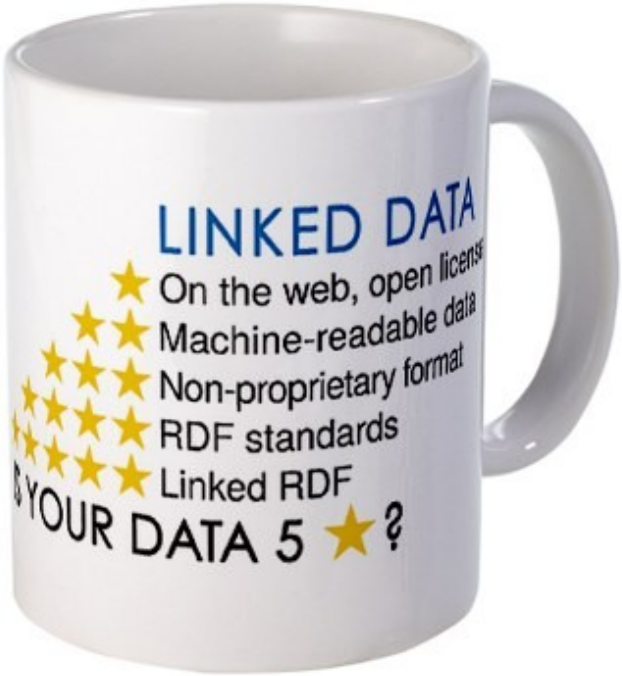
Linked Data

Tim Berners-Lee
Date: 2006-07-27, last change: $Date: 2009/06/18 18:24:33 $
Status: personal view only. Editing status: imperfect but published.

Traducere în limba română de Nicolaie Constantinescu,
Februarie-Martie 2011
www.kosson.ro
kosson@gmail.com

Web-ul semantic nu este vorba doar de a pune datele de pe web. Este despre a face legături, astfel încât o persoană sau o mașină să poată explora web-ul de date. Cu date legate, atunci când ai câteva, poți găsi alte date legate.



Ca și web-ul hypertext, web-ul datelor este construit cu documente pe web. Cu toate acestea, spre deosebire de web-ul hipertext, unde link-urile sunt ancore relaționale în documente hypertext scrise în HTML, în cazul datelor legăturile între lucruri arbitrare descrise de RDF, URI-urile identifică orice fel de obiect sau concept. Dar pentru HTML sau RDF sunt aceleași așteptări pentru ca web-ul să crească:

1. Folosiți URI-urile ca nume pentru lucruri
2. Folosiți URI-uri HTTP astfel ca oamenii să le priceapă
3. Atunci când cineva se uită la un URI, oferiți informații utile folosind standardele (RDF, SPARQL)
4. Includeți linkuri către alte URI-uri astfel ca acestea să poată descoperi mai multe lucruri.

Simplu. Realitatea este că o cantitate surprinzătoare de date nu este conectată în 2006 (momentul scrierii articolului n.n.), din cauza unuia sau a mai multora dintre pași. Acest articol discută soluțiile la aceste probleme, detaliile de implementare și factorii care afectează opțiunile privind cum să publici datele.

Cele patru reguli

Mă voi referi la pașii de mai sus ca la niște reguli, dar acestea sunt de fapt așteptări privind comportamentul. Nerespectarea lor nu distruge nimic, dar conduce la pierderea oportunității de a face datele să fie interconectate. Astfel, acest lucru limitează modalitățile în care acestea pot fi reutilizate în moduri neașteptate. Reutilizarea în moduri neașteptate este de fapt valoarea adăugată de rețea.

Prima regulă de a identifica lucrurile cu URI-uri este înțeleasă în mare parte de toți cei care lucrează în

domeniul tehnologiei web-ului semantic. Dacă nu folosește setul de simboluri universale de URI-uri, nu se poate numi Web Semantic.

