BY NC ND

Xarxa Punt TIC

MÒDUL 1 NIVELL AVANÇAT
Recerca i recuperació de la informació a Internet (avançat)
Internet invisible

plan avanza2 · Gobierno de España · Plan E · Ministerio de Industria, Turismo y Comercio · Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información · Generalitat de Catalunya

➔ ÍNDEX

➔ E. Internet invisible

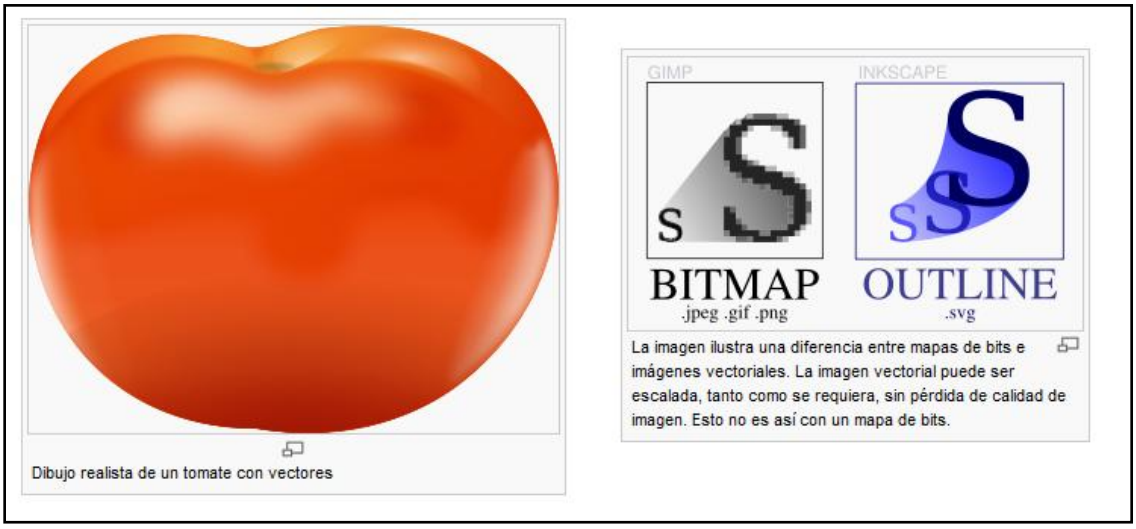
Definició i reptes

Lluís Codina és professor titular de Ciències de la Documentació a la Universitat Pompeu Fabra i membre de l'Observatori de la Comunicació Científica. És sense dubte uns dels millors experts en Anàlisi i Mètodes en Ciències de la Documentació. Podeu trobar molta documentació a la seva web http://www.lluiscodina.com/. Lluis Codina afirma que:

"Internet invisible és un nom clarament inadequat per referir-se al sector de llocs i de pàgines web que no poden indexar els motors de cerca d'ús públic com Google o Altavista. Malgrat el nom, afortunadament, el Web invisible és perfectament visible ja que els continguts d'aquestes pàgines i llocs web es poden veure o bé mitjançant un navegador convencional o bé mitjançant un navegador complementat amb algun programa addicional (*plugin*). Per aquest motiu, s'hauria de denominar, en realitat, el web "no indexable", un terme molt més adequat, però clarament allunyat de la capacitat suggeridora del terme "invisible". Atès que, tot i així, és el terme més habitual fins i tot en la bibliografia tècnica, usarem en aquest treball el terme Web o Internet invisible per referir-nos a la informació publicada en servidors web que per diversos motius no pot ser indexada i, per tant, no pot ser trobada pels motors de cerca convencionals."

Vegem ara per què hi ha continguts no indexables al Web. Hi ha almenys tres motius. En un ordre no significatiu, podem dir que el primer motiu són els formats dels documents. Els motors de cerca van ser creats originalment per descarregar, llegir i indexar pàgines HTML. Qualsevol altre format era il·legible, és a dir, invisible per a aquests motors. Tots coneixem la proliferació de formats no HTML en el Web (que no obstant això s'integren amb tota facilitat en el navegador). És el cas, per exemple, dels cada vegada més abundants documents en format .pdf (documents Acrobat) i fins i tot en format .doc (documents Word). En la mesura que una part dels continguts del Web està formada per documents no HTML, aquesta part és candidata a ser Internet invisible.

Figura 1. Part d'un document –un dibuix d'un tomàquet– en format no HTML (SVG) vist en un navegador.



El format de gràfics vectorials SVG, Scalable Vectorial Graphics, és un format estàndard per fer dibuixos, plànols, esquemes, etc, per pàgines web, en aquest cas el dibuix d'un tomàquet molt realista. SVG està basat en XML i per tant és indexable pels cercadors. El format Flash (animacions) ja és indexat també per Google. Més informació a http://es.wikipedia.org/wiki/.svg.

