La web semàntica

PID_00272092

Blas Torregrosa García

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Blas Torregrosa García

Enginyer en Informàtica i màster universitari en Seguretat de les Tecnologies de la Informació i de les Comunicacions (MISTIC) per la Universitat Oberta de Catalunya (UOC). Especialitzat en ciberseguretat. Professor col·laborador del màster de Ciència de Dades de la UOC i professor associat a la Universitat de Valladolid (UVA).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Ferran Prados Carrasco (2020)

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

En els seus anys de vida, la World Wide Web (WWW), més coneguda com **la web**, ja s'ha convertit en una eina indispensable per a la vida quotidiana de les persones i ha arribat a ser el principal mitjà de comunicació mundial d'informació. Des dels seus inicis, al voltant de l'any 1990 i fruit de l'existència de la xarxa internet, la web ha experimentat un creixement gairebé exponencial en termes de contingut.

Aquest contingut no és solament utilitzat per persones, sinó que cada vegada hi ha una major necessitat que pugui ser consumit per mecanismes automàtics. En aquest context i per aquest motiu els avenços tecnològics dirigits al tractament automàtic del contingut són cada vegada més necessaris i habituals. Per realitzar tractaments automàtics de dades i d'informació cal identificar el significat de les dades de forma explícita, amb l'objectiu d'ampliar la web amb informació semàntica i convertir-la en una web de dades.

1. Evolució de la web

L'any 1989 la World Wide Web (WWW, o simplement, la web) sorgeix a partir d'una proposta de Tim Berners-Lee per utilitzar l'hipertext com a mecanisme per intercanviar informació. Actualment la web és un dels serveis més coneguts i populars d'internet.

La **web** és una xarxa de pàgines escrites en hipertext i connectades entre si per mitjà d'enllaços. Aquestes pàgines estan allotjades en diferents servidors connectats entre si i que utilitzen un protocol (*HyperText Transfer Protocol* o HTTP) que permet descarregar i consultar les pàgines d'hipertext i els recursos que enllacen: imatges, vídeo, àudio, documents, etc.

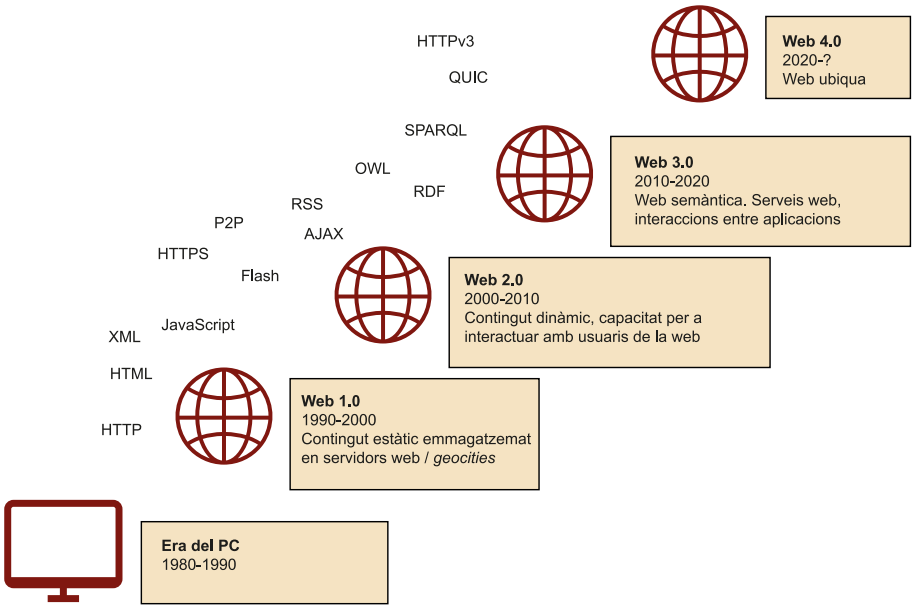
Quan es vol consultar una pàgina web cal utilitzar un programari especial per a aquesta funció denominat **client web** o **navegador**,[1] que permet visualitzar les pàgines d'hipertext i guiar la navegació per mitjà dels diferents enllaços. En aquest sentit, l'**hipertext** inclou diferents característiques que defineixen com s'ha de presentar la informació d'una pàgina.

(1)En anglès, *browser*.

El **contingut** de les pàgines web convencionals està dissenyat per ser llegible per persones, per la qual cosa no és adequat per a un processament automàtic i és molt poc eficient quan es busca informació relacionada. Els conjunts de dades a la web són repositoris de dades aïllades que no estan vinculades entre si. Aquesta limitació es pot abordar organitzant i publicant dades, utilitzant formats que agreguin l'estructura i dotin de significat el contingut de les pàgines web, vinculant les dades que estiguin relacionades entre si. Els sistemes informàtics poden «comprendre» millor aquest tipus de dades, la qual cosa permet automatitzar les tasques. Precisament aquest és un dels reptes de la web semàntica.

Per entendre l'evolució del contingut de la web, veurem quina ha estat l'evolució del seu ús i de la tecnologia subjacent.

