JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Linked data su larga scala: alcune sfide tecnologiche, ingegneristiche e sociali nell'ambito delle digital humanities, delle biblioteche digitali e dei beni culturali

Michele Barbera

Negli ultimi dieci anni la quantità di dati disponibili nella cosiddetta linked data cloud ha continuato a crescere molto rapidamente. Malgrado sia oggi disponibile un enorme ammontare di dati anche nei settori dei beni culturali e delle *digital humanities*, scarseggiano ancora gli esempi di riuso creativo di tali dati sia nell'ambito dell'industria culturale, sia in quello accademico. La carenza potrebbe non dipendere soltanto dagli attuali limiti tecnologici, ma anche da alcuni fattori sociali e culturali che, a parere di chi scrive, richiedono maggiore attenzione e comprensione. Per sfruttare appieno l'enorme potenziale offerto dai linked data sarà necessario affrontare la sfida di far maturare una nuova cultura, una cultura in grado di rivoluzionare i processi di produzione, gestione e pubblicazione dei dati. In settori come le *digital humanities*, ancora dominati da una forte tradizione proveniente dal mondo bidimensionale della carta stampata, affrontare queste sfide sarà più difficile e più urgente che in altri settori.

Traduzione italiana a cura dell'Autore.

La prima sezione introduce brevemente le sfide tecnologiche e ingegneristiche da affrontare più urgentemente nel mondo dei linked data. La sezione successiva si concentra sull'assunto che, per sfruttare appieno le potenzialità offerte dalle nuove tecnologie, è indispensabile comprendere meglio e assimilare alcuni principi fondamentali del semantic web e dei linked data. La parte finale accenna al potenziale beneficio che queste tecnologie potrebbero portare all'industria culturale e alla comunità scientifica. La generazione e il mantenimento di un efficiente ecosistema economico, capace di bilanciare il ruolo dei soggetti pubblici e delle organizzazioni private, è una delle precondizioni necessarie per realizzare il modello proposto dai linked data e dal web semantico. Sebbene non sia evidentemente possibile in questa sede analizzare a fondo tutti i fattori in gioco, questa introduzione mira a offrire alcuni utili spunti di approfondimento.

Sfide tecniche e ingegneristiche

Il web dei dati si distingue da altri sistemi di organizzazione e condivisione della conoscenza per la sua natura universale, per la dimensione virtualmente infinita e per l'eterogeneità dei dati in esso contenuti. Alcune di queste caratteristiche distintive sono alla base dei principi ispiratori dell'architettura del web e influenzano profondamente – o dovrebbero farlo – l'architettura delle applicazioni che producono e consumano i contenuti del web dei dati.
In primo luogo, a causa dell'enorme quantità di dati disponibili nella linked open data cloud e nei repository di aziende e organizzazioni, è impossibile pensare di centralizzare i dati e la computazione in un singolo punto. Si pensi che Sindice.com,[1] il più grande repository semantico di dati pubblici esistente, contiene oltre 80 miliardi di triple,

[1] http://sindice.com.

che rappresentano soltanto una piccola frazione di quelli disponibili nel web dei dati. Sindice.com si basa sull'utilizzo di un cluster di calcolo distribuitomap-reduce molto vasto (implementato utilizzando Apache Hadoop), i cui costi di gestione superano ampiamente le possibilità delle piccole e medie aziende e di molte istituzioni di ricerca.

Una delle caratteristiche peculiari dell'approccio linked data consiste nello sfruttare la natura eterogenea dei dati, in modo tale che le applicazioni siano in grado di inferire nuova conoscenza manipolando dati il cui modello non è noto a priori. Per non sacrificare questa importante caratteristica, è necessario trovare modi efficienti di tagliare on-demand i dataset più grandi in piccoli pezzi, che sono gestibili più facilmente. Questo approccio è noto da tempo nel settore dei database relazionali, dove prende talvolta il nome di dataspace. I dataspace dei linked data sono essenzialmente una materializzazione transitoria e calcolata on-demand di una porzione del grafo originario. Malgrado esistano già alcune pionieristiche implementazioni di linked dataspace, non è ancora stato definito alcuno standard né esiste alcuna implementazione di riferimento.