El segon motiu són les pàgines que es generen de forma dinàmica; típicament, a través de la consulta a una base de dades. Per exemple, si utilitzem All Movie (www.allmovie.com) per buscar informació sobre un film obtindrem una URL com aquesta:

http://www.allmovie.com/search/work/star+trek/results

Els motors de recerca no poden indexar continguts que es generen d'aquesta manera. Abans de llançar la cerca, el contingut existeix en el format binari (i propietari) d'alguna base de dades. Solament després de la consulta, i com a resultat d'executar una instrucció com la que mostra la figura anterior, es crearà una pàgina en format HTML. El lector pot fer la prova, si copia la URL de la figura anterior (que conté una consulta a una base de dades) i la introdueix com a adreça en un navegador, obtindrà una pàgina HTML que li informarà sobre un film determinat. Abans, però, aquesta pàgina no existia. En el cas de bases de dades com l'anterior, els motors de cerca poden proporcionar accés a la pàgina d'inici (*home page*).

És a dir, podem accedir a les pàgines principals dels llocs web que proporcionen accés a bases de dades, perquè aquestes pàgines principals són pàgines HTML convencionals, però no podem accedir a la resta del lloc a través del motor de cerca; i la resta del lloc pot ser (en ocasions) una enorme base de dades. Per exemple, si llancem la consulta 2001 a Google, en cap dels resultats obtenim la fitxa del film corresponent d'All Movie. De fet, obtindrem

una diversitat de resultats que reflecteix que el terme 2001, fora de context, té molts significats i no necessàriament el de títol principal d'un film de Kubrick.

Finalment, forma part del Web invisible el conjunt de llocs o de pàgines web que, de forma expressa, s'exclouen de l'activitat indicadora dels motors de cerca. Alguns servidors exclouen als motors de recerca de totes o de part de les seves carpetes i directoris mitjançant l'ús d'un protocol d'exclusió que, en general, respecten els programes rastrejadors (*spiders* o *crawlers*) d'aquests motors de recerca. Aquest protocol consisteix en un petit nombre de valors que pot adquirir l'atribut *content* com a part d'una etiqueta meta i l'altre atribut, *name*, obté el valor "robots". Aquestes indicacions es guarden en un simple arxiu de text de nom robots.txt que se situa en el servidor de pàgina web i que se suposa que llegeixen i respecten els rastrejadors (robots). La figura següent mostra l'ús d'aquest protocol per a indicar als robots dels motors que no indexin la pàgina en qüestió ni segueixin cap dels enllaços que pugui contenir.

```
<meta name="ROBOTS" content="noindex,nofollow">
```

A més del protocol que acabem de veure, hi ha altres raons per les quals els motors no poden entrar en un lloc. En general, qualsevol lloc web que demani l'ús de contrasenyes o *passwords* quedarà fora de la capacitat indexadora dels motors. Aquests llocs poden ser Extranets o serveis que demanin, no només una subscripció prèvia, sinó que el pagament d'una quantitat en concepte d'abonament, etc. Els motors també tenen dificultats per interpretar els llocs que usen marcs (*frames*), encara que són dificultats d'un altre tipus i no les considerarem aquí.

La qüestió és que, en total, alguns analistes assenyalen que el Web invisible pot ser fins a 500 vegades més gran que el Web visible (Bergman, 2001). Des del punt de vista de l'accés al coneixement i de la classe de cerca i obtenció de la informació que ens interessa aquí, no hi ha cap problema amb que una part del Web invisible segueixi sent invisible.

Per exemple, no és cap tragèdia per al desenvolupament de la ciència o del coneixement humà que l'Extranet o la Intranet d'una corporació sigui invisible als motors de recerca. No només no és un problema, sinó que és desitjable que segueixi sent així. Ningú vol que els motors de cerca puguin indexar documents administratius particulars o informacions confidencials.

Per tant, de les tres raons per les quals tenim una Internet invisible, una d'elles no és cap problema, però les altres dues sí. Recordem: documents amb format no HTML i pàgines generades dinàmicament (típicament a través de bases de dades).

Amb la impossibilitat d'indexar documents no HTML tenim, efectivament, un autèntic problema. Molts informes i estudis que contenen informació valuosa estan publicats i disponibles al Web de forma pública i oberta; tanmateix, si no

són indexades de forma adequada, són inaccessibles a gairebé tot el món i a tots els efectes pràctics.

D'altra banda, no deixa de ser un problema que, tot i disposar d'un client universal d'accés a la informació –el navegador web–, no existeixi, en canvi, alguna cosa similar a una interfície universal d'accés a la informació des del moment que, per a cadascuna de les vàries desenes de milers de bases de dades existents a Internet sigui necessari: primer, un accés diferenciat, i segon, un sistema de consulta (en part) diferent.

En aquest últim cas, cal observar que les barreres al coneixement són dues: el coneixement de les fonts i el domini de la interfície d'usuari de cada font. En efecte, en primer lloc, perquè un usuari pugui beneficiar-se dels continguts d'una base de dades és necessari, almenys, que sàpiga que existeix. Però, suposant que sàpiga que existeix, llavors haurà de tenir habilitats d'ús d'aquesta base de dades, i cada base de dades, no només presenta una interfície d'usuari diferent, sinó un conjunt de funcions diferents.

