Open data: introduzione ai profili tecnici

Pierluigi Feliciati
pierluigi.feliciati@unimc.it

CONVEGNO

OPEN DATA E PROTEZIONE DEI DATI PERSONALI
NEL CONTESTO DELL'AGENDA DIGITALE ITALIANA

GIOVEDÌ 30 MARZO 2017 / ORE 16:00 – 19:00
AULA A / PIAZZA STRAMBI 1 / MACERATA

UNIVERSITÀ DI MACERATA
l'umanesimo che innova

La PA (e le imprese) producono, raccolgono e archiviano informazioni diverse:

- **Dati necessari alla gestione delle attività** (*anagrafici*, provenienti da altri Enti, per la gestione dei procedimenti amministrativi...)
- **Dati prodotti nella gestione delle attività** (relativi ad es. all'attività degli eletti, per certificare le azioni della macchina amministrativa, dati relativi alle prestazioni erogate dalle ASL...)
- **Dati prodotti come risultato delle attività** (relativi all'inquinamento ambientale, relativi all'incidenza della criminalità sul territorio, inerenti i risultati scolastici prodotti dalle scuole e dai provveditorati, riferiti al mercato immobiliare, sul tessuto imprenditoriale prodotti dagli organismi camerali, economico finanziari e bilanci...)

Che dati (pubblici e non)?

http://okfn.org/opendata/

Che dati (pubblici e non)?

Cosa significa open?

- **indicizzabile dai motori di ricerca:** *If it can't be spidered or indexed, it doesn't exist*

- disponibile in un **formato standardizzato e leggibile** da una applicazione informatica: *If it isn't available in open and machine readable format, it can't engage*

- rilasciati attraverso **licenze** libere che **non impediscano la diffusione e il riutilizzo creativo:** *If a legal framework doesn't allow it to be repurposed, it doesn't empower*

David Eaves, *The three laws of open government data*, http://eaves.ca/2009/09/30/three-law-of-open-government-data/

I dati aperti sono dati che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di condividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati

The Open Definition, Version 2.1, http://opendefinition.org/, a project of Open Knowledge International – Source Code

Cosa significa open?

Cosa significa open?

- **Disponibilità e accesso:** i dati devono essere disponibili nel loro complesso, per un *prezzo non superiore ad un ragionevole costo di riproduzione*, preferibilmente mediante scaricamento da Internet. I dati devono essere disponibili in un *formato utile e modificabile*.
- **Riutilizzo e ridistribuzione:** i dati devono essere forniti a condizioni tali da permetterne il riutilizzo e la ridistribuzione. Ciò comprende la *possibilità di combinarli con altre basi di dati*.
- **Partecipazione universale:** tutti devono essere in grado di usare, riutilizzare e ridistribuire i dati. Non ci devono essere *discriminazioni né di ambito di iniziativa né contro soggetti o gruppi*. Ad esempio, la clausola 'non commerciale', che vieta l'uso a fini commerciali o restringe l'utilizzo solo per determinati scopi (es. quello educativo) o non è ammessa.

Come fare open data? 3 regole

Ci sono tre regole fondamentali che il manuale dell'**Open definition** consiglia di seguire nell'apertura dei dati:

- **Scegliere la semplicità**. Cominciare con un progetto piccolo, semplice e veloce. Non è necessario aprire tutti i dati in una sola volta.
- **Coinvolgere gli utenti fin dall'inizio e coinvolgerli spesso.** Cercare presto e spesso il confronto con i potenziali utilizzatori dei dati fra cittadini, imprese o sviluppatori
- **Affrontare i timori e le incomprensioni diffuse.** Questo è importante soprattutto se lavori in o con grandi organizzazioni come le istituzioni governative.

quattro passi principali per rendere i dati aperti:

- **Scegliere i dataset**. Scegliere ciò che si intende rendere aperto, ricordando che si può (o potrebbe essere necessario), rivedere questo passaggio se si incontrano problemi nelle fasi successive.
- **Utilizzare una licenza open**. Determinare quali sono i diritti di proprietà intellettuale che insistono sui dati. Applicare una adeguata licenza che copra tutti i diritti identificati. Se ciò non è possibile, si ritorni al punto 1 e riprovare con una banca dati diversa.
- **Rendere i dati disponibili** in gran quantità e in un formato utile. Si possono prendere in considerazione anche metodi alternativi come la distribuzione attraverso API.
- **Rendere disponibile un catalogo.** Organizzare e offrire un catalogo dove elencare l'insieme dei dati aperti (usabilità).