Cea de-a doua regulă de a utiliza URI-uri HTTP este înțeleasă și ea în mare parte. De la debutul web-ului, singura deviație a fost tendința oamenilor de a inventa noi scheme URI (și sub-scheme în cadrul urn:) cum ar fi LSIDs și handle-uri și XRI-uri și DOI-uri și tot așa din diferite motive. De regulă, acestea implică împotrivirea sistemului Domain Name System (DNS) existent în ceea ce privește delegarea de autoritate și de a controla ceva sub un control separat. De multe ori acest lucru se leagă de faptul că nu se înțelege că URI-urile HTTP sunt nume (nu adrese) și că mecanismul de căutare al acestora prin HTTP constituie un set complex și puternic de seturi de standarde. Această problemă este discutată mai pe larg în altă parte și nu ne permite să ne ocupăm de ea aici. [@@ref TAG finding, etc])

Cea de-a treia regulă care spune că trebuie oferite informații suplimentare la un URI este, în 2006, urmată corect de cele mai multe dintre ontologii, dar, din anumite motive nu pentru unele seturi de date mari. În generale, cineva ar putea să se uite la proprietățile și clasele găsite în date și să obțină informații de la ontologiile RDF, RDFS și OWL incluzând relațiile dintre termenii ontologiei.

Aici, formatul de bază pentru RDF/XML împreună cu populara sa alternativă, este serializarea N3 (sau Turtle). Seturi mari de date oferă un serviciu de interogare prin SPARQL, dar datele de bază legate ar trebui să fie oferite de asemenea.

Multe proiecte de cercetare sau de evaluare desfășurate în cei câțiva ani ai tehnologiilor Web-ului Semantic au produs ontologii și acumulări de date semnificative, dar, datele, dacă sunt și disponibile, sunt îngropate într-o arhivă zip pe undeva decât să fie accesibile pe web ca date conectate. Proiectul Biopax, datele privind cercetătorii din domeniul științei computerelor și a proiectelor CSAktive sunt numai două exemple. [În acest moment, 2007, datele CSAktive sunt disponibile ca date conectate]

Mai există și un volum în creștere de URI-uri aparținând datelor care nu au ontologii, dar care pot fi descoperite. Wikiurile semantice reprezintă unul din exemple. Ontologiile „Friend of a Friend" (FOAF) și *Description of a Project* (DOAP) sunt folosite pentru a construi rețele sociale pe internet. Portalurile pentru rețele sociale des întâlnite nu oferă linkuri către alte site-uri și nici nu-și expun datele într-o formă standard.

LiveJournal și Opera Community sunt două portaluri care publică datele pe web folosindu-se de RDF. (Plaxo are o schemă de urmărire și nu sunt sigur că suportă linkuri de tip *cunoaște*). Acest lucru înseamnă că eu pot scrie în fișierul meu FOAF faptul că-l cunosc pe Håkon Lie folosind URI-ul său luat din datele existente în Opera Community iar o mașină sau o persoană care parcurge acele date poate urmări apoi acel link și poate găsi toți prietenii săi. Ei bine, chiar toți prietenii? Nu chiar: doar prietenii care sunt în comunitatea Opera Community. Sistemul încă nu permite stocarea URI-urile persoanelor din sisteme diferite. Astfel că rețeaua permite interogarea venită de pe linkuri externe, fiind navigabilă la nivel intern, totuși nu permite linkuri adresate exteriorului.

Cea de-a patra regulă, de a face linkuri peste tot este necesară pentru a conecta datele pe care le avem pe un segment de web neconectat în care cineva poate găsi diferite tipuri de lucruri exact ca și pe web-ul de hypertext pe care l-am construit.

Pe web-ul site-urile hypertext, în general, sunt considerate ca având o reputație proastă dacă nu se leagă la materiale externe. Valoarea propriei informații este considerată mai mult o funcție a ceea ce se leagă la ea, după cum este și valoarea implicită pe care pagina web o prezintă. Astfel, și în acest caz acestea

se află în Web-ul Semantic.

Să ne uităm deci la modalitățile de a conecta datele, pornind de la cea mai simplă cale de a face un link.

Baza căutării pe web

Cea mai simplă cale de a conecta datele este de a utiliza într-un singur fișier un URI care trimite către conținutul altuia.

Atunci când scrii un fișier RDF, să zicem <http://example.org/smith>, poți folosi identificatori locali din interiorul fișierului precum `#albert`, `#brian` și `#carol`. Folosind N3 poți compune:

```
<#albert> fam:child <#brian>, <#carol>
```

sau utilizând RDF/XML

```
<rdf:Description about="#albert">
      <fam:child rdf:Resource="#brian">
      <fam:child rdf:Resource="#carol">
</rdf:Description>
```

În acest moment arhitectura WWW oferă un identificator global "http://example.org/smith#albert" pentru Albert. Acesta este un lucru bun să-l faci de vreme ce toată lumea de pe planetă știe cum să folosească un identificator pentru a se referi la Albert și pentru a oferi mai multe informații.

