PROGETTI CONCLUSI 2018	
Titolo progetto	**Versione reingegnerizzata sperimentale del portale LOD**
Descrizione	Il progetto si propone costruire un sistema che permetta di pubblicare dati in formato Linked Open Data (LOD), interfacciandosi direttamente con le fonti dati, rendendoli fruibili ai sistemi interni ed esterni all'Istituto. Nell'ambito della modernizzazione, al fine di fornire al cittadino un servizio sempre più completo e qualitativamente significativo, sono state studiate numerose tecnologie, tra le quali il Linked Open Data. Il Linked Open Data permette di fruire di dati e tabelle come si farebbe con un sito web, organizzando e presentando le informazioni durante la navigazione secondo una base semantica. Un portale che integrasse questa tecnologia permetterebbe di presentare all'utente un contenuto creato in base alle sue richieste in maniera adattiva online ed in real time. Il vantaggio del linked open data è nella possibilità di definire in maniera scientifica sia il significato del dato, il contesto della sua utilizzazione e il dettaglio del suo collegamento verso il mondo. Open, nel nostro caso, significa piuttosto organizzare, definire, standardizzare in maniera che il dato sia descritto in un contesto comune e condiviso e che quindi possa essere fruito con regole chiare e significato preciso, organizzato e standardizzato. In questo ambito c'è uno sforzo comune a vari livelli nella pubblica amministrazione Italiana, e nelle istituzioni sovranazionali a livello europeo (Eurostat) e internazionale (Unece) per definire un framework che favorisca l'interoperabilità dei dati e processi. La standardizzazione, infatti, impone che i dati siano descritti in termini accettati e condivisi, che il dato sia facilmente linkabile e navigabile ed inseribile nel giusto contesto in base ad un'attenta metadatazione, nascondendo, nel contempo, i dettagli strutturali. Inoltre, il Linked open data apre le porte del web semantico, dove la navigazione è fatta attraverso il significato e non attraverso la forma. L'interfaccia unica permette l'integrazione tramite protocolli machine to machine, permettendo una navigazione che, per la prima volta, si estende anche all'automa ed alle macchine. Un primo passo, quindi permette di riformattare il dato in questo nuovo standard partendo dalle effettive tavole dei dati. Un connettore apposito fa da interfaccia tra le sorgenti dati che si desidera pubblicare ed il consumatore, che accede ai dati attraverso query in un linguaggio standard denominato SparQL.

Obiettivi	L'Istituto ha partecipato ad un progetto ESSNET finalizzato alla definizione di una pipeline per la pubblicazione dei Linked Open Data secondo standard condivisi e riconosciuti a livello Comunitario. Sebbene il progetto si sia concluso con la definizione della pipeline, e con la pubblicazione di una serie di strumenti per implementarla, restava da valutarne l'effettiva utilità e realizzabilità nei singoli istituti di Statistica. Nell'ambito del Laboratorio dell'Innovazione è stato quindi realizzato un prototipo modulare per la pubblicazione di dati interni dal formato tabellare a Linked Open Data, con le seguenti finalità: • Costruzione di un dispositivo modulare integrando tecnologie proposte dall'ESSNET, Interne all'Istituto o disponibili su piattaforme proprietarie o open source • Costruzione di un dispositivo scalabile in grado di essere utilizzato efficientemente in regime di produzione integrandolo con le tecnologie esistenti. • Valutazione degli strumenti disponibili in vista dell'evoluzione della piattaforma interna di Data Warehouse e collegamento verso il portale di Linked Open Data. • Valutazione delle prestazioni, specialmente in relazione agli strumenti di memorizzazione dati "Triplestore" impiegati Stardog (proprietario) e Blazegraph (open source) • Implementazione su un dato di esempio pratico: Nella fattispecie sono state utilizzate tabelle provenienti dalle tabelle di pubblicazione del Censimento Industria e Servizi.
Metodologia	Alla base della pubblicazione dei dati in formato LOD c'è la definizione dell'ontologia. L'ontologia è una rappresentazione astratta formale e condivisa che descrive il dominio di interesse, garantendo l'interoperabilità semantica non solo al livello di diffusione dati umana ma anche rendendolo fruibile e processabile al livello automatico. L'ontologia consente di dedurre nuova conoscenza mediante l'inferenza, cioè permette di dedurre informazioni celate nel dato stesso calcolandolo in maniera esplicita. Quindi, oltre ad un fortissimo potere descrittivo, ci fornisce anche di un potente strumento di analisi e di alta qualità. L'ontologia fornendo una descrizione astratta dal dettaglio implementativo, permette di integrare a livello semantico molteplici basi di dati sulle quali si mappa l'ontologia, il che significa che se si collega la

descrizione standard di una certa area di conoscenza ai dettagli di diverse basi dati, non concepite originariamente per interoperare fra di loro. L'ontologia permette di costruire un ponte tramite il quale la comunicazione e lo scambio di informazioni possono avvenire facilmente. L'utenza, a prescindere dal proprio grado di conoscenza della statistica, si trova un sistema che non è passivo, ma che permette di essere scoperto, elaborato e che è, a tutti gli effetti un contenuto attivo.