Accedir als continguts d'Internet invisible

Formats no html

Malgrat tot això, es pot accedir cada vegada més a parts més grans del Web invisible. Examinem primer el cas dels formats de documents. Afortunadament, en aquest aspecte, les fronteres del Web invisible no fan més que retrocedir.

Google té capacitat per a localitzar una gran varietat de documents en diferents formats (pdf, excel, word, access, flash, rtf, postscript, i molts més). L'últim format incorporat més destacable és el dels arxius swf confeccionats en Flash.

En aquest sentit, sembla que la tendència és clara: a poc a poc, la major part dels formats de documents significatius en el món científic i cultural seran indexats pels motors de cerca i, per tant, aquesta zona del Web invisible deixarà de ser-ho aviat. A més, hi ha dos factors més que conflueixen en aquest aspecte: d'una banda, els navegadors cada vegada incorporen amb major facilitat documents no HTML. És exemplar, en aquest sentit, la integració de les últimes versions dels navegadors i el format pdf. D'altra banda, el progressiu ample de banda disponible per als usuaris fa que aquesta integració sigui transparent.

D'aquesta manera, si els motors tendeixen al que podríem anomenar una "indexació universal" i els navegadors (o agents d'usuari) tendeixen a poder mostrar qualsevol tipus de document, podem concloure que aquest aspecte del Web invisible està destinat a ser marginal.

Ara bé, de vegades les solucions als problemes aporten també problemes nous. A mesura que formats com pdf i Word s'integren al Web amb major naturalitat, per a benefici dels usuaris, descendeix el grau de connectivitat general del Web. És a dir, una de les virtuts del Web és la facilitat amb la qual es poden publicar pàgines web (o llocs sencers) interconnectats de forma interna, així com la facilitat per connectar pàgines i llocs web remots. Però part d'aquestes facilitats desapareixen amb formats com pdf i Word. És cert que un

document pdf, per exemple, pot contenir enllaços interns o externs, però en la pràctica, es publiquen documents pdf com una forma fàcil d'obtenir una publicació de qualitat tipogràfica amb un mínim esforç. En la pràctica, per tant, la immensa majoria de documents pdf estan interconnectats de manera molt pobre.

Bases de dades

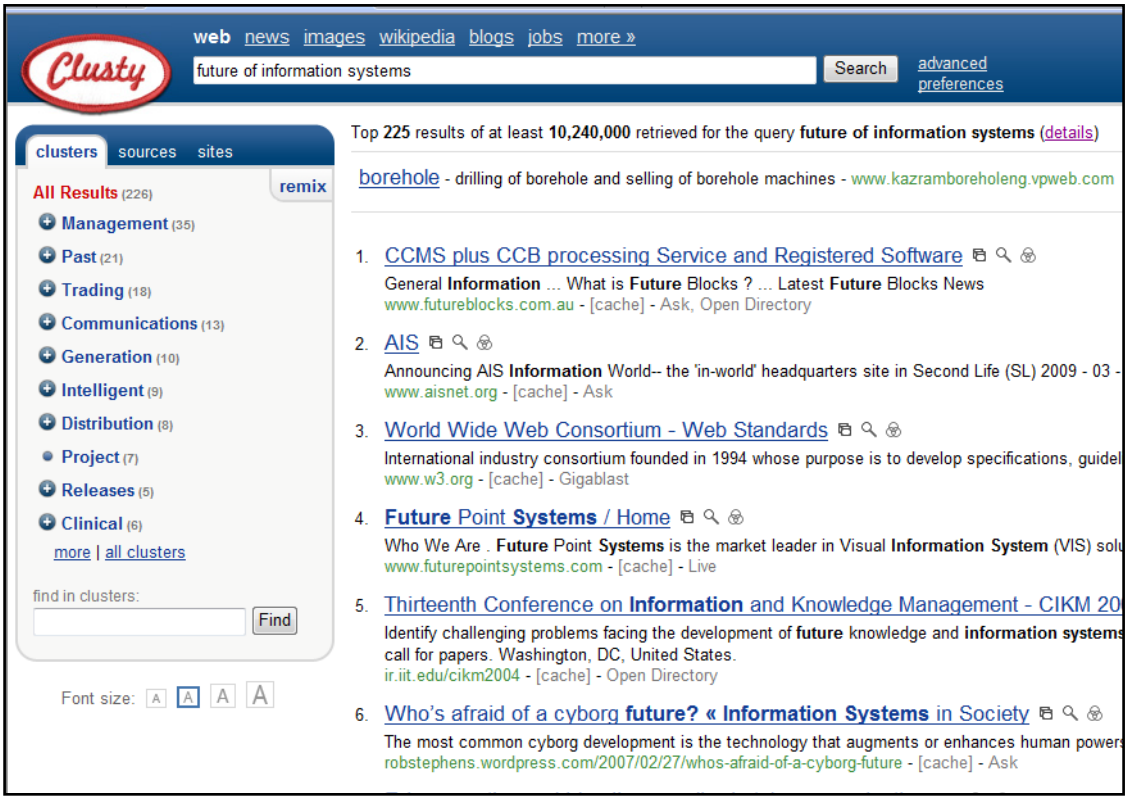
També tenim indicis de solució al segon gran "problema" del Web invisible: l'accés al contingut de les bases de dades, però des de motors convencionals.

