COMMISSIONE DI COORDINAMENTO SPC

LINEE GUIDA PER L'INTEROPERABILITÀ SEMANTICA ATTRAVERSO I LINKED OPEN DATA

Agenzia per l'Italia Digitale

INDICE

INDICE DELLE FIGURE

INDICE DELLE TABELLE

EXECUTIVE SUMMARY

L'Agenda Digitale Europea prescrive, agli Stati Membri, di allineare i propri framework nazionali di interoperabilità a quello Europeo (EIF) entro il 2013 (azione 26). Come definito dal CAD, il framework italiano di interoperabilità è il Sistema Pubblico di Connettività e Cooperazione (SPC). Nell'ambito del programma ISA, programma che ha il compito di attuare a livello europeo gli obiettivi di interoperabilità del pilone 2 della Digital Agenda, il National Interoperability Framework Observatory (NIFO) ha valutato i vari framework nazionali di interoperabilità tra i quali SPC. La valutazione è stata effettuata sulla base di cinque ambiti: "principi", "governance", "modello concettuale", "accordi d'interoperabilità", e "livelli di interoperabilità". Dall'analisi, pubblica, emerge come l'Italia [75] sia piuttosto bene allineata sui primi tre ambiti, evidenziando una conformità pari quasi al 100%, mentre risulti più carente sugli accordi di interoperabilità e, soprattutto, sui livelli di interoperabilità garantiti.

Tra i livelli di interoperabilità definiti dal modello EIF rientra **l'interoperabilità semantica,** ovvero la possibilità, offerta alle organizzazioni, di elaborare informazioni da fonti esterne o secondarie senza perdere il reale significato delle informazioni stesse nel processo di elaborazione. Questa definizione coglie il ruolo centrale che tale dimensione assume nella collaborazione e nell'interscambio di informazioni tra istituzioni, base quindi di ogni processo di e-government innovativo.

D'altra parte, anche sotto la spinta della promozione a livello dell'Unione Europea, molte amministrazioni hanno intrapreso la strada di pubblicare Open Data per favorire la trasparenza e per rendere a cittadini e imprese quell'enorme patrimonio di informazioni pubbliche che le Pubbliche Amministrazioni raccolgono e detengono in virtù dei propri ruoli istituzionali. In tale ambito, è utile disporre di linee guida per la produzione di open data interoperabili.

Nella cornice descritta, l'obiettivo principale del presente documento, redatto da un gruppo di lavoro istituito dalla Commissione di Coordinamento SPC nell'ambito della definizione delle infrastrutture condivise SPC, è quello di proporre delle linee guida per l'interoperabilità semantica, cogliendo quindi gli obiettivi della Digital Agenda e, in ambito nazionale, fornendo un quadro di riferimento per la produzione di Open Data interoperabili.

Il lavoro si è articolato partendo da un'attenta disamina dello scenario attuale, sia nazionale che internazionale, nell'ambito della gestione dei dati del settore pubblico. Da tale analisi è emerso come un profondo cambiamento in questo settore sia in atto, grazie allo sviluppo del nuovo paradigma degli Open Data. I dati delle Pubbliche Amministrazioni, tipicamente "nascosti" in applicazioni o basi di dati, e da loro gelosamente custoditi nella maggior parte dei casi, sono sempre più resi accessibili a chiunque con la consapevolezza che, di fatto, essi rappresentano un patrimonio della collettività, e non di singole istituzioni, e un importante strumento per la trasparenza, responsabilità e possibile sviluppo economico. Tuttavia, l'analisi dello scenario ha evidenziato altresì che, per sfruttare pienamente i suddetti benefici, è necessario favorire la facilità d'uso dei dati, così come il loro reperimento e consumo sia da parte degli esseri umani che, soprattutto, da parte dei software attivabili anche in maniera automatica.

Il gruppo di lavoro ha quindi analizzato le tipologie di dati in possesso delle Pubbliche Amministrazioni e lo stato dell'arte relativo ad alcune iniziative volte a raggiungere i suddetti obiettivi in tale ambito, e ha identificato nelle tecnologie standard del Web semantico, e in particolare nel modello dei **Linked Open**

Data, gli strumenti imprescindibili per dare ai dati (aperti o non) un'identità e per renderli collegabili tra loro e soprattutto interoperabili. In altre parole, il gruppo di lavoro, all'unanimità, ha ritenuto che, per abilitare lo sviluppo di una concreta interoperabilità semantica tra Pubbliche Amministrazioni a livello nazionale e transfrontaliero sia necessario adottare il modello Linked Open Data.

Le presenti linee guida vogliono, quindi, essere un aiuto per le Pubbliche Amministrazioni offrendo un approccio metodologico all'interoperabilità semantica attraverso il modello Linked Open Data accompagnato dall'uso di ontologie condivise, e un'attenta analisi dell'insieme di standard, tecnologie di base e strumenti che consentono di implementare l'approccio. A tal riguardo, sono state definite una serie di raccomandazioni, ben evidenziate all'interno del documento, allo scopo di focalizzare l'attenzione del lettore su azioni concrete che l'esposizione dei concetti non rende così immediate.

A completamento dello studio, il gruppo ha posto particolare attenzione agli aspetti legali derivanti dalle licenze d'uso associate ai dati pubblicati, e ai modelli di business che possono essere abilitati, evidenziando aspetti di interoperabilità anche per le licenze, sostenibilità e governance dell'approccio Linked Open Data.

Lo studio è stato poi contestualizzato nell'ambito del framework nazionale d'interoperabilità SPC, al fine di individuare lo specifico ruolo che, sia le infrastrutture condivise sia i servizi e-government per la PA possono assumere per la concreta acquisizione e attuazione dell'approccio proposto e quindi dell'interoperabilità semantica. Dallo studio ne è risultato che SPC, e in particolare il servizio "Catalogo Schemi e Ontologie", come definito nel DPCM 1° aprile 2008 recante regole tecniche e di sicurezza per SPC, può essere profilato come il servizio Linked Open Data SPC (il Web dei Dati SPC), che consente di produrre Linked Data a partire da dati generati e scambiati in SPC, di collegare tali dati ad altri dati delle PA (centrali e locali), e di arricchire i dati con opportuni metadati semantici per stabilire uno standard di qualità a livello di pubblicazione, di utenza e di interoperabilità nella PA. Il servizio è pensato sia per la gestione dei meri dati pubblici che per la gestione di tutti quei dati relativi alle funzioni di back end delle PA.

Il documento evidenzia, infine, che l'elenco e le descrizioni delle tipologie di dati individuate e delle tecnologie non vogliono essere esaustivi ma rappresentano il punto di convergenza dei membri del gruppo durante il periodo di operatività dello stesso. Il gruppo, infatti, ha convenuto che un ulteriore studio nei prossimi mesi è necessario in questo ambito, che coinvolga altre tipologie di dati e che consenta di aggiornare opportunamente le linee guida sulla base degli sviluppi di standardizzazione in questo settore.

1. PREFAZIONE

1.1. Componenti del gruppo di lavoro

Francesco Tortorelli (Coordinatore)	Agenzia per l'Italia Digitale
Giorgia Lodi	Agenzia per l'Italia Digitale
Antonio Maccioni	Agenzia per l'Italia Digitale
Alfio Raia	Agenzia per l'Italia Digitale
Daniele dell'Osso	Agenzia per l'Italia Digitale
Antonio Rotundo	Agenzia per l'Italia Digitale

Felice Balsamo	Comune di Napoli
Moira Benelli	Anci
Giuliana Bonello	CSI Piemonte
Michele Bordi	Comune di Macerata
Giampiero Zaffi Borgetti	Anci
Dario Bottazzi	Regione Emilia Romagna
Enrico Cammarata	Ministero Difesa
Paolo Castiglieri	Comune di Genova
Daniela Cavaldesi	INPS
Anna Cavallo	CSI Piemonte
Filippo D'Angelo	INPS
Stefano De Francisci	ISTAT
Massimo Fustini	Regione Emilia Romagna
Aldo Gangemi	ISTC-STLab
Giovanni Gentili	Regione Umbria
Vania Corelli Grappadelli	Lepida SpA
Mario Grassia	Comune di Livorno
Silvia Losco	ISTAT
Giovanni Malesci	Ministero dell'Istruzione Università e Ricerca

Giovanni Menduni	Comune di Firenze
Andrea Nicolini	CISIS
Dario Piermarini	Ancitel
Alessandra Potrich	Fondazione Bruno Kessler
Antonio Putignano	Ministero dell'Istruzione Università e Ricerca
Alessandro Radaelli	Comune di Prato
Angelo Rossi	Comune di Padova
Michele Trainotti	Fondazione Bruno Kessler
Francesco Paolo Valente	Comune di Pescara
Gianluca Vannuccini	Comune di Firenze
Clementina Villani	Comune di Roma

1.2. Ringraziamenti

Il gruppo di lavoro desidera ringraziare l'ing. Michele Osella e il dott. Enrico Ferro, dell'Istituto Superiore Mario Boella, e il prof. Paolo Pasini, di SDA Bocconi, per la definizione della sezione che analizza gli impatti socio economici degli Open Data. I contenuti di quella sezione sono stati sviluppati sulla base di riflessioni e spunti del team di lavoro, e integrando i risultati dello studio esplorativo "Modelli di Business nel Riuso dell'Informazione Pubblica "[24] condotto dall'Istituto Superiore Mario Boella per conto dell'Osservatorio sulle ICT di Regione Piemonte.

Il gruppo ringrazia altresì il dott. Luciano Serafini della Fondazione Bruno Kessler, per il contributo dato alla sezione relativa all'approccio metodologico, e Michele Barbera di SpazioDati per il contributo dato alla sezione relativa allo stato dell'arte su Linked Open Data e interoperabilità semantica.

Il gruppo infine ringrazia Enrico Castagnoli, Emanuele Geri, Giacomo Innocenti, Elena Marassini, Leonardo Ricci del Comune di Firenze per il loro contributo dato all'esperienze italiane e alla sezione "Standard, tecnologie di base e strumenti".

1.3.Modifiche Documento

Descrizione Modifica	Edizione	Data
Prima bozza indice	v 0.1	22-03-2012
Seconda bozza indice	v 0.3	23-03-2012
Scopo documento, indice e candidature definitive	v 0.4	05-04-2012
Inserimento alcuni contributi	v. 0.5	30-04-2012
Prima bozza documento	v. 0.7	10-05-2012
Revisione completa dell'indice e revisione ulteriore di alcuni contenuti del documento	v. 0.8	15-05-2012
Integrazione contributi specifici dei partecipanti	v. 0.9	04-06-2012
Integrazione ulteriori contributi specifici dei partecipanti e bibliografia	v. 1.0	15-06-2012
Integrazione revisioni generali	v. 1.1	10-07-2012
Correzioni finali	v. 1.2	20-07-2012
Bozza per consultazione pubblica	v. 1.3	30-07-2012
Recepimento commenti/suggerimenti consultazione pubblica; adeguamento al decreto n. 83 convertito in legge n. 134 del 7 agosto 2012 , al decreto n. 95 convertito in legge n 135 del 7 agosto 2012 e al decreto crescita 2.0; recepimento di riferimenti forniti dalla Commissione Europea in merito a lavori sviluppati nel contesto ISA. Inserimento licenza d'uso.	v. 2.0	12-11-2012

1.4.Acronimi

Acronimo	Definizione
API	Application Programming Interface
B2B	Business to Business
CAD	Codice dell'Amministrazione Digitale
CC	Creative Commons
CdC	Commissione di Coordinamento
CMS	Content Management System
DBMS	Data Base Management System
DM	Decreto Ministeriale
DPCM	Decreto Presidente del Consiglio dei Ministri
EAV	Entity Attribute Value
EIF	European Interoperability Framework
ER	Entity-Relationship
ETL	Extract, Transform, Load
FOAF	Friend Of A Friend
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
INSPIRE	INfrastructure for SPatial InfoRmation in Europe
IPA	Indice delle Pubbliche Amministrazioni
IRI	Internationalized Resource Identifier
ISA	Interoperability Solutions for public Administrations
JSON	JavaScript Object Notation
LOD	Linked Open Data
NIF	National Interoperability Framework
OD	Open Data
ORDBMS	Object Relational Data Base Management System
OGC	Open Geospatial Consortium
OGD	Open Government Data
OKF	Open Knowledge Foundation

OWL	Ontology Web Language
PA	Pubblica Amministrazione
PaaS	Platform as a Service
PSI	Public Sector Information
R&S	Ricerca e Sviluppo
RDBMS	Relational Data Base Management System
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
REST	Representational State Transfer
RIF	Rule Interchange Format
RNDT	Repertorio Nazionale Dati Territoriali
SDI	Spatial Data Infrastructure
SDMX	Statistical Data Metadata eXchange
SICA	Servizi Infrastrutturali di Cooperazione Applicativa
SKOS	Simple Knowledge Organization System
SOAP	Simple Object Access Protocol
SPC	Sistema Pubblico di Connettività e Cooperazione
SPARQL	Simple Protocol And RDF Query Language
SQL	Structured Query Language
UML	Unified Modeling Language
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WoD	Web of Data
XML	eXtensible Markup Language
XSD	XML Schema Definition
XSLT	eXtensible Stylesheet Language Transformations

1.5. Glossario

APPS: termine utilizzato per indicare le applicazioni software sviluppate per operare su terminali mobili.

CHIAVE NATURALE: un valore reale (ad esempio un codice fiscale) con cui s'identificano univocamente dati (tuple) in una base di dati (relazionale).

CLOUD COMPUTING: Dal NIST, l'istituto nazionale statunitense per gli standard e le tecnologie, il cloud computing è definito come un modello che abilita in rete l'accesso pratico e su richiesta (on demand) a un pool condiviso di risorse computazionali configurabili (e.g., reti, server, storage, applicazioni e servizi) che possono essere ottenute ed erogate rapidamente con il minimo sforzo di gestione e con un'interazione limitata con il fornitore.

COPYRIGHT: è il modo con cui viene chiamato il diritto d'autore nei paesi di common law.

CSV (COMMA SEPARATED VALUES): formato di file di testo che consente di rappresentare dati alfanumerici di una tabella mediante la separazione dei singoli valori con un apposito carattere separatore (virgola).

DATA MINING: l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di un sapere o di una conoscenza a partire da grandi quantità di dati (attraverso metodi automatici o semi-automatici) e l'utilizzo scientifico, industriale o operativo di questo sapere.

DATASET: una collezione di dati, generalmente riguardanti una stessa organizzazione, che vengono erogati e gestiti congiuntamente; insieme di dati strutturati in forma relazionale.

DATI: rappresentazione fisica di informazioni atta alla comunicazione, interpretazione ed elaborazione da parte di essere umani o mezzi automatici.

DATI A CONOSCIBILITA' LIMITATA: dati la cui conoscibilità è riservata per legge o regolamento a specifici soggetti o categorie di soggetti.

DATI GREZZI (RAW DATA): dati raccolti che non hanno subito alcuna modifica, manipolazione o

aggregazione.

DATI PERSONALI: Qualsiasi informazione che riguardi persone identificate o che possano essere identificate anche attraverso altre informazioni, ad esempio, attraverso un numero o un codice identificativo. Sono dati personali: nome e cognome o indirizzo; codice fiscale; ma anche una foto, la registrazione della voce di una persona, la sua impronta digitale o vocale. La persona può essere infatti identificata anche attraverso altre informazioni (ad esempio, associando la registrazione della voce di una persona alla sua immagine, oppure alle circostanze in cui la registrazione è stata effettuata: luogo, ora, situazione).

DATI PUBBLICI: dati conoscibili da chiunque.

DATI SENSIBILI: dati personali che, per la loro delicatezza, richiedono particolari cautele. Sono dati sensibili quelli che possono rivelare l'origine razziale ed etnica, le convinzioni religiose o di altra natura, le opinioni politiche, l'adesione a partiti, sindacati o associazioni, lo stato di salute e la vita sessuale delle persone.

DATI STATISTICI: dati provenienti da uno studio o da una sorgente amministrativa, utilizzati per produrre statistiche e/o dati che comprendono tali statistiche.

E-GOVERNMENT: gestione digitalizzata dei processi e delle comunicazioni che riguardano la funziona pubblica e la funzione amministrativa.

HTTP (HYPER-TEXT TRANSFER PROTOCOL): protocollo standard per la trasmissione di informazione su Web. Tipicamente e nativamente usato per il trasferimento di iper-testi e iper-media.

INFERENZA: il processo con il quale una proposizione viene dedotta dal contenuto di altre proposizioni.

INTEROPERABILITÀ: in ambito informatico, la capacità di sistemi differenti e autonomi di cooperare e di scambiare informazioni in maniera automatica, sulla base di regole comunemente condivise.

INTEROPERABILITÀ SEMANTICA: la capacità di elaborare informazioni da fonti esterne o secondarie senza perdere il reale significato delle informazioni stesse nel processo di elaborazione.

IRI (Internazionalized Resource Identifier): una generalizzazione dell'URI in cui la stringa di caratteri che identifica una risorsa su Internet include anche caratteri Unicode/ISO 10646.

LICENZA: in ambito informatico, lo strumento con il quale si governano legalmente le condizioni d'uso e di distribuzione del software e dei dati.

KNOWLEDGE STORE: una base di conoscenza.

MASHUP: un processo informatico in cui si integrano contenuti, dati e informazioni provenienti da fonti differenti.

ONTOLOGIA: in ambito informatico, una rappresentazione formale e condivisa dei concetti e delle mutue relazioni che caratterizzano un certo dominio di conoscenza.

OPEN GOVERNMENT ("governo aperto"). un nuovo concetto di governance a livello centrale e locale, basato su modelli, strumenti e tecnologie che consentono alle amministrazioni di essere "aperte" e "trasparenti" nei confronti dei cittadini. Tutte le attività dei governi e delle amministrazioni dello stato devono essere aperte e disponibili per favorire azioni efficaci e garantire un controllo pubblico sull'operato.

OPEN GOVERNANCE: il modello di amministrazione che abilita la totale apertura e trasparenza dell'operato di governo (a livello centrale e locale) nei riguardi dei cittadini.

OPEN SOURCE: una modalità con cui il software viene fornito. Si realizza attraverso la concessione a terzi del diritto di accedere liberamente al codice sorgente, talvolta consentendo la possibilità di effettuare modifiche e prender parte alle decisioni progettuali sul software in questione.

PSI (Public Sector Information): nella direttiva Europea 2003/98/EC è definita come l'informazione della Pubblica Amministrazione. L'informazione pubblica si caratterizza per essere di tipo statico o di tipo dinamico. L'informazione statica è rappresentata dal contenuto informativo in possesso della Pubblica Amministrazione (ad esempio gli archivi dei beni culturali); l'informazione dinamica è invece prodotta dalle istituzioni pubbliche nello svolgimento dei propri compiti istituzionali (ad esempio i dati di bilancio di un ente).

RIUSO DI DATI: qualsiasi uso dei dati diverso da quello per il quale sono stati prodotti o raccolti originariamente.

SEGRETO DI STATO: un vincolo giuridico che determina l'esclusione dalla divulgazione *"di atti, documenti, notizie, attività e ogni altra cosa la cui diffusione sia idonea a recare danno all'integrità della Repubblica, anche in relazione ad accordi internazionali, alla difesa delle istituzioni poste dalla Costituzione a suo fondamento, all'indipendenza della Stato rispetto ad altri Stati e alle relazioni con essi, alla preparazione e alla difesa militare dello Stato"*, ponendo delle sanzioni nei confronti di chi violi l'obbligo di non divulgazione.