Figura 1. Evolució de la web



1.1. La web 1.0

La web va néixer al començament dels anys noranta com la unió de dues tecnologies ja existents: l'**hipertext** i **internet**.

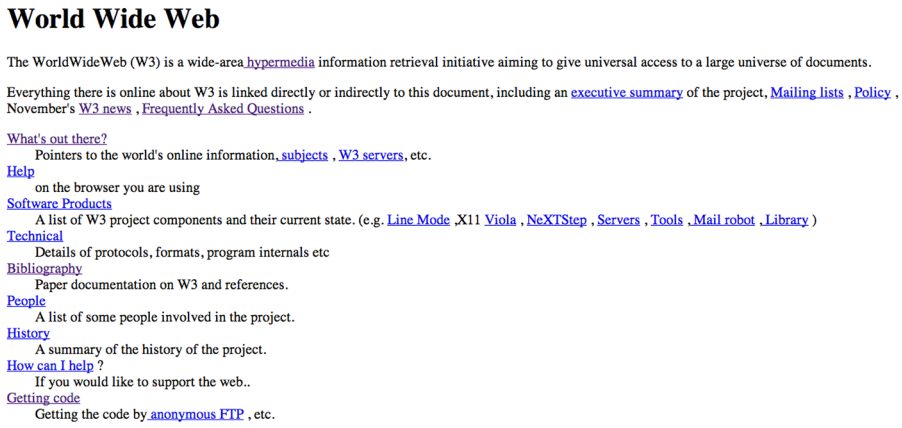
> **Web 1.0** és el terme amb el qual se solen denominar les primeres pàgines web, caracteritzades per oferir informació estàtica de manera unidireccional.

És a dir, l'usuari que accedeix al contingut únicament pot llegir-lo de manera passiva, sense la possibilitat de contribuir, en cap cas, a la seva ampliació o correcció. Per tant, l'usuari és un simple consumidor d'una web enfocada a la lectura.

Tecnològicament apareix el llenguatge de marques HTML (*HyperText Mark-Up Language*), llenguatge emprat per a la creació de pàgines web especialment idoni per enllaçar continguts web per mitjà dels hipervincles (*hyperlinks*). També apareix un nou protocol, l'HTTP (*HyperText Transfer Protocol*), i el sistema de localització de recursos URL (*Uniform Resource Locator*).

La popularitat aconseguida per la web a partir de 1994 va depassar les fronteres de l'intercanvi científic i va permetre l'inici de les primeres aplicacions comercials. Moltes de les companyies de comerç electrònic, com ara Amazon, eBay o Yahoo!, es van fundar entre 1994 i 1995.

Figura 2. Primera pàgina web



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

Font: home.cern/science/computing/birth-web

Figura 3. Primeres versions de Google, creada el 1996

L'any 1995 Microsoft llança el sistema operatiu Windows 95 que inclou el navegador Internet Explorer. La competència entre Internet Explorer i Netscape (navegador en aquells dies més popular) dona lloc a la primera guerra de navegadors, en què els navegadors competeixen entre si incloent cada vegada més característiques i forçant els límits del llenguatge HTML.

1.2. La web 2.0

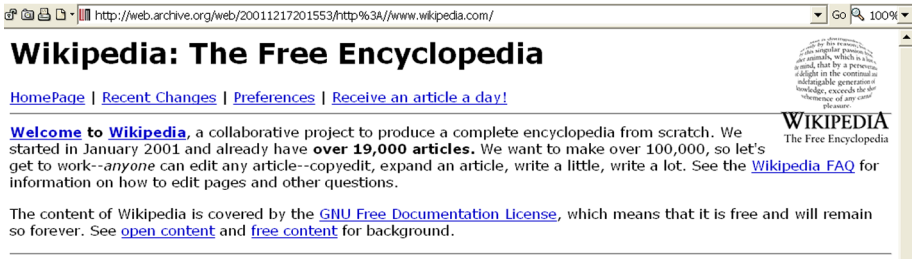
Gairebé una dècada més tard apareix el concepte de **web 2.0**, també coneguda com a **web col·laborativa**. Es considera una evolució de la web 1.0, amb l'objectiu de permetre una major interacció amb el contingut, és a dir, dotar les pàgines de mecanismes per a la col·laboració dels usuaris en la transformació i creació de nou contingut. D'aquesta manera, l'usuari també passa a ser productor i s'aconsegueix una comunicació bidireccional amb la web.

Per fer això possible, cal que tecnològicament es permeti la modificació de continguts amb el mínim coneixement tècnic. Per això apareixen eines web com els blogs o els wikis, entre d'altres. Això representa la convergència entre el mitjà de comunicació i el contingut, i en aquest context apareixen les primeres xarxes socials, com ara MySpace, YouTube, etc.

Figura 4. Disseny original i nom de Thefacebook el 2004

Font: Wikipedia.

Figura 5. Disseny de Wikipedia el 2001



Font: Wikipedia.

La **web 2.0** té dues característiques essencials: la consolidació de les xarxes socials i l'acostament al temps real.

La web social es va popularitzar gràcies a l'èxit de Facebook i d'altres xarxes socials. I la web en temps real podria estar representada per Twitter, que va néixer l'any 2006 com un sistema de missatges curts, semblants als SMS (menys de 140 caràcters), per a la web.