Streaming linked data

L'erogazione in streaming di linked data è un argomento che è stato ad oggi affrontato solo parzialmente dalla comunità scientifica[2] (Barbieri e Della Valle; Le-Phuoc et al.; Sequeda e Corcho) e non esiste alcuna implementazione commerciale di livello industriale. Benché possa inizialmente apparire irrilevante, questo è un tema di fondamentale importanza: si pensi, ad esempio, alla crescente disponibilità di dati prodotti in tempo reale da reti di sensori o da dispositivi di telefonia mobile.

[2]LarKC: The Large Knowledge Collider, http://www.larkc.eu.

Versioning

La capacità di gestire le diverse versioni di grafi RDF offre la possibilità di identificare cambiamenti atomici e di poterli annullare, riportando il grafo al suo stato originario. Sebbene esistano in letteratura diversi approcci a questo problema, non sono ad oggi ancora disponibili sul mercato tecnologie mature; in particolare, le soluzioni offerte dai graph store open source[3] sono ancora molto instabili e poco efficienti. Le sorgenti dati presenti nel web dei dati sono per definizione altamente eterogenee ed è quindi importantissimo essere in grado di separare, fondere e mescolare differenti sorgenti dati, ognuna con un proprio livello di autorevolezza (che è una misura soggettiva di qualità) e con una propria frequenza di aggiornamento. Di conseguenza, la mancanza di solide tecnologie per la gestione del *versioning* è un gap che deve essere colmato nel più breve tempo possibile. Sebbene tutte le sfide citate precedentemente siano certamente importanti, l'innovazione tecnologica resta una condizione necessaria ma non sufficiente per sfruttare appieno il potenziale del web dei dati. Come discuteremo nella prossima sezione, l'innovazione tecnologica ha bisogno di essere accompagnata da una comprensione critica del cambiamento culturale introdotto dal modello architetturale del semantic web.

Sfide sociali

In un famoso blog post del 2007, (Berners-Lee) ha utilizzato l'espressione "thinking in the graph" e introdotto il concetto di "giant global graph", in contrapposizione con il web dei documenti. Berners-Lee riassume alcuni dei cambiamenti più significativi che sono attual-

[3] La disponibilità di software open source di buona qualità può essere un elemento determinante nella crescita del settore.

mente in atto, ma che non sono stati del tutto compresi nel settore delle pubblicazioni elettroniche; scrive: "The less inviting side of sharing is losing some control. Indeed, at each layer — Net, Web, or Graph — we have ceded some control for greater benefits", and "It is about getting excited about connections, rather than nervous". Per comprendere la portata dei cambiamenti in atto, introdurremo brevemente tre principi fondanti del web dei dati: il cosiddetto "principio AAA", la natura del giant global graph e la Open World Assumption (OWA). Uno dei pilastri del web dei dati, parimenti valido per il web tradizionale, è il principio noto come AAA, un acronimo che sta per "Anyone can say Anything, Anywhere". Letteralmente: chiunque può dire qualsiasi cosa, in ogni luogo (digitale). Ancora una volta, questo principio implica un fondamentale cambiamento di paradigma nell'industria editoriale e dei media. Nell'era del web dei documenti e ancor prima dell'avvento del web, il modello dominante si basava su attori diversi che agivano come produttori, come gatekeeper, o come distributori d'informazione. Raramente un attore ricopriva contemporaneamente i tre ruoli. I consumatori d'informazione (i lettori) non erano coinvolti né nella produzione né nella disseminazione di informazione. Con l'avvento del cosiddetto web 2.0, i lettori sono divenuti soggetti attivi nella produzione di informazione, tuttavia il ruolo di gatekeeper (filtro e selezione dell'informazione) è rimasto in mano agli editori o, al più, agli aggregatori di contenuti. Il web dei dati – un termine che qui utilizzo impropriamente come sinonimo del semantic web e dei linked data – e il principio AAA propongono nuovi modelli di produzione, consumo e disseminazione dell'informazione. Nel nuovo paradigma l'elemento determinante nel ciclo di vita dell'informazione si sposta dalla produzione alla ricombinazione di molteplici sorgenti informative personalizzate. In questo nuovo scenario individui e organizzazioni ricoprono allo stesso tempo il ruolo di produtto-