SEGRETO STATISTICO: i dati raccolti nell'ambito di rilevazioni statistiche comprese nel programma statistico nazionale da parte degli uffici di statistica non possono essere esternati se non in forma aggregata, in modo che non se ne possa trarre alcun riferimento relativamente a persone identificabili e possono essere utilizzati solo per scopi statistici.

SPC: l'insieme di infrastrutture tecnologiche e di regole tecniche, per lo sviluppo, la condivisione, l'integrazione e la diffusione del patrimonio informativo e dei dati della Pubblica Amministrazione, necessarie per assicurare l'interoperabilità di base ed evoluta e la cooperazione applicativa dei sistemi informatici e dei flussi informativi, garantendo la sicurezza, la riservatezza delle informazioni, nonché la salvaguardia e l'autonomia del patrimonio informativo di ciascuna Pubblica Amministrazione. SPC è definito nel D. Lgs 7 marzo 2005 n. 82, negli artt. dal 72 all'87.

TASSONOMIA: una classificazione gerarchica di concetti ed elementi che consente di rappresentare un dominio di conoscenza.

UNICODE: lo standard per la codifica di caratteri che copre un vastissimo repertorio di caratteri comprendente tutti i sistemi di scrittura.

URI (Uniform Resource Identifier): stringa di caratteri che identifica univocamente una risorsa (pagina web, documento, immagine, file, ecc).

URL (Uniform Resource Locator): stringa di caratteri che identifica una risorsa su Internet, ne specifica formalmente la collocazione e indica il protocollo da utilizzare per accedervi. È un tipo specifico di URI.

VOCABOLARIO DEI DATI: l'insieme dei possibili valori che le entità di una classe possono assumere all'interno di un dominio di conoscenza o di una ontologia.

W3C (WORLD WIDE WEB CONSORTIUM) [80]: il consorzio internazionale che ha lo scopo di definire gli standard aperti per il Web.

WEB SEMANTICO: insieme di modelli e standard Web in cui le risorse vengono descritte e correlate fra loro in modo formale attraverso l'uso opportuno di metadati. In questo modo si abilitano gli agenti automatici a comprendere il significato dei dati e delle informazioni

XML (EXTENSIBLE MARKUP LANGUAGE): meta-linguaggio di markup (rappresentazione) di testi che è stato standardizzato dal W3C. Costituisce inoltre una delle forme sintattiche per RDF e OWL.

1.6. Licenza

2. SCOPO, DESTINATARI E STRUTTURA DEL DOCUMENTO

Scopo. Il presente elaborato rappresenta un documento di LINEE GUIDA per la "*promozione dell'evoluzione del modello organizzativo e dell'architettura tecnologica del SPC in funzione del mutamento delle esigenze delle pubbliche amministrazioni e delle opportunità derivanti dalla evoluzione delle tecnologie*", come specificato all'art. 79, comma 2, lettera c) del CAD, che definisce i compiti della Commissione di Coordinamento SPC.

In linea con tali compiti, il documento propone e approfondisce un approccio metodologico per la produzione di open data interoperabili attraverso cui garantire l'interoperabilità semantica, così come definita dall'European Interoperability Framework (EIF). Il documento mira inoltre ad analizzare i relativi aspetti di interrogazione, pubblicazione, ricerca dei dati nonché gli aspetti legati alle licenze d'uso e ai modelli di business che si possono abilitare dalla produzione e dalla distribuzione di open data interoperabili.

Le linee guida si concentrano, in coerenza con il mandato SPC, su un indirizzo tecnico e metodologico utile alla creazione di Linked Open Data (LOD), e rispondono alle esigenze di interoperabilità, autenticità e qualità del dato nello specifico contesto delle infrastrutture condivise SPC.

Il documento sarà aggiornato alla luce dei possibili cambiamenti normativi e tecnologici in materia di (Linked) Open Data e interoperabilità semantica.

Destinatari. Ai sensi dell'art. 75 del CAD, che regolamenta la partecipazione al Sistema Pubblico di Connettività (SPC), il presente documento è destinato a tutte le amministrazioni, così come definite all'art. 2 comma 2 del CAD, nonché a tutti i gestori di servizi pubblici e ai soggetti che perseguono finalità di pubblico interesse (art. 75 comma 3-bis del CAD). In virtù della connotazione metodologica e tecnica dei contenuti, le presenti linee guida si rivolgono principalmente a tutte quelle figure professionali, sia dei soggetti prima citati sia di quelli esterni a loro supporto, in possesso di competenze tecnico-informatiche e competenze specifiche relative ai dati (ad esempio, direttori dei sistemi informativi delle pubbliche amministrazioni, tecnici di fornitori qualificati e consulenti tecnici).

Struttura del documento. Le sezioni del documento sono così strutturate. La sezione 3 introduce lo scenario, la normativa di riferimento e il contesto d'interoperabilità semantica come introdotto dall'European Interoperability Framework (EIF). La sezione 4 illustra alcuni ambiti di interesse per l'apertura dei dati (ad esempio, si analizzano le peculiarità e criticità di dati quali i dati territoriali, i dati elettorali, i dati scolastici, ecc.). La sezione 5 presenta lo stato dell'arte citando i principali lavori di riferimento condotti in campo internazionale e nazionale. La sezione 6, sulla base dell'analisi dello stato dell'arte e di alcune esperienze italiane, cui è dedicata una sezione sul sito Web istituzionale di DigitPA [81], presenta una metodologia generale per l'interoperabilità semantica basata sulla produzione di LOD. La sezione 0 illustra e classifica le tecnologie di base utilizzate per modellare, creare e utilizzare LOD, mentre la sezione 8 descrive le problematiche relative agli aspetti legali (licenze per l'uso dei dati) e i modelli di business legati alla qualità e al riuso del dato. La sezione 9 inquadra le tematiche affrontate nelle precedenti parti nel contesto del SPC indicando il ruolo specifico delle infrastrutture condivise

SPC e quello dei servizi e-government per la PA nell'attuazione dell'approccio metodologico proposto. La sezione 10, infine, presenta un'analisi della sostenibilità della metodologia proposta e della governance LOD nella Pubblica Amministrazione. In alcune sezioni, sono evidenziate le raccomandazioni individuate dal gruppo di lavoro.

Il documento, in una fase successiva, sarà corredato di un'appendice che illustrerà la "roadmap" di un'esperienza pratica di collegamento di LOD già pubblicati da alcune amministrazioni del gruppo di lavoro. L'esperienza ha l'ambizione di avviare le attività necessarie per la creazione della nuvola LOD SPC.

3. INTRODUZIONE

Grazie allo sviluppo delle tecnologie digitali e al crescente livello di informatizzazione della Pubblica Amministrazione (PA), in questi ultimi anni si è fatta strada la consapevolezza che i dati del settore pubblico (Public Sector Information - PSI) possano svolgere un ruolo importante non solo ai fini della trasparenza amministrativa e della partecipazione pubblica ma anche sul piano economico. Recenti studi europei e nazionali mirano a stimare il valore dell'informazione del settore pubblico in termini di partecipazione, trasparenza, e valore economico che si può ottenere grazie a un più facile accesso e a un più ampio uso di tale informazione [2] (ulteriori approfondimenti in merito agli aspetti legali e al trattamento dei dati del settore pubblico lettori si trovano anche in [83], [84], [85]).

Le leggi approvate dalla Comunità Europea si sono focalizzate inizialmente sul diritto del cittadino di accedere ai dati per una questione di trasparenza. Successivamente, l'attenzione si è estesa alla questione della partecipazione e della messa a disposizione dei dati secondo modalità che permettessero di riutilizzarli liberamente, in contesti diversi e innovativi rispetto a quelli nei quali essi sono stati raccolti. Questa tendenza, già presente in documenti strategici quali l'Agenda Digitale, è stata ulteriormente e ripetutamente confermata. Anche nel corso della conferenza stampa tenuta dalla commissaria Neelie Kroes il 12 dicembre 2011, è stato ribadito che i dati della PA sono in grado di generare ricchezza e posti di lavoro.

Queste norme comunitarie, recepite dai vari Stati e dalle Amministrazioni locali, cominciano a essere finalmente messe in atto. Stimoli molto forti in questa direzione provengono dal memorandum di Obama del 2009[1] (questo più attento alla trasparenza e alla partecipazione) e le successive decisioni del governo inglese prese in materia di riuso di dati pubblici e riassetto della PA. L'esempio fornito da questi grandi attori a livello internazionale ha fatto sì che molte pubbliche amministrazioni abbiano superato alcune delle riserve importanti nei confronti di quello che è divenuto un vero e proprio movimento, qual è appunto quello degli Open Data. Si osservava infatti, e si osserva in parte tuttora, il timore delle pubbliche amministrazioni di contravvenire a qualche legge, di essere accusati di svendere beni pubblici e di esporre dati di bassa qualità o inaffidabili. In coloro che hanno cominciato a pubblicare i dati si osserva ancora talvolta un eccesso di prudenza (ad esempio nella scelta della licenza d'uso) che finisce per limitare in modo sostanziale il riuso in contesti commerciali.

In generale, quando si parla di **dati** s'intende rappresentazioni elementari, spesso codificate, di "cose", avvenimenti o altro. Questi normalmente sono parti di una informazione o conoscenza strutturata che può essere codificata e archiviata in un formato digitale. Esempi pratici di dati in possesso delle pubbliche amministrazioni sono gli atti ufficiali, le spese e i dati di bilancio, le presenze e gli stipendi dei dipendenti, i quali hanno un impatto sulla trasparenza e la partecipazione democratica. Ma anche le informazioni geografiche, le statistiche, i dati ambientali, le informazioni economiche e giuridiche rappresentano una risorsa preziosa per il tessuto economico di un territorio perché utilizzabili per la creazione di servizi innovativi e di nuovi prodotti.

Con il termine **Open Data** si introduce un nuovo paradigma nella gestione dei dati: questi, tipicamente

1 http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment

"nascosti" in applicazioni o basi di dati, sono resi accessibili a chiunque abbattendo, per quanto possibile e ragionevole, le restrizioni tecnologiche ed imponendo vincoli legali minimi al riuso dei dati. Sono naturalmente esclusi da questa categoria i dati vincolati da leggi e norme quali ad esempio quelle del diritto alla privacy (Sezione 3.1), della tutela della proprietà intellettuale, del segreto di Stato, del segreto statistico, della flora e fauna protetta, ecc.

Non esiste una definizione unica di Open Data; quella cui normalmente si fa riferimento è stata promossa da Open Knowledge Foundation che, nella sua versione breve, recita "un contenuto o un dato si definisce aperto se chiunque è in grado di utilizzarlo, ri-utilizzarlo e ridistribuirlo, soggetto, al massimo, alla richiesta di attribuzione e condivisione allo stesso modo" [3]. Altre definizioni sono state introdotte per ambiti più specifici come i "Panton principles" in ambito scientifico o declinati in base alla provenienza, come nel caso gli Open Government Data con cui si indicano gli open data della PA.

La definizione di **Open Government Data**, elaborata per la prima volta da un gruppo di lavoro riunitosi a Sebastopoli, California nel 2007, definisce 8 principi in base ai quali si può valutare se un dato della PA è da considerarsi aperto [4]. Un'altra definizione è quella fornita dalla Open Knowledge Foundation per la quale gli "Open Government Data" sono dati che 1) "possono essere liberamente usati, riusati e distribuiti da chiunque (come definito al sito della Open Definition) 2) "sono prodotti o commissionati dalla PA o da entità controllate".

Tutte le definizioni presentano un'ispirazione comune: il dato deve essere accessibile, preferibilmente via Internet, eventualmente a un costo marginale e in un formato modificabile; deve essere libero da vincoli tecnologici che ne limitino di fatto la più ampia diffusione; eventuali vincoli legali non devono pregiudicare la possibilità di distribuzione e riuso; non devono esserci discriminazione all'uso contro persone o gruppi, campi di indagine e destinazione. In talune istanze, le definizioni si soffermano su alcune specificità: secondo quella data a Sebastopoli "i dati devono essere pubblicati come sono stati raccolti alla fonte, con il livello di granularità più fine possibile, non in forme aggregate o modificate" e "i dati devono essere resi disponibili il prima possibile per preservarne il valore".

La disponibilità e il rilascio di Open Data è, come detto, un patrimonio prezioso per la società civile e per le imprese, ma affinché si possa valorizzare del tutto l'informazione, l'apertura da sola non basta. È desiderabile rendere gli Open Data autodescrittivi e poter inferire conoscenza dall'aggregazione e correlazione di dataset differenti. Occorre inoltre favorirne la facilità di uso, il reperimento e il consumo sia per gli esseri umani che per i software automatici. In tempi recenti diversi sforzi di ricerca e sviluppo hanno identificato nelle tecnologie del Web Semantico [5], e in particolare nel modello dei **Linked Open Data**, interessanti opportunità per superare le limitazioni dei modelli Open Data. Mentre in generale gli Open Data abbattono le barriere culturali, legali ed economiche al riuso, il movimento Linked Data si concentra piuttosto sulla messa a punto di strumenti che permettono di dare ai dati (aperti o non) un'identità e di renderli collegati tra loro e interoperabili. Il problema dell'interoperabilità deriva in parte proprio dall'identità: se non si sa come "risolvere" l'identità di una "cosa" fra i diversi sistemi che ne "parlano", è molto difficile aggregare le informazioni ad essa relative. Il problema dell'integrazione fra dati di sistemi diversi è reso ancora più complesso dal fatto che ogni sistema si basa tipicamente su infrastrutture eterogenee: diversi linguaggi, formati, protocolli, ecc. È possibile, infatti, che a una stessa "cosa" sistemi differenti assegnino identità differenti. Pertanto, queste andrebbero allineate per continuare a garantire l'interoperabilità tra sistemi eterogenei.

I Linked Data sono stati proposti nel 2006 da Tim Berners-Lee come metodo elegante ed efficace [6] per semplificare e omogeneizzare le soluzioni proprio ai problemi di identità e interoperabilità. Il

metodo consiste dei seguenti quattro principi base:

1. Usare indirizzi Web (URI) come nomi per le "cose";
2. Usare URI utili al protocollo HTTP in modo che sia possibile cercare e risolvere quei nomi;
3. Quando qualcuno cerca una URI, fornire un'informazione utile;
4. Includere link ad altre URI, così da permettere a chi cerca di scoprire nuovi collegamenti.

Adottare modelli, tecnologie e standard aperti di Linked Data (e.g., RDF), sfruttando le migliori esperienze maturate nell'ambito del Web Semantico [5], offre benefici di sicuro interesse per utenti e sviluppatori. I primi acquisiscono la possibilità di riferirsi a entità specifiche, anziché a posizioni all'interno di un database, e di navigare tra i dati; i secondi possono realizzare applicazioni, anche complesse, che combinano i dati della PA con altri, aprendoli anche alla possibilità di arricchimento automatico attraverso il cosiddetto "ragionamento automatico".

Questo documento di linee guida si propone di essere quindi un aiuto alle pubbliche amministrazioni per compiere quest'ulteriore passo verso i Linked Data (classificati a livello 5 nello schema di Tim Berners-Lee [6]) e per far sì che i dati della PA italiana, entrino a far parte del cosiddetto "Web of Data" (Web dei dati). In questa visione, il Web diventa uno spazio dati distribuito in cui molteplici informazioni, rispetto ad entità e fatti, sono connesse fra di loro tramite collegamenti semantici così da facilitare lo sviluppo di applicazioni innovative e il supporto ad interrogazioni che coinvolgono una molteplicità di basi di dati distribuite. Questa facilità di accesso e riuso amplifica il valore dei dati stessi in quanto essi non saranno utilizzati più da una sola applicazione o da un solo sito Web, ma da tantissime possibili applicazioni e per gli usi più diversi.

Da qualche tempo un numero crescente di amministrazioni, rendendosi conto del valore aggiunto dei Linked Data (anche in termini di interoperabilità interna alla PA stessa), si sta facendo carico non solo del lavoro necessario per pubblicarli come dati aperti ma anche di quello di pubblicarli direttamente in modalità "linked". L'esempio per eccellenza in Europa in questo settore è quello del governo inglese [7], che è stato un precursore nel pubblicare i suoi dati come Linked Data.

Prende corpo dunque uno scenario nel quale la PA, anche in virtù della sua possibilità di definire norme, può svolgere un ruolo da protagonista in un sistema complesso di aziende, comunità e cittadini.

3.1. Normativa di riferimento

Al momento della prima stesura del presente documento, in Italia non esisteva ancora una norma primaria dello Stato specifica per il trattamento degli Open Data. Diversi dibattiti, all'interno di alcuni gruppi di lavoro della Cabina di Regia per l'Agenda Digitale Italiana (ad esempio quello su e-government e Open Data), avevano sollevato la necessità e l'opportunità di introdurla. Come risultato di tali dibattiti, tra Giugno e Ottobre 2012, sono state approvate due norme che trattano nello specifico di Open Data. La prima è il decreto legge n. 83, convertito nella legge n. 134 del 7 agosto 2012, che all'art. 18 – Amministrazione aperta, sancisce per la prima volta in Italia l'obbligo di pubblicare online, entro il 1° gennaio 2013, i dati relativi a "*concessione delle sovvenzioni, contributi, sussidi ed ausili finanziari alle imprese e l'attribuzione dei corrispettivi e dei compensi a persone, professionisti, imprese ed enti privati e comunque di vantaggi economici di qualunque genere […]*". Tali dati "*devono essere resi di facile consultazione, accessibili ai motori di ricerca ed in formato tabellare aperto che ne consente l'esportazione, il trattamento e il riuso ai sensi dell'articolo 24 del*

decreto legislativo 30 giugno 2003, n. 196".