D'altra banda, en paral·lel s'avança en tecnologies per al client: es consolida JavaScript en els navegadors i les tècniques d'accés asíncron com AJAX (amb *frameworks* com jQuery), i també l'estàndard d'intercanvi de dades JSON.

1.3. La web 3.0

La **web 3.0** es coneix generalment com la **web semàntica** o també **la web de lectura-escriptura-execució**.

La web 3.0 descentralitza els serveis com ara la cerca, les xarxes socials i les aplicacions de missatgeria instantània que depenen d'una sola organització per funcionar. La web semàntica i els serveis web són els principals components de la web 3.0.

L'objectiu de la web semàntica és transformar l'actual web sintàctica (una web de documents), en què la unitat d'informació és el document, en una **web de dades**, en la qual la unitat d'informació sigui la dada. Per aconseguir-ho, primer cal dotar de significat els recursos web, és a dir, s'ha de classificar, estructurar i anotar semànticament cada recurs perquè pugui ser interpretat per les aplicacions que l'hauran de processar.

Els serveis web són un mètode de comunicació entre els sistemes informàtics per mitjà d'una xarxa de comunicacions i aquesta comunicació es realitza de manera estandarditzada (mitjançant XML, JSON, SOAP, etc.) que permeti la integració d'aplicacions web heterogènies.

Des de l'any 2011 es detecta un ús massiu de dispositius mòbils per accedir a la web.

1.4. La web 4.0

La **web 4.0** és el proper gran avenç que se centrarà a oferir un comportament més intel·ligent i es basarà a explotar tota la informació que ara mateix conté, però d'una forma més natural i efectiva.

Actualment, els cercadors segueixen sent elements essencials. El que proposa la web 4.0 és millorar aquesta experiència mitjançant l'ús de tecnologies que permetrien un nivell d'interacció més complet i personalitzat, usant tota la informació que donem i existeix a la web. Tot això es fonamentarà en quatre pilars:

1) La comprensió del llenguatge natural i tecnologies de conversió de veu en text i viceversa.

2) Sistemes de comunicació de màquina a màquina (M2M).

3) Ús de la informació de context com el posicionament GPS o el ritme cardíac detectat per un *wearable*, dispositiu mòbil, etc.

4) Model millorat d'interacció amb l'usuari.

La web 4.0 seria la unió de la web semàntica, la intel·ligència artificial i la veu com a forma de comunicació. L'objectiu és una web ubiqua el propòsit primordial de la qual serà el d'unir les intel·ligències que es comuniquen entre si per generar la presa de decisions.

2. La web semàntica

2.1. Què és la web semàntica?

> «La web semàntica és una extensió de l'actual web en què la informació té un significat ben definit i permet que les persones i els ordinadors cooperin».
>
> Tim Berners-Lee, James Hendler, Ora Lassila (2001). *The Semantic Web, Scientific American* (vol. 5, núm. 284, pàg. 34-43).

Una definició de web semàntica comença amb la definició de la **semàntica**, és a dir, el significat. Les pàgines web estan plenes de dades i etiquetes associades. La majoria de les etiquetes representen instruccions de format, com ara `<H1>` per indicar un encapçalament. Semànticament sabem que les paraules envoltades amb `<H1>` són més importants que un altre text que no ho estigui.

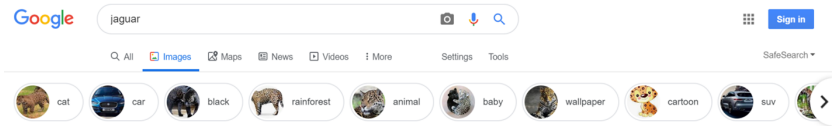
Semàntica

La paraula *semàntica* també s'usa a la web en més contextos. Per exemple, en HTML5 hi ha elements estructurals semàntics (com ara `section` que permet agrupar per temes), encara que aquesta expressió es refereix al «significat» dels elements. És a dir, no s'ha de confondre la semàntica dels elements d'etiquetatge amb la semàntica (la capacitat de processament de les màquines) de les anotacions utilitzades a la web semàntica.

Les pàgines web es basen en llenguatges d'etiquetatge (HTML) per determinar l'estructura del document, els fulls d'estil (CSS), per a l'aparença, i *scripts* (JavaScript) per al comportament, encara que el contingut segueix estant orientat a ser comprensible bàsicament per a les persones.

> Si es busca «Jaguar» a la web, per exemple, els algorismes dels motors de cerca no saben si ens referim a una marca d'automòbils de luxe o a un felí depredador sud-americà.

Figura 6. Cerca



Font: Google.

Per defecte, els encapçalaments, textos, enllaços i altres elements de les pàgines web no tenen sentit per als ordinadors. Els navegadors simplement mostren els documents web, encara que solament el cervell humà pot interpretar el significat.

El concepte de dades entendibles per les màquines no és nou i no es limita a la web. Per exemple, les targetes de crèdit o els codis de barres contenen dades llegibles pels humans i per les màquines.