De exemplu, în documentul <http://example.org/jones> cineva ar putea scrie?

```
<#denise> fam:child<#edwin>, <smith#carol>
```

sau în RDF/XML

```
<rdf:Description about="#denise">
      <fam:child rdf:Resource="#edwin">
      <fam:child rdf:Resource="http://exmple.org/smith#carol">
</rdf:Description>
```

În mod clar este rezonabil pentru cei care întâlnesc identificatorul „http://example.org/smith#carol" să:

1. Formeze URI-ul documentului prin trunchierea acestuia înainte de hash
2. Să acceseze documentul pentru a obține informații despre `#carol`

Numim acest lucru dereferențierea URI-ului. Prin acest lucru înțelegem baza web-ului semantic. Există câteva variații.

Variație: URI-urile fără slash-uri și HTTP 303

Sunt câteva circumstanțe în care dividerea identificatorilor în documente separate nu lucrează prea bine. La nivel logic poate să existe un simbol global pe fiecare document dar există o reținere la a include un # în URI precum în cazul
http://wordnet.example.net/antiestablishmentarianism#word
Din punct de vedere istoric primele forme ale vocabularelor Dublin Core și FOAF nu au avut # în URI-urile lor. În orice caz atunci când URI-urile HTTP fără hash-uri sunt utilizate pentru concepte abstracte și când există documente purtătoare de informații despre ele, atunci:

1. O cerere HTTP GET asupra unui URI-ului conceptului, returnează un 303 See Also și pune în headerul Location: URI-ul documentului.
2. Documentul este adus complet

Această metodă are avantajul că URI-urile pot avea diferite forme. Are dezavantajul că o cerere HTTP

mBrowse-ableust poate fi operată pentru fiecare dintre ele. În cazul Dublin Core, de exemplu, dc:title și dc:creator sunt, de fapt, oferite de același document ontologic, dar cineva nu ar putea ști acest lucru până când acestea nu au fost aduse unul câte unu și au for redirectate prin HTTP.

Variație: FOAF și rdfs:seeAlso

Convenția Friend-of-a-Friend folosește un fel de link pentru date, dar nu folosește niciuna dintre cele două forme menționate mai sus. Pentru a te referi la o altă persoană într-un fișier FOAF, convenția a fost de a da două proprietăți, una care trimite la documentul în care sunt descrise iar celaltă pentru identificarea acestora în interiorul documentului.

```
<#i>  foaf:knows  [
       foaf:mbox <mailto:joe@example.com>;
                  rdfs:seeAlso <http://example.com/foaf/joe> ].
```

Se citește „Știu cine are emailul joe@example.com și câtă informație mai e în <http://example.com/foafjoe>”.

De fapt, din motive de confidențialitate, de cele mai multe ori, oamenii nu-și pun pe net direct adresele de email, ci oferă un cod hash unidirecțional (SHA-1) a adresei de email. Acest truc inteligent permite oamenilor care-și cunosc deja adresa de email să se identifice fără a da adresa de email altora.

```
<#i>  foaf:knows  [
       foaf:mbox_sha1sum "2738167846123764823647";  # @@ dummy
  rdfs:seeAslo <http://example.com/foaf/joe> ].
```

Acest sistem de linking a fost un succes care a condus la formarea rețelei sociale în expansiune și care domina, în 2006, datele conectate disponibile pe web.

Totuși, sistemul face figura că nu oferă URI-uri oamenilor și astfel conexiuni normale către aceștia nu se pot face.
Recomand (de ex. în cazul blogurilor să luați în considerare Linkuri în Webul Semantic, Ia-ți un URI și Linkurile înainte și înapoi sunt deopotrivă importante în RDF) ca cei care fac un fișier FOAF să-și ia un URI personal și să urmeze o convenție FOAF. În mod similar, atunci când vă referiți la un fișier FOAF care oferă un URI către o persoană, folosiți-l atunci când vă referiți la acea persoană astfel ca cei care vor folosi acel URI și nu cunosc convenția FOAF, să poată urma linkul.