La solució aquí prové d'aquest enfocament: si bé és difícil o impossible indexar per part dels motors de cerca el contingut de bases de dades alienes, no hauria d'haver molta dificultat a generar interfícies de consulta unificades que enviessin una mateixa consulta a diferents bases de dades des de, per exemple, una mateixa pàgina web. El model en aquest cas són els multicercadors, també (mal) anomentas metcercadors.

Un multicercador és un sistema que accepta com a entrada la pregunta d'un usuari i retorna en una resposta unificada les respostes de diversos motors de cerca.

Un bon exemple de multicercador és http://clusty.com. Una cerca en Clusty pels tremes *future of information systems* mostra com a resultat una compilació de la informació oferida per diversos buscadors.

Figura 8: El resultat d'una cerca en Clusty:



Compilar informació, en el cas de Clusty, significa que no es limita a bolcar els resultats que envia cada cercador, sinó que: (a) unifica resultats (o sigui, elimina duplicats); i (b) distribueix els resultats per grups o pseudocategories que el sistema d'agrupació (*clustering*) de que és capaç de generar de manera automàtica.

Però el que ens interessa aquí examinar és la següent idea: Clusty no intenta explotar directament els índexs dels diferents motors de cerca. En el seu lloc, fa una cosa més viable: envia la pregunta a diversos motors i processa els resultats abans d'oferir-los a l'usuari. Aquesta operació li permet oferir un resultat unificat les fonts del qual, però, tenen procedències molt diverses.

Multicercadors de segona generació

Un altre exemple molt interessant i bona mostra del que, probablement, ens espera en els propers anys és el motor de cerca Scirus (www.scirus.com). És aviat encara per saber si Scirus serà un experiment efímer, com tants altres projectes esperançadors en el Web (esperem que aquesta vegada no) o solament un avançament d'una nova generació de sistemes de cerca en línia que trenqui d'una vegada per sempre les barreres del Web invisible.

Scirus és un projecte d'una important editorial científica, Elsevier, que ha produït un motor que és capaç d'enviar les preguntes dels usuaris a les bases de dades que indica la taula de la Figura 10.

Figura 10. Bases de dades que pot interrogar Scirus de forma simultània:

- Medline
- Sciencedirect
- Uspto
- Beilstein Abstracts
- E-Print Arxiv
- Nasa Technical Reports
- Cogprints
- Biomed Central
- Mathematics Preprint Server
- Chemistry Preprint Server
- Computer Science Preprint Server

A més, Scirus indexa gairebé 90 milions de pàgines web, és a dir, documents en format HTML publicats en servidors de pàgines web convencionals, però sempre vinculats amb institucions acadèmiques o científiques. D'aquesta manera, l'usuari de Scirus, típicament un investigador o un professional, quan realitza una cerca en aquest motor, obté dos tipus de resultats: (1) pàgines o llocs web relacionats amb la ciència, la universitat, etc.; (2) articles de revista o registres referenciales procedents de bases de dades de ciència i tecnologia (és a dir, una part del Web invisible).

Scirus, per tant, és un dels millors exemples que tenim ara al nostre abast del que poden ser els futurs sistemes d'informació en línia: una interfície unificada d'informació a fonts diverses.

Podem concloure, en relació a aquest apartat, que les barreres d'Internet invisible probablement cediran, una a una, fins que els continguts no indexables d'Internet siguin exactament els que han de ser: porcions del web que els seus administradors o propietaris, en ús legítim de les seves prerrogatives, no desitgen que siguin indexats.

En canvi, els continguts de la Internet invisible corresponents a formats no HTML i part del contingut que es troba en el format binari de diferents bases de dades, seran accessibles des de motors de cerca públics, del tipus Google o Scirus.

El que significa això últim és que els productors de bases de dades haurien de començar a plantejar-se si desitgen, per dir-ho d'alguna manera, sindicar els seus continguts als motors de cerca. Un model pot ser el que representa Scirus. Els productors de bases de dades poden decidir que entra en els seus interessos permetre la recepció de consultes i l'enviament consegüent de resultats a un o més motors de cerca, conscients que els usuaris finals sempre persegueixen, d'una forma o altra, la idea (en part utòpica) de la interfície de consulta universal. Naturalment, sindicació de continguts implica també un model de negoci. Implica que els motors de cerca com Google estiguin disposats a retribuir als productors de les bases de dades, o bé que, a partir d'un moment donat, una part dels resultats oferts pel sistema sigui d'accés lliure i una altra sigui d'accés condicionat al pagament d'una certa quantitat o a la condició de ser abonat o subscriptor.