La seconda norma è l'ancor più recente decreto legge Crescita 2, non ancora convertito in legge al momento della pubblicazione del presente documento v.2.0, che alla Sezione II e in particolare all'art 9 – dati di tipo aperto e inclusione sociale, indica le modifiche apportate al CAD relative al dato pubblico (in particolare gli artt. 52 e 68 così come sotto riportato) definendo nello specifico cosa si intende per formato di dati di tipo aperto ("*reso pubblico, documentato esaustivamente e neutro rispetto agli strumenti necessari per la fruizione dei dati stessi*") e identificando le caratteristiche dei dati di tipo aperto ovvero:

1) "*disponibili con una licenza che ne permette l'utilizzo da parte di chiunque, anche per finalità commerciali*";
2) "*accessibili attraverso le tecnologie dell'informazione e della comunicazione in formati aperti, adatti all'utilizzo automatico da parte di programmi per elaboratori e provvisti dei relativi metadati";*
3) "*resi disponibili gratuitamente attraverso le tecnologie dell'informazione e della comunicazione, oppure resi disponibili ai costi marginali sostenuti per la loro riproduzione e divulgazione".*

Infine, sempre l'art. 9 del suddetto decreto stabilisce che *"l'Agenzia definisce e aggiorna annualmente le linee guida nazionali che individuano gli standard tecnici, compresa la determinazione delle ontologie dei servizi e dei dati, le procedure e le modalità di attuazione […]".*

Il CAD (D. Lgs. n. 82/2005 e s.m.i) riserva diversi articoli sui temi della disponibilità, accesso, pubblicazione e riuso dei dati delle PA. L'art. 50 afferma concetti base sulla conservazione, disponibilità accessibilità e sul riutilizzo dei dati sottolineando che "*i dati delle pubbliche amministrazioni sono formati, raccolti, conservati, resi disponibili e accessibili con l'uso delle tecnologie dell'informazione e della comunicazione che ne consentano la fruizione e riutilizzazione, alle condizioni fissate dall'ordinamento, da parte delle altre pubbliche amministrazioni e dai privati; restano salvi i limiti alla conoscibilità dei dati previsti dalle leggi e dai regolamenti, le norme in materia di protezione dei dati personali ed il rispetto della normativa comunitaria in materia di riutilizzo delle informazioni del settore pubblico*". L'art. 52, ai commi 1 e 1-bis, regolamenta l'accesso telematico ai dati che deve essere messo in atto seguendo le disposizioni dello stesso Codice e della normativa vigente in materia di protezione dei dati personali (i.e., D.Lgs. n. 196/2003 e deliberazione del 88/2011 dell'Autorità Garante per la protezione dei dati personali) evidenziando la valorizzazione e la fruizione dei dati pubblici di cui le amministrazioni sono titolari mediante la promozione di "*progetti di elaborazione e di diffusione degli stessi anche attraverso l'uso di strumenti di finanza di progetto, assicurando […] b) la pubblicazione dei dati e dei documenti in formati aperti di cui all'articolo 68, commi 3 e 4".* Quest'ultimo art. 68, ai commi 2, 3 e 4, afferma che le pubbliche amministrazioni "*consentono la rappresentazione dei dati e documenti in più formati, di cui almeno uno di tipo aperto, salvo che ricorrano motivate ed eccezionali esigenze",* definendo al comma 3 il formato aperto "*un formato dati reso pubblico e documentato esaustivamente"* e assegnando a DigitPA il ruolo di "*istruire e aggiornare, con periodicità almeno annuale, un repertorio dei formati aperti utilizzabili nelle pubbliche amministrazioni e delle modalità di trasferimento dei formati".*

Infine l'art. 54, che disciplina i contenuti dei siti pubblici delle pubbliche amministrazioni, rafforza al comma 3 e al comma 4 concetti quali l'immediata fruibilità dei dati pubblici e l'autenticità della fonte affermando che i "*dati pubblici contenuti nei siti delle pubbliche amministrazioni sono fruibili in rete gratuitamente e senza necessità di identificazione informatica"* e che le pubbliche amministrazioni "*garantiscono che le informazioni contenute sui siti siano conformi e corrispondenti alle informazioni contenute nei provvedimenti amministrativi originali dei quali si fornisce comunicazione tramite il sito".*

Il concetto di interoperabilità semantica è richiamato più volte nel DPCM del 1° Aprile 2008 sulle

regole tecniche del SPC. L'art. 10 sull'economicità nell'utilizzo dei servizi di rete, di interoperabilità e di supporto alla cooperazione applicativa afferma che si devono attuare misure che favoriscono "*l'integrazione delle informazioni attraverso una rappresentazione semantica condivisa*" e l'art. 15 elenca l'insieme delle infrastrutture condivise SPC per l'interoperabilità e la cooperazione applicativa introducendo al comma 4 lettera b) tra i servizi infrastrutturali di cooperazione applicativa (SICA) il Servizio di Catalogo Schemi/Ontologie. Il servizio consente la descrizione degli elementi semantici associati ai servizi applicativi e alle informazioni gestite dalle PA, "*anche ai fini dell'individuazione automatica dei servizi disponibili per l'erogazione delle prestazioni richieste, e la condivisione tra le Amministrazioni cooperanti degli schemi di dati e metadati, nonché delle ontologie di dominio*".

Se per i "dati pubblici" il CAD e le opportunità aperte dagli Open Data hanno rappresentato una svolta radicale verso nuove regole di accesso e diffusione, in ambito statistico l'attenzione alla fornitura, diffusione e accesso alle informazioni da parte degli utenti preesiste rispetto al concetto dell'Open Data, pur se in origine con specifico riferimento al concetto di dato aggregato.

I dati statistici rimandano direttamente ai concetti di "collettivo", "unità statistiche" e "carattere" e possono essere definiti in chiave tecnica come "*il risultato dell'operazione di determinazione della modalità con cui un carattere è presente in ciascuna unità statistica del collettivo*[2]" e in chiave istituzionale come "*dati provenienti da uno studio o da una sorgente amministrativa, utilizzati per produrre statistiche e/o dati che comprendono tali statistiche*"[3].

In questo quadro, nell'ambito dei dati statistici esistono da tempo specifiche normative di riferimento, quali ad esempio il D. Lgs. 6 settembre 1989, n. 322 - Norme sul Sistema statistico nazionale e sulla riorganizzazione dell'Istituto nazionale di statistica, nel quale viene evidenziata e ripresa più volte la necessità di interconnettere, a fini statistici, i sistemi informativi delle Pubbliche Amministrazioni e degli enti facenti parte del Sistema statistico nazionale, garantendo l'accesso alle informazioni prodotte dalle singole amministrazioni. Tale specifico riferimento è presente nell'art. 6, del D. Lgs. n. 322/1989 che disciplina i compiti degli uffici di statistica che: *"a) promuovono e realizzano la rilevazione, l'elaborazione, la diffusione e l'archiviazione dei dati statistici che interessano l'amministrazione di appartenenza, nell'ambito del programma statistico nazionale; b) forniscono al Sistema statistico nazionale i dati informativi previsti del programma statistico nazionale relativi all'amministrazione di appartenenza, anche in forma individuale ma non nominativa ai fini della successiva elaborazione statistica; c) collaborano con le altre amministrazioni per l'esecuzione delle rilevazioni previste dal programma statistico nazionale; d) contribuiscono alla promozione e allo sviluppo informatico a fini statistici degli archivi gestionali e delle raccolte di dati amministrativi."*

Gli uffici inoltre *"attuano l'interconnessione ed il collegamento dei sistemi informativi dell'amministrazione di appartenenza con il Sistema statistico nazionale."*.

Per i compiti suddetti, *"gli uffici di statistica hanno accesso a tutti i dati statistici in possesso dell'amministrazione di appartenenza, salvo eccezioni relative a categorie di dati di particolare riservatezza espressamente previste dalla legge. Essi possono richiedere all'amministrazione di appartenenza elaborazioni di dati necessarie alle esigenze statistiche previste dal programma statistico nazionale"*. Inoltre, in base poi all'art. 10 che regolamenta l'accesso ai dati statistici, il decreto prevede che: *"I dati elaborati nell'ambito delle rilevazioni statistiche comprese nel programma statistico nazionale sono patrimonio della collettività e vengono distribuiti per fini di studio e di ricerca a coloro che li richiedono […]. Sono distribuite altresì, ove disponibili, su richiesta motivata e previa autorizzazione del presidente dell'ISTAT,*

[2] http://www3.istat.it/servizi/studenti/binariodie/CorsoExcel/Glossario.htm
[3] http://stats.oecd.org/glossary/detail.asp?ID=2543

collezioni campionarie di dati elementari, resi anonimi e privi di ogni riferimento che ne permetta il collegamento con singole persone fisiche e giuridiche. […]. Enti od organismi pubblici, persone giuridiche, società, associazioni e singoli cittadini hanno il diritto di accedere ai dati […] facendone richiesta agli uffici […]. I dati, se non immediatamente disponibili, vengono consegnati ai richiedenti nel tempo strettamente necessario per la riproduzione, con rimborso delle spese, il cui importo è stabilito dall'ISTAT […].

A livello locale, diverse regioni (e.g., Piemonte, Lazio, Puglia, ecc.), province autonome (e.g., provincia di Trento) e comuni (e.g., Firenze, Bologna, Venezia, ecc.) recentemente hanno emanato delibere locali allo scopo di regolamentare e istituzionalizzare la produzione e la pubblicazione di dati aperti delle PA. Una delle prime regioni italiane è stata la Regione Piemonte che ha avviato la predisposizione di strumenti per promuovere il riuso professionale dei dati regionali da parte di privati. Nel 2005, in particolare, la Regione Piemonte ha stipulato il Protocollo d'Intesa per la condivisione, la valorizzazione e la diffusione del Patrimonio Informativo Regionale, [8]. Il protocollo sottolinea l'importanza della valorizzazione del patrimonio informativo, al fine di creare le condizioni di mercato più favorevoli e competitive, e suggerisce modalità di condivisione con gli operatori economici per stimolare la creazione di nuovi servizi basati sui contenuti digitali. Nel giugno 2009, con Delibera di Giunta 31 - 11679 del 29 giugno 2009 la stessa Regione ha definito le linee guida [9] regionali per i processi di riuso, relative alla definizione delle licenze standard da concedere in funzione della tipologia dei dati regionali messi a disposizione e delle categorie di utenza. Con la Delibera di Giunta regionale 36 - 1109 del 30 novembre 2010, è stata approvata una nuova versione delle linee guida [10]: con esse diventano riusabili, con licenza CC0 (Sezione 8.1), tutti i dati di tipo aggregato/anonimo, o senza vincoli di privacy, in possesso della Regione. Tutti i dati pubblicati e scaricabili dal sito istituzionale, o da altri canali istituzionali della Regione, sono inoltre rilasciati con licenza standard CC0.

Nel contesto europeo, sia l'Agenda Digitale Europea [1], relativamente al pilastro del mercato unico (Action 3), che la conseguente revisione della direttiva 2003/98/CE del Parlamento europeo e del Consiglio stabiliscono norme minime per l'apertura e il riutilizzo dell'informazione del settore pubblico nell'Unione europea. Tali iniziative e direttive incoraggiano gli Stati membri a spingersi oltre tali norme per adottare politiche sempre meno restrittive in materia di dati e che consentano così un più ampio utilizzo di dati in possesso degli organismi del settore pubblico. In linea con le iniziative già avviate in ambito europeo, l'Agenda Digitale Italiana ha dedicato particolare attenzione a questo ambito costituendo un gruppo di lavoro su e-government e open data. I principali risultati di tale gruppo di lavoro saranno resi noti il 30 settembre prossimo con la pubblicazione della relazione strategica dell'Agenda Digitale Italiana [11].

3.2. Interoperabilità semantica nell'European Interoperability Framework

Secondo quanto previsto dall'Agenda Digitale Europea [1], ogni paese membro dell'UE deve allineare, entro il 2013, il proprio framework di interoperabilità nazionale (nel caso italiano, come definito dal CAD, il Sistema Pubblico di Connettività e Cooperazione) al framework europeo di interoperabilità (i.e., l'EIF - European Interoperability Framework).

Il modello EIF si basa su tre livelli di interoperabilità: *l'interoperabilità tecnica, l'interoperabilità semantica e l'interoperabilità organizzativa*; a questi si aggiungono altre due dimensioni, ossia il contesto politico e

l'interoperabilità legale. Il presente documento focalizza la sua attenzione sul concetto di interoperabilità semantica; per ulteriori dettagli relativamente a ciascun livello d'interoperabilità si rimanda pertanto al documento in [12].

L'EIF evidenzia come l'interoperabilità semantica offra alle organizzazioni la possibilità di elaborare informazioni da fonti esterne o secondarie senza perdere il reale significato delle informazioni stesse nel processo di elaborazione; il significato viene preservato affinché persone, istituzioni e applicazioni tutte possano efficacemente comprenderlo. In altre parole, l'interoperabilità semantica coinvolge, all'interno di settori specifici, la definizione di insiemi comuni di schemi di dati e protocolli.

Nella pratica, l'EIF indica quali sono le azioni che gli Stati membri devono intraprendere per garantire l'interoperabilità semantica nella cooperazione e nell'interscambio tra organizzazioni. Tali azioni prevedono la definizione di:

- strumenti generali, e specifici di settore, a supporto della condivisione delle informazioni;
- principi base chiari e condivisi per la gestione d'informazioni detenute dai governi;
- un insieme comune di schemi di dati e protocolli, e di tassonomie per settori specifici; a tal riguardo i framework nazionali di interoperabilità dovrebbero considerare anche la natura transfrontaliera dell'interoperabilità semantica nella definizione e nello sviluppo di tali tassonomie e adottare quindi rappresentazioni standard dei dati che possano essere utilizzate e comprese non solo in un contesto nazionale ma anche in un contesto di interscambio internazionale;
- protocolli per la condivisione e il riuso delle informazioni attraverso settori pubblici e privati;
- processi per la gestione del ciclo di vita delle informazioni delle organizzazioni.

Per attuare tali azioni, l'EIF pone sì l'accento sulla necessità di una dimensione verticale, specifica per settori della PA, ma anche su una dimensione orizzontale, costituita da un insieme di servizi comuni infrastrutturali. Alcuni di questi servizi infrastrutturali possono essere progettati per raccogliere le tassonomie e i vocabolari dei dati, ed erogare strumenti di rappresentazione a supporto dell'interoperabilità semantica e della fruibilità dei dati anche in processi eterogenei che coinvolgono diversi attori.

Nell'ambito ISA, il programma della Commissione Europea che mira a fornire soluzioni che possano abilitare una collaborazione transfrontaliera efficace ed efficiente tra pubbliche amministrazioni europee, sono state avviate una serie di "action" sui temi dell'interoperabilità. In particolare, l'"action 1.1" si è concentrata sullo studio delle problematiche legate all'interoperabilità semantica come precedentemente definita nel contesto EIF, arrivando a formulare un insieme di raccomandazioni e a rendere disponibili schemi comuni, vocabolari di base, e licenze relativamente a metadati e asset semantici forniti nel contesto della pubblica amministrazione. Per dare al lettore una visione più dettagliata delle attività svolte in questo ambito, i risultati finora raggiunti nell'"action" suddetta sono riassunti nelle successive sezioni 5 e 8 del presente documento.

4. AMBITI DI APPLICAZIONE

Sin dall'inizio del movimento globale Open Government, l'utilizzo dei dati esposti dalle PA è stato promosso e stimolato anche attraverso l'organizzazione di competizioni e "civic hackathons", eventi che riuniscono aziende IT, programmatori e progettisti nel campo dello sviluppo di software chiamati a sfidarsi per realizzare le applicazioni più interessanti. Il mondo degli Open Data, tuttavia, comprende fenomeni molto più complessi e dai risvolti difficilmente prevedibili. Si pensi, per esempio, alle ricadute sul piano della formazione scolastica, in cui s'insegna agli studenti a interagire con la PA attraverso strumenti innovativi di lettura del suo operato, non solo grafici, quindi, e che permettono anche la navigazione dei dati pubblicati da un determinato Ente. Si possono poi innescare meccanismi per cui i cittadini stessi possono proporre "apps", o elaborazioni ed interpretazioni originali dei dati, potenzialmente di grande interesse e rilevanza per la società.

Un aspetto interessante da considerare, solitamente collegato all'apertura dei primi Open Data di una PA, è la possibilità per quest'ultima di pubblicare dati che già erano presenti all'interno del proprio sito istituzionale, in forme o in aree diverse del sito stesso. Va sottolineato come questo non rappresenti una duplicazione della pubblicazione del dato. Se da un lato è vero, infatti, che la pubblicazione di dati sulla trasparenza, la performance o il bilancio risponde a esigenze discendenti da precisi obblighi di legge, e a criteri definiti dalle Linee Guida sui Siti Web della PA [16], è altrettanto vero che è utile che essi siano esposti in formati aperti e facilmente processabili, per scopi e utenze diverse, grazie all'individuazione di specifiche licenze di riuso (sezione 8.1), tipicamente non associate ai dati pubblicati per motivi di trasparenza sui siti istituzionali. I portali di aggregazione regionali, nazionali o trans-nazionali indicizzano e diffondono infatti i diversi dataset esposti da singole amministrazioni a partire dalle loro sezioni dedicate agli Open Data, e non dai loro interi portali istituzionali. Inoltre, la distribuzione dei dati associata a una licenza promuove e semplifica il riuso dei dati poiché tutela l'utilizzatore permettendogli di pubblicare a sua volta i risultati di analisi originali, o lo sviluppo di servizi e applicazioni di interesse per la società.

Sulla base di queste considerazioni e seguendo la definizione di dati aperti così come riportata nell'introduzione del presente documento, possono essere individuate diverse tipologie di dati in possesso delle pubbliche amministrazioni, con le quali le PA possono efficacemente rispondere ai requisiti di trasparenza, responsabilità e di sviluppo di applicazioni e servizi.

L'elenco e la descrizione delle tipologie di dati di seguito riportati non intendono essere esaustivi ma rappresentano il punto di convergenza del gruppo di lavoro. Sono state individuate le seguenti tipologie:

- Dati territoriali;
- Dati ambientali;
- Dati relativi al personale della PA;
- Dati scolastici e universitari;
- Dati della ricerca e competenze;

- Classificazioni e dati statistici.

La scelta delle suddette tipologie è motivata da ragioni diverse: alcuni dati sono esplicitamente previsti nell'elenco di dati pubblici da rendere disponibili attraverso i siti istituzionali della PA (ad esempio, dati relativi al personale della PA, secondo quanto sancito dall'art. 54 del CAD), altri sono inclusi in banche dati di interesse nazionale (e.g., i dati territoriali), e altri ancora sono considerati di particolare interesse in virtù dell'impatto sociale che potrebbero avere, come nel caso dei dati scolastici, universitari e della ricerca, le classificazioni e i dati statistici.

Per ognuna delle categorie sopra elencate si riportano le relative peculiarità, gli standard e i formati di rappresentazione che vengono ad oggi utilizzati per la loro rappresentazione, eventuali restrizioni normative sull'uso, i vantaggi derivanti dall'apertura dei dati, e le eventuali criticità riscontrabili con riferimento alla loro trasformazione in formati Linked.

Dati territoriali:

L'art. 59 del Codice dell'Amministrazione Digitale (CAD) definisce i dati territoriali come "*qualunque informazione geograficamente localizzata*"; una definizione, questa, coerente con quella riportata all'art. 3 della Direttiva INSPIRE[4] secondo cui i dati territoriali sono "*i dati che attengono, direttamente o indirettamente, a una località o un'area geografica specifica*".

Le PA, a tutti i livelli di governo (centrale, regionale e locale) acquisiscono e trattano in modo sistematico, nell'esercizio delle proprie funzioni, una gran mole di dati di natura geografica. I dati territoriali sono infatti fondamentali in tutte le attività di pianificazione e gestione del territorio, nonché di coordinamento. Esempi di dati territoriali includono: le cartografie tematiche (geologica, idrogeologica, geomorfologica), i limiti amministrativi (regionali, provinciali, comunali, municipali, unità urbanistiche, sezioni di censimento), le carte delle aree soggette a vincoli o carte del rischio (aree d'interfaccia boschiva, aree percorse dal fuoco, aree a rischio incendio, aree esondabili ed allagate, zone a rischio frane), lavori pubblici, barriere architettoniche.

Un elenco delle tipologie di dati territoriali è fornito nell'allegato 1 del DM 10 novembre 2011 relativo alle regole tecniche del Repertorio Nazionale dei Dati Territoriali (RNDT).

