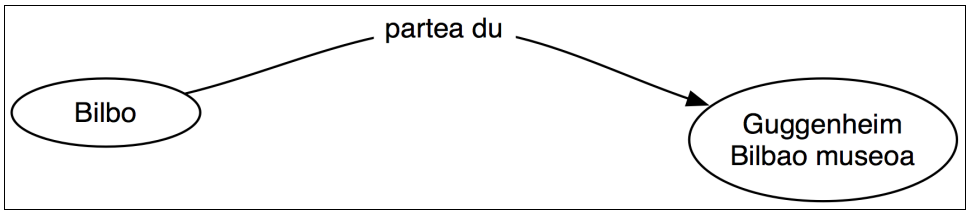
# DATUEN AMARAUNA

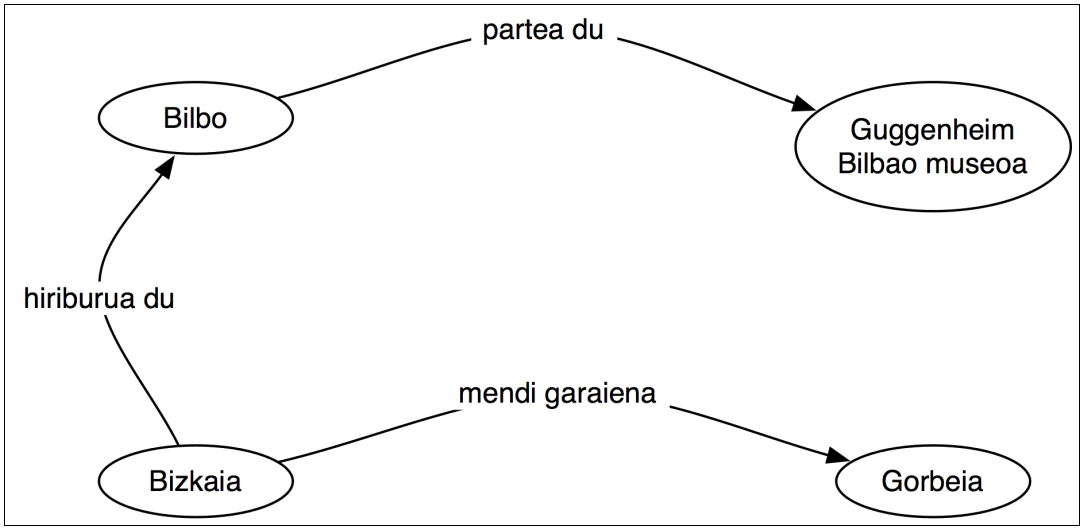
## 1.- INFORMAZIOA INTERNETEN ZUZENEAN ARGITARATU

Web sarean gabiltzanean, gure xedea informazioa aurkitzea da. Hala ere, informazioa eskuratu ordez, informazioa gordetzen duten web orrialdeetan hura bilatzen dugu, denbora galduz. Izan ere, web orrialdeetako testuak irakurri eta beharrezko informazioa erauzten dugu, informaziora zuzenean jo beharrean. Datu estekatuen teknologiak (*Linked Data*) konponbidea eskaintzen digu [1]: datu estekatuak interneten zuzenean argitara ditzakegu, web orrialdeak eraiki beharrik gabe, beste erabiltzaile batzuk datuok nahi bezala erabil ditzaten.

Datu estekatuen funtsa RDF (*Resource Description Framework*) lengoaia da. RDFk informazioa egitura jakin batean adierazten du: RDF triplea (1 irudia). Hiru elementuk osatzen dute RDF triplea: subjektua, predikatua eta objektua. Predikatuak subjektua eta objektua lotzen ditu. Adibidez, *Bilbok parte du Guggenheim Bilbao museoa* triplean *Bilbo* subjektua *parte du* predikatuaren bidez *Guggenheim Bilbao museoa* objektuarekin erlazionatzen da. Beraz, predikatuak bi entitateen arteko erlazioak dira: entitate batek (Bilbok), propietate jakin batean (Zein parte duen) beste entitate bat du (Guggenheim Bilbao museoa). Horrelako tripleak lotuz informazio konplexua adierazten duten RDF sareak eratzen dira (2 irudia).




1 irudia: RDF triple baten adibidea (Subjektua: *Bilbo*; Predikatua: *partea du*; objektua: *Guggenheim Bilbao museoa*). Triple egituran edozein motako informazioa adieraz daiteke.