Això és el que fa Scirus. Quan un usuari llança una cerca a Scirus pot trobar tres tipus de resultats: (1) documents d'accés totalment lliure, per exemple, un estudi publicat com una pàgina web en un servidor web convencional i d'accés lliure; (2) documents als quals té accés perquè la seva institució posseeix una subscripció a la publicació corresponent, per exemple un article d'una revista subscrita per la biblioteca de la seva institució; i (3) documents als quals té accés mitjançant pagament amb targeta de crèdit.

El Web semàntic

Definicions

Vegem primer la definició oficial de Web semàntic (*semantic web*), segons el W3 Consortium (l'organisme promotor de la idea):

"The Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming."

Dues coses sobre la definició anterior: en primer lloc, com es pot observar, no diu absolutament res. Què significa que alguna cosa és "la representació de dades en la *World Wide Web*"? Res. La resta de la suposada definició és pitjor. Abandona clarament l'intent de dir el que és el Web semàntic (veient l'antecedent, tal vegada sigui el millor) i es limita a assenyalar, entre d'altres coses summament informatives, "que integra una varietat d'aplicacions"(!).

La segona cosa que cal assenyalar és que el Web semàntic no existeix. No sabem si el Web semàntic serà realitat algun dia, però ara per ara, ni existeix ni se l'espera (almenys de manera imminent). Tot i així, s'ha de reconèixer en aquest concepte una autèntica idea-força, en el sentit que és una idea que ja ha estat capaç de mobilitzar moltes energies (i moltes il·lusions) i que, sens dubte, no deixarà de presentar resultats durant els propers anys, perquè segurament seguirà mobilitzant energies.

És una idea, per dir-ho d'alguna manera, semblant als viatges que tenen sentit per si mateixos, independentment de la destinació prevista. Diuen els experts en narrativa que tota autèntica aventura és en realitat un viatge en el qual, al final, el protagonista ha patit alguna transformació (se suposa que per a bé). El Web semàntic pot veure's com un viatge que inicia ara la *World Wide Web* i tal vegada no arribi mai (del tot) a la seva destinació, però que, mentre, la transformarà profundament.

Si haguéssim de proposar una definició de Web semàntic, nosaltres començaríem amb aquesta:

Definició: El Web Semàntic és un conjunt d'iniciatives, tecnològiques en la seva major part, destinades a crear una futura *World Wide Web* en la qual els ordinadors puguin processar la informació, és a dir, representar-la, trobar-la, gestionar-la, com si els ordinadors tinguessin intel·ligència

A continuació, intentarem presentar una aproximació a la idea del Web semàntic; per fer-ho, ens hem basat en un treball previ (Codina, 2003), però, sobretot, en la informació que sobre el Web semàntic pot trobar-se en el ja esmentat organisme promotor de la idea, el W3 Consortium (www.w3.org/2001/sw/), i en un famós i citadíssim article publicat a *Scientific American* (Berners-Llig, 2001). Hem consultat també altres autors que s'indiquen a la bibliografia.

Estat actual

Si el Web semàntic no existeix, què és en aquests moments? De moment, és el nom d'una aspiració; el nom d'un objectiu molt ambiciós que, de complir-se, canviaria de forma radical el Web tal com el coneixem avui. En què consisteix aquesta aspiració? Ni més ni menys es tracta d'aconseguir que les pàgines que formen el Web deixin de ser simples cadenes de caràcters per als ordinadors i

es converteixin en textos amb sentit, és a dir, text proveït de semàntica, tal com, de fet, ho és per als éssers humans.

Perquè un objectiu com aquest? Tal com es codifiquen les pàgines web actuals, principalment mitjançant el llenguatge HMTL, tenen molt poc sentit per a les màquines. En efecte, si veiem el codi font d'una pàgina web actual, trobem, per exemple, un tros de codi com aquest:

…Superar la bretxa digital…

Quan l'ordinador l'interpreti, a través del programa navegador, apareixerà com un text en negreta i cursiva com aquest:

…***Superar la brecha digital***…

Amb això gairebé s'acaba tot el que és capaç de fer un ordinador amb les pàgines HTML. Com saben els informàtics i documentalistes, una altra cosa que poden fer els ordinadors és construir índexs amb les paraules que apareixen en les pàgines web. Després, quan algú envia una pregunta a un motor de cerca, el que fa aquest últim és comparar les paraules de la pregunta amb les paraules del seu índex. Per exemple, suposem que el responsable d'un programa de govern sobre el problema de la bretxa digital decideix indagar a Internet per veure si troba estudis o informes sobre la bretxa digital.