Fins i tot els documents XML, que tenen unes regles sintàctiques rigoroses, tenen les seves limitacions quan són processats per les màquines. Per exemple, si es defineix una entitat XML amb l'etiquetatge `<CMD>` i `</CMD>`, què significa realment CMD? Es pot referir a un comandament del sistema o a un centre de comandament distribuït o a una cabina doble.

Per fer que els continguts siguin processables sense ambigüitats per les aplicacions informàtiques cal afegir a les pàgines web dades organitzades (estructurades), com a anotacions o com a metadades, que facin que estiguin vinculades a altres dades estructurades relacionades.

L'avantatge d'afegir anotacions semàntiques a les pàgines web és que les persones podran seguir navegant pels documents web, mentre que les aplicacions informàtiques podran processar aquestes anotacions per classificar les entitats de dades, descobrir relacions lògiques entre entitats, crear índexs, etc.

La **web semàntica** és un conjunt d'estàndards i de bones pràctiques que permet compartir dades a la web i la seva semàntica per al seu ús per les aplicacions informàtiques.

Web semàntica

La web semàntica està impulsada pel Consorci World Wide Web (W3C).

Taula 1. La web tradicional enfront de la web semàntica

Característiques	Web tradicional	Web semàntica
Component fonamental	Contingut no estructurat, llenguatge natural	Contingut estructurat
Audiència	Persones	Màquines/Aplicacions
Enllaços	Indiquen localització	Indiquen localització i significat
Basada en	Sintaxi	Semàntica

2.2. Comprenent el contingut de la web

Quan s'accedeix a un contingut de la web, no importa en quin idioma estigui, una persona o un sistema informàtic ha de considerar les qüestions següents:

- Quina informació és important i com puc saber-ho?
- Què és informació i què és publicitat?
- Quina informació està relacionada amb el contingut?
- Què significa la informació?

Les persones tenen coneixement del context i experiència per resoldre aquestes qüestions. Per als sistemes informàtics és una tasca complexa filtrar i obtenir informació de la web per determinar el que és important. És a dir, per entendre el significat de la informació de la web.

2.2.1. La importància del significat

Significat, comprensió i enteniment són tres conceptes interrelacionats. L'**enteniment** és la capacitat de comprendre el significat de la informació, mentre que la **comprensió** és fer propi el que s'entén i assimilar-ho. La informació s'envia mitjançant un missatge des d'un remitent fins a un receptor usant un llenguatge concret.

El receptor del missatge entén la informació si el receptor interpreta correctament la informació. La **correcta interpretació de la informació** depèn de:

- **Sintaxi**.[2] En gramàtica, és l'estudi dels principis i processos pels quals les oracions estan ben construïdes en un determinat idioma. En llenguatges formals, la sintaxi és un conjunt de regles pel qual es generen expressions correctes (ben formades) amb un conjunt de símbols (alfabet). En informàtica, la sintaxi regula l'estructura de les dades, és a dir, les regles del que està permès i el que no.

(2)Del grec, *sin* (junts) i *taxis* (ordre) que significa 'coordinació'.

- **Semàntica**.[3] És la part de la lingüística centrada en el sentit i significat del llenguatge i és l'estudi de la interpretació dels símbols usats per una comunitat en unes circumstàncies particulars o context. La semàntica també utilitza regles de sintaxis per determinar el sentit i significat de conceptes complexos que deriven de conceptes simples. La semàntica d'un missatge depèn del context i de la pragmàtica.

(3)Del grec, *semantikos* ('explicació del que significa').

- **Context**.[4] En una expressió, denota tot l'adjacent a un símbol (concepte) respecte a les relacions amb expressions (conceptes) adjacents i altres elements relacionats. El context es refereix a tots els elements de qualsevol tipus de comunicació que determinen la interpretació del contingut comunicat. Podem distingir entre contextos **generals** (lloc, temps, etc.) i contextos **personals o socials** (relació entre el remitent i el receptor).

(4)De llatí, *contextus* ('connexió', 'entreteixit').

- **Pragmàtica**.[5] Reflecteix la intenció de l'ús del llenguatge per comunicar un missatge, és a dir, la finalitat del remitent. En lingüística, és l'estudi de l'ús de l'idioma en diferents situacions. La pragmàtica estudia les formes en què el context contribueix al significat.

(5)Del grec, *pragmatikos* ('acció').

- **Experiència**.[6] Considera tota la informació apresa i posada en un context, és a dir, coneixement mundial o sentit comú que influeix en com s'interpreta la informació.

(6)Del llatí, *experientia* ('esdeveniment viscut').

En resum, per a una comunicació eficaç cal que la informació sigui correctament transmesa (sintaxi) i el significat (semàntica) de la informació transmesa sigui interpretat correctament. L'enteniment del missatge dependrà del context del remitent i del receptor, a més de la intenció del remitent. I el context dependrà de l'experiència que tinguin el remitent i el receptor.

2.2.2. De la web tradicional a la web de dades

El problema de la web tradicional és que no té una semàntica explícita, ja que la majoria de la informació que es transfereix a la web està codificada en llenguatge natural o dins de continguts multimèdia, com ara imatges, vídeos, àudios, etc. Per tant, el significat està ocult i això dificulta que els sistemes informàtics puguin entendre el contingut de la web.