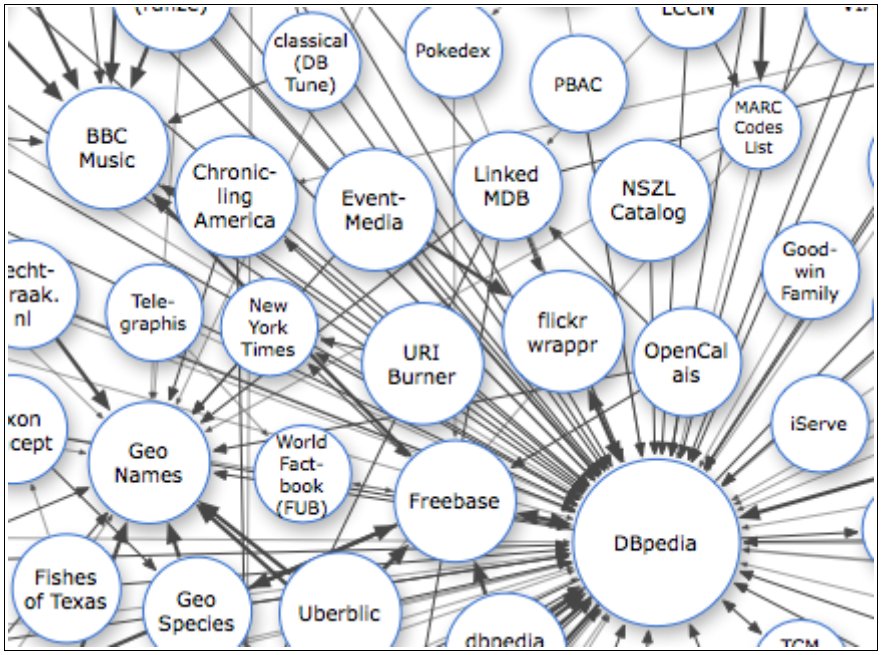



2 irudia: RDF tripleak lotuz eratzen den RDF sarea (Triple baten subjektua beste triple baten objektua izan daiteke: *Bilbo*, adibidez). RDF sarea informazio konplexua adierazteko erabiltzen da.

RDF erabiliz edozein informazio interneten modu egituratuan (Triple egiturari jarraituz) plazaratu dezakegu. Horrek hainbat informazio iturri elkartzea ahalbidetzen du: RDF sare baten konektaturiko elementuak (Subjektuak, predikatuak edo objektuak) interneteko hainbat baliabidetik etor daitezke.

Gainera, predikatuak semantikoak dira. Hau da, *izena* erlazioak subjektuaren izenera eramango gaitu, *kokapena* erlazioa bere kokapenera, *parte du* erlazioak bere parteetara eta abar. Beraz, zerbaiten kokapena jakin nahi badugu, *kokapena* predikatua jarraitu besterik ez dugu. Ondorioz, informazioa erabiltzen duena (Gizakia edo ordenagailua) informazioa "ulertzeko" gai da, esteka semantikoak egitura jakin batean (Triplean) adierazten baitira.

Horretan datza datu estekatuen indarra: edozein erabiltzailek edo ordenagailuk datuak kudea ditzake, erlazio semantikoen bidez entitate batetik bestera "salto eginez", internet osoa datu-base erraldoia izango balitz bezala (3 irudia). Orobat, SPARQL lengoaia erabiliz, datu-base erraldoi horri edozein galdera egin diezaiokegu (5 irudia). Datu-base erraldoia datuen amarauna da, hainbat RDF sarez osatua. Datuen amaraunak gaur egun web sarea erabiltzean pairatzen dugun informazio uholdea kudeatzen laguntzen digu.




3 irudia: Datuen amarauna. Bertan Wikipediaren (Dbpedia) eta BBCren (BBC Music) RDF bertsioak daude, beste hainbat RDF baliabiderekin batera. Wikipediako entitate batetik (Led Zeppelin taldea, adibidez) BBCko entitate batzuetara (Led Zeppelinen abestiak) esteka semantikoak daude. Hau da, esaterako, triple batek *Led Zeppelin* subjektua (Wikipedia), *Abestia sortu zuen* predikatua (Wikipedia) eta *Immigrant song* objektua (BBC Music) izango ditu. Gezi bakoitzak hainbat triple adierazten du. [http://richard.cyganiak.de/2007/10/lod/]