Suposem que accedeix a Google i entra la següent pregunta: "bretxa digital". El que farà Google és comparar les paraules de la seva pregunta amb les paraules del seu índex. Si troba un document que tingui la "bretxa digital", el retornarà com a resposta. Això és gairebé tot el que poden fer els ordinadors que tingui a veure amb processament d'informació en pàgines web.

Amb aquestes limitacions, la cerca a Internet, com tothom sap, està plena de frustracions. Si algú cerca per "cavalls", no trobarà res que tracti sobre "egües". Si algú cerca sobre com evitar la guerra, no trobarà un document sobre com aconseguir la pau, etc. El Web semàntic vol solucionar això. Això sona a intel·ligència artificial. Per tant, encara que no vulguin anomenar-ho així, amb el Web semàntic s'està buscant el mateix objectiu, és a dir, que els ordinadors entenguin que un document sobre "egües" pot ser molt rellevant per a una necessitat d'informació sobre "cavalls", i que la semàntica de la pregunta "és possible evitar la guerra?" és la mateixa que la de la pregunta "és possible aconseguir la pau?".

A més, s'espera que els ordinadors puguin desenvolupar tasques de gestió que demanin interpretar informació i prendre decisions adaptant-les al context. Es tracta, ni més ni menys, d'un objectiu que la informàtica ha denominat fins ara intel·ligència artificial.

Infraestructura

Els mitjans amb els quals se suposa que s'aconseguirà el Web semàntic són els següents: primer, un nou llenguatge de codificació de pàgines, un nou llenguatge de marcatge. Aquest llenguatge, com és sabut, es denomina XML. Amb XML es poden dissenyar llenguatges de marcatge molt estructurats i molt explícits en els quals, en lloc d'etiquetes com <b> i <i>, hi haurà etiquetes com <títol>, <subtítol>, <capítol>, <subcapítol>, <autor>, <institució>, <ciutat>, etc.

Com que faran falta etiquetes específiques per a cada tipus d'informació –per exemple, les pàgines web de les companyies aèries necessitaran etiquetes com <vol>, <hora de sortida>, <destinació>, etc.– s'ha creat una especificació, una espècie de metallenguatge, XML, que permet definir llenguatges específics, és a dir, conjunts d'etiquetes específics per a cada necessitat d'informació. Per exemple, els editors de diaris disposen ja del seu propi conjunt d'etiquetes, així com els matemàtics per a expressar equacions, etc.

El segon element amb el qual es compta són les metadades. Com saben molt bé els documentalistes, les metadades són informació sobre la informació i són, en realitat, una antiga fórmula. Els catàlegs de les biblioteques són metadades. La venerable norma ISBD és una norma sobre metadades, els descriptors assignats a un document són metadades, els tesaurus i les classificacions són el que ara en l'argot de les metadades es denominen també *schemes*, etc.

La qüestió és que les pàgines web ja tenen metadades. Almenys, solen tenir la metadada títol, en forma d'etiqueta <*title*> en una zona de les pàgines web invisible per a les persones, però visible per als ordinadors. A més, algunes pàgines, molt poques, solen tenir altres metadades, com <*keyword*>, <*description*>, etc.

Com ja se sap, existeix una ambiciosa norma d'abast internacional, *Dublin Core*, que proporciona una llista unificada i normalitzada de fins a quinze metadades del tenor dels ja comentats, perquè els editors i autors que ho desitgin els incloguin en les seves pàgines web. La idea és simple: si les pàgines web tinguessin metadades del tipus <títol>, <autor>, <tema>, <lloc de publicació>, etc., els usuaris podríem fer preguntes molt més precises als motors de cerca. Podríem, per exemple, fer peticions d'informació d'aquest tipus: "busca'm documents publicats a tal lloc i que tractin d'aquest i aquest tema, sota aquest punt de vista".

Però les metadades actuals no tenen ni semàntica ni sintaxi ni estan unificades sota una norma comuna que agrupi la diversitat de plataformes de metadades existents.

Per dotar-les d'aquestes tres coses, s'han desenvolupat altres normes. La més important és la RDF (*Resource Description Framework*). Aquesta norma especifica una gramàtica lògica perquè els autors de pàgines web puguin descriure les propietats semàntiques dels documents en una notació estàndard i comuna per a qualsevol tipus de metadades. Es tracta d'una notació basada en nocions fonamentals. Bàsicament: hi ha objectes, com ara pàgines web, i els objectes tenen propietats, com un responsable intel·lectual, una data de publicació o un contingut expressat en paraules clau, etc. Així mateix, hi ha

relacions entre els objectes, com ara una pàgina web que forma part d'una sèrie o és una versió en altra llengua d'altra pàgina web, etc.

Per a descriure el contingut semàntic i altres propietats d'una pàgina web, es pot utilitzar la norma RDF mitjançant el procediment d'etiquetatge XML per expressar els temes d'un document, entre d'altres coses.