I dati territoriali, oltre a consentire lo sviluppo di nuovi servizi, concorrono a formare decisioni utili per lo sviluppo di quasi tutte le attività economiche.

Per quanto riguarda la condivisione e l'accesso ai dati territoriali, non meno importanti, inoltre, sono gli aspetti legati alla trasparenza, alla partecipazione e alla democrazia che derivano dalla possibilità, da parte dei cittadini, di conoscere gli elementi e gli aspetti fondamentali del proprio territorio e di divenire soggetti attivi nelle politiche e nei processi decisionali e di pianificazione urbanistica e territoriale.

Se negli ultimi anni le Amministrazioni hanno notevolmente sviluppato i sistemi per la gestione dei propri dati territoriali, la loro azione è oggi maggiormente orientata a rendere tale patrimonio informativo accessibile e disponibile in modo generalizzato e interoperabile. Non mancano tuttavia le difficoltà: la complessità dei dati territoriali, gli elevati costi associati ai processi di acquisizione dei dati

[4] Direttiva 2007/2/CE del Parlamento Europeo e del Consiglio del 14 marzo 2007 che istituisce un'Infrastruttura per l'Informazione territoriale nella Comunità Europea (INSPIRE).

stessi e alla realizzazione di sistemi evoluti di fruizione costituiscono sicuramente un fattore critico che ne rallenta l'apertura. Inoltre, i dati territoriali, anche simili ma trattati da amministrazioni diverse, per la realizzazione, ad esempio, di carte tecniche, piani regolatori, carte di vincolo, sono spesso disomogenei per contenuto, precisione, formati e documentazione. I potenziali utenti, infine, quali cittadini, piccole imprese, professionisti e anche enti locali di piccole dimensioni, spesso non sanno bene a chi rivolgersi per ottenerli o come interpretare i dati ricevuti.

Per risolvere le problematiche suddette, il recentissimo decreto spending review n. 95, convertito in legge n. 135 del 7 agosto del 2012, ha regolamentato all'art. 23 comma 12-quaterdecies la fruibilità di dati geospaziali acquisiti con risorse pubbliche. Inoltre, il CAD, già nel 2005, ha istituito un apposito Comitato[5] con il compito di "*definire le regole tecniche per la realizzazione delle basi dei dati territoriali, la documentazione, la fruibilità e lo scambio dei dati stessi tra le pubbliche amministrazioni centrali e locali in coerenza con le disposizioni del presente decreto che disciplinano il sistema pubblico di connettività*".

Il CAD ha istituito, inoltre, il Repertorio Nazionale dei Dati Territoriali (RNDT) allo scopo di "*agevolare la pubblicità dei dati di interesse generale, disponibili presso le pubbliche amministrazioni a livello nazionale, regionale e locale*". Il Repertorio rappresenta il catalogo nazionale dei metadati su dati territoriali e relativi servizi e si configura, altresì, come registro pubblico di tali dati certificandone l'esistenza. Esso è basato sugli Standard ISO 19115, 19119 e TS 19139 ed è coerente con il Regolamento INSPIRE sui metadati[6].

Le prime regole tecniche, elaborate dal Comitato, sono state pubblicate con i Decreti del Ministro per la pubblica amministrazione e l'innovazione, di concerto con il Ministro dell'ambiente e della tutela del territorio e del mare del 10 novembre 2011[7] . Esse hanno riguardato:

- formazione, documentazione e scambio di ortofoto digitali alla scala nominale 1:10.000;
- adozione del sistema di riferimento geodetico nazionale;
- definizione delle specifiche di contenuto dei database geotopografici;
- definizione del contenuto del Repertorio nazionale dei dati territoriali, nonché delle modalità di prima costituzione e di aggiornamento dello stesso.

Le problematiche esistenti nel contesto italiano, riconducibili a mancanza di disponibilità, qualità, organizzazione, accessibilità e condivisione dei dati territoriali, sono comuni anche agli altri Paesi dell'Unione europea. Tanto che nel 2007 è stata emanata la citata Direttiva INSPIRE per la creazione di un'infrastruttura per l'informazione territoriale finalizzata a garantire l'interoperabilità attraverso l'armonizzazione, l'accesso e il riuso dei dati prodotti e/o gestiti dalle PA.

L'infrastruttura europea si basa sulle infrastrutture istituite a livello nazionale; in Italia, la Direttiva è stata recepita con il D. Lgs. n. 32/2010, che ha istituito l'infrastruttura nazionale per l'informazione territoriale. La Direttiva si applica agli insiemi di dati di cui alle categorie tematiche indicate negli allegati I, II e III. Secondo la Direttiva INSPIRE per interoperabilità si intende la possibilità di combinare dati e servizi provenienti da diverse fonti in tutta la Comunità Europea in modo consistente. Tale interoperabilità può essere ottenuta o attraverso l'armonizzazione dei dati esistenti o attraverso la loro trasformazione tramite i servizi previsti per utilizzarli nella infrastruttura europea. A tale scopo, quando

[5] Comitato nazionale per le regole tecniche sui dati territoriali delle Pubbliche Amministrazioni

[6] REGOLAMENTO (CE) N. 1205/2008 DELLA COMMISSIONE del 3 dicembre 2008 recante attuazione della direttiva 2007/2/CE del Parlamento europeo e del Consiglio per quanto riguarda i metadati

[7] Gazzetta Ufficiale n. 48 del 27 febbraio 2012 - Supplemento ordinario n. 37

possibile, si fa riferimento agli Standard e alle specifiche tecniche internazionali, in particolare agli Standard ISO della serie 19100.

Un apposito Regolamento[8], emanato in applicazione della Direttiva, fornisce le indicazioni per l'interoperabilità degli insiemi di dati e dei servizi. Tale Regolamento indica che:

- per garantire l'interoperabilità e l'armonizzazione tra le categorie tematiche di dati territoriali, è opportuno rispettare i requisiti in materia di tipi di dati comuni, individuazione degli oggetti territoriali, metadati per l'interoperabilità, modello generico di rete e altri concetti e norme che si applicano a tutte le categorie tematiche di dati territoriali;
- al fine di garantire l'interoperabilità e l'armonizzazione all'interno di una categoria tematica di dati territoriali, è opportuno utilizzare le classificazioni e le definizioni degli oggetti territoriali, i relativi attributi chiave e relazioni, i tipi di dati, gli ambiti dei valori e le norme specifiche che si applicano alla categoria tematica di dati territoriali interessata.

Inoltre, sono stati elaborati:

- un framework di sviluppo che, oltre a riassumere la metodologia da utilizzare per le specifiche dei dati, fornisce anche un insieme coerente di requisiti e raccomandazioni per conseguire l'interoperabilità;
- per ciascuna categoria tematica, una specifica dei dati coerente con la struttura definita nello Standard ISO 19131 "Geographic Information – Data product specifications". Le specifiche comprendono la documentazione tecnica dello schema di applicazione, i tipi di oggetti territoriali con le loro proprietà ed altre indicazioni sulle categorie tematiche, utilizzando sia il linguaggio naturale sia il linguaggio formale dello schema concettuale.

Garantire l'interoperabilità di dati territoriali e dei relativi servizi implica anche incrementarne e migliorarne l'utilizzo, anche in prospettiva di renderli aperti. In linea di massima, infatti, tutti i dati territoriali possono essere resi aperti, ma, per aumentarne il valore aggiunto, sempre in termini di opportunità di riutilizzo e di integrazione nelle infrastrutture e/o sistemi di gestione, la traduzione anche in Linked Data implica una valutazione e una selezione dei dati in funzione delle informazioni, anche non esplicitamente territoriali, che a essi possono essere relazionate.

È da tenere presente, infatti, che il dato territoriale, per sua natura, si presta a essere utilizzato per 'collegare', secondo il paradigma dei 'Linked Data', informazioni provenienti da fonti diverse, arricchendole, rendendole interoperabili e più facilmente fruibili nel contesto del Web Semantico, anche in sistemi che prevedono comunicazioni tra macchine. Il dato territoriale permette, in estrema sintesi, di collegare dati che afferiscono a domini differenti.

Attualmente, nell'ambito dei dati territoriali il formato più diffuso è lo Shapefile, che, pur essendo nato come formato proprietario, è diventato uno standard *de facto* essendo ormai utilizzato in modo aperto. Anche altri formati come KML, GML, GeoRSS, GeoJSON, e GeoTIFF sono utilizzati, aumentando il

[8] REGOLAMENTO (CE) N. 1089/2010 DELLA COMMISSIONE del 23 novembre 2010 recante attuazione della direttiva 2007/2/CE del Parlamento europeo e del Consiglio per quanto riguarda l'interoperabilità dei set di dati territoriali e dei servizi di dati territoriali

ventaglio di soluzioni applicative possibili che ne fanno uso. Ancora poco numerose sono le iniziative mirate all'apertura in formato Linked di tale tipologia di dati. Alcuni esempi di rilievo si registrano sempre in campo inglese [17][18] mentre in Italia sono state da poco avviate iniziative in tale contesto. In particolare in Regione Emilia Romagna sono stati creati componenti software e ontologie che rendono disponibili come Linked Open Data dati e metadati gestiti dalla propria infrastruttura di dati territoriali tramite un'applicazione riusabile[9].

Dati ambientali:

L'Art. 2 della direttiva 2003/4/CE definisce come "informazione ambientale" qualsiasi informazione disponibile riguardante lo stato degli elementi dell'ambiente quali ad esempio l'aria, l'atmosfera, l'acqua, il suolo, il territorio, il paesaggio ma anche parametri di misura della qualità dell'ambiente quali ad esempio le sostanze chimiche di origine non naturale, l'energia, il rumore, le emissioni inquinanti, che incidono, o possono incidere, sugli elementi dell'ambiente. Sono inoltre considerate informazioni ambientali le misure, quali le politiche, le disposizioni legislative, i piani, i programmi, gli accordi e le attività che incidono, o possono incidere, sugli elementi dell'ambiente e sui parametri di misura della qualità dell'ambiente, nonché le azioni e le attività intese a proteggere i suddetti elementi. Alcuni esempi di tale tipologia di dati sono:

- le rilevazioni sulla qualità dell'aria, effettuate dalle centraline dislocate nel territorio, riguardanti le concentrazioni di agenti inquinanti, come gli ossidi di azoto, l'ozono e le polveri sottili (es. PM10, PM2,5). Questi dati sono espressi nell'unità di misura: *numero di ore di superamento di una data soglia critica, definita in microgrammi al metro cubico, per uno specifico inquinante;*
- la quantità di acque prelevate, immesse nella rete idrica e consumate, nell'anno solare. Questo dato è espresso nell'unità di misura: *metri cubi all'anno;*
- i consumi idrici civili totali, nell'anno solare, distinti in utenze domestiche e di servizio. Questi dati sono espressi nell'unità di misura: *metri cubi all'anno, distinti tra utenze domestiche, di servizio e usi pubblici (es. giardini, fontane, scuole);*
- la depurazione delle acque reflue, coi dettagli su eventuali fermi d'impianto, recanti i valori della domanda chimica di ossigeno (COD). L'unità di misura della domanda chimica d'ossigeno – dato medio – è milligrammi di ossigeno per litro d'acqua da trattare (parametro misurato prima dell'ingresso al depuratore) e per quella già trattata (parametro misurato all'uscita dal depuratore). I dettagli sui fermi d'impianto si misura nel *numero di giorni di fermo impianto nell'anno solare.*

I dati ambientali sono alla base dell'informazione ambientale. In materia di informazione ambientale, nel 2005 è stato emanato il D. Lgs. n. 195/2005, norma di recepimento della citata Direttiva europea, con l'obiettivo di "*garantire il diritto d'accesso all'informazione ambientale detenuta dalle autorità pubbliche (…) e garantire, ai fini della più ampia trasparenza, che l'informazione ambientale sia sistematicamente e progressivamente messa a disposizione del pubblico e diffusa, anche attraverso i mezzi di telecomunicazione e gli strumenti informatici, in forme o formati facilmente consultabili, promuovendo a tale fine, in particolare, l'uso delle tecnologie dell'informazione e*

[9] http://blog.planetek.it/2012/06/20/verso-i-linked-open-data-geografici/

della comunicazione" (art. 1). Per raggiungere tale obiettivo, la norma individua anche alcuni strumenti:

- cataloghi pubblici dell'informazione ambientale contenenti l'elenco delle tipologie dell'informazione ambientale detenuta;
- banche dati elettroniche facilmente accessibili al pubblico tramite reti di telecomunicazione pubbliche, da aggiornare annualmente, in cui, oltre a relazioni, trattati, accordi, convenzioni, autorizzazioni e studi, siano inseriti tutti i dati o le sintesi di dati ricavati dal monitoraggio di attività che incidono o possono incidere sull'ambiente.

L'indice dei cataloghi pubblici dell'informazione ambientale di cui sopra rappresenta, inoltre, ai sensi del D. Lgs. n. 32/2010, norma di recepimento della Direttiva INSPIRE, una componente dell'infrastruttura nazionale per l'informazione territoriale e del monitoraggio ambientale istituita dal medesimo Decreto.

L'art. 8 del D. Lgs. n. 195/2005, inoltre, ribadisce che l'autorità pubblica "*renda disponibile l'informazione ambientale detenuta avvalendosi, ove disponibili, delle tecnologie di telecomunicazione informatica e delle tecnologie elettroniche disponibili*".

Dal contesto delineato, si evince che l'apertura dei dati ambientali va nella direzione dell'applicazione delle norme indicate e, quindi, nella direzione di garantire il diritto all'accesso e la diffusione dell'informazione ambientale.

È evidente, infatti, che la pubblicazione aperta di dati ambientali permetterebbe alla cittadinanza di monitorare i progressi della PA rispetto a temi di pubblico interesse quali, ad esempio, il livello di inquinamento delle diverse zone territoriali di competenza delle amministrazioni, discriminante da non sottovalutare in relazione alle politiche di urbanizzazione del territorio. Tuttavia, il valore dei dati ambientali è strettamente legato alla contestualizzazione degli stessi, poiché tipicamente gli indicatori dello stato dell'ambiente sono realmente significativi solo quando vengono correlati a dati territoriali. L'adozione di formati LOD nel settore dei dati ambientali è quindi fondamentale a tale scopo in quanto, oltre a contestualizzare gli stessi dati in una particolare prospettiva, ne arricchisce di fatto il loro significato. Sarebbe quindi possibile, ad esempio, porre in relazione le rilevazioni di fattori inquinanti con le suddivisioni territoriali dell'amministrazione (quartieri e contrade di un Comune), evidenziando le zone più bisognose di interventi migliorativi. Ugualmente interessanti sarebbero le relazioni di dati ambientali relativi alla qualità dell'aria con i dati territoriali rappresentanti il grafo stradale comunale, dato particolarmente utile all'Amministrazione per adottare, ed eventualmente giustificare, determinate scelte sulla viabilità urbana (Costituzione/attivazione di ZTL, ordinanze di limitazione eccezionale del traffico urbano, ecc.). Infine, un altro processo notevole di contestualizzazione consisterebbe nel mettere in relazione gli indicatori di inquinamento, come ad esempio quelli dell'atmosfera e dei corpi idrici che insistono sul territorio d'interesse, coi dati relativi alla dislocazione nel territorio di attività agricole, industriali, o di cava, o di natura mineraria oppure di smaltimento dei rifiuti, evidenziando il reale impatto delle stesse sulle rilevazioni ottenute.

Una peculiarità fondamentale dei dati ambientali sta nell'adozione di molteplici unità di misura, richiedendo quindi un trattamento particolare in sede di definizione delle relative ontologie. Allo scopo di rendere questi dati espressivamente potenti, ed incentivare la loro associazione con dataset esterni, l'Amministrazione dovrebbe assicurare la memorizzazione dei dati con il maggiore livello di granularità o dettaglio possibile (dati atomici), evitando cioè forme aggregate che potrebbero comunque essere sempre derivate, se ritenute d'interesse, in sede di elaborazione successiva a mezzo software.

Ad esempio, tornando ai dati relativi alle acque prelevate nell'anno solare, va sicuramente preferita la

memorizzazione di rilevazioni riferite alle singole utenze (ad esempio, di servizio e pubbliche) piuttosto che di valori aggregati in una delle tre macro categorie. Altro esempio: relativamente ai dati sulla presenza nell'aria di agenti inquinanti, sarebbero preferibili dati mirati alle misurazioni delle concentrazioni in termini assoluti (comunque espressi in microgrammi al metro cubo) per ogni centralina rivelatrice distribuita sul territorio, piuttosto che memorizzare le ore di superamento di una data soglia, altra informazione derivabile comunque da rilevazioni di tipo atomico.

Un altro aspetto che non andrebbe sottovalutato, ai fini di un'implementazione realmente efficace del principio di trasparenza, è il fattore del tempo. I dati ambientali perdono gran parte della loro potenza espressiva quando decontestualizzati dalla variazione degli stessi nel tempo, potendo solo quest'ultimo evidenziare una tendenza riferibile ai dati scelti come indicatori. Aggiungere una variabile temporale a questi dati permetterebbe quindi di estrapolare a mezzo software informazioni rappresentative della qualità ed efficacia delle politiche ambientali dell'Amministrazione nel corso della propria gestione.

Indipendentemente dalle politiche che potranno essere adottate in futuro dalle singole PA per la raccolta dei dati, e che appare plausibile possano orientarsi sempre più verso l'adozione di formati standard atti a presentare dati di tipo LOD, la problematica principale da affrontare nel presente è rappresentata sicuramente dalla difficoltà di convertire i dati ambientali già raccolti, presenti attualmente in una varietà di formati: talvolta aperti e "machine readable", ma non identificati da un URI (ad esempio file CSV); spesso "machine readable", ma strutturati in formati chiusi (ad esempio file Excel); spesso chiusi e non strutturati, come nel tipico esempio dei file PDF, sebbene siano ancora disponibili nella maggioranza dei casi i documenti in formati strutturati da cui sono stati originariamente esportati e memorizzati.

Dati relativi al personale della PA:

L'impianto normativo vigente prevede che tutte le pubbliche amministrazioni debbano rendere note alcune informazioni relative al proprio personale.

In particolare, la Legge n. 69 del 18 giugno 2009 recante Disposizioni per lo sviluppo economico, la semplificazione, la competitività nonché in materia di processo civile impone, all'art. 21 comma 1, che tutte le PA rendano note informazioni relative ai propri dirigenti quali i curricula vitae, la retribuzione, i recapiti istituzionali, unitamente ai tassi di assenza e di presenza, aggregati per ciascun ufficio.