La web de dades o web semàntica es considera una actualització de la web de documents tradicional.

La web de dades considera la web com una enorme base de dades descentralitzada (base de coneixement) amb dades accessibles per als sistemes informàtics.

Per aconseguir una web de dades fa falta una condició prèvia: el contingut de la web ha de ser llegit i interpretat correctament (entès) per les màquines. Hi ha dos enfocaments per obtenir això:

1) Motors de cerca amb processament del llenguatge natural[7] que utilitzen les tecnologies de recuperació d'informació que tracten d'entendre el contingut de la web mitjançant estadístiques i l'aprenentatge automàtic. L'objectiu és extreure la semàntica implícita a la web.

(7) *Natural Language Processing* (NLP).

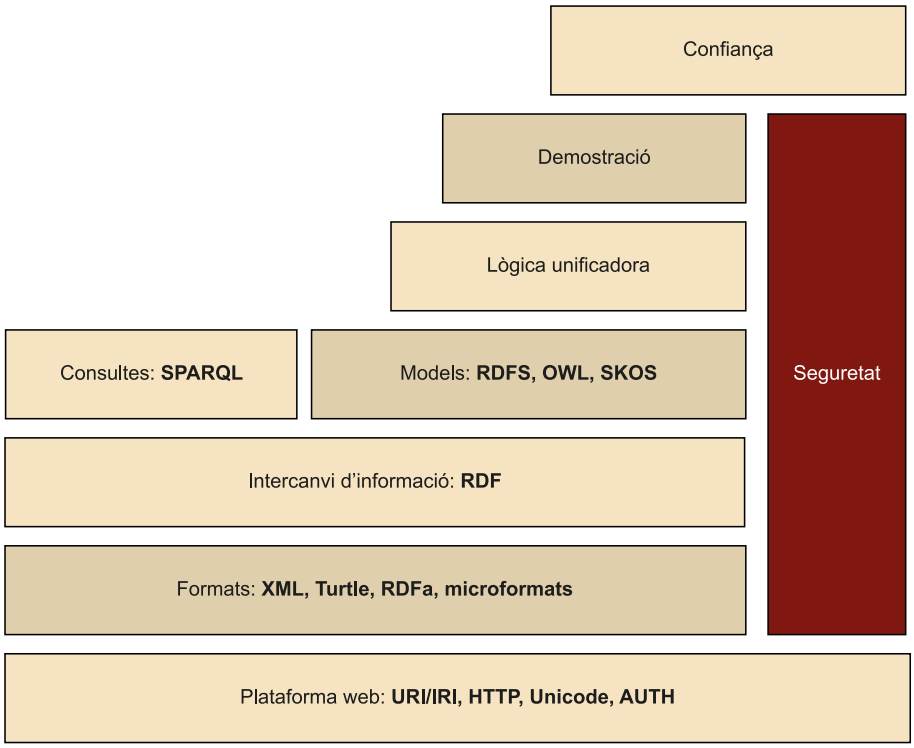
2) Tecnologies de la web semàntica en què el contingut en llenguatges naturals s'anota explícitament amb metadades semàntiques. Les metadades semàntiques codifiquen el significat del contingut i, llavors, pot ser llegit i interpretat correctament per les màquines.

2.3. Arquitectura de la web semàntica

L'any 2000, Tim Berners-Lee va proposar un model de capes per esquematitzar el desenvolupament futur de la web semàntica. El seu model es va batejar com **el pastís de la web semàntica**.[8] Més endavant, el 2006, es va publicar una nova versió del pastís.

(8) En anglès, *Semantic Web Layer Cake*.

Figura 7. El pastís de la web semàntica



En aquest model es distingeixen les capes següents:

1) Plataforma web: la primera capa es refereix a les tecnologies d'infraestructura més bàsiques, com ara l'HTTP (*HyperText Transfer Protocol*), Unicode per a la codificació de caràcters internacionals, URI per identificar recursos i mecanismes d'autenticació per a la seguretat.

2) Formats: la capa següent es refereix als formats per representar o serialitzar la informació de les capes superiors. Es pot representar tot en XML, Turtle, RDFa o microformats.

3) Intercanvi d'informació: és el bloc principal que desenvolupa la infraestructura que permetrà la descripció dels recursos. En aquesta capa el llenguatge de referència és RDF (*Resource Description Framework*).

4) Models: en aquesta capa d'organització de la informació s'agrega semàntica a les dades definides amb RDF mitjançant vocabularis definits en RDFS (RDF *Schema*) o SKOS (*Simple Knowledge Organization System*), o mitjançant la definició d'ontologies per als diferents dominis amb OWL (*Web Ontology Language*), que és el llenguatge recomanat per a la definició de les ontologies. Aquests models són representacions del coneixement.

5) Consultes: es proposa un llenguatge de consultes per a la web de dades que exerceixi un paper similar al del llenguatge SQL per a les bases de dades relacionals. Aquest llenguatge es denomina **SPARQL** (*SPARQL Protocol and RDF Query Language*).

6) Lògica/Demostració: en aquestes dues capes s'identifiquen les qüestions d'inferència que cal desenvolupar per generar nou coneixement a partir de la web de dades.