Grafuri navigabile

În acest moment am privit la modurile în care se pot face linkuri, dar să ne uităm la alternativele pe care le avem atunci cînd facem un link.

Un tipar important este un set de date care pot fi explorate parcurgându-le link cu link aducând date. Ori de câte ori cineva caută după un URI pentru un nod dintr-un graf RDF, serverul returnează informații despre arcurile de ieșire și arcurile intrărilor acelui nod. Cu alte cuvinte, returnează orice declarații RDF în care apare termenul, fie ca subiect, fie ca obiect.

În mod formal, numiți un graf G ca fiind *navigabil* dacă pentru fiecare URI a oricărui nod din G, dacă facem o căutare după un URI va fi returnată informația care descrie nodul, unde descrierea unui nod înseamnă:

1. returnarea tuturor declarațiilor acolo unde nodul este un subiect sau un obiect și

2. descrierea tuturor nodurilor goale atașate nodului printr-un singur arc.

(Sub-graful returnat a fost referit ca fiind „minimum Spanning Graph (MSG [@@ref])" sau moleculă RDF [@@ref], asta dacă nodurile sunt considerate a fi identificate și dacă poate fi exprimată ca o cale a funcției sau proprietăți funcționale inverse. O descriere concisă care doar urmează linkurile de la subiect la obiect, nu poate exista.)

În practică, atunci când datele sunt stocate în două documente acest lucru înseamnă că oricare declarație RDF care pune în legătură lucrurile existente în cele două fișiere trebuie să se repete în cele două. Astfel, de exemplu, în pagina mea FOAF menționez că sunt un membru al grupului DIG iar acea informație este repetată și în grupul datelor DIG. Astfel, pornind de la conceptul de grup poate afla că și eu fac parte ca membru. De fapt, cineva care pornește de la URI-ul meu poate găsi toți oamenii care sunt în același grup.

Limitări privind datele navigabile

Deci, declarațiile care leagă datele din cele două documente trebuie să se repete în amândouă. Acest lucru este contrar primei reguli a stocării datelor: nu stoca aceleași date în două locuri diferite. Apar probleme privind consistența. Acest lucru prezintă reale probleme în ceea ce privește datele navigabile. Setul de date navigabile complete cu linkuri în ambele direcții trebuie să fie consistente mai ales dacă autori diferiți sau programe diferite sunt implicate.

Putem acea date complete navigabile acolo unde sunt generate automat.

De exemplu, serverul dbview oferă documente virtuale navigabile care conțin date din oricare baze de date arbitrare.

Atunci când avem date din surse multiple atunci suntem supuși compromisurilor. De regulă, acestea sunt soluționate în mod normal punând întrebarea:

„Dacă cineva are URI-ul acelui lucru, atunci de care relații care se stabilesc la care alte obiecte ar trebui să avem habar?"

Uneori problemele de natură socială determină răspunsul. Avem relații în fișierul FOAF care indică faptul că am contacte cu diferiți oameni. În general, ceilalți nu repetă aceste informații în propriile lor fișiere FOAF. Cineva ar spune că mă cunoaște, care este o afirmație care, prin prisma convenției FOAF, ține de acesta să o facă iar cei care citesc sunt puți în fața deciziei de a crede sau nu.

Uneori numărul arcurilor face acest lucru impracticabil. Punctele unui traseu GPS poate oferi de mii de ori identic la ce latitudine și longitudine sunt. Fiecare persoană care încarcă fișierul meu FOAF se așteaptă să obțină informațiile din cartea mea de vizită dar nu toate acele puncte de traseu. Este normal să avem un reper din traseu (sau chiar fiecare dintre puncte) care să trimită la persoana a cărei locație este reprezentată și nu invers.

Un model este să ai linkuri pentru o anumită proprietate în câte un document separat. Pagina inițială a unei persoane nu afișează toate publicațiile acestuia, ci în loc pune un link către acestea într-un document separat care le listează. Există o convenție prin care foaf:made returnează o anumită lucrare, dar foaf:pubs trimite către un document care returnează o listă de lucrări. Astfel, cineva care navighează după linkul foaf:made ar face bine să urmeze linkul foaf:pubs. Ar fi util să se formalizeze o declarație ca și:

```
foaf:made  link:listDocumentProperty foaf:pubs
```

într-o antologie.