La circolare n. 3/2009 del Dipartimento della Funzione Pubblica prescrive altresì l'obbligo per le amministrazioni di: "*[…] pubblicare per ogni ufficio o unità organizzativa di livello dirigenziale:*

- *i dati mensili relativi alle percentuali di assenza del personale calcolati, in modo indifferenziato, tutti i giorni di mancata presenza lavorativa a qualsiasi titolo verificatasi (malattia, ferie, permessi, aspettativa congedo obbligatorio, ecc.), del personale dell'ufficio o unità organizzativa (compreso il dirigente);*
- *il dato relativo alla presenza dovrà emergere dal rapporto percentuale tra il numero dei giorni lavorativi complessivamente prestati dal personale dell'ufficio o unità organizzativa (compreso il dirigente) e il numero dei giorni lavorativi del mese di riferimento.*"

L'art. 54 del CAD dispone, come di seguito specificato, l'obbligo per le PA di pubblicare alcuni dati relativi al personale avente funzione di responsabile di procedimento, oltre che l'elenco completo delle caselle di posta elettronica istituzionali. Infatti, al comma 1, lettera *b)* si legge: "*l'elenco delle tipologie di*

procedimento svolte da ciascun ufficio di livello dirigenziale non generale, il termine per la conclusione di ciascun procedimento ed ogni altro termine procedimentale, il nome del responsabile e l'unità organizzativa responsabile dell'istruttoria e di ogni altro adempimento procedimentale, nonché dell'adozione del provvedimento finale, come individuati ai sensi degli articoli 2, 4 e 5 della Legge 7 agosto 1990, n. 241". Alla lettera *d)* dello stesso comma si legge: "*l'elenco completo delle caselle di posta elettronica istituzionali attive, specificando anche se si tratta di una casella di posta elettronica certificata di cui al decreto del Presidente della Repubblica 11 febbraio 2005, n. 68*."

L'art. 24 della Legge n. 183 del 04 novembre 2010 e la circolare n. 2/2011 del Dipartimento della Funzione Pubblica, recanti modifiche alla disciplina in materia di permessi per l'assistenza alle persone con disabilità (ex Legge n. 104 del 5 febbraio 1992), e diversi altri strumenti normativi inerenti la misurazione qualitativa e quantitativa delle agevolazioni per il personale, rendono d'interesse i dati relativi alla tipologia di permessi usufruiti dal personale. Tali dati confluiscono infatti in banche dati centrali del Dipartimento della Funzione pubblica della Presidenza del Consiglio dei Ministri, popolate sulla base delle comunicazioni fornite dalle singole PA. Tali comunicazioni contengono, tra l'altro: il nominativo del dipendente; la tipologia del permesso; e il contingente complessivo di giorni e ore di permessi fruiti.

L'apertura dei dati relativi al personale contribuisce in maniera significativa a garantire il principio della trasparenza sull'operato della PA; facilita la valutazione qualitativa e quantitativa dei risultati da essa raggiunti, e di quelli ottenuti dai propri dirigenti; permette la comunicazione diretta con i titolari degli uffici e i responsabili dei procedimenti; consente l'analisi dei trend prestazionali delle PA e può rappresentare un forte stimolo al miglioramento continuo della performance organizzativa e di quella individuale dei dirigenti. Se opportunamente collegati a dati relativi ai risultati ottenuti dalle PA e ai dati di bilancio e di spesa, è possibile in ultima analisi disegnare un quadro più chiaro della PA italiana, evidenziandone con esattezza aree di eccellenza e carenze, la cui conoscenza è indispensabile per progettare interventi premiali e correttivi mirati.

Alcune PA, sia centrali che locali, hanno recentemente pubblicato i dati sotto forma principalmente di file CSV o fogli Excel. Per quanto a conoscenza del gruppo di lavoro, non risultano iniziative di apertura di questa tipologia di dati in formati LOD.

Dati scolastici e universitari:

In riferimento alla definizione di dato delle pubbliche amministrazioni, di cui all'art. 1 comma 1, lettera *m)* del CAD, per dato scolastico si intende qualsiasi dato creato o comunque trattato dal Ministero dell'Istruzione, Ricerca e Università e/o dalle istituzioni scolastiche, afferente al mondo dell'istruzione.

A titolo non esaustivo, alcuni esempi di dati scolastici includono:

- i riferimenti delle istituzioni scolastiche (e.g., denominazione, tipologia, anagrafica e riferimenti telefonici, di posta elettronica, ecc.);
- le attrezzature multimediali (e non) presenti nelle scuole del territorio (e.g., laboratori, numero di pc e laptop, reti WiFi e LAN, Lavagne Interattive Multimediali, attrezzature per le palestre, attrezzature per lo studio della lingua straniera);
- i dati relativi al personale docente e non docente. Tali dati possono includere quelli relativi ai docenti per tipologia di contratto, ai pensionamenti, ai trasferimenti, alle assenze, ecc.;
- i dati relativi agli alunni. Tali dati possono includere quelli relativi al numero di alunni per classe

e per indirizzo di studio, alle iscrizioni, ai trasferimenti, ai ripetenti, agli abbandoni, agli esiti degli scrutini, agli ammessi alle classi successive e ai diplomati con relativa fascia di voto;
- i dati relativi alle biblioteche scolastiche;
- i dati relativi ai fondi per progetti di alfabetizzazione per alunni stranieri.

Una delle criticità che emerge nel trattamento di questi dati è la natura in taluni casi personale degli stessi. In generale, nel vademecum del garante per la protezione dei dati personali relativamente ai dati scolastici [19] si afferma che "*le scuole pubbliche non sono tenute a chiedere il consenso per il trattamento dei dati personali degli studenti. Gli unici trattamenti permessi sono quelli necessari al perseguimento di specifiche finalità istituzionali oppure quelli espressamente previsti dalla normativa di settore*". Tuttavia, alcuni dati relativi agli studenti e alle famiglie, come per esempio i dati sensibili, devono essere comunque trattati con estrema cautela valutando l'indispensabilità degli stessi nel perseguire finalità pubbliche. In questo caso, la pubblicazione del dato in forma aggregata può essere preferibile rispetto alla pubblicazione di dati atomici/elementari. In generale, nel trattamento dei dati degli studenti è bene considerare i principi di necessità, pertinenza e non eccedenza del trattamento così come sancito all'art. 11 del d.lgs n. 96 del 6 giugno 2003 - Codice della Privacy, ed evitare di diffondere online dati idonei a rilevare lo stato di salute degli studenti (art. 22 del d.lgs n. 96 del 6 giugno 2003 - Codice della Privacy).

L'apertura di dati scolastici comporta diversi vantaggi in termini di: trasparenza dello specifico settore della scuola; pianificazione incisiva delle attività di intervento nella scuola grazie a un maggior controllo sulla dispersione di energie o sulla duplicazione di risorse; controllo del percorso educativo offerto che può essere arricchito con opportune politiche laddove i dati evidenziano carenze. Nell'apertura dei dati, la produzione nello specifico di Linked Data per l'area scolastica è auspicabile nell'ottica di collegamento con dati relativi a servizi collaterali che concorrono a garantire un buon servizio scolastico (e.g., dati relativi alle mense) e un buon livello formativo (e.g., dati relativi a progetti musica, a progetti musei).

Ad oggi un insieme di dati scolastici relativi all'anagrafe delle strutture, al personale e agli alunni è stato pubblicato sul sito del Ministero dell'Istruzione, dell'Università e della Ricerca nell'ambito del progetto "La scuola in chiaro" [20]. I dati pubblicati, tuttavia, sono unicamente in formato Excel.

Una specializzazione di tale tipologia di dati è rappresentata dal dato universitario relativo a tutte le attività connesse alla didattica universitaria e al mondo della ricerca (per quest'ultimo caso si rimanda alla categoria di seguito descritta).

Dati della ricerca e competenze

I dati della ricerca comprendono sia le informazioni anagrafiche dei professori, dei ricercatori, degli assegnisti e dei borsisti, dei tecnici e degli amministrativi della ricerca pubblica, sia i dati relativi alle strutture (ad esempio, istituti, dipartimenti, laboratori) e ai prodotti della ricerca (pubblicazioni, brevetti, progetti, ecc.). Con riferimento a questi dati è anche possibile definire o estrarre le *competenze* di ricercatori e strutture, l'impatto della ricerca, la sua distribuzione geografica e il collegamento con le richieste del mercato e delle istituzioni.

Gli utenti potenziali dei dati includono:

- imprese e istituti finanziari interessati a tecnologie, brevetti, soluzioni di ricerca;
- laboratori, ricercatori e studenti per la ricerca di competenze in un ambito di ricerca;
- decisori interessati all'analisi della ricerca e delle potenzialità di trasferimento tecnologico.

Più in generale, i dati possono essere usati anche da organizzazioni che producono o utilizzano dati eterogenei come enti, imprese e consulenti che vogliono far incontrare domanda e offerta di qualsiasi natura: "expert finding", "competence matching", centri interinali, oppure sviluppatori di prodotti e servizi per l'analisi organizzativa, ad esempio cruscotti aziendali, conoscenza adattiva.

I vantaggi nell'apertura di tali tipi di dati sono molteplici:

- accessibilità e trasparenza dei dati della ricerca;
- sostegno alle attività di trasferimento tecnologico e in generale della conoscenza interna a un'organizzazione;
- arricchimento automatico dei dati già disponibili mediante analisi automatica del testo e inferenze automatiche;
  interoperabilità con dati già disponibili in Italia e nel resto del mondo.

Classificazioni e dati statistici

I dati relativi a classificazioni ufficiali sono particolarmente importanti per supportare l'interoperabilità semantica di diversi sistemi. Infatti, la conformità a una classificazione definita consente di: superare il problema dell'eterogeneità semantica, connesso alla comprensione del significato dei dati pubblicati e/o scambiati con altri soggetti; supportare l'interoperabilità mediante il collegamento ("linking") di dati aperti. In particolare, l'utilizzo di un'unica classificazione di riferimento per varie fonti di dati consente il collegamento delle stesse.

Un insieme di classificazioni ufficiali è reperibile in [21]. Nell'ambito di tale insieme, alcune classificazioni sono individuabili come particolarmente rilevanti:

1. **Attività economica**: la classificazione delle attività economiche è l'Ateco 2007, congiuntamente adottata dall'Istat, dai Ministeri interessati, dagli Enti che gestiscono le principali fonti amministrative sulle imprese (mondo fiscale e camerale, enti previdenziali, ecc.) e dalle principali associazioni imprenditoriali. Tale classificazione costituisce la versione nazionale della nomenclatura europea, Nace Rev.2, pubblicata sull'Official Journal il 20 dicembre 2006 (Regolamento (CE) n.1893/2006 del Parlamento europeo e del Consiglio del 20/12/2006). L'Ateco 2007 è un importante risultato d'integrazione dati, in quanto, per la prima volta, il mondo della statistica ufficiale, il mondo fiscale e quello camerale utilizzano la stessa classificazione delle attività economiche;
2. **Professioni**: la classificazione ufficiale di riferimento per le professioni è la CP2011, frutto di un lavoro di aggiornamento della precedente versione (CP2001) e di adattamento alle novità introdotte dalla International Standard Classification of Occupations - Isco08. La CP2011 riprende il formato della Nomenclatura e Classificazione delle Unità Professionali (NUP06), costruita dall'Istat in partnership istituzionale con l'Isfol, prevedendo, per ciascun livello classificatorio, una descrizione che traccia i contenuti e le caratteristiche generali del lavoro;

3. **Titolo di studio**: L'Istat ha predisposto, per la prima volta in Italia, la classificazione dei titoli di studio. L'obiettivo è di ricostruire l'insieme dei titoli di studio emessi in Italia e potenzialmente in possesso della popolazione. In assenza di fonti di tipo normativo-istituzionale, la classificazione è stata costruita principalmente a partire da fonti statistiche. Si è in particolare fatto riferimento: alle rilevazioni sulle scuole secondarie di secondo grado, di competenza Istat, MIUR e INVALSI; alle indagini sull'istruzione universitaria, di competenza Istat e MIUR-URST; ai Censimenti della popolazione italiana;
4. **Unità istituzionali del settore pubbliche amministrazioni**: Sulla base del Sec95, il sistema europeo dei conti, l'Istat predispone l'elenco delle unità istituzionali che fanno parte del settore delle PA (Settore S13), i cui conti concorrono alla costruzione del Conto economico consolidato delle PA[10]. I criteri utilizzati per la classificazione sono di natura statistico-economica, indipendenti dal regime giuridico che governa le singole unità istituzionali. Ai sensi dell'art. 1, comma 3 della Legge n.196 del 31 dicembre 2009 - Legge di contabilità e di finanza pubblica, e s.m.i., l'Istat è tenuto, con proprio provvedimento, a pubblicare annualmente l'elenco sulla Gazzetta Ufficiale. L'ultimo elenco è stato pubblicato sulla Gazzetta Ufficiale - Serie Generale n. 228 il 30 settembre 2011.

Altre classificazioni d'interesse nel contesto della PA includono: i codici dei comuni, delle province e delle regioni; l'Ateco 2007; le Unità legali; gli Stati esteri; le professioni e le attività di apprendimento. Ad oggi, tali classificazioni vengono fornite utilizzando soltanto formati Open Data quali CSV e SDMX [22]. Visto il ruolo che possono svolgere nel contesto dell'interoperabilità semantica, è auspicabile che anche la trasformazione di tali dati in formato Linked Open sia prevista.

Oltre alle classificazioni ufficiali, una tipologia di dati per la quale i requisiti di accessibilità e fruibilità sono particolarmente importanti è costituita dai dati pubblicati nell'ambito della *statistica ufficiale*, ovvero dati aggregati che costituiscono il prodotto finale delle rilevazioni e delle elaborazioni condotte nell'ambito del Programma statistico nazionale. La fornitura di questi dati in formato aperto ha diversi vantaggi: tra questi, ad esempio, la possibilità di integrare fonti statistiche distinte (ad esempio a livello europeo), la possibilità di confrontare in modo diretto tali dati con elaborazioni condotte su microdati disponibili presso altri soggetti, ecc.

[10] http://www.istat.it/it/archivio/6729

Questa tipologia è prodotta da Istat ed è diffusa a vari livelli di aggregazione e principalmente in modalità CSV, SDMX e JSON-stat [23]. Il gruppo di lavoro ha identificato inoltre altre tipologie di dati aperti di complessa gestione la cui diffusione avrebbe comunque un considerevole impatto socio-economico. A tal riguardo, il gruppo di lavoro ritiene opportuno avviare uno studio più approfondito di tali dati dopo la pubblicazione delle presenti linee guida. Alcune categorie d'interesse sono: i dati di bilancio della PA; i dati di spesa della PA; i dati sanitari; i dati di sicurezza pubblica; i dati giudiziari e disciplinari, i dati relativi all'infomobilità e i dati relativi ai laboratori pubblici di analisi accreditati (e.g., ARPA, IIZZSS, Università, VVFF, SSC, ENEA, ecc.) e le rispettive analisi di accreditamento. Infine, altre categorie possono essere individuate attraverso un ulteriore approfondimento.

5. STATO DELL'ARTE SU LINKED OPEN DATA E INTEROPERABILITÀ SEMANTICA

Questa sezione è dedicata alla presentazione di un quadro di riferimento su LOD e interoperabilità semantica sia a livello internazionale che nazionale.

5.1. Lavori e iniziative internazionali

Un buon punto di partenza per studiare e approfondire la tematica e l'utilità dei Linked (Open) Data è il libro in [25]; da un punto di vista pratico invece, il progetto W3C "Linking Open Data" [38] costituisce certamente il principale riferimento per l'ambito dei LOD. Questo ha molto contribuito per la loro diffusione e si è posto l'obiettivo di identificare e diffondere "best practices" e linee guida per la pubblicazione di dataset eterogenei per tipo e provenienza, collegati tra di loro utilizzando standard e tecnologie del Web Semantico. Il progetto, che ha riunito ricercatori, ingegneri e professionisti operanti in diverse istituzioni, ha rappresentato un importante stimolo in questo ambito: da un lato è stata promossa la formazione di una comunità di pratica articolata a livello internazionale la quale costituisce, oggi, il principale riferimento nel settore; dall'altro, il progetto ha condotto allo sviluppo, al consolidamento, e alla promozione di strumenti e tecnologie per la gestione dei dati in modalità Linked.

Il successo del progetto è rappresentato da quella che viene chiamata nuvola LOD (LOD Cloud) [26]. La nuvola visualizza il risultato di censimenti periodici dei dataset Linked prodotti dalla comunità del progetto e da altri individui e organizzazioni. All'ultimo censimento del settembre 2011 (si veda la Figura 1) la nuvola contava 504 milioni di concetti e 31 miliardi di connessioni [27].

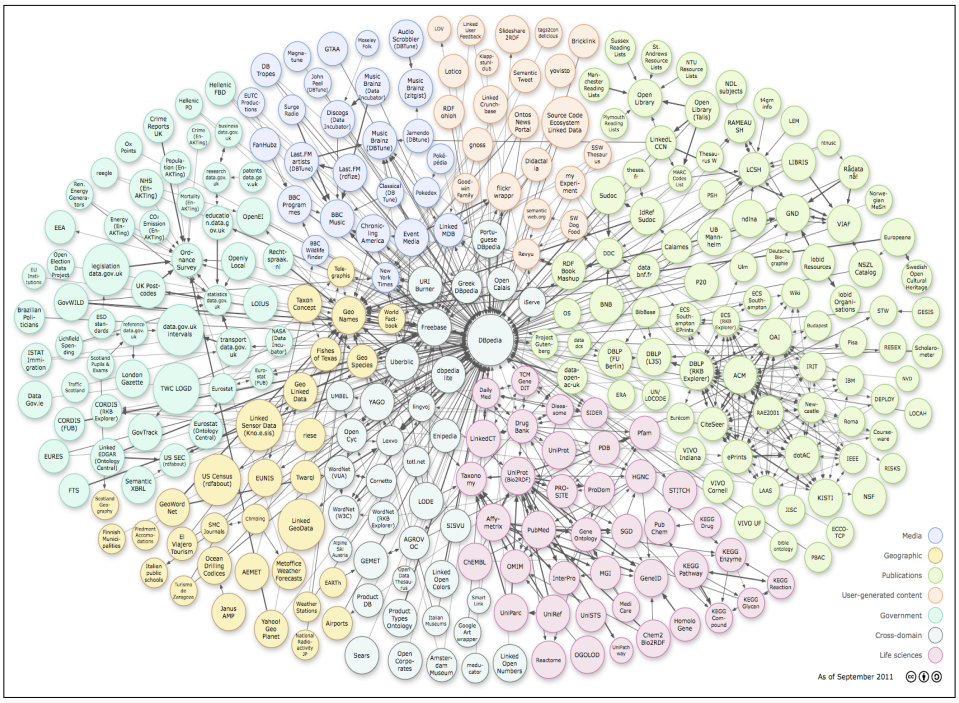

Figura 1: Basi di dati pubblicate nella LOD Cloud [26].

In Figura 1, i nodi rappresentano i singoli dataset pubblicati e gli archi rappresentano la presenza di connessioni semantiche tra essi. Fra i numerosi dataset si osserva come DBpedia [28], la versione RDF dei dati contenuti in Wikipedia, occupi una posizione centrale nella nuvola in virtù della vastità e varietà dei dati che essa contiene. DBpedia costituisce pertanto un importante accentratore attraverso il quale è possibile la navigazione tra un gran numero di dataset diversi.

Per una maggior chiarezza, i vari domini della LOD Cloud sono visualizzati in Figura 1 con colori diversi. Così, per esempio, il rosa indica l'insieme di dataset scientifici (dominio a cui appartiene per esempio Uniprot [29] un grande database bioinformatico per le sequenze proteiche di organismi viventi e virus); il colore giallo indica i dataset geografici (dominio cui appartengono, per esempio, LinkedGeoData [30] che rende disponibili in RDF le informazioni del progetto OpenStreetMap [31], GeoNames [32] che rende disponibili più di 6 milioni di toponimi, ecc.). Il colore azzurro chiaro sulla sinistra rappresenta i dataset relativi al dominio delle pubbliche amministrazioni e vede un considerevole contributo del Regno Unito. In generale, il sempre crescente numero di dataset presenti nella LOD Cloud evidenzia l'elevato livello di maturità dell'approccio e i diversi esempi dimostrano i benefici che si possono conseguire.