7) Confiança: si es pretén que es puguin realitzar tasques de forma autònoma a partir de la informació publicada a la web, cal desenvolupar una infraestructura que permeti assegurar la seva fiabilitat. Aquesta infraestructura es recolzaria en signatures digitals que permetrien identificar l'autoria de les publicacions.

Aquest model en capes ha rebut crítiques. El més controvertit del pastís de la web semàntica és que es pretén que cada capa d'un nivell es basi en el nivell anterior. D'una banda, hi ha algunes dificultats tècniques per assegurar que cada capa superior es basi en la capa inferior. I, de l'altra, algunes tecnologies de la web semàntica encara segueixen en desenvolupament.

Per tot això, el pastís no s'ha d'entendre de forma literal, sinó com un mapa conceptual de les tecnologies implicades.

2.3.1. Nomenant les coses: URL, URI i URN

Els **recursos web** es poden localitzar mitjançant adreces IP que són úniques. No obstant això, les adreces IP són força difícils de recordar. Per això s'utilitzen els **noms de domini**, que segueixen unes regles de sintaxis. Els noms de domini convencionals no poden contenir caràcters accentuats o no alfanumèrics. Amb la introducció dels noms de domini internacionals (IDN)[9] és possible usar noms en diversos idiomes i amb diferents alfabets.

[9] Acrònim de l'anglès *Internationalized Domain Names*.

IP

IPv4 utilitza adreces de 32 bits (4 nombres de 8 bits separats per punt) que limiten l'espai de les adreces a 4.294.967.296 (2^{32}) adreces. Mentre que una adreça IPv6 està formada per 128 bits.

Els **URI** (*Uniform Resource Identifier*), el subconjunt més conegut dels quals són els URL (*Uniform Resource Locators*), proporcionen el mecanisme per identificar de forma unívoca qualsevol recurs de la web: documents, imatges, vídeos, etc. A la web semàntica, els URI tindran a més la funció d'identificadors d'objectes del món real. Qualsevol persona o objecte podrà ser identificat mitjançant un URI.

Regles de sintaxis

Regles de sintaxis definides en les RFC 1035, RFC 1123 i RFC 2181.

Figura 8. Format d'URI

Els URI tenen l'estructura següent:

- **Esquema**: és un nom que es refereix a l'especificació per assignar identificadors com owl: o rdf:. I també pot identificar el protocol d'accés al recurs, com ara http:, ftp:, etc.

- **Autoritat**: és un nom jeràrquic que representa el nom del domini a internet. Comença pel símbol //.

- **Ruta**: és una seqüència de segments separats per / i en forma jeràrquica, similar als directoris en sistemes de fitxers.

- **Consulta**: és una part que comença per ?, aporta informació opcional, normalment mitjançant parells atribut=valor separats pel símbol &.

- **Fragment**: també és una part opcional que comença per # i permet identificar una part o recurs secundari dins d'un recurs principal.

Moltes vegades s'utilitza els URI com a sinònim d'URL, encara que URI és un terme més ampli. Els URI es poden classificar com a URL, com a **URN** (*Uniform Resource Names*), o tots dos. Un URN defineix la identitat d'un recurs, mentre que l'URL proporciona un mètode per trobar-lo (incloent el protocol i la ruta).

Com que la notació dels URI pot ser massa llarga, també hi ha els CURIE (*Compact* URI), que és una notació abreujada pels URI. La sintaxi consta de tres parts:

1) El prefix (dbpedia com a prefix d'http://dbpedia.org/resource/)
2) El caràcter dos punts (:)
3) La referència

Planeta Venus

Per exemple, com a referència al planeta Venus hi ha l'URI http://dbpedia.org/resource/venus que es pot abreujar com dbpedia:Venus.

A més, hi ha l'**IRI** (*Internationalized Resource Identifier*), que és una extensió de l'URI, que permet utilitzar el joc de caràcters Unicode, amb el qual es poden incloure caràcters xinesos, kanji japonesos, àrabs o ciríl·lics per identificar els recursos.

Moltes vegades ocorre que els recursos canvien de localització o de domini. Encara que les adreces web es poden redirigir (generalment usant la redirecció amb HTTP 302) a la nova adreça, això pot causar problemes. Una opció és usar una localització persistent a la xarxa. Per a això, caldrà utilitzar **PURL** (*Persistent Uniform Resource Locator*).

2.3.2. Formats

En un nivell superior trobem els documents i la seva estructuració lògica. Originalment, XML (*eXtensible Markup Language*) constitueix la base sintàctica de la web semàntica i sobre la qual es recolzen la resta de capes. XML és un metallenguatge que permet definir diferents llenguatges d'etiquetatge validant-los mitjançant definicions de documents o DTD (*Document Type Definitions*) i també amb XML *Schemas*. No obstant això, té l'inconvenient de ser un llenguatge molt verbós i complicat de llegir per les persones.

Un altre format habitual és Turtle (*Terse RDF Triple Language*), que va ser desenvolupat expressament per ser llegible per les persones. Es tracta d'un format de text en què apareixen codificades les tripletes del llenguatge RDF.

