

Linked Data – La Web Semàntica pràctica

Com s'ha comentat en [posts anteriors](http://griho.blogs.udl.cat/2008/05/05/34/) [<http://griho.blogs.udl.cat/2008/05/05/34/>] sobre la Web Semàntica, es tracta d'una [iniciativa](http://griho.blogs.udl.cat/2008/05/05/34/) [<http://griho.blogs.udl.cat/2008/05/05/34/>] bastant ambiciosa i que degut a la seva complexitat (metadades semàntiques, ontologies, regles lògiques, motors d'inferència, enginyeria del coneixement, etc.) està sent desplegada a una ritme inferior al que [molts esperaven](http://wjbsemweb.blogspot.com/2006/05/semantic-web-revisited.html) [<http://wjbsemweb.blogspot.com/2006/05/semantic-web-revisited.html>], sobretot [comparant-la amb la Web 1.0 o la 2.0](http://griho.blogs.udl.cat/2008/05/26/conclusions-de-la-sessio-web-30-web-semantica-al-bdgtal-global-congress/) [<http://griho.blogs.udl.cat/2008/05/26/conclusions-de-la-sessio-web-30-web-semantica-al-bdgtal-global-congress/>].

És per això que recentment s'ha posat en marxa la [iniciativa Linked Data](http://linkeddata.org/) [<http://linkeddata.org/>], que vindria a ser una versió “Lite” o simplificada de la Web Semàntica. Aquesta és menys ambiciosa, l'objectiu no és desplegar dades amb una gran riquesa semàntica (basades en ontologies complexes i sofisticats mecanismes de raonament), sinó assolir una massa crítica mitjançant un gran volum de dades menys sofisticades.

Per tant, l'objectiu és aconseguir que un gran volum de dades amb semàntica molt senzilla estiguin vagin omplint la Web Semàntica. A més a més, és fonamental que els diferents conjunts de dades que es publiquin estiguin molt interconnectats entre ells i no siguin illes, ja que això [multiplica el valor de les dades](http://griho.blogs.udl.cat/2008/05/05/34/) [<http://griho.blogs.udl.cat/2008/05/05/34/>].

Per tal de facilitar que les dades que es publiquin estiguin fàcilment a l'abast i ben interconnectades, [Tim Berners-Lee](http://en.wikipedia.org/wiki/Tim_Berners-Lee) [http://en.wikipedia.org/wiki/Tim_Berners-Lee], un dels principals impulsors, [va definir quatre principis bàsics](http://www.w3.org/DesignIssues/LinkedData.html) [<http://www.w3.org/DesignIssues/LinkedData.html>]:

- Fer servir [URI](http://en.wikipedia.org/wiki/Uniform_Resource_Identifier) [http://en.wikipedia.org/wiki/Uniform_Resource_Identifier]s per identificar les “coses” que s'exposin a la Web com a [recursos](http://en.wikipedia.org/wiki/Resource_%28Web%29) [http://en.wikipedia.org/wiki/Resource_%28Web%29].
- Fer servir URIs [HTTP](http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol) [http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol] per tal de facilitar que la gent pugui localitzar i recuperar ([desreferenciar](http://en.wikipedia.org/wiki/Dereferenceable_Uniform_Resource_Identifier) [http://en.wikipedia.org/wiki/Dereferenceable_Uniform_Resource_Identifier]) aquests recursos.
- Proporcionar informació útil sobre el recurs quan la seva URI es desreferencii.
- Incloure enllaços a altres URIs [Linked Data](http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol) relacionades per tal de facilitar el descobriment de dades a la Web.

Si a aquests principis bàsics s'afegeix el de que les dades siguin obertes, tenim [Linked Open Data](http://en.wikipedia.org/wiki/Linked_Open_Data) i la materialització d'aquesta iniciativa en quelcom tangible, un gran volum de dades obertes basades en tecnologies de la Web Semàntica (fonamentalment les més bàsiques) i que són el resultat de traduir conjunts de dades existents a aquests formats, seguint el principis [Linked Data](http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol).

En aquest moment la mida del graf de [Linked Open Data](http://en.wikipedia.org/wiki/Linked_Open_Data) és aproximadament de [4.500 milions](http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics) [<http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics>] d'arestes (tripletes en l'argot de la Web Semàntica). La Figura 1 mostra quins conjunts de dades formen part actualment d'aquest graf i com estan interconnectats entre ells aquests conjunts.

[

http://griho.blogs.udl.cat/files/2009/03/lod-datasets_2009-03-05.png]

Fig 1. Conjunts de dades integrats a la Web de les dades seguint els principis de Linked Open Data

Alguns d'aquests conjunts de dades són:

- **DBpedia** [<http://en.wikipedia.org/wiki/DBpedia>]: conté informació extreta de **Wikipedia** [<http://en.wikipedia.org/wiki/Wikipedia>]; aproximadament 2,18 milions de termes descrits per 218 milions de tripletes, incloent resums en 11 idiomes diferents.
- **DBLP Bibliography** [<http://www4.wiwiw.fu-berlin.de/dblp/>]: proporciona informació bibliogràfica sobre articles científics i és generada a partir de la base de dades de DBLP; uns 800.000 articles, 400.000 autors i aproximadament 15 milions de tripletes.
- **GeoNames** [<http://en.wikipedia.org/wiki/GeoNames>]: proporciona descripcions de més de 6.500.000 punts geogràfics de tot el món, per exemple [aquest](http://www.geonames.org/6619295/escola-superior-politecnica.html) [<http://www.geonames.org/6619295/escola-superior-politecnica.html>]
- **UMBEL** [<http://www.umbel.org/>]: una estructura lleugera de referència de uns 20.000 temes i les seves interrelacions derivat d'**OpenCyc** [<http://en.wikipedia.org/wiki/OpenCyc>].

GRIHO [<http://griho.udl.cat/>] també vol contribuir a aquesta iniciativa de dades semàntiques obertes. Ens hem centrat en les dades financeres que ja hi ha disponibles públicament en format **XBRL** [<http://es.wikipedia.org/wiki/XBRL>], format que es fa servir fonamentalment pels informes financers que les empreses estan obligades a enviar a les entitats reguladores de cada país. Per exemple, a la **SEC** [http://en.wikipedia.org/wiki/U.S._Securities_and_Exchange_Commission] dels Estats Units en el marc del programa **EDGAR** [<http://www.sec.gov/edgar.shtml>].

XBRL, tot i ser un format estàndard basat en XML, no facilita la integració de dades de diferents empreses, i fins i tot dificulta la integració de dades de la mateixa empresa però seguint versions lleugerament diferents dels esquemes definits per XBRL. Per tant, les tecnologies de la Web Semàntica i els principis de Linked Data semblen una bona aproximació per fer aquesta informació més útil i transparent, i potser fins i tot fer més difícils escàndols financers com els que hem estat veient recentment.

Aquest esforç per part de GRIHO per tal de transformar dades XBRL en Linked Open Data està començant a tenir els seus primers fruits amb **SemanticXBRL** [<http://rhizomik.net/semanticxbrl/>]. Per ara s'ha transformat més de 500 informes enviats per les principals empreses de la borsa dels Estats Units com a XBRL, donant com a resultat un graf de dades semàntiques amb més d'un milió de tripletes. Aquest resultat es presentarà el proper 20 d'Abril en el marc del segon taller sobre Linked Open Data (**LODW 2009** [<http://events.linkedata.org/ldow2009/#programme>]) que és celebra durant el congrés internacional de la Web a Madrid, **WWW 2009** [<http://www2009.org/>].

Roberto García [<http://rhizomik.net/~roberto>]