Negli ultimi tempi molti paesi hanno adottato politiche di Open Data istituendo portali che fungono da collettori per i dataset della PA. Il primo e più famoso è stato quello degli Stati Uniti "data.gov" [7] a cui poi ne sono seguiti molti altri. La prassi di esporre inizialmente i dati in modalità Open lasciando poi ad altri soggetti l'onere di renderli Linked comincia ad essere superata. Da qualche tempo, infatti, un numero crescente di amministrazioni, rendendosi conto del valore aggiunto dei Linked Data, si sta facendo carico del lavoro necessario per pubblicare gli Open Data direttamente in modalità Linked. In

Europa questa è stata la decisione presa dal governo britannico che ha creato un portale ricco di dataset e di applicazioni. Negli Stati Uniti, il governo ha deciso di avviare una collaborazione con la Tetherless World Constellation [33], dopo che questa autonomamente aveva cominciato a trasformare i dati pubblicati da "data.gov" in Linked Data.

Una lista (parziale) di paesi e di progetti di governi nazionali e locali che hanno adottato una strategia Open Data si può trovare in [34][35][36]. Un elenco invece di dataset pubblicati in modalità Linked, non esclusivamente appartenenti alla PA, è accessibile alla pagina [37] e sul sito del W3C [39].

Nonostante il moltiplicarsi di siti che si occupano di Linked Data, il W3C rappresenta comunque il punto di riferimento privilegiato per il Web Semantico, in generale, e per i Linked Data, in particolare [5]. A tal riguardo, sono stati istituiti diversi gruppi di lavoro dal W3C. I risultati dei gruppi sull'RDF [40], sul Semantic Web Deployment [41], e sui Government Linked Data [42] sono da considerarsi di riferimento nel contesto della PA ai fini dell'apertura interoperabile dei dati pubblici. Quest'ultimo in particolare ha per scopo istituzionale quello di fornire standard e altre informazioni per aiutare le amministrazioni a pubblicare i loro dati in modalità Linked mediante ricorso a tecnologie del Web semantico. Il gruppo si prefigge un insieme di obiettivi: costruire e manutenere una "Community Directory"; raccomandare "best practices" per quanto riguarda l'approvvigionamento di prodotti e servizi; selezionare vocabolari; costruire URI; gestire le varie versioni; garantire la stabilità; gestire i dati legacy e altre soluzioni a problemi specifici; raccomandare quali vocabolari usare in aree di concetti comuni.

Una forma importante d'interoperabilità è quella relativa alla provenienza dei dati. Con la nascita del "Web of Data" e la crescita del numero di dataset pubblicati in modalità Linked Data è divenuta più pressante l'esigenza di disporre e saper trattare la "provenance" o provenienza (o origine) dei dati e delle risorse in generale. La facilità con cui i Linked Data permettono di combinare informazioni e dati amplifica la necessità di sapere da chi, quando, dove e come sono state prodotte le informazioni. Conoscere questi elementi permette di valutare, anche con strumenti automatici, la qualità e altre caratteristiche delle informazioni e, conseguentemente, di prendere decisioni più consapevoli. Le applicazioni che nel "Web of Data" utilizzano e combinano dati provenienti da fonti diverse, siano esse nell'ambito scientifico, della PA, o altro, devono disporre di queste informazioni per attribuire il credito agli autori originari, per sapere se stanno utilizzando informazioni aggiornate, o violando un diritto d'autore, o per attribuire una misura di affidabilità al risultato. Attualmente non esiste alcuna indicazione pratica di come rappresentare la provenienza dei dati; uno dei modi è attraverso l'uso di specifiche licenze (Sezione 8.1) che presenta però lo svantaggio di non poter verificare e processare automaticamente queste informazioni.

Al riguardo, il W3C ha costituito un gruppo di lavoro di riferimento, il W3C Provenance Interchange. Come evidenziato nel suo rapporto finale [43], quello di rappresentare e utilizzare la provenienza associata alle informazioni è un problema affrontato in molti ambiti e discipline e molte sono le tecnologie che sono state sviluppate per risolverlo. Manca però un modello standard di provenienza che permetta di esprimere, richiedere e utilizzare queste informazioni in un ambito eterogeneo e aperto come il Web. Alcuni casi di studio sono stati analizzati a tale scopo e, dall'analisi, ne è emersa una varietà di esigenze e soluzioni possibili; pertanto, il gruppo ha ritenuto poco praticabile proporre lo sviluppo di un modello standard che rispondesse cioè a tutte le esigenze, e ha raccomandato di sviluppare un modello base di provenienza basato su 17 concetti ben identificati che costituiscono il riferimento per lo scambio di informazioni di provenienza fra sistemi eterogenei.

Anche la Commissione Europea ha deciso di investire in tale contesto finanziando progetti nell'ambito del 7° programma quadro. Uno di questi, importante per le finalità delle presenti Linee guida, è il progetto Linked Open Data, LOD2 [44], che ha messo a punto uno "stack" tecnologico per la gestione del ciclo di vita dei Linked Data. Il risultato è una distribuzione integrata di strumenti e tecnologie, alcune delle quali saranno descritte in maniera più approfondita nella Sezione 0; queste sono, o stanno diventando, soluzioni di riferimento per la gestione dei Linked Data.

Per completezza, e per fornire spunti di approfondimento, va aggiunto che il W3C e la Comunità Europea hanno di fatto costruito l'infrastruttura e la cultura del Web Semantico fin dal 1999, attraverso svariati gruppi di lavoro, standard e diversi progetti finanziati. Per quanto riguarda il W3C, oltre ai già citati gruppi di lavoro su RDF e sul Semantic Web Deployment, meritano di essere citati i gruppi di lavoro su OWL (Web Ontology Language) [45] e su RIF (Rule Interchange Format) [46] che hanno creato le condizioni per poter fare ragionamento automatico su schemi e dati nel Web; il gruppo su SPARQL [47] che ha permesso di realizzare un linguaggio di interrogazione molto flessibile e ricco per RDF; il gruppo su SKOS (Simple Knowledge Organization System) [48] che ha definito una semplice ontologia per esportare strutture di metadati (classificazioni, tesauri, ecc.) in RDF; il gruppo su RDFa [49] che ha definito formati per "incorporare" informazione semantica direttamente in HTML, così abilitando di fatto annotazioni semantiche del testo interoperabili con RDF.

Per quanto riguarda la Comunità Europea, oltre a LOD2, tra i progetti finanziati sul Web Semantico vanno citati tra i più importanti e in ordine cronologico: OnToKnowledge (1999-2001, sui primi linguaggi e metodi per il Web Semantico dopo RDF), WonderWeb (2002-2004, che ha creato l'infrastruttura teorica e tecnologica per OWL), KnowledgeWeb (2004-2007, ancora su metodi e tecnologie per il Web Semantico), SUPER (2005-2008, sulle infrastrutture e i linguaggi per i servizi sul Web Semantico), NeOn (2006-2010, sui modelli e strumenti per la progettazione di ontologie distribuite), LARKC (2008-2011, sulle infrastrutture e i componenti scalabili per RDF e OWL), IKS (2009-2012, sulle infrastrutture e i componenti per arricchire e connettere i CMS al Web Semantico), Robust (2010-2013, sui componenti per connettere i linguaggi di business con il Web Semantico).

I siti di questi progetti contengono moltissimo materiale di formazione, ricerca, componenti riusabili, casi d'uso, ed esperienze. Dai risultati di questi progetti si capisce come l'evoluzione del Web Semantico, da una pura visione supportata da buoni risultati accademici, si sia orientata verso la realizzazione di una piattaforma per l'interoperabilità semantica, utilizzabile da tutti i sistemi informativi.

Sempre nell'ambito della Commissione Europea, una serie di risultati sono stati raggiunti nel contesto dell'"action 1.1" del programma ISA menzionata in sezione 3.2. In particolare, è stato prodotto un documento [90] in cui gli "asset" di interoperabilità semantica sono definiti come quei metadati ampiamente riutilizzabili e quei dati di riferimento (e.g., vocabolari, tassonomie) utilizzati per lo sviluppo di sistemi di e-government. Il documento individua il livello di maturità dei metadati, con un approccio simile a quello adottato da Tim-Berners Lee per lo schema di definizione degli open data, fornendo in conclusione delle raccomandazioni, come nel caso del presente documento, sull'uso e la gestione di metadati in sistemi di e-government.

Partendo da tali raccomandazioni, sono state poi intraprese iniziative volte alla definizione di specifiche per implementare una federazione di "repository" di "asset" semantici sul sito Joinup [91] dalla Commissione Europea. Tali specifiche, che vanno sotto il nome di Asset Description Metadata Schema [92], sono attualmente sottoposte al processo di standardizzazione W3C [93] e si propongono come

modello per facilitare la cooperazione e la federazione di repository esistenti [94] agendo a livello di strato comune standard tra i repository che devono scambiarsi dati. Inoltre, in tale contesto, alcuni vocabolari sono stati proposti a livello europeo per la descrizione di persone, luoghi ed entità legali, i cosiddetti Core Person, Core Location e Core Business [95] rispettivamente. I vocabolari a loro volta riutilizzano ontologie e vocabolari diffuse allo stato dell'arte (e.g., FOAF, Dublin Core); essi sono disponibili anche in modalità LOD e condivisi da un'ampia pletora di attori europei. Quest'ultima caratteristica quindi li rende particolarmente efficaci per poter avviare un processo di apertura di dati interoperabili transfrontalieri, in un'ottica di fornitura di servizi e-government su scala europea e non solo nazionale.

Infine, alcuni studi recenti analizzano lo scenario (Linked) Open Data a livello globale. Il lavoro [50] nel quale si valutano le strategie adottate in Australia, Danimarca, Regno Unito, Spagna, e Stati Uniti, mette in rilievo come dietro il termine Open Data ci siano, in realtà, piani e strumenti diversi. L'articolo riporta anche quali sono, nelle diverse realtà prese in esame, le barriere e gli stimoli all'introduzione degli Open Data.

Lo studio in [51] condotto dal Dipartimento d'Informatica dell'Università di Dresden, in collaborazione con OKF, su 50 piattaforme Open Data a livello regionale, nazionale, internazionale, comunitario e di organizzazioni internazionali evidenzia alcune carenze importanti delle piattaforme e dei dataset analizzati, quali: la presenza di riferimenti che diventano obsoleti o di riferimenti a pagine html anziché a dataset scaricabili; la varietà di formati, non sempre aperti; la mancanza di metadati standardizzati che rendano i dati effettivamente riusabili. Queste carenze si osservano peraltro anche in piattaforme importanti come data.gov.uk (UK) e data.gov (US). Per entrambe le piattaforme, infatti, solo nel 79% dei casi i metadati inseriti sono serviti a scaricare i dati e solo rispettivamente nel 77% e nel 42,2% dei casi è stato possibile aprirli con "parser" standard. Sempre secondo questo studio il 43% dei portali Open Data non implementa alcuna API o endpoint di interrogazione, richiedendo quindi un download manuale o una soluzione non standard o di tipo non Linked.

Dall'analisi delle esperienze in campo internazionale emerge quindi che i Linked Data rappresentano una soluzione economicamente e funzionalmente competitiva per implementare l'interoperabilità semantica per la PA.

5.2. Iniziative nazionali

In Italia, diverse PA hanno di recente intrapreso iniziative di apertura dei propri dati pubblici. Alcune di queste iniziative sono state raccolte dal gruppo di lavoro all'interno di una sorta di "catalogo" pubblicato sul sito di DigitPA al fine di offrire ulteriori esempi concreti a beneficio di altre PA.

Dall'esame delle iniziative si possono evidenziare alcuni elementi comuni: nella maggior parte dei casi non è chiara la governance dell'iniziativa e la tendenza prevalente sembra concentrarsi più sull'apertura di dati pubblici sotto forma di Open Data piuttosto che sul ricorso ai LOD, limitando così i livelli di interoperabilità semantica ottenibili.

Nonostante questo, esistono comunque alcune esperienze/iniziative degne di nota che sono in linea con lo stato dell'arte internazionale. Il gruppo di lavoro è unanime nel ritenere che prendendo esempio anche da queste iniziative sia possibile incentivare, nel medio termine, politiche e strategie di apertura in formato Linked dei dati della PA italiana.

Una di queste è "data.cnr.it" [52], il progetto di dati aperti del Consiglio Nazionale delle Ricerche, progettato e mantenuto dal Laboratorio di Tecnologie Semantiche dell'ISTC CNR, condiviso con l'unità Sistemi Informativi del CNR. Inoltre, altri Linked Open Data sono stati pubblicati dal comune di Firenze [53], da DigitPA [54], dalla regione Piemonte [55], dalla Camera dei Deputati [56].

Per quanto riguarda data.cnr.it, realizzato nel 2010, si tratta di un progetto motivato non solo dalla volontà di pubblicare dati aperti della ricerca, ma anche dalla necessità di rispondere a requisiti di trasferimento tecnologico richiesti dal CNR. Il requisito principale era quindi quello di far incontrare l'offerta di risultati della ricerca con l'eventuale domanda. Questo requisito è analogo a quelli che sussistono in molte altre amministrazioni, sia nei rapporti con il cittadino, sia con le aziende o con altre amministrazioni. Sul piano tecnico, il progetto è stato sviluppato dedicando particolare attenzione alla completezza e all'accuratezza degli schemi (ontologie) e all'arricchimento dei dati con l'integrazione di tecnologie d'inferenza e di analisi del linguaggio naturale. Analogamente, particolare attenzione è stata posta all'implementazione di interfacce di navigazione ed esplorazione dei dati innovative, accessibili dal portale Semantic Scout [57]. I metodi usati sono descritti in [58] ed esemplificano abbastanza bene i metodi di trasformazione, analisi, bonifica, linking dei dati ecc., sintetizzati nella Sezione 6 di questo documento.

Un'altra iniziativa innovativa è costituita dal Geo-catalogo semantico sviluppato dalla Provincia Autonoma di Trento. L'obiettivo del progetto è favorire l'interoperabilità e la collaborazione tra le strutture che si occupano di pianificazione e gestione economico-produttiva e ambientale del territorio trentino attraverso la pubblicazione ordinata e aperta di dati territoriali linkabili. Nello specifico, il progetto offre un servizio che consente di interrogare il dataset dei dati territoriali attraverso il Portale Geocartografico Trentino [59]. Il Geocatalogo semantico [60] implementa il paradigma LOD ed è conforme alle indicazioni contenute nella Direttiva europea INSPIRE, e alle direttive nazionali IntesaGIS e RNDT. L'infrastruttura è in questo ambito intesa come un insieme di politiche, accordi istituzionali, tecnologie, dati e persone che rende possibile una condivisione efficiente ed efficace dei dati geo-referenziati. I benefici attesi dal servizio sono, da un lato, il miglioramento dell'accuratezza dei risultati di ricerca delle geo-informazioni e, dall'altro, l'avanzamento verso la società della conoscenza attraverso lo sviluppo della cultura dell'uso di geo-informazioni nella vita quotidiana.

Un'iniziativa promettente nel contesto dei Linked Data nella PA è quella avviata da DigitPA che ha deciso di aprire le banche dati di interesse nazionale riferite al contesto SPC. Al riguardo, DigitPA ha scelto di partire dall'Indice delle Pubbliche Amministrazioni (IPA) [61] in quanto base di dati contenente quelle informazioni che identificano in modo univoco tutte le PA italiane, e potenzialmente collegabile con altri dati di tipo Linked pubblicati dalle PA (Sezione 9).

Altra l'iniziativa interessante è Linked PA [62], di Ablativ, Linkalab e Istos, che a partire dal coinvolgimento della società civile ha consentito lo sviluppo di un'ontologia della PA. L'iniziativa si propone come punto di aggregazione e proposta per lo sviluppo del (Linked) Open Data nel nostro paese. L'ontologia proposta rappresenta un'opzione di scelta interessante che si affianca alle diverse ontologie della PA già disponibili.

Infine, un altro progetto che si ritiene opportuno evidenziare in questa sezione è la versione italiana di DBpedia recentemente pubblicata [63]. DBpedia Italiana, frutto di una collaborazione tra SpazioDati srl e la Fondazione Bruno Kessler, estrae dati dalla versione italiana di Wikipedia e li espone sotto forma di Linked Data attraverso un meccanismo di "content negotiation" e mediante l'uso di uno SPARQL endpoint pubblico.

La versione internazionale di DBpedia [28] contiene soltanto le voci italiane che hanno un corrispondente nella Wikipedia inglese. Si tratta di circa 400.000 entità, estratte da un piccolo sottoinsieme della Wikipedia Italiana, che conta circa 1 milione di pagine. Inoltre, nella DBpedia internazionale, tutti i dati sono estratti dalla versione inglese di DBpedia (eccetto l'abstract e poco altro) che, su voci "italiane" tendono a contenere meno dati o ad essere meno aggiornate. DBpedia Italiana contiene invece circa 1,5 milioni di entità estratte direttamente dalla versione italiana di Wikipedia.

A parte l'interesse della base di conoscenza in sé, DBpedia Italiana rappresenta una risorsa pubblica di fondamentale importanza come nodo cruciale per i Linked Data italiani. Grazie alla sua natura enciclopedica, essa contiene infatti risorse appartenenti a diversi domini della conoscenza e può quindi essere utilizzata come nodo a cui collegare altre basi di conoscenza sia a fini di disambiguazione delle entità, sia per facilitare la scoperta di risorse collegate nella nuvola LOD.

RACCOMANDAZIONI

R1: È preferibile conformarsi agli strumenti sviluppati dalla Commissione Europea in ambito ISA per definire gli "asset" di interoperabilità semantica, in un'ottica di abilitazione di servizi e-government transfrontalieri.

6. APPROCCIO METODOLOGICO ALL'INTEROPERABILITÀ SEMANTICA TRAMITE LINKED OPEN DATA

Volendo riassumere quanto detto sin qui, l'interoperabilità semantica conferisce alle organizzazioni la capacità di condividere il significato delle informazioni che esse si scambiano. Dall'esame del contesto di riferimento, tanto sul piano normativo quanto su quello dello stato dell'arte e delle tendenze in atto a livello internazionale, il gruppo di lavoro è unanime nel ritenere i LOD uno strumento necessario ed efficace per abilitare lo sviluppo di una concreta interoperabilità semantica tra PA, sia a livello nazionale che a livello transfrontaliero.

Questa sezione propone quindi un approccio metodologico generale per l'apertura interoperabile di dati pubblici attraverso i LOD. La metodologia proposta tiene nella dovuta considerazione le informazioni ricavate dallo studio delle esperienze descritte nella sezione precedente, ma aggiunge a queste alcune "best practices" a livello internazionale.

L'approccio che si propone vuole essere sufficientemente flessibile per essere adattato alle specifiche esigenze delle singole PA e, più in generale, a quelle di qualsiasi "produttore" di dati. L'obiettivo è quello di generare dataset delle PA utilizzabili come Linked Data, ovvero dataset in formato RDF e contenenti connessioni fra loro e con dataset esterni alle PA.