ri, gatekeeper e consumatori di informazione, in un ecosistema in continuo mutamento e in perpetua riconfigurazione. Nel mondo dell'editoria e dei media tradizionali, l'informazione (e i dati) sono strutturati in modo tale da massimizzarne l'usabilità da parte dei consumatori. Questo risultato viene ottenuto cercando di prevedere in che modo gli utenti utilizzeranno i dati. Nel mondo del web dei dati, invece, cercare di prevedere come i dati verranno utilizzati, combinati, arricchiti e riconfigurati per produrre beni informativi è impossibile (e persino sbagliato). Questo è uno dei principi fondamentali alla base dell'approccio linked data. Tuttavia, questo aspetto è uno dei più difficili da accettare dato che ha un impatto sia sulla fase di produzione dei dati, sia sulla fase di consumo dei dati (si pensi, ad esempio, che le applicazioni che consumano linked data dovrebbero essere programmate in modo da sapersi adattare a manipolare dati modellati secondo schemi ignoti a priori). Dal lato della produzione, il fatto di non conoscere in quale modo i dati verranno utilizzati, dovrebbe spingere i produttori a modellare i dati secondo schemi il più possibile flessibili e universali. Questa flessibilità ha un costo: comporta l'abbandono di ogni tentativo di ottimizzazione del design dell'informazione in funzione dell'usabilità o delle performance, rendendo più complicato per gli utenti finali comprendere e consumare i dati, che devono essere continuamente rimodellati. Anche la natura a grafo del giant global graph ha profonde implicazioni sociali che impattano sul modo in cui l'informazione viene prodotta e consumata. Grazie allo straordinario e duraturo successo dei database relazionali e dei fogli di calcolo, siamo abituati a crearci modelli mentali dei dati in forma tabulare. Per responsabili IT e sviluppatori, information e data manager, scienziati, esperti di marketing, educatori e molti altri attori coinvolti nel ciclo di vita dei dati è normale pensare i dati all'interno di uno specifico contesto e da un punto di vista individuale e soggettivo.

Thinking in the graph non è semplice e come il mio collega Gradmann sottolinea spesso, è ancora più difficile nel mondo dei beni culturali e delle digital humanities, che hanno una forte tradizione di pensiero bidimensionale, tipico del mondo della carta stampata. Il mondo digitale è spesso caduto nella trappola di copiare il vecchio mondo, invece di cercare di rivoluzionarlo. Ancora una volta il problema viene aggravato da fattori tecnologici: in primo luogo la moltitudine dei linked data oggi disponibili nella LOD cloud culturale è stato generata automaticamente a partire da formati tabellari; in secondo luogo molti corsi universitari di informatica sono ancora fortemente incentrati su strutture dati tabulari e relazionali. Un altro motivo che genera incomprensioni deriva dall'utilizzo di URI e IRI per identificare sia risorse cosiddette *informational* sia *non-informational*. Questo modello è spesso fonte di difficoltà di comprensione anche all'interno della comunità di esperti del settore (si vedano, a riguardo, le frequenti discussioni su "http-range-14").[4] Il problema noto come "HTTP Range 14 problem" riguarda i meccanismi da utilizzare per distinguere tra statement che riguardano pagine web o risorse digitali e statement che riguardano oggetti del mondo reale oppure concetti. Lasciando da parte le considerazioni più tecniche, quello che è importante è che la distinzione tra informational resources e non informational resources spesso non è chiara né ai produttori né ai consumatori di dati. Questa incomprensione fa sì che molte delle applicazioni esistenti presentino inconsistenze che derivano da una errata interpretazione del problema. Sebbene questo non costituisca di per se un problema critico, diventa rilevante allorché i dati inconsistenti vengono utilizzati per fare inferenza.