Servicii de interogare

Uneori volumul mare de date face posibilă oferirea a cât mai multe fișiere posibile, dar este îngreunată interogarea eficientă a setului de date de la distanță. În acest caz este mai rezonabil să fie pus la dispoziție un serviciu de interogare SPARQL. Pentru a face conexiuni eficiente între date, cineva care are doar URI-ul unui lucru trebuie să fie capabil să-și găsească calea prin SPARQL.

Aici poate fi utilizat răspunsul HTTP 303 pentru a îndrepta interogatorul la un document cu metadate privind care punct de interogare poate fi oferit, ce informații pot fi oferite și despre care clase ale URI-urilor.

Vocabularele care permit acest lucru nu au fost încă standardizate.

Sunt datele tale de 5 stele?

(Adăugat 2010). În acest an pentru a încuraja oamenii &emdash; mai ales deținătorii datelor guvernamentale &emdash; să pășească pe calea datelor bine conectate, am elaborat acest sistem de evaluare prin steluțe.

★	Sunt disponibile pe web (indiferent de format), dar cu o licență deschisă
★★	Disponibile ca date structurate pentru mașină (ex. excel în loc de imaginea scanată a unui tabel)
★★★	La fel ca la (2) plus formate libere (ex. CSV în loc de excel)
★★★★	Toate cele de mai sus plus utilizarea standardelor deschise de la W3C (RDF și SPARQL)
★★★★★	pentru identificarea lucrurilor astfel ca oamenii să poată face referințe la propriile lucruri
★★★★★★	Toate cele de mai sus plus: conectarea datelor la datele altor oameni pentru a oferi un context

Cât de bine stau datele tale? Poți achiziționa căni cu cinci stele, tricouri și stikere de la magazinul W3C din cafepress: folosiți-le pentru a-i face pe colegi și pe cei care merg la conferințe să se gândească la datele conectate. (Profiturile ajută și W3C :-).

Acum în 2010 în ceea ce privește datele guvernamentale, oamenii m-au împins să adaug o nouă cerință iar aceasta este ca să existe metadate pentru datele în sine iar ca metadatele să fie fie oferite printr-un catalog central. Orice date deschise (sau chiar seturi de date care nu sunt dar ar trebui să fie deschise) pot fi înregistrate la ckan.net. Seturile de date din Marea Britanie și Statele Unite ale Americii ar trebui să fie înregistrate la data.gov.uk și la data.gov. Mă aștept ca și alte țări să-și dezvolte propriile registre. Da, trebuie să existe metadate despre setul propriu de metadate. Acesta poate fi subiectul unei noi note din această serie. Permiteți-mi acum să concluzionez prin a spune aceasta.

Concluzie

Datele conectate sunt esențiale pentru a conecta efectiv web-ul semantic. Acest lucru este chiar ușor de a face acest lucru cu o mică analiză, devenind o a doua natură.

Clientul Tabulator (rulat într-un browser corespunzător) îți permite să navighezi prin datele conectate folosind convențiile de mai sus și poate fi folosit pentru a verifica dacă datele conectate funcționează.

Referințe

[Ding2005] Li Ding, et. al., *Tracking RDF Graph Provenance using RDF Molecules*, UMBC Tech Report TR-CS-05-06

Ca urmare

2006-02 Rob Crowell adapts Dan Connolly's DBView (2004) which maps SQL data into linked RDF, adding backlinks.

2006-09-05 Chris Bizer et al adapt D2R Server to provide a linked data view of a database.

2006-10-10 Chris Bizer et al produce the Semantic Web Client Library, "Technically, the library represents the Semantic Web as a single Jena RDF graph or Jena Model." The code fetches web documents as needed to answer queries.

2007-01-15 Yves Raimond has produced a Semantic Web client for SWI prolog with similar functionality.

Am o prelegere la conferința din 2009 O'Reilly eGovernment 2.0 în Washington DC, unde voi vorbi despre „Doar o punguță cu chipsuri” @@ref și voi vorbi despre schema cu cinci stele. Ca urmare, InkDroid a scris pe blog (plus CSS) de schema aceasta de 5 stele adoptată aici.