La metodologia proposta si articola in sette distinte fasi:

1. individuazione e selezione dei dataset;
2. bonifica;
3. analisi e modellazione;
4. arricchimento;
5. linking esterno (interlinking);
6. validazione;
7. pubblicazione.

Un possibile piano di rilascio è mostrato in Figura 2. Si noti come alcune delle fasi sopra elencate possano in realtà essere svolte contemporaneamente. Questo vale, in particolare, per le fasi analisi e modellazione e bonifica, e per le fasi arricchimento e linking esterno. Per quanto riguarda la fase di pubblicazione, essa può essere intesa come una serie di rilasci successivi e incrementali che, a partire dai *raw data*, che possono essere rilasciati subito a valle della trasformazione in RDF, arrivano fino ai Linked Data, rilasciati via via che vengono validati.

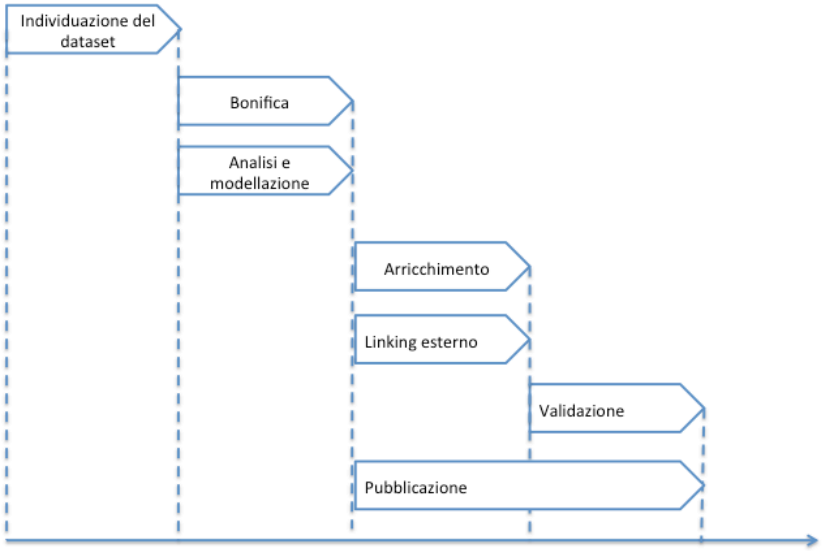

Figura 2: Le fasi dell'approccio metodologico all'interoperabilità semantica attraverso LOD in un possibile piano di rilascio

6.1. Individuazione e selezione dei dataset

Il processo di individuazione dei dati è sicuramente il punto di partenza del processo di apertura e linking dei dati. Il patrimonio informativo di cui dispongono le PA è molto vasto e una pubblicazione massiva dei dati in formato aperto spesso non è praticabile. Nasce quindi l'esigenza di selezionare un sottoinsieme dei dati che ha senso aprire e trasformare in Linked Data. È quindi necessario organizzare il lavoro di selezione seguendo criteri generali di valutazione opportunamente predefiniti, al fine di rendere quanto più fluido possibile l'intero processo, e che consentano di mediare tra la domanda di dati (interna o esterna) e lo sforzo richiesto per trattarli o la perdita di controllo che ne può derivare.

In primo luogo è bene selezionare i dati sulla base della domanda territoriale, in particolare cercando di individuare quei dataset per cui gli impatti potenziali sull'amministrazione o sulla collettività del territorio sono più evidenti. E' chiaro che dati di interesse generale possono considerarsi, valutando il livello di interesse, rispondenti a domande di più territori. Qui è importante sottolineare come i dati legati alla domanda dell'amministrazione, delle imprese, o dei cittadini non sempre coincidono (si veda la sezione 8.2). È fondamentale quindi, nella fase di selezione dei dataset candidati, avere ben chiaro quali sono gli obiettivi che l'amministrazione intende raggiungere con l'apertura e la pubblicazione dei propri dati.

In secondo luogo è bene individuare i vincoli all'apertura dei dati presenti, sia da un punto di vista normativo, sia organizzativo. Il livello di confidenzialità dei dati e la loro modalità di accesso sono due fattori chiave da considerare per una stima preliminare; in particolare è fondamentale considerare almeno tre aspetti distinti:

1. la segretezza di alcuni dati che non possono, per volontà del legislatore, essere oggetto di comunicazione e divulgazione a terzi;

2. il rispetto della riservatezza dei cittadini e delle imprese[11];
3. il rispetto del diritto d'autore[12].

In terzo luogo è bene avviare una fase di pre-analisi del dominio di riferimento in modo da acquisire informazioni preliminari utili per le fasi successive di trasformazione e identificare altri elementi-chiave per la misura della complessità.

Dal punto di vista del metodo, la raccolta di *casi d'uso*, *scenari*, o direttamente *requisiti* è la strada maestra per arrivare a una decisione efficace e consapevole. La specifica dei casi d'uso è molto utile per evidenziare le relazioni tra dataset e fruitore/i. È inoltre fondamentale definire dei criteri per valutare la precedenza tra i diversi blocchi di dati da liberare; per esempio potrebbe essere interessante associare a ogni dataset un valore di opportunità che misura l'eventuale interesse che un utilizzatore ha per uno specifico insieme di dati, ad esempio, sulla base della sua disponibilità momentanea, della specificità dei dati contenuti, di motivi di sperimentazione, ecc.

RACCOMANDAZIONI

R2: Selezionare i dati sulla base della domanda avendo chiari gli obiettivi che l'amministrazione vuole ottenere con l'apertura dei dati.

R3: Prestare attenzione a possibili vincoli sull'apertura dei dati, tanto sul piano normativo (tutela del segreto, protezione dei dati personali, rispetto del diritto d'autore) quanto su quello organizzativo.

R4: Privilegiare, ove possibile, l'apertura dei dati più atomici rispetto alle forme aggregate degli stessi, grazie alla maggiore potenza espressiva dei primi, mediante i quali è comunque possibile ricavare i secondi tramite un processo di elaborazione.

6.2. Bonifica

I dati all'interno dei sistemi informativi o degli archivi di un Ente sono spesso "sporchi" (talvolta sono stati concepiti per essere funzionali, attraverso le applicazioni informatiche, ai processi di business interni) e non immediatamente pronti per la pubblicazione o per le dovute elaborazioni. La qualità del dato rappresenta quindi un aspetto particolarmente importante perché un dataset sporco può rendere inefficienti, o addirittura impraticabili, alcune operazioni di confronto, di similitudine e di aggregazione sui dati.

Dovranno quindi con molta probabilità essere effettuate campagne di bonifica che consentano al dato di acquisire la qualità richiesta. Sebbene non sia possibile elencare o prevedere tutte le tecniche necessarie per bonificare i dati, si può almeno indicare alcuni dei problemi più diffusi: incompletezza dei dati, formati diversi, significati ambigui, datatype inconsistenti, mancanza di corrispondenza fra

11 D.Lgs. n. 196/2003 e Deliberazione n. 88/2011 del Garante della Privacy
12 Legge n. 633/1941 – Legge a protezione del diritto d'autore e di altri diritti connessi al suo esercizio

nomi usati negli schemi fisici e dati effettivamente contenuti (tipicamente per ragioni storiche), ecc.

La bonifica va pianificata immaginando il formato RDF e le pratiche del Web Semantico. Si tenga conto che, l'incompletezza (o la mancanza di integrità) dei dati non è un problema altrettanto grave di come viene percepito nel contesto della progettazione delle basi di dati. Il Web e il Web Semantico non hanno infatti pretesa di integrità e di rigorosità; è la numerosità crescente delle informazioni (documenti e/o dati) e il loro utilizzo a decretarne qualità, valore, e completezza.

6.3. Analisi e modellazione

Gli obiettivi di questa fase sono di formalizzare il modello concettuale e dare una rappresentazione coerente con esso del dataset di riferimento. Questa fase rappresenta una ristrutturazione logica e concettuale dei dati; in analogia con le metodologie di analisi e progettazione del software, possiamo dire che viene attivato un processo di *reingegnerizzazione* e *refactoring* della base informativa.

Sia il modello concettuale sia il dataset alla fine di questa fase saranno rappresentati in RDF. In RDF i dati e le relazioni tra essi sono rappresentate attraverso delle proposizioni ("triple") della forma `<soggetto> <predicato> <oggetto>`.

Gli elementi devono avere dei nomi appropriati, devono seguire semplici convenzioni e devono rispettare gli *schemi URI* dei Linked Data. In pratica, ogni elemento deve avere un indirizzo Web univoco che dev'essere "dereferenziabile", per esempio

`<http://spcdata.digitpa.gov.it/Amministrazione/PCM>` oppure `<http://spcdata.digitpa.gov.it/Nazione/Italia>`.

Per una definizione corretta degli URI si rimanda alla guida [64].

Il processo di analisi e modellazione viene svolto eseguendo un insieme di passi in modo non necessariamente sequenziale:

(a) analizzando prima la struttura di dati usata nelle sorgenti (es. relazionale, XML, spreadsheet, ecc.) e gli schemi definiti entro quella struttura (es. schemi logici in un relazionale);

(b) applicando una ricetta di reingegnerizzazione che li trasformi in un linguaggio standard, RDF, serializzato in almeno uno tra XML, N3, N-Triple, Turtle, che genera un sistema di grafi interconnessi, condizione preliminare per una semplice integrazione incrementale con altri dataset dell'organizzazione o esterni e il raggiungimento di un grado elevato di interoperabilità.

(c) portando la struttura dei dati a un livello concettuale: cioè analizzando lo schema logico dei dati sorgente e ricostruendo (è una forma di *reverse engineering*) lo schema concettuale originale. Questa ricostruzione può avvenire nei casi ideali semplicemente trasformando gli schemi concettuali (e.g., ER, UML) esistenti in linguaggi come RDFS o OWL (Ontology Web Language) (Sezione 0). Laddove non c'è più (o non c'è mai stato) allineamento fra schemi concettuali e logici, nei casi semplici può essere sufficiente un dialogo fra un ingegnere della conoscenza e i responsabili dei dati-sorgente. Nei casi standard, occorre usare un processo più sofisticato, riassunto in (d);

(d) acquisendo una parte della cosiddetta ontologia di dominio dell'amministrazione mediante

studio dei requisiti (sezione 6.1), interrogazione degli esperti, analisi di documenti e applicazioni precedenti, per esempio siti Web, interpretando diagrammi concettuali quando disponibili, ecc. L'ontologia consisterà di concetti ("classi"), relazioni ("proprietà") e vincoli di cardinalità ("restrizioni"), oltre a commenti, annotazioni, ecc. I nomi di questi elementi devono essere appropriati, seguire semplici convenzioni e rispettare gli *schemi URI* dei Linked Data come per i dati;

(e) l'ontologia dell'amministrazione così ricostruita diventa ora lo schema di rappresentazione dei dati, che vengono quindi rifattorizzati in funzione degli elementi definiti nell'ontologia. Anche in questo caso, ogni dato diventa un'entità con un indirizzo specifico: un'entità riceve il suo significato dalle relazioni che ha con altre entità o valori simbolici in un grafo. Per esempio, il dato relazionale [DigitPA | Viale Marx 33 | Roma] estratto da una tabella "amministrazione" può diventare:

```
<http://spcdata.digitpa.gov.it/Amministrazione/cnipa>

<http://www.w3.org/2000/01/rdf-schema#label> "DIGITPA" .

<http://spcdata.digitpa.gov.it/Amministrazione/cnipa>
<http://www.geonames.org/ontology#locatedIn>
<http://spcdata.digitpa.gov.it/Comune/H501> .

<http://spcdata.digitpa.gov.it/Amministrazione/cnipa>
<http://spcdata.digitpa.gov.it/indirizzo> "viale Marx 31/49" .
```

Se il livello di qualità è considerato accettabile rispetto ai casi d'uso previsti, i dati possono essere pubblicati già al termine della fase (b). È possibile anche accelerare il processo di formalizzazione di una prima versione dell'ontologia attraverso un approccio automatico che descriva i dati modellati secondo il loro puro schema logico. La differenza, rispetto alla fase (b), indotta dalle attività descritte in (c)(d)(e) è costituita dal maggior grado di precisione e dalla predisposizione all'arricchimento automatico e all'interliking. La modellazione descritta in (d) ha infatti l'obiettivo di definire in modo più preciso e formale il dominio di riferimento. Nella fase di modellazione è bene verificare se esistono ontologie (o schemi di riferimento) standard per esprimere la semantica dei concetti e delle relazioni individuati. Se tale ontologia esiste, è opportuno riutilizzarla, in quanto questo permette di aumentare il livello di interoperabilità semantica e l'integrazione con il Web dei Dati. Tuttavia, laddove questa ontologia evidenziasse elementi ambigui, mancasse di elementi o ne avesse in eccesso, quando applicati al dominio di riferimento, è possibile usare uno o più metodi di raffinamento: (1) aggiunta di ulteriori vincoli all'ontologia per risolvere le ambiguità; (2) riuso solo di parte dell'ontologia per evitare di importare elementi dei quali non si condivide il senso o i vincoli esistenti; (3) riuso di diverse ontologie, possibilmente aggiungendo elementi e/o vincoli che le colleghino fra loro; (4) aggiunta di elementi in un ulteriore modulo, possibilmente collegati alle ontologie riusate. Va tenuto presente che il raffinamento richiede una certa esperienza di modellazione e va effettuato dopo un'attenta analisi dei requisiti. Quest'ultima va infatti condotta per verificare che le esigenze di modellazione nel caso in esame siano soddisfatte dalle soluzioni di modellazione esistenti nelle ontologie da riusare, oppure per verificare se occorre un raffinamento. Per esempio, si consideri un caso riguardante dati fiscali sulle sanzioni da erogare nei confronti di un contribuente per un contributo dovuto dove è importante associare il tempo entro il quale quella sanzione va erogata, e dove l'ontologia da riusare non contiene il design pattern di modellazione adeguato (per esempio una classe "erogazione_sanzione" con relazioni a

sanzioni, contributi, contribuenti e tempo), bensì solo classi distinte e non correlate fra loro, oppure meno classi di quelle necessarie. In questo caso è necessario aggiungere nuove classi, nuove relazioni e/o nuovi vincoli, cioè relazioni possibili, con cardinalità adeguate, fra quelle classi.

Nel caso si scelga di produrre un'ontologia (o uno schema di riferimento), è utile svilupparla coinvolgendo gli stakeholder che rappresentano i possibili fruitori dei dati in modo da creare le basi per il riutilizzo della stessa. Dato che spesso i concetti di base hanno un significato intrinsecamente legato al dominio di riferimento e che gli stessi potrebbero subire modifiche sostanziali nel corso del tempo, nella fase di definizione dell'ontologia può essere utile sfruttare il principio di modularità. Applicando questo principio, i concetti sono sviluppati attraverso l'uso di moduli separati che sono poi richiamanti da un'ontologia di raccordo; in questo modo è più semplice estendere, modificare e sostituire sezioni intere dell'ontologia qualora si renda necessario.

RACCOMANDAZIONI

R5: Utilizzare chiavi naturali, ove possibile, per la creazione degli URI, evitando di usare valori di tipo posizionale all'interno di un documento o di una base di dati**.**

R6: Definire ontologie, a livello di servizio o a livello di singola base di dati, per i diversi dataset che si vogliono pubblicare.

R7: Utilizzare RDFS e OWL per la definizione delle ontologie.

R8: Sviluppare nuove ontologie solo se strettamente necessario, privilegiando invece l'adozione di ontologie e vocabolari condivisi e largamente utilizzati a livello nazionale, europeo e internazionale.

R9: Evitare di definire ontologie estese che mirino a modellare in modo "monolitico" tutte le tipologie di informazioni che caratterizzano i dati gestiti dalla PA, privilegiando una costruzione incrementale e modulare.

R10: Identificare pattern di modellazione semplici nella formalizzazione dell'ontologia.

R11: È preferibile l'allineamento tra ontologie per armonizzare le informazioni gestite in diverse basi di dati, facilitando così la loro gestione nel tempo.

R12: Rendere univoci, mediante l'aggiunta di opportuni vincoli, elementi dell'ontologia che evidenzino ambiguità nel dominio di riferimento.

6.4. Arricchimento

In questa fase i dati, precedentemente bonificati e modellati, sono arricchiti attraverso l'esplicitazione di informazioni di contorno (*metadatazione*) che ne semplificano il riutilizzo e/o attraverso la derivazione di contenuto informativo aggiuntivo tramite l'utilizzo di tecniche di estrazione automatica dell'informazione o di ragionamento automatico (*inferenza).*

6.4.1. Metadatazione

I metadati arricchiscono il contenuto informativo dei dati esplicitandone delle proprietà che semplificano il processo di fruizione dei dati stessi facilitandone la ricerca, il recupero, la composizione e di conseguenza il riutilizzo. Così come i dati, anche i metadati sono espressi attraverso RDF che nasce proprio come standard per l'annotazione semantica di pagine Web per poi divenire lo strumento base per la creazione di Linked Data. Alcuni dei metadati più utili ai fine dell'interoperabilità e del riutilizzo dei dati sono: le informazioni sulla semantica dei dati, le informazioni di contesto e le informazioni di provenienza.

Le informazioni sulla semantica (per esempio commenti, etichette, definizioni, ecc.) servono per chiarire il significato dei dati e facilitarne quindi l'interpretazione, cioè servono per descrivere il significato di attributi, proprietà e classi di entità. Anche attraverso questo tipo di metadati è possibile legare il contenuto informativo con la semantica di riferimento cosi come definita nelle fasi precedenti. Una cosa importante da tenere presente è che uno stesso dato potrebbe essere descritto con informazioni semantiche diverse; questo deriva dalla possibilità di classificare il dato in modo multidimensionale.

Le informazioni di contesto permettono di descrivere i confini di validità del dato o del set di dati di riferimento. Il contesto viene espresso in termini di intervalli di tempo, regioni di spazio, argomenti ed eventuali altri parametri specifici. Anche in questo caso è possibile procedere inizialmente con una modellazione informale e poi valutare se utilizzare un'ontologia (o uno schema) standard o svilupparne una proprietaria, sempre tenendo presente i vantaggi d'interoperabilità derivanti dal riuso. Alcune meta-informazioni di contesto possono essere espresse ricorrendo allo standard consolidato Dublin Core (Standard ISO15836) che individua un insieme di elementi essenziali (titolo, autore, oggetto, editore, ecc.) per la descrizione di qualsiasi materiale digitale accessibile via rete informatica, e di cui esiste una versione RDF [65]. Anche le informazioni sulla licenza rientrano nella categoria dei metadati di contesto; è buona pratica, quando possibile, inserire informazioni sulle licenze all'interno dei dati stessi, quindi assegnare licenze, permessi e vincoli tramite triple RDF. Un esempio può essere quello della licenza Creative Commons per cui è stato definito una schema RDF specifico [66]. Per un approfondimento sulle licenze si veda la Sezione 8.1.