En síntesi, la gran esperança del Web semàntic es basa, almenys, en tres coses: XML per fer els documents més explícits; metadades (expressades també en XML) per fer els documents més fàcils de representar, indexar i buscar i, finalment –es desprèn de l'anterior, encara que sol obviar-se–, una nova generació de programari (*software*) –programes i mètodes de representació del coneixement– que sàpiga explotar les dues coses anteriors.

La representació del coneixement necessitarà, alhora, procediments normalitzats, ja sigui per representar coneixement complex o de sentit comú. Aquestes representacions solen denominar-se ontologies, un camp interdisciplinari on solen confluir diverses disciplines cognitives, des de la intel·ligència artificial fins a la lingüística.

Ara bé, en l'esquema del Web semàntic se suposa que les metadades les posen principalment els propis autors dels documents. Quin és el problema? En primer lloc, els autors no solen estar entrenats per posar metadades i es necessita molta formació per saber triar bones paraules clau.

En segon lloc, els autors –no tots, ni de bon tros– menteixen. Així de simple. Volen que les seves pàgines web quedin molt alt en els cercadors, de manera que col·loquen trenta vegades la mateixa paraula, amb petites variants, perquè quedin a la part de dalt dels rànquings dels motors de cerca per als temes que a ells els interessa, encara que la seva pàgina no tingui en realitat molt (o gens) a veure amb aquest tema.

En tercer lloc, les persones ens equivoquem, i els autors de les pàgines web s'equivoquen: s'obliden de posar metadades, les posen malament, les posen en unes pàgines sí i en unes altres no, s'equivoquen en l'ortografia, etc.

Conclusió: gairebé cap motor de cerca es fia de les metadades per generar els resultats dels seus rànquings.

Possibilitats reals a curt i mig termini

El lector ja haurà deduït que, almenys segons l'opinió de qui escriu això, les possibilitats a curt i mig termini del Web semàntic són reduïdes. Efectivament. Una cosa és que es tracti d'un objectiu que val la pena perseguir i l'altra que es tracti d'un objectiu factible. Permeteu-me un exemple molt significatiu: sens dubte és un bon objectiu (almenys, molts ho creiem així) acabar amb la pobresa al món. És un exemple d'una fita lloable, amb la qual tots hauríem de comprometre'ns. Però que sigui un objectiu magnífic i molt desitjable, no ho converteix automàticament en assolible; almenys no en la seva totalitat i no a mig o a curt termini. S'ha d'abandonar per això? Ni de bon tros. Tot el contrari. Cal perseguir-la amb afany, perquè és l'única forma d'aconseguir progressos en aquest terreny, encara que siguin parcials.

El problema amb el Web semàntic, tal com la presenten alguns dels seus defensors (sobretot el W3 Consortium, que sembla haver-se especialitzat a llançar confusió sobre tots els seus projectes recents) és la immensa quantitat d'ingenuïtat o d'ignorància que exhibeix. En comparació, els programes contra la pobresa i a favor dels drets humans són obres mestres de pragmatisme (i saviesa).

Seguim, per exemple, amb les metadades: si gairebé ningú utilitza metadades ara, per quina raó, de sobte, tot el món posarà metadades en les seves pàgines? A més, si els autors de pàgines web han demostrat la seva incapacitat per utilitzar una norma relativament simple com era la primera versió de *Dublin Core*, per què ho faran ara que ha dut la seva complexitat al límit del que és impracticable?

Finalment, respecte a les ontologies i la seva explotació mitjançant motors d'inferència o sistemes experts, si la intel·ligència artificial suma ja diverses dècades de fracassos, almenys en relació a la hipòtesi forta, és a dir en relació al seu objectiu declarat a so de bombo i platerets d'assolir que els ordinadors pensin, per què tindrà èxit ara?

Per tant, les possibilitats que el Web semàntic sigui una realitat tal com la presenta el W3 Consortium, sense que es produeixi abans, almenys, un canvi de paradigma en les ciències de la computació, són ridícules. A més, necessitarem en paral·lel canvis no menys importants en altres àrees, incloent, per descomptat, les ciències de la documentació.

Tanmateix, no ens enganyem, l'objectiu del Web semàntic és magnífic, produirà importants avenços en alguns o en tots els terrenys relacionats amb la representació i l'accés al coneixement i, al meu entendre, des de les ciències de la documentació, hauria d'obtenir tot el nostre suport.