Con un unico sistema è possibile quindi pubblicare qualunque contenuto informativo e far in modo che, con strumenti facili e customizzati, l'utente sia in grado di interrogare il sistema e di trovarsi da solo i dati rilevanti.
A livello Eurostat ci sono diversi gruppi di lavoro con tema LOD o diversamente correlato. Il tutto finalizzato alla costruzione di una piattaforma comune per l'interscambio dei dati in formato LOD.
A livello Istat c'è da sempre lo sforzo di riorganizzare i propri dati secondo un sistema standardizzato di metadati che sono la base della concettualizzazione del LOD.
Si vede quindi che l'ambito di utilizzo di questo paradigma è omni-pervasivo e interseca tutte le fasi del ciclo di vita del dato, dalla progettazione, alla elaborazione, alla diffusione, alla fruizione.
Nell'ambito della ESSNET "Linked Open Statistics" si sono poste le basi per la costruzione del framework comune di produzione del LOD a livello comunitario.

C'è però da notare che lo sforzo si è concentrato soprattutto al livello operativo, cioè l'ontologia è stata utilizzata maggiormente per descrivere la struttura tabellare del modello dimensionale piuttosto che l'area di conoscenza del dato stesso, traducendo quindi le tabelle di dati in formato Linked Open Data. Esiste infatti una meta ontologia che descrive molto accuratamente i dati aggregati in generale, prescindendo dalle relative aree di conoscenza.
Successivamente questo dato vien caricato su un dispositivo di storage detto triple-store al quale si può accedere mediante un endpoint SparQL.

Sono stati installati e configurati gli strumenti utili alla sperimentazione. Oracle per contenere i dati provenienti dal data warehouse dell'istituto con cui alimentare la pipeline.

- Stardog e Blazegraph, i triplestore.
- JUMA, come strumento di mapping, software open source del Derilinx-ADAPT-Insight consortium.
- MySQL per supportare il funzionamento di JUMA.
- J- r2rml per la conversione del dato in formato rdf.
- Pubby, per la navigazione dei grafi.

Una volta installati e configurati i componenti del sistema, è stato necessario integrarli per far funzionare il convertitore. A questo

	proposito, è stata necessaria una fase di studio delle API disponibili per interfacciare i componenti. E' stata sviluppata una **web application di test** per l'interrogazione diretta degli endpoint sparql.
Risultati ottenuti	Il progetto del Laboratorio di Innovazione è servito per testare l'effettiva implementazione della pipeline LOD ESSNET con strumenti utilizzabili in Istat, sia open source, sia proprietari, basandosi sulle effettive tabelle disponibili in istituto. Sono stati testati diversi software alternativi seguendo un approccio standardizzato per l'integrazione verificandone le prestazioni ed il grado di integrabilità. La nostra proposta consiste nell'adattare la pipeline studiata a livello comunitario, realizzando un connettore modulare tra le tabelle disponibili convertendole in RDF fruibile tramite un triplestore. E' stata necessaria una particolare attenzione per integrare il mapper (JUMA) nella pipeline: Sono state necessarie diverse modifiche al software per adattarlo efficientemente. Sono ora disponibili vari tool per la pubblicazione di dati in formato LOD, per l'utilizzo con i triplestore Blazegraph e Stardog, ma che possono anche essere facilmente estesi all'utilizzo con altri triplestore. Tali tool o l'intera pipeline potranno essere utilizzati anche in alternativa agli strumenti attualmente utilizzati in Istituto per la gestione dei LOD. Il progetto ha mostrato come sia possibile sviluppare una pipeline dinamica, scegliendo di integrare strumenti diversi a scelta. Il prototipo può essere integrato negli esistenti sistemi di diffusione dati e permette di mettere in comunicazione il datawarehouse con il portale linked open data. A tutti gli effetti, il sistema si comporta da connettore dati, in grado di alimentare il portale LOD previa costruzione dei mapping adeguati. Un successivo strato di generalizzazione permetterebbe di collegare anche la parte dei metadati tematici ai dizionari tematici linkabili tramite il mapper all'inizio del processo di mappatura, e non a valle del processo di diffusione come semplice link del metadato in formato LOD. In sostanza l'intero sistema si compone di quattro componenti di cui il connettore è centrale, corredato di un integratore per i metadati, ancora in fase di studio, di una parte applicativa di navigazione e di fruizione del dato, integrabile nel portale LOD ed in un sistema di interfaccia verso le fonti dati, fonti alle quali ad oggi non è possibile sempre accedere direttamente, ma presuppongono il download in formato csv per la pipeline.

	Qualunque di questi quattro componenti è indipendente dagli altri e può essere integrato in vista di un'evoluzione del sistema di diffusione dati dell'Istituto. Oppure può semplicemente essere usato come proof of concept nella realizzazione di un substrato comune per l'interscambio dei dati a livello europeo e fornire idee e strumenti per le evoluzioni future.