Le informazioni di provenienza descrivono come e da chi i dati sono stati prodotti. Ancora una volta la metadatazione può essere fatta attraverso un'ontologia (o un modello) standard o attraverso una definizione proprietaria. Una valida opzione in prospettiva potrebbe essere quella di considerare l'adozione delle raccomandazioni cui sta lavorando il Provenance Interchange Working Group [43] del W3C. Inoltre, esistono specifici vocabolari/ontologie, alcuni dei quali utilizzati come base dal Provenance WG, creati allo scopo di definire informazioni di provenienza, e che quindi possono essere riusati nel trattare questa tipologia di metadato [96], [97], [98]. Infine, si può notare come anche Dublin Core includa, tra gli altri, informazioni di questo tipo.

6.4.2. Inferenza ed estrazione automatica dell'informazione

Nella fase di arricchimento si possono derivare nuovi dati a partire da quelli esistenti. Il modo

tradizionale è usare la conoscenza già espressa per inferirne altra implicita. Grazie ai dati noti è possibile inferire (per deduzione) nuove informazioni e nuova conoscenza tramite "query" costruttive nel linguaggio di interrogazione per RDF, i.e., SPARQL [47], o tramite ragionatori automatici (Sezione 0) basati su OWL o regole. L'esempio più banale è quello delle relazioni inverse: se sappiamo che

```
<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://www.geonames.org/ontology#locatedIn>
<http://spcdata.digitpa.gov.it/Comune/H501>
```

possiamo inferire automaticamente che

<http://spcdata.digitpa.gov.it/Comune/H501>

è la città di

<http://spcdata.digitpa.gov.it/Amministrazione/PCM>.

Un esempio più interessante è l'inferenza della classe di appartenenza (tipo) quando la relazione di un'ontologia è vincolata a quel tipo. Per esempio, se sappiamo che

<http://www.geonames.org/ontology#locatedIn>

ha come vincolo del codominio

<http://spcdata.digitpa.gov.it/Comune>,

possiamo inferire:

```
<http://spcdata.digitpa.gov.it/Comune/H501> <rdf:type>
<http://spcdata.digitpa.gov.it/Comune> .
```

Un altro modo di derivare nuovi dati è invece l'estrazione di conoscenza da dati testuali, per esempio descrizioni contenute nei record di una base di dati relazionale o testi tratti da documenti amministrativi. Grazie a tecniche di analisi delle lingue naturali è possibile inferire per induzione nuove informazioni, come persone, luoghi, eventi, relazioni, argomenti, che sono poi rappresentate in RDF.

È importante sottolineare, tuttavia, che queste tecniche non sono di facile utilizzo e che per una loro corretta applicazione sono necessarie competenze specifiche. Tecnologie abilitanti sono in fase avanzata di sperimentazione per facilitarne l'utilizzo.

6.5. Linking esterno (interlinking)

In questa fase il contenuto informativo locale viene legato ad altre informazioni, che possono essere altri dataset prodotti dalla stessa amministrazione, oppure insiemi di dati già presenti nel Web dei Dati, Questo consente di garantire una più facile navigazione e un accesso a un più ampio insieme di dati, e di fornire dati attestati a livello 5 della classifica di qualità dei LOD [6].

Il linking esterno consiste nella produzione di triple RDF in cui il soggetto e l'oggetto sono entità appartenenti a dataset diversi (interni o esterni all'amministrazione). In pratica, si tratta di allineare entità di diversi dataset. Per esempio, nel caso si pubblichino dati che fanno riferimento a entità geografiche, è certamente utile aggiungere un collegamento alla stessa entità geografica presente in GeoNames [32]. Questa operazione è necessaria perché data l'unicità degli URI, le due entità che sono concettualmente

la stessa, non lo sono semanticamente finché non si crea, per esempio, una tripla:

```
<http://spcdata.digitpa.gov.it/Comune/H501> <owl:sameAs>
<http://sws.geonames.org/3169070/>.
```

L'operazione di linking, tuttavia, non è sempre banale come può sembrare. Per aiutare l'utente finale in questo processo, si possono utilizzare strumenti specifici che consentono di creare tali collegamenti esterni in maniera agevole. Questi tipi di strumenti rientrano nella categoria degli strumenti di *record linkage* (Sezione 0).

RACCOMANDAZIONI

R15: Verificare la possibilità di collegare i propri dati a DBpedia e/o ad altra base di dati contenuta nella nuvola LOD per garantire interoperabilità semantica anche transfrontaliera.

R16: Preferire collegamenti che rappresentano legami forti tra entità (ad esempio, di uguaglianza, d'inclusione, ecc.).

6.6. Validazione

In generale, possono essere eseguite tre tipi di validazione: quella sintattica, quella logica e quella concettuale. Nella **validazione sintattica** viene verificato che il contenuto dei dati rispetti i formati standard del W3C. A tale scopo esistono diversi strumenti che possono essere direttamente utilizzati; uno di questi è messo a disposizione dal W3C [67]. Nella **validazione logica**, mutuando i metodi di *unit testing* usati nell'ingegneria del software, individua un insieme di casi di test che devono essere soddisfatti. I casi di test possono essere costituiti da un insieme di domande di cui si vorrebbe conoscere la risposta (chiamate a volte "competency questions") e di cui si sa esserci i dati. Tali domande, tradotte in interrogazioni espresse in un linguaggio per l'interrogazione dei dati (tipicamente SPARQL; si veda la Sezione 0 per le relative tecnologie di interrogazione), sono processate e i risultati comparati con quelli attesi. Nel caso in cui alcuni risultati non collimino è possibile dover ricontrollare i passi di analisi e modellazione per identificare eventuali errori commessi. Le query ovviamente possono anche essere negative, cioè possono tentare di trovare direttamente errori di modellazione. Per esempio, è possibile chiedere se esiste qualche entità che è sia una città che una persona: in caso di risposta positiva vi è un buon indizio per cercare errori di progettazione. Infine, nella **validazione concettuale** viene verificata l'adeguatezza dell'ontologia ai requisiti e all'intuizione degli esperti. In generale, se uno solo dei tre tipi di analisi fallisce, è necessario ricontrollare le fasi precedenti.

Un repertorio utile di *design pattern* per la costruzione e la validazione di ontologie e dati è disponibile in [68].

6.7. Pubblicazione

La pubblicazione è qui intesa come un processo incrementale che potenzialmente segue il corso del

progetto rilasciando progressivamente dati sempre più raffinati. È infatti ragionevole pensare che subito a valle del processo di reingegnerizzazione si possano già rilasciare i primi LOD, e che successivamente gli stessi siano prima affiancati e poi sostituiti da versioni concettualmente più adeguate, inter-linked, e arricchite.

Uno degli aspetti principali di questa fase è la selezione della piattaforma di pubblicazione (si veda la Sezione 0). È molto importante che la piattaforma tecnologica metta a disposizione delle funzionalità che facilitino il riutilizzo e l'interoperabilità: per questo è importante mettere a disposizione (direttamente o tramite "hosting") una funzionalità di ricerca (SPARQL endpoint) che consenta di interrogare agevolmente i dataset pubblicati. Inoltre, è altrettanto importante individuare una soluzione che consenta un'integrazione leggera e flessibile con il sistema informativo e organizzativo istituzionale. In questo modo è possibile garantire la sostenibilità del progetto nel tempo e soprattutto l'aggiornamento costante dei dataset pubblicati. Se è infatti ragionevole pensare che un'amministrazione compia un sforzo iniziale importante per il rilascio dei primi dati, è altrettanto evidente che tale impegno non possa essere mantenuto a lungo nel tempo. Per questo motivo è irrinunciabile dotarsi di strumenti tecnologici e organizzativi tali da rendere sostenibile nel tempo (Sezione 10) lo sforzo richiesto dal processo di pubblicazione e aggiornamento dei dati.

RACCOMANDAZIONI

R17: È preferibile pubblicare pochi dati ma di buona qualità e in modalità Linked, anziché pubblicare grosse quantità di dati non interoperabili.

R18: Fare attenzione a non pubblicare dati con URI inconsistenti, soprattutto quando i dati sono di tipo dinamico. Le URI di una stessa entità non devono variare al cambiare della versione del dataset.

R19: Individuare una soluzione tecnologica per la pubblicazione che permetta un'integrazione leggera e flessibile con il sistema informativo e organizzativo dell'amministrazione.

R20: Non limitare il servizio di pubblicazione al semplice download dei Linked Open Data, ma consentire accessi puntuali ai dati con standard come SPARQL.

7. STANDARD, TECNOLOGIE DI BASE E STRUMENTI

Questa sezione affronta la natura più tecnica del processo di pubblicazione di LOD, descrivendo alcuni standard che regolano il mondo dei LOD e alcune delle tecnologie utili a supportare la metodologia proposta nella precedente sezione. Si noti che la trattazione rappresenta la convergenza su questi temi del gruppo di lavoro e pertanto non può considerarsi esaustiva.

7.1. Standard per i Linked Open Data

I LOD ereditano gli standard definiti dal W3C e impiegati nel contesto del Web Semantico. Il Web Semantico è l'evoluzione del Web dei documenti (un grande contenitore di documenti collegati tra loro), verso il Web delle entità. Il Web dei dati è la parte del Web Semantico riguardante i dati e quindi anche i LOD.

Nel Web Semantico, le entità (risorse, dati, cose) hanno un'identità attraverso una URI[13] non ambigua, per esempio `<http://dbpedia.org/resource/Rome>`, sono semanticamente descritte dallo schema (ontologia, vocabolario) usato e collegate tra loro mediante relazioni o "link". Al contrario, nel Web dei documenti, le entità sono solo "riferite" da termini o dati che non hanno un'identità esplicita.

Un'entità può avere identità (URI) diverse e quindi una delle attività più importanti per la preparazione dei dati è la scoperta/definizione di collegamenti fra URI che esprimono la stessa entità, per esempio `<http://dbpedia.org/resource/Rome>` e `<http://spcdata.digitpa.gov.it/Comune/H501>` .

Spesso si fa riferimento allo stack di linguaggi del Web Semantico, rappresentato in

Figura 3. Per ognuno degli elementi dello stack si è cercato di definire un linguaggio standard, cosa avvenuta finora solo ai livelli bassi e intermedi dello stack (oggetto di analisi della presente sezione).

[13] Ultimamente il W3C sta promuovendo l'adozione degli IRI, una generalizzazione degli URI.

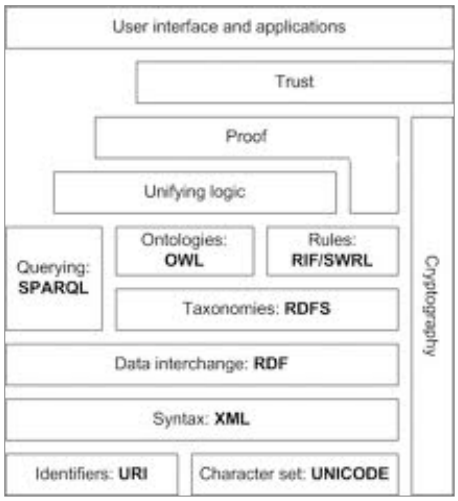

Figura 3: Stack del Web Semantico [5]

I principali standard con i quali è possibile, in modo efficace e omogeneo, rappresentare, ragionare e interrogare, rispettivamente, i Linked Data sono RDF(S), OWL e SPARQL.

RDF (Resource Description Framework)

RDF è un linguaggio relativamente semplice che permette di rappresentare dati e metadati attraverso la definizione di asserzioni, dette triple, secondo lo schema `<soggetto> <proprietà> <oggetto>`.

Gli elementi fondamentali del linguaggio sono le risorse, identificate univocamente per mezzo di URI, che possono comparire in una delle tre posizioni di una tripla. Una risorsa in posizione di proprietà mette in relazione due risorse in posizione <soggetto> e <oggetto>. Una proprietà può anche mettere in relazione una risorsa e un "literal", cioè un'espressione puramente simbolica: numero, stringa, ecc., presa nell'ambito dei "datatype" definiti nello schema XSD. RDF genera così un grafo di nodi interconnessi, chiamato anche grafo RDF.

Esempi di triple RDF possono essere:

```
<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://spcdata.digitpa.gov.it/Amministrazione> .

<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://spcdata.digitpa.gov.it/eroga>
<http://spcdata.digitpa.gov.it/Servizio/pcm-DRP-1> .

<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://www.geonames.org/ontology#locatedIn>
<http://spcdata.digitpa.gov.it/Comune/H501> .
```

```
<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://spcdata.digitpa.gov.it/indirizzo> "Piazza Colonna, 370" .
```

Una tripla può anche comparire come `<subject>` o `<object>` di un'altra tripla.

RDF è quindi una struttura di dati ricorsiva, analogamente alla struttura grammaticale più astratta delle lingue occidentali (Soggetto-Verbo-Oggetto).

RDF ha anche un'estensione, chiamata RDF Schema (RDFS). RDFS permette di definire semplici schemi. Una proprietà molto usata è `rdfs:subClassOf`, che definisce strutture tassonomiche basate sulla relazione di sotto-insieme ed è molto utile per gestire la cosiddetta eredità dei tipi e delle restrizioni (descritte sotto).

Consideriamo degli esempi con delle ipotetiche triple. Se `<Città> <rdfs:subClassOf> <Luogo>` e `<Roma> <rdf:type> <Città>`, allora è possibile inferire che `<Roma> <rdf:type> <Luogo>`.

Altre due proprietà importanti di RDFS sono `rdfs:domain` e `rdfs:range`, che permettono di definire il dominio e il codominio (restrizioni "globali") di una proprietà RDF, per esempio `<natoA> <rdfs:domain> <Cittadino>` e `<natoA> <rdfs:range> <Comune>`. Queste proprietà permettono di inferire per esempio che se `<MarioRossi> <natoA> <Roma>`, allora si inferisce anche che `<MarioRossi> <rdf:type> <Cittadino>` e `<Roma> <rdf:type> <Comune>`. Due altre proprietà molto utilizzate sono `rdfs:label` e `rdfs:comment`, che permettono di aggiungere etichette e commenti a ogni elemento di un dataset o di un'ontologia. Per esempio:

```
<http://spcdata.digitpa.gov.it/Amministrazione/PCM>
<http://www.w3.org/2000/01/rdf-schema#label> "Presidenza del Consiglio
dei Ministri" .
```

Il modello RDF è supportato da diverse rappresentazioni sintattiche, quali RDF/XML, N3, N-Triple e Turtle. La scelta tra queste diverse soluzioni sintattiche, anche dette serializzazioni di RDF, deve essere compiuta sulla base di requisiti voluti, quali la compattezza, lo spazio fisico utilizzato, la leggibilità, ecc. Le serializzazioni sono comunque fra loro inter-traducibili.

OWL (Web Ontology Language)

RDF permette di assegnare un tipo a qualsiasi entità, quindi anche i tipi stessi (che sono sempre entità) o le proprietà; per esempio, l'entità `spcdata:Amministrazione` menzionata nella tripla precedente è a sua volta una `rdfs:Class`:

```
<http://spcdata.digitpa.gov.it/Amministrazione>

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

<http://www.w3.org/2000/01/rdf-schema#Class>
```

mentre `spcdata:eroga` è a sua volta una `rdf:Property`:

```
<http://spcdata.digitpa.gov.it/eroga>
      <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
```

```
< http://www.w3.org/1999/02/22-rdf-syntax-ns#Property>
```

Le entità usate come tipi (classi) o proprietà formano il vocabolario (chiamato anche schema o ontologia) usato da un dataset.

Nei casi migliori, il vocabolario è scritto in OWL, una famiglia di linguaggi per la rappresentazione della conoscenza mediante ontologie. OWL è diventato lo standard W3C per la rappresentazione di ontologie su Web e può essere rappresentato come un'estensione di RDF(S). OWL permette di esprimere le ontologie in maniera più dettagliata e precisa di RDF(S), garantendo anche la possibilità di verificare automaticamente la correttezza logica di ciò che si rappresenta.

OWL permette inoltre l'uso di ragionatori automatici per le cosiddette logiche descrittive, per derivare inferenze logiche dalla struttura dei dati; per esempio, un ragionatore automatico genera le inverse delle triple, le triple simmetriche, le triple che valgono transitivamente (laddove una regola transitiva sia stata introdotta nel vocabolario), l'eredità delle caratteristiche di una classe (laddove queste caratteristiche siano state definite nel vocabolario), l'appartenenza a una classe, date certe condizioni, ecc.

SPARQL (Sparql Protocol And RDF Query Language)

È un linguaggio con una sintassi simile a quella SQL per l'interrogazione di dati RDF e un protocollo di comunicazione basato su HTTP. Uno SPARQL client può quindi interrogare un endpoint SPARQL con interrogazioni ("query") riguardanti un grafo RDF. Le query esprimono le caratteristiche che un sottografo (un insieme di connessioni tra risorse di un certo tipo e con certe caratteristiche) del dataset RDF deve avere. Le risposte alla query sono tutti quei sottografi del grafo RDF che soddisfano le caratteristiche volute. In termini più astratti, si può dire che SPARQL consente di fare "graph pattern matching" all'interno di dati RDF. Per esempio supponiamo di voler conoscere l'indirizzo della Presidenza del Consiglio dei Ministri:

```
select ?indirizzo ?comune_uri
   where {
     ?pcm ?label "Presidenza del Consiglio dei Ministri" .
     ?pcm <http://spcdata.digitpa.gov.it/indirizzo> ?indirizzo .
     ?pcm <http://www.geonames.org/ontology#locatedIn> ?comune_uri .
    }
```

SPARQL può essere utilizzato sia per estrarre dati, sia per costruire viste sul Web dei Dati, usando delle "query" di tipo "construct" invece di "select". Si supponga, per esempio, di voler costruire le triple di una nuova proprietà "nomeComune" a partire dalle triple. La query SPARQL sarà:

```
construct { ?indirizzo nomeComune ?comune }
   where {
     ?pcm ?label "Presidenza del Consiglio dei Ministri" .
     ?pcm <http://spcdata.digitpa.gov.it/indirizzo> ?indirizzo .
     ?pcm <http://www.geonames.org/ontology#locatedIn> ?comune_uri .
     ?comune_uri ?label ?comune .
```

}

ALTRI STANDARD

Nell'ambito dei dati statistici, un gruppo di lavoro tecnico dell'SDMX ha avviato un insieme di attività che mirano a integrare lo standard SDMX con RDF. Il risultato di questa integrazione è rappresentato dal Data Cube vocabulary [76] che consente di pubblicare sul Web dati multi-dimensionali in RDF e conformi all'ISO SDMX. I dati pubblicati tramite RDF Data Cube possono essere combinati in modo flessibile anche a dati non statistici. Si può, ad esempio, pensare a un'interrogazione in cui si richiedano "tutte le scuole religiose in zone agricole con alti valori di indicatori nazionali che misurano la tolleranza religiosa". L'integrazione di tali standard fa sì che anche i dati statistici diventino parte integrante del Web dei Dati.

7.2. Tecnologie a supporto dell'approccio metodologico LOD

Nel variegato mondo dell'Open Data e dei LOD esiste un'ampia gamma di tecnologie utilizzabili per differenti scopi. La moltiplicazione di queste tecnologie è direttamente riconducibile ai numerosi standard che, in tempi recenti, si stanno affermando. Le tecnologie a supporto degli standard non sono ancora del tutto mature e interoperabili tra loro, così come non esistono, ancora, piattaforme "all-in-one".

A seconda degli input al processo di generazione dei LOD, degli output che da esso si desiderano ottenere e delle risorse disponibili, le PA e i vari produttori di dati sono chiamati a svