitar aquest servei, fa que sigui d'interès la concepció d'una arquitectura distribuïda per al procés de RI, de manera que no hi hagi necessitat de repetir una vegada i una altra la mateixa tasca. Aquesta feina no és més que un exemple senzill de com es poden portar a terme serveis web de RI d'utilitat per als desenvolupadors d'aquest tipus de sistemes i que facilitin als documentalistes sistemes de normalització conceptual basats en tesaurus.

6 Referències bibliogràfiques

Alistair, Milers; Rogers, Nikki; Beckett, Dave (2004). *SKOS-Core 1.0 guide: an RDF schema for thesauri and related knowledge organisation systems*. SWAD-Europe Thesaurus Activity, W3C. <http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/> [Consulta: 15/9/2004].

Alistair, Milers; Rogers, Nikki; Beckett, Dave (2004). *SKOS-Core 1.0 guidelines for migration: guidelines and case studies for generating RDF encodings of existing thesauri*. SWAD-Europe Thesaurus Activity, W3C . <http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/> [Consulta: 15/9/2004].

Gil Urdiciain, Blanca (1999). *Manual de lenguajes documentales*. Madrid: Noesis.

McBride, Brian (2002). *An introduction to RDF and the Jena RDF API*. <http://jena.sourceforge.net/tutorial/RDF_API/> [Consulta: 15/9/2004].

Méndez Rodríguez, Eva María (2002). *Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales*. Gijón: Trea.

Ogbuji, Uche (2002). *Using RDF with SOAP: beyond remote procedure calls*. <http://www-106.ibm.com/developerworks/webservices/library/ws-soaprdf/> [Consulta: 15/9/2004].

Rogers, Nikki; Beckett, Dave (2004). *SWAD-Europe: use cases for a thesaurus service (draft document)*. <http://www.w3.org/2001/sw/Europe/200311/thes/Use_cases_Thes_Service.html> [Consulta: 15/9/2004].

W3C (2004). *RDF primer: W3C recommendation 10 February 2004*. Brian McBride (series editor). <http://www.w3.org/TR/rdf-primer/>

[Consulta: 15/9/2004].

Annex. Aplicació informàtica

Data de recepció: 28/07/2004. Data d'acceptació: 3/10/2004.

Notes

[1] Una primera versió d'aquest text es va presentar al taller "Introducción al uso de la web semántica" organitzat per SWAD-Europe a Madrid el 13 de juny de 2004 (http://www.w3.org/2001/sw/Europe/events/200406-esp/.

[2] Pot consultar-se a: http://pci204.cindoc.csic.es/tesauros/SpinTes/Spines.htm.

[3] Pot consultar-se a: http://www.w3.org/TR/rdf-syntax-grammar/.

[4] A la pàgina http://www.w3c.rl.ac.uk/SWAD/deliverables/8.2.html#4.1, hi ha un repàs excel·lent de l'evolució de les diverses propostes d'etiquetatge per a tesaurus des de l'any 2000.

[5] Pot consultar-se a: http://ceres.ca.gov/thesaurus/RDF.html.

[6] Pot consultar-se a: http://www.niso.org/standards/standarddetail.cfm?stdid=518.

[7] Pot consultar-se a: http://ceres.ca.gov/thesaurus/.

[8] Pot consultar-se a: http://www.w3schools.com/rdf/rdf_owl.asp.

[9] Recordem que el significat dels termes i les seves relacions és el que denominem *ontologia*.

[10] Pot consultar-se a: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[11] Pot consultar-se a: http://www.mindswap.org/2003/CancerOntology/.

[12] Eva M. Méndez Rodríguez, *Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales* (Gijón: Trea, 2002).

[13] En tots els exemples s'utilitza, de manera abreujada, el marc de descripció de recursos, RDF, ja que és una denominació més comprensible i llegible.

[14] FOAF és un vocabulari destinat a la descripció de tot tipus de recursos web, en aquest cas concret es tracta d'imatges.

[15] Per obtenir-ne més informació, es pot visitar l'adreça següent: http://zthes.z3950.org/.

[16] Pot consultar-se a: http://www.topicmaps.org/.

[17] La paraula *classe* s'ha heretat de la metodologia orientada a objectes que s'ha utilitzat per a la programació de l'agent de gestió de tesaurus. L'aplicació ha estat desenvolupada en Java, de manera que és interessant que el lector repassi alguns conceptes d'aquest llenguatge per a la comprensió total del text que segueix a continuació.

[18] Un *String* és un tipus de dada que es refereix a una cadena de caràcters.

[19] Un *ArrayList* és una de les implementacions que ofereix Java per a les llistes dinàmiques.

[20] També hi ha l'opció d'utilitzar Hibernate (http://www.hibernate.org/) per dur a terme aquestes tasques, però no s'ha usat causa del fet que el sistema descrit està donant bons resultats amb tesaurus de grandària mitjana.

[21] Fins i tot admetria com a entrada un text complet, ja que desenvolupa un tractament de cadenes de paraules prou complex com per tractar cadenes que continguin diversos conceptes al mateix temps.

[22] La generació de RDF es fa mitjançant l'ús de Jena 2.1.

[23] Blanca Gil Urdiciain, *Manual de lenguajes documentales* (Madrid: Noesis, 1996), p. 215–220.

[24] L'estàndard FIPA-RDF especifica com dur a terme aquesta tasca.

[25] És un servei web que ofereix Barnes and Noble.

[26] Pot consultar-se a: http://www.google.com/apis/.

[27] Els exemples de serveis web i la introducció han estat extrets del web http://web-services.bankhacker.com/.

[28] Uche Ogbuji, *Using RDF with SOAP: beyond remote procedure calls*, <http://www-106.ibm.com/developerworks/webservices/library/ws-soaprdf/> [Consulta: 15/9/2004].