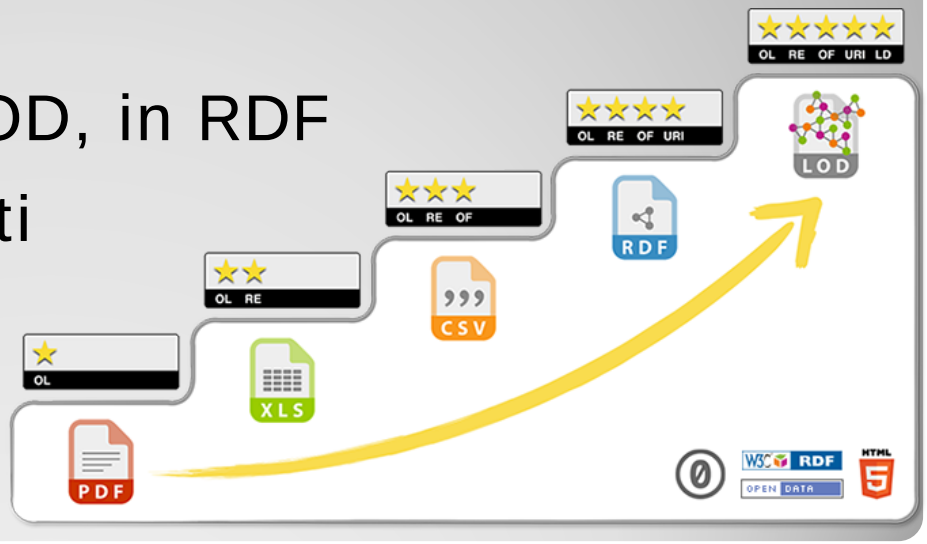
Come fare open data? 4 steps

Lo sviluppo verso *the next Web*

Tim Berners Lee ha lanciato nel 2006 e nel 2009, a 20 anni dall'idea originaria del Web, una prospettiva di sviluppo che denominò **Linked Open Data**, ovvero un ripensamento del *Web delle informazioni* 1.0 verso il **Web dei dati** (o **Web semantico**), connessi semanticamente tra loro e interpretabili da agenti software

5 livelli di «potenza» dei dati aperti nel *next web*, secondo TBL:

1. File per la lettura umana, come il PDF
2. File con dati leggibili automaticamente, come XLS
3. File con dati in forma di tabella aperta, ad esempio CSV
4. Singole asserzioni LOD, in RDF
5. Dataset di LOD aperti verso risorse esterne



Le 5 stelle di sir Berners Lee

Per raggiungere questo obiettivo TBL ha proposto quattro regole, che non tradiscono i fondamenti del Web 1.0:

- **Usare URI** per identificare oggetti.
- Usare **HTTP URI** in modo che questi oggetti possano essere **referenziati e cercati** sia da persone che da *user agents*.
- Fornire **informazioni utili sull'oggetto** quando la sua URI non è referenziata, usando formati **standard** come XML-RDF.
- Includere **link ad altre URI** relative ai dati esposti per migliorare la ricerca di altre informazioni relative nel Web.

Lo sviluppo verso i LOD

Come funziona RDF?

L'RDF Data Model si basa su tre principi chiave:

- Qualunque cosa può essere identificata da un URI
- *The least power*: va utilizzato il linguaggio meno espressivo per definire qualunque cosa
- Qualunque cosa può dire qualunque cosa su qualunque cosa (tutto è collegabile).

Qualunque cosa descritta da RDF è detta **risorsa**, tendenzialmente reperibile sul web. L'unità base di informazione in RDF è lo **statement**, una tripla del tipo ***Soggetto – Predicato – Oggetto*** dove il soggetto è una risorsa, il predicato è una **proprietà** e l'oggetto è un **valore** (e quindi, potendo, un URI che punta ad un'altra risorsa).