Últimament hi ha dos formats que han estat d'interès per al desenvolupament de la web semàntica: el JSON-LD i els microformats. JSON-LD (*JavaScript Object Notation for Linked Data*) és un format més llegible per a les persones i que utilitza JSON de conjunts de parells atribut-valor. Un microformat (abreujat com a µF) és un mètode per agregar semàntica usant HTML amb un etiquetatge específic.

2.3.3. El llenguatge RDF

RDF (**Resource Description Framework**) és un llenguatge d'etiquetatge, creat mitjançant XML, que defineix un model de dades per descriure recursos (qualsevol objecte identificable per un URI). Per a això, utilitza enunciats en forma de tripletes subjecte-predicat-objecte (recurs-propietat-valor), en què el subjecte i el predicat són URI i l'objecte pot ser un URI o un valor literal. El predicat (propietat) descriu la relació entre el subjecte i l'objecte. El llenguatge RDF és l'equivalent a HTML (*HyperText Markup Language*) a la web convencional.

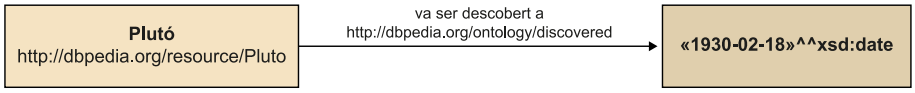
Si volguéssim expressar la frase en llenguatge natural «Plutó va ser descobert el 1930» mitjançant tripletes RDF seria:

Taula 2. Exemple de tripletes RDF

	Model de dades RDF	**Tripleta RDF**	**Tipus**
Subjecte	Plutó	http://dbpedia.org/resource/pluto	URI
Predicat	Descobert	http://dbpedia.org/ontology/discovered	URI
Objecte	1930	«1930»	Literal

RDF també és un **graf** dirigit, en què els nodes són el subjecte i l'objecte, i els arcs són els predicats que els connecten.

Figura 9. Exemple de graf RDF



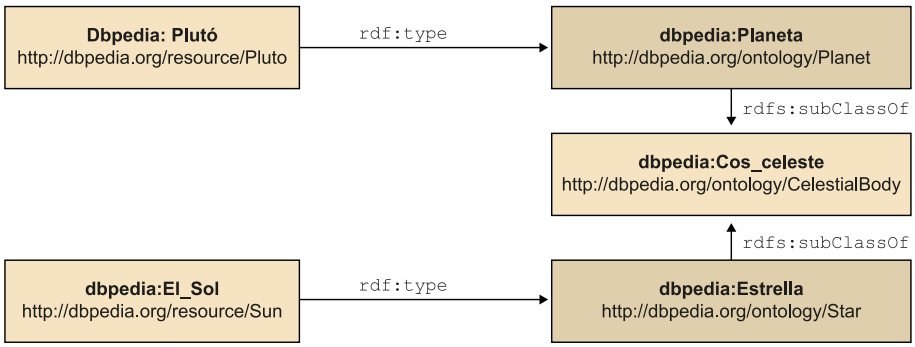
2.3.4. RDF Schema

RDF **Schema** (RDFS) és un vocabulari RDF que ens permet descriure recursos mitjançant una orientació a objectes similar a la de molts llenguatges de programació com Java. Per a això proporciona un mecanisme per definir classes, objectes i propietats, a més de relacions entre les classes i les propietats. També restriccions de domini i rang sobre les propietats.

RDFS permet la definició de classes amb `rdfs:Class` i la instanciació de classes en RDF amb `rdf:type`. Hi ha més definicions com `rdf:Property`, `rdfs:domain`, `rdfs:range`, `rdfs:Literal` o `rdfs:Resource`.[10] RDFS també defineix relacions jeràrquiques amb `rdfs:subClassOf` i `rdfs:subPropertyOf`.

(10)Tot en el model RDF són recursos.

Figura 10. Exemple d'esquema amb RDFS



2.3.5. Ontologies

Les ontologies s'utilitzen com a model de **representació de coneixement** en el camp de la intel·ligència artificial.

> Es pot definir **ontologia** com una especificació explícita i formal d'un concepte dins d'un determinat domini d'interès.

En el nostre cas, una ontologia s'utilitza com un artefacte que defineix:

- Un **vocabulari compartit** que descriu un determinat domini.
- Un **conjunt d'hipòtesis** sobre els termes d'aquest vocabulari. Generalment s'utilitza un **llenguatge formal** manipulable automàticament.

Les ontologies es poden usar per buscar, consultar, indexar i administrar metadades, i per millorar la interoperabilitat de les aplicacions i les bases de dades.

OWL (*Web Ontology Language*) és un llenguatge que estén RDF i RDFS, i que permet la construcció de classes complexes a partir d'altres definicions de classes, i també l'encadenament de propietats.

3. DBpedia

La DBpedia és un projecte per extreure de la Wikipedia dades estructurades.

La versió en anglès de la base de coneixement de la DBpedia de juny de 2018 descriu 6,6 milions d'entitats, de les quals 4,9 milions tenen resums, 1,9 milions tenen coordenades geogràfiques i 1,7 milions de representacions.