Bolcadors, mapadors i altres eines de localització d'informació

Les eines de cerca de Segona Generació són programes client que automatitzen processos de localització, recerca i recuperació d'informació. Classificació:

- Bolcadors
- Multicercadors
- Traçadors
- Indexadors
- Mapadors de ports
- Continguts de la Infranet: catàlegs de biblioteques, bases de dades bibliogràfiques, obres de referència, estadístiques i bases de dades numèriques, o Bases de dades Textuals. Els agents de la Infranet són clients Z39.50, amb mecanismes per a la realització automàtica de cerques de forma simultània i que sol permetre el bolcat dels registres. Entre els directoris més interessants, destaquen:

- Directori de recursos Z39.50, a nivell internacional. http://www.ilrt.bris.ac.uk/discovery/z3950/resources/
- Directorio espanyol de recursos Z39.50. http://www.absysnet.com/recursos/recz3950.html. Mereix un especial esment Bookwhere, aplicació de cerca, recuperació i exportació de la informació que utilitza el protocol Z39.50 d'Internet, i que té com a objectiu facilitar l'accés a registres bibliogràfics i a text complet via Internet. (Versió Demo a: http://www.web-clarity.com/products/overviewbookwhere.html).
- Continguts del Web invisible: pàgines òrfenes (sense connexió hipertextual); pàgines no textuals (com fitxers multimèdia i executables); pàgines amb accés mitjançant passarel·les (com pàgines amb paraula clau d'accés, ja siguin gratuïtes o de pagament; dipòsits de documents; revistes electròniques, etc.) o pàgines dinàmiques. Algunes adreces que permeten accés als continguts d'Internet invisible són: www.internetinvisible.com i www.completeplanet.com

Conclusions

En el futur dels sistemes d'informació hi ha una llarga llista d'innovacions a les quals val la pena parar atenció. Assenyalarem les que són més importants, segons la nostra opinió, pel seu impacte en les Ciències de la Documentació:

1. **Internet invisible**. S'ha produït un gran avenç en la varietat de formats que poden indexar els motors de recerca. D'altra banda, és possible que motors de cerca com Scirus siguin només un exemple de la classe de sistemes d'accés a la informació que podem esperar en el futur. Tot i així, hi ha diversos fronts en què hauríem de començar a col·locar les nostres energies i esforços. D'una banda, els documents no HTML són potencials enemics de la hipertextualidad. Hauríem de considerar si els avenços per una banda, no són retrocessos per una altra. En aquest cas, hauríem de considerar què fer o, almenys, considerar què fer en el camp de la investigació i les polítiques d'informació. Segur que tenim un ampli i bonic programa d'investigació per aquest costat. D'altra banda, les interfícies de consulta dels motors de cerca estan a anys llum de les possibilitats reals i del *know-how* sobre el tema. Un altre punt sobre el qual cal pensar o, millor encara, actuar.

2. **Web semàntic**. Encara que sigui amb mentalitat d'ONG, què podem fer a favor del Web semàntic si creiem en els seus beneficis a escala social encara que, ara com ara, aporti escassos beneficis individuals? Els organismes vinculats al món de la promoció del coneixement i la ciència i el patrimoni cultural (universitats, arxius, biblioteques, centres d'investigació, museus, etc.) s'haurien d'interessar-se pel Web semàntic. Per tant, a curt i mig termini, les organitzacions vinculades amb el món de la ciència, la cultura, el patrimoni, l'educació, etc., haurien de sentir-se obligades a: (1) interessar-se almenys per

coses tan aparentment innocents com el llenguatge XHTML juntament amb els fulls d'estil (CSS) i (2) estudiar polítiques de metadades en relació a totes les seves publicacions digitals.

3. **Què ens ensenya el Web semàntic?** Al meu entendre, ens ensenya el que, en realitat, ja sabíem: si agafes un conjunt de dades i les etiquetes sistemàticament i exhaustivament, tens el més semblat a la intel·ligència. Si les bases de dades exhibeixen un notable grau d'intel·ligència en comparació al Web és perquè en una base de dades tots les dades estan "etiquetades", és a dir, formen part dels valors d'un camp. Cada camp, al seu torn, té uns atributs ben definits: és un camp de text o és un camp numèric, o lògic, etc. Finalment, tots les dades en una base de dades estan sistematitzades: cada registre respon a la mateixa estructura, així que la posició (la sintaxi) genera sentit (semàntica). Així que, el que és (genialment) nou en el Web semàntic és la idea de convertir tota el Web en la més gegantesca base de dades que la humanitat hagués somiat mai.

Iniciatives de patrimoni digital

Projecte de la Carta de la UNESCO per a la Preservació del Patrimoni Digital.

Diu la UNESCO en el seu document "DIRECTRIUS PER A LA PRESERVACIÓ DEL PATRIMONI DIGITAL":

"Gran part de la ingent quantitat d'informació que es ***produeix*** en el món és d'origen digital i existeix en una gran varietat de formats: text, bases de dades, enregistraments sonors, pel·lícules, imatges. Per a les institucions culturals que tenen al seu càrrec la ***recol·lecció*** i la preservació del patrimoni cultural, definir quins elements han de conservar-se per a les generacions futures i com procedir en la seva selecció i conservació, s'està tornant un problema urgent. L'enorme tresor d'informació digital produïda avui dia en pràcticament totes les àrees de les activitats humanes i concebuda per a ser consultada amb computadores, podria perdre's si no s'elaboren tècniques i ***polítiques específiques per a seva*** conservació**.**"

Podeu llegir el text complet en:

http://unesdoc.unesco.org/images/0013/001300/130071s.pdf