Per superare questo modo di rappresentare la conoscenza (duple):

- 1. "Oreste Signore è l' autore del DocumentoX"
- 2. "L' autore del DocumentoX è Oreste Signore"

Si adotta questo modello:

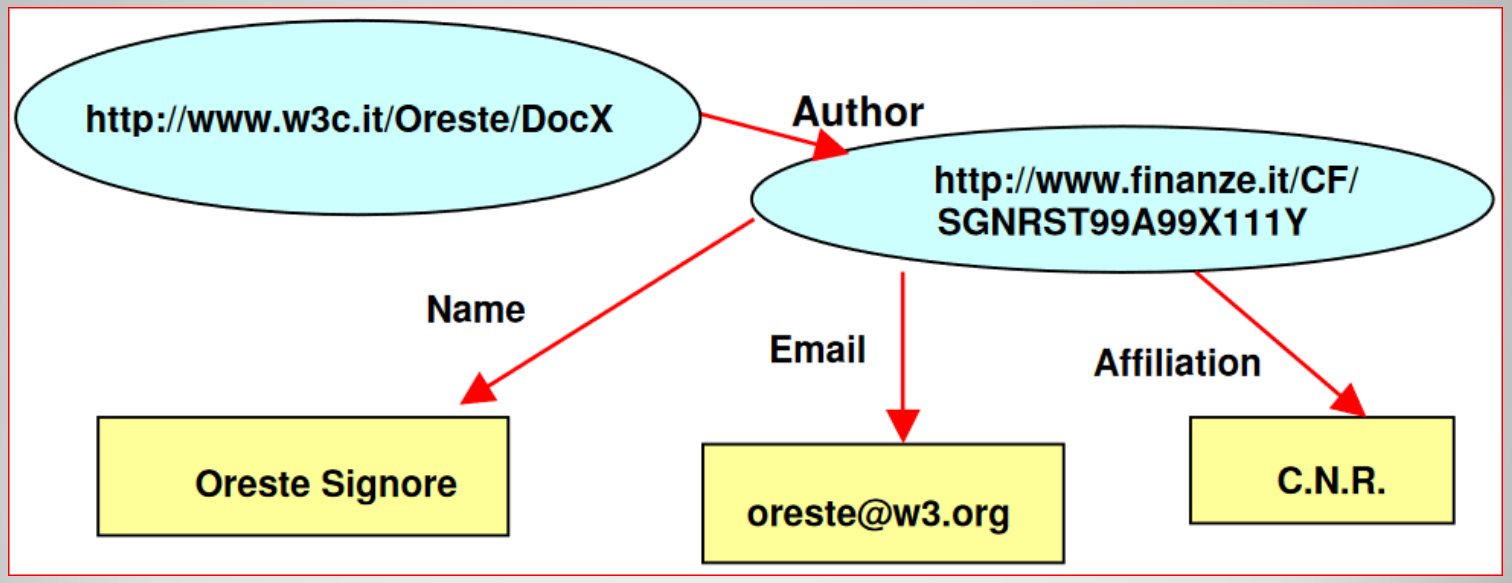
- *Resource:* http://www.w3c.it/Oreste/DocX
- *Property:* author
- *Value:* Oreste Signore

Quindi:

La risorsa http://www.w3c.it/Oreste/DocX **has Author** «Oreste Signore»

RDF: un esempio

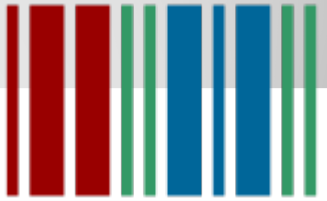
O, meglio ancora, aggiungendo ulteriori proprietà e connessioni, dovendo esprimere:
Oreste Signore, la cui email è oreste@w3.org, il c.f. /SGNRST99A99X111Y, lavora presso il C.N.R., è *l'autore* del DocumentoX



RDF: un esempio

Linked Open Data

- L'obiettivo del progetto Linking Open Data del W3C è di estendere il Web pubblicando diversi **open dataset** come RDF sul Web e impostando link RDF tra i dati di differenti risorse.
- Nell'ottobre del 2007, i dataset contenevano più di due miliardi di triple RDF, collegate da più di due milioni di link RDF.
- Da maggio 2009 sono cresciuti a 4,2 miliardi di triple RDF, collegate da circa 142 milioni di link RDF.
- Oggi, vedi *Linking Open Data cloud diagram* 2017
- La fonte maggiore di LOD è **Wikidata** (o Dbpedia), (https://www.wikidata.org/), l'estrazione di dati RDF da Wikipdia, che assomma oggi a circa 30 milioni di elementi



Grazie dell'attenzione!

pierluigi.feliciati@unimc.it

unimc
UNIVERSITÀ DI MACERATA
l'umanesimo che innova