Infine è opportuno ricordare che il web dei dati è costruito sulla base della cosiddetta Open World Assumption (OWA)[5], secondo la

[4] http://www.w3.org/2001/tag/group/track/issues/14.

[5] Open World Assumption, in Wikipedia, http://en.wikipedia.org/wiki/Open_

quale il valore di verità di un'asserzione è indipendente dalle conoscenze dell'osservatore. In altre parole, questo significa che se un osservatore non sa se un'asserzione è vera non può inferire che essa sia falsa (come invece accadrebbe invece in un mondo in cui vale la Closed World Assumption). L'assunzione di mondo aperto rappresenta un'altra importante deviazione rispetto al mondo dei database relazionali che sono invece basati sulla Closed World Assumption. Se le implicazioni logiche della OWA vanno al di là degli scopi di questa introduzione, è tuttavia interessante analizzarne alcune implicazioni sociali. La scelta di operare in logica di mondo aperto è giustificata dal fatto che i mondi aperti sono particolarmente adatti quando si ha a che fare con informazione incompleta e frequenti eccezioni. Le caratteristiche dei mondi aperti sono molto adatte a sistemi universali come il web. Tuttavia questa scelta pone anche alcuni problemi non banali. Ad esempio alcuni problemi sono intrinsecamente legati a mondi chiusi e, soprattutto, la quasi totalità degli strumenti informatici oggi disponibili sono progettati per operare in mondi chiusi. Infine, per gli stessi motivi esposti in precedenza, la maggioranza delle persone sono state educate a pensare in sistemi chiusi.

Stimolare lo sviluppo di un'economia dei linked data

Nel corso dell'ultimo decennio l'Europa ha fatto grandi investimenti sulle tecnologie semantiche, che hanno generato idee brillanti, conoscenza scientifica e numerosissime implementazioni prototipali. Sfortunatamente non siamo stati in grado di trasferire la ricerca all'industria per produrre strumenti di qualità industriale facilmen-

world_assumption.

te utilizzabili dagli utenti finali. Non esiste a oggi un Microsoft Excel o un Apple iTunes dei linked data, come non esiste ancora un MySPARQL o un Apache HTTPD per pubblicare linked data in streaming: quello che ancora manca è una fiorente economia dei linked data. È forse giunto il momento di investire in innovazione che sia in grado di sfruttare l'enorme bagaglio di conoscenze accumulate in anni di ricerca e il vasto ammontare di dati prodotti per creare un circolo virtuoso in grado di generare un ecosistema sostenibile e in continua evoluzione.

Recentemente sono state rese pubbliche importanti notizie che potrebbero avere un impatto importante nella nascita di una linked data economy: in primo luogo Google ha annunciato la pubblicazione del Google *Knowledge Graph*, una sorta di linked closed data cloud industriale; Google ha anche parlato dell'acquisizione di Freebase, uno dei nodi più importanti della LOD cloud; in secondo luogo, è stata stipulata una coalizione tra alcuni dei più grandi motori di ricerca (tra cui Google, Yahoo e Bing) che ha prodotto un insieme di tecnologie e incentivi economici e sociali con lo scopo di indurre i produttori di contenuti ad arricchire le proprie pagine web con markup semantico; infine, molte grandi organizzazioni pubbliche e private si stanno avvicinando al web dei dati, anche modificando profondamente i loro modelli di business e processi di produzione, oppure creando le proprie closed linked data clouds (es: molte grandi aziende farmaceutiche stanno costruendo le proprie linked cloud aziendali). Da una parte alcuni di questi annunci pongono certamente questioni socio-economiche relative al rischio di impoverimento della cosa pubblica e al generarsi di posizioni di monopolio (si veda su questi temi l'interessante analisi pubblicata da (Tennison) sul suo blog). Il mercato europeo è caratterizzato dalla presenza di una moltitudine di PMI che rappresentano la forza trainante nell'innovazione e nella crescita economica. In questo scenario, è fondamentale