En total, 5,5 milions de recursos es classifiquen en una ontologia consistent, que consta d'1,5 milions de persones, 840.000 llocs (incloent 513.000 llocs habitats), 496.000 obres (incloent 139.000 àlbums de música, 111.000 pel·lícules i 21.000 videojocs), 286.000 organitzacions (incloses 70.000 companyies i 55.000 institucions educatives), 306.000 espècies, 58.000 plantes i 6.000 malalties.

3.1. De la Wikipedia a la DBpedia

La **Wikipedia** s'ha convertit en una font d'informació de referència sobre conceptes, fets, ciència i cultura utilitzada per milions d'usuaris. El projecte **DBpedia** sorgeix sobre la base de la formalització del coneixement dels articles de Wikipedia.

Els articles de Wikipedia no solament inclouen contingut textual. Una bona part d'aquests també conté una gran quantitat d'informació estructurada mitjançant fitxes descriptives (*infoboxes*).

Una ***infobox*** és una plantilla wiki en què es defineix una estructura de dades comuna i la seva representació visual per a determinats tipus d'articles (persones, ciutats, pel·lícules, etc.).

Figura 11. Pàgina de Wikipedia amb una *infobox* destacada

Plutón (planeta enano)

Plutón, designado **(134340) Pluto**, es un planeta enano del sistema solar situado a continuación de la órbita de Neptuno. Su nombre se debe al dios mitológico romano Plutón (Hades según la mitología griega). En la Asamblea General de la Unión Astronómica Internacional celebrada en Praga el 24 de agosto de 2006 se creó una nueva categoría llamada plutoide, en la que se incluye a Plutón. Es también el prototipo de una categoría de objetos transneptunianos denominada plutinos. Plutón posee una órbita excéntrica y altamente inclinada con respecto a la eclíptica, que recorre acercándose en su perihelio hasta el interior de la órbita de Neptuno. Asimismo posee también cinco satélites: Caronte, Nix, Hidra, Cerbero y Estigia,[3] [4] los cuales son cuerpos celestes que comparten esa misma categoría.

Su gran distancia al Sol y a la Tierra, unida a su reducido tamaño, impide que brille por encima de la magnitud 13,8 en sus mejores momentos (perihelio orbital y oposición), por lo cual solo puede ser apreciado con telescopios a partir de los 200 mm de abertura, fotográficamente o con cámara CCD. Incluso en sus mejores momentos aparece como astro puntual de aspecto estelar, amarillento, sin rasgos distintivos (diámetro aparente inferior a 0,1 segundos de arco). No fue hasta el año 2015 cuando la sonda espacial New Horizons pasó sobre el planeta y permitió apreciar por primera vez de forma nítida su aspecto real.

Plutón fue descubierto el 18 de febrero de 1930 por el astrónomo estadounidense Clyde William Tombaugh (1906-1997) desde el Observatorio Lowell en Flagstaff, Arizona, y fue considerado el noveno y más pequeño planeta del sistema solar por la Unión Astronómica Internacional y por la opinión pública desde entonces hasta 2006, aunque su pertenencia al grupo de planetas del sistema solar fue siempre objeto de controversia entre los astrónomos. Durante muchos años existió la creencia de que Plutón era un satélite de Neptuno que había dejado de ser satélite por el hecho de alcanzar una segunda velocidad cósmica. Sin embargo, esta teoría fue rechazada en la década de 1970.[5]

Tras un intenso debate, y con la propuesta de los astrónomos uruguayos Julio Ángel Fernández y Gonzalo Tancredi ante la Asamblea General de la Unión Astronómica Internacional en Praga, República Checa, se decidió por

Font: Wikipedia.

D'això precisament sorgeix DBpedia, intentant convertir el coneixement de les *infoboxes* de la Wikipedia en coneixement formalitzat mitjançant una ontologia que aplica els principis i les tecnologies de dades obertes i enllaçades.

3.2. Wikidata

Wikidata és un concentrador de dades estructurades en què cada entitat es representa mitjançant un IRI, en què també es recullen els enllaços als articles equivalents de Wikipedia en diferents idiomes.

Wikidata

Així, l'entitat Q339 fa referència al planeta Plutó del qual hi ha articles en anglès, espanyol, català i gallec.

Encara que pot semblar que DBpedia i Wikidata són projectes molt semblants, que produeixen dades estructurades derivades dels articles de Wikipedia, hi ha diferències, des de la identificació dels recursos (URI/IRI per la DBpedia, mentre que Wikidata usa identificadors numèrics independents de l'idioma) fins a l'estructura interna (RDF per la DBpedia, mentre Wikidata desenvolupa el seu propi model de dades).

Bibliografia

Allemang, D.; Hendler, J. (2011). *Semantic Web for the working ontologist* (2a. ed.). Massachusetts: Morgan Kaufmann.

Saorín, T.; Pastor-Sánchez, J. A. (2018). «Wikidata y DBpedia: viaje al centro de la web de datos». *Anuario ThinkEPI* (núm. 12, pàg. 207-214).

Sikos, L. F. (2015). *Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data*. Apress.

Taylor, J.; Evans, C.; Segaran, T. (2009, juliol). *Programming the Semantic Web*. O'Reilly Media.