# 2.- DATUEN AMARAUNAREN ERABILERAK

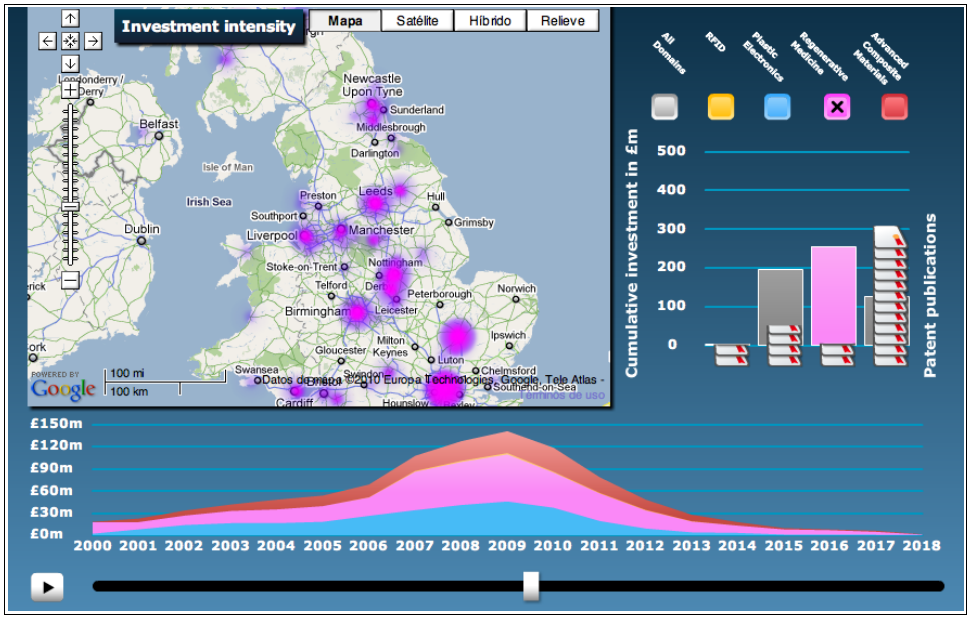
Datu estekatuetan dagoen informazioa erabiltzeko software programen ekosistema aberatsa sortu da, batez ere bi arrazoiengatik. Batetik, ohizko web sarearekin alderatuz datu estekatuen bidez informazio zehatzagoa arinago lor dezakegu. Bestetik, informazio horrekin lan egingo duen softwarea aise garatzea dago, informazioa RDF bidez egituratua egongo denez, software garatzaileak aldez aurretik jakingo duelako nolakoa den informazio horren egitura (Hainbat RDF triplez osatutako RDF sarea).

Datuen amaraunari datu berriak eranstea erraza da: Triple egitura jarraitzea besterik ez dugu, eta gure tripleak predikatu bidez beste norbaiten tripleekin estekatu. Ondorioz, datuen amarauna egunero hedatzen da. Erabiltzaileek datu berri horiekiko elkarrekintza interesgarriak izan dezaten software programak garatzen dira egunero ere. Datu gehiago ez ezik, datu mota berriek software programa are interesgarriagoak garatzea eskatuko dute etorkizunean.

## 2.1.- GOBERNUEN DATU IREKIAK

Hiritarrok gobernuak gordetzen dituen datuekin lan egiteko eskubidea dugu. Hala ere, orain arte datu horiek datu-base itxietan gorde izan dira, hiritarron eskutik at. Gaur egun, RDFk datuok argitaratzeko teknologia paregabea eskaintzen du: datuak beste edozein daturekin uztartzeko moduan argitara daitezke, eta hiritarrok, datuak aztertu ez ezik, software programak eraiki ditzakegu datuokin lan egiteko. Ondorioz, gobernuek gero eta datu publiko gehiago argitaratzen dituzte datu estekatuen ereduari jarraiki, hiritarrekiko gardentasuna bermatzeko.

Erresuma Batuko gobernua[1] aitzindaria izan da datu estekatuak eskaintzeko prozesuan, eta ekimen horren inguruan software programa interesgarriak sortu dira. Adibidez, Talis enpresak ikerketa guneen mapa elkarreragilea garatu du[2]: mapa horretan Erresuma Batuko hiritarrek goi-mailako ikerketa guneak aurki ditzakete, lortutako diru eta patenteen arabera, kronologikoki ordenatuak (4 irudia).