riuscire a concepire una strategia oculata capace allo stesso tempo di proteggere il patrimonio culturale pubblico e di offrire incentivi alle PMI per riuscire a generare una linked data economy vitale e sostenibile. D'altra parte, stiamo assistendo a enormi passi avanti del web dei dati, il cui valore economico viene valorizzato grazie alla massa critica (di utenti, investimenti, tecnologia, visibilità nei confronti dei media e stimolo della domanda) che le aziende leader del web sono capaci di smuovere. In un recente post, (Dodds) suggerisce che il rumore mediatico creato dall'annuncio del Google Knowledge Graph – che è ancora per la gran parte costituito da commons – può rappresentare una grande opportunità per le PMI, che possono sfruttare gli stessi commons per soddisfare la crescente domanda di linked data in settori verticali e specializzati.

Conclusioni

Dopo aver presentato alcune delle sfide tecnologie per il pieno utilizzo del linked open data web, questo articolo ha messo in evidenza come tale innovazione debba procedere di pari passo con una nuova comprensione dei cambiamenti culturali insiti nel web di dati. Ciò, comunque, non è sufficiente. Il web di dati necessita di un vivace ambiente economico nel quale crescere e svilupparsi intorno al proprio potenziale. Come fare tutto questo? Credo che le linee di intervento che ho cercato di suggerire siano ben espresse dalla Commissione Europea, che scrive:

> The volume of data being digitally stored and exchanged is growing exponentially. [...] Obviously, these data generate the potential for many new types of products and services. The accessibility of public services can be improved for open and linked data, smart traffic and

cities can improve mobility, products can report their life cycles, monitoring their provenance and quality, social trends can be recognized and turned into services, and products can come closer to meeting consumers' needs. We foresee a whole new industry implementing services on top of large data streams. The impact of this emerging economic sector - the data economy - may soon outrange the current importance of the software industry. The gist of the matter is to turn large streams of data into added value for the public and private sector. This industry can help to increase the efficiency of processes working with these data, it can provide transparency, support well-informed decision making, and enable new services not possible today (e.g., smart cities, interactive trend analysis or seamless data flows along value creation chains). Clearly, research, engineering, policy making for the Data Economy and the exploitation of the unprecedented wealth of data have become keys to the Future of Europe.[6]

Works cited

Barbieri, Davide e Emanuele Della Valle. «A Proposal for Publishing Data Streams as Linked Data - A Position Paper». *Proceedings of the Linked Data on the Web (LDOW2010) Workshop, co-located with WWW2010*. 2010. (Cit. a p. 3).

Berners-Lee, Tim. «Giant Global Graph». http://dig.csail.mit.edu/breadcrumbs/node/215.

Dodds, Leigh. «Welcome to the Knowledge Graph». http://talis-systems.com/2012/05/welcome-to-the-knowledge-graph/.

[6] http://2012.data-forum.eu/about.

Sequeda, Juan F. e Oscar Corcho. «Linked Stream Data: A Position Paper». *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09)*. Washington DC, USA, 2009. (Cit. a p. 3).

Tennison, Jeni. «Schema.org and the Responsibility of Monopoly». http://www.jenitennison.com/blog/node/157.

MICHELE BARBERA, Net7; Spazio Dati.
info@netseven.it

Barbera, M. "Linked (open) data at web scale: research, social and engineering challenges in the digital humanities". *JLIS.it*. Vol.4, n.1 (Gennaio/January 2013): Art: #6333. DOI: 10.4403/jlis.it-6333. Web.

ABSTRACT: The amount of data available in the Linked Data Cloud has grown enormously in the last years in several domains, including Cultural Heritage and digital humanities. However creative reuse of data both within the scholarly community and within the cultural industry is still very limited. This depends on a mixture of technical and social problems that needs to be adressed in research and whithin the industry. The talk will explore some of these challenges with a focus on the digital humanities.

KEYWORDS: Web semantico; Library linked data

Submission: 2012-06-12
Accettazione: 2012-08-31 2012-08-31
Pubblicazione: 2013-01-15 2013-01-15