4 irudia: Talis enpresak ikerketa gune garrantzitsuenak aurkitzeko sortutako mapa. Emaitzak ikerketa arloaren arabera sailkatzeko aukera dago. [http://bis.clients.talis.com/]

1 http://data.gov.uk
2 http://bis.clients.talis.com/

## 2.2.- SENDAGAIAK AURKITZEKO BIDE BERRIAK

Biologian informazio asko argitaratzen da interneten zehar barreiaturiko baliabideetan. Horrek esan nahi du ikerlari batek informazioa jaso eta bateratu behar duela galdera konkretuak ebazteko. Datu estekatuekin, aldiz, ikerlariak behar duen informazioa sistema bakarretik jaso dezake. Horrelako sistemetako bat Neurocommons da [2]: RDF bidez Alzheimer gaixotasunari buruzko informazio bateratua eskaintzen du. Ikerlariek Neurocommons sistema farmako berrientzako jomuga posibleak aurkitzeko erabiltzen dute. Adibidez, CA1 piramide neuronak oso kaltetuak daude Alhzeimer gaixotasuna pairatzen duten pertsonetan, eta ikerlariek badakite neurona horietan seinale-transdukzioa garrantzitsua dela. Beraz, seinale-transdukzioa burutzen duten eta piramide neuronetan dauden proteinak jakinez gero, proteina horiek farmako berrientzako jomuga posibleak izango dira.

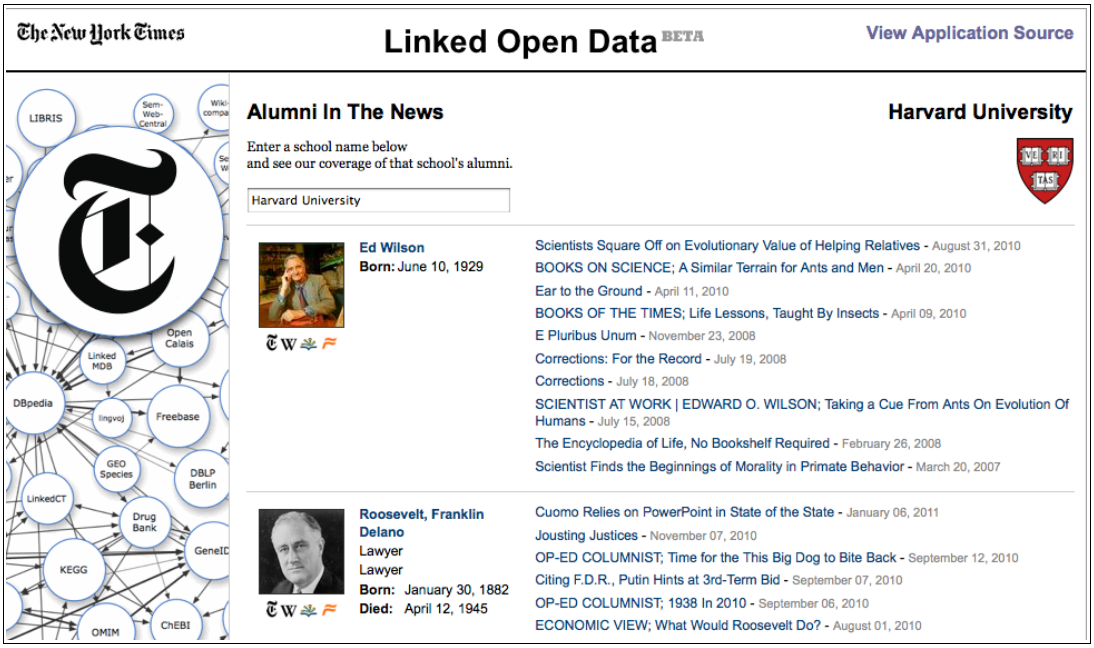
Galdera hori Googlen eginez gero (*Signal transduction in pyramidal neurons*) esanahi gabeko erantzunez jositako zerrenda handiegia lortuko dute ikerlariek. Aitzitik, Neurocommons sistema erabiliz (5 irudia), osagaientzako jopuntu gutxi eta baliotsuak lortzen dituzte, Alzheimer gaixotasuna sendatzeko balizko tratamendu berriak, alegia.

```
1.   prefix go: <http://purl.org/obo/owl/GO#>
2.   prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3.   prefix owl: <http://www.w3.org/2002/07/owl#>
4.   prefix mesh: <http://purl.org/commons/record/mesh/>
5.   prefix sc: <http://purl.org/science/owl/sciencecommons/>
6.   prefix ro: <http://www.obofoundry.org/ro/ro.owl#>
7.   SELECT ?gene_name ?process_name
8.   WHERE
9.   { ?pubmed_record ?related mesh:D017966 .
10.    ?article sc:identified_by_pmid ?pubmed_record.
11.    ?gene_record sc:describes_gene_or_gene_product_mentioned_by ?article.
12.    ?protein rdfs:subClassOf ?restriction.
13.    ?restriction owl:onProperty ro:has_function.
14.    ?restriction owl:someValuesFrom ?restriction2.
15.    ?restriction2 owl:onProperty ro:realized as.
16.    ?restriction2 owl:someValuesFrom ?process.
17.    { { ?process ro:part_of go:GO_0007165 }
18.        UNION
19.      { ?process rdfs:subClassOf go:GO_0007165 }}
20.    ?protein rdfs:subClassOf ?parent.
21.    ?parent owl:equivalentClass ?restriction3.
22.    ?restriction3 owl:onProperty sc:is_protein_gene_product_of_dna_described_by.
23.    ?restriction3 owl:hasValue ?gene_record.
24.    ?gene_record rdfs:label ?gene_name.
25.    ?process rdfs:label ?process_name.
26.  }
```

5 irudia: Neurocommons sistemari SPARQL erabiliz egindako galdera. Erantzunak hainbat jatorriko informazioa elkarlotzen du, eta beharrezko informazio zehatza soilik ematen du, gehiegizko informazioa baztertuz. [http://bib.oxfordjournals.org/content/10/2/193/F6.expansion.html]

## 2.3.- HEDABIDEAK HEDATU

NYTk[3] (New York Times) eta BBCk[4], sona handiko hedabideek, datu estekatuak bi helburuekin erabiltzen dituzte. Alde batetik, irakurleei informazioa igorri aurretiko datuen barne kudeaketa hobetu dute[5]. Bestetik, haien datuak ireki dituzte[6], irakurleek datuok erraz erabili ditzaten, beste datuekin uztartu ditzaten edo datuen inguruan aplikazioak eraiki ditzaten[7]. NYTk, adibidez, pertsona ospetsuak zein unibertsitateetara joan ziren aurkitzeko bilatzailea[8] garatu du (6 irudia). NYTk argitaratutako datu estekatuen inguruan oso aplikazio interesgarriak sortuko dira datozen urteetan, egunkari horrek gordetzen duen informazioaren kalitatea eta garrantzia kontuan hartuta.



6 irudia: NYTk eskaintzen duen zerbitzua, pertsona famatuen jatorrizko unibertsitateak aurkitzeko. Bilaketaren emaitzetan pertsona bakoitza aipatzen duten albisteak ere agertzen dira.
[http://data.nytimes.com/schools/schools.html]

3 http://data.nytimes.com/
4 http://www.bbc.co.uk/
5 http://www.bbc.co.uk/blogs/bbcinternet/2010/07/the_world_cup_and_a_call_to_ac.html
6 http://www.bbc.co.uk/nature/feedsanddata
7 http://open.blogs.nytimes.com/2010/03/30/build-your-own-nyt-linked-data-application/
8 http://data.nytimes.com/schools/schools.html

# 3.- DATUEN AMARAUNAREN ETORKIZUNA: WEB SEMANTIKOA

Datuen amarauna beste proiektu orokorrago baten lehen pausua da: web semantikoa[9]. Web semantikoaren muina ordenagailuak informazioa "ulertzeko" gai izatea da, erabiltzaileak web sareko informazioaren kudeaketan denbora aurrez dezaten.

Datuen amarauna web semantikoa eraikitzeko oinarri paregabea da, datuen amaraunean datuak RDF lengoaian adierazten baitira. Izan ere, RDF lengoaia semantikoa da (Datuen esanahia deskribatzen du), eta web semantikorantz lehenengo pausua ahalik eta informazio gehien ordenagailuek kudea dezaketen lengoaia semantikoan adieraztea da. Hurrengo pausua datu horien gainean "ezagutza" sortzea da, ordenagailuek datu horiekin logikoki arrazonatzeko gai izan daitezen, ordenagailuek informazioa ulertuko balute bezala erabiltzaileontzat lan egin dezaten.

Baliteke web semantikoa inoiz ez ezartzea, baina web semantikora bidean oso teknologia erabilgarriak sortuko ditugu. Datuen amarauna adibide deigarria da, artikulu honetan aurkeztutako aplikazioek erakusten dutenez.

# ERREFERENTZIAK

[1] Bizer C, Heath T, Berners-Lee T (2009). Linked Data - The Story So Far. *IJSWIS* 5(3), 1-22

[2] Ruttenberg A, Rees J, Samwald M, Marshall M (2009). Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief. Bioinformatics* 10, 193-204

9 http://www.w3.org/standards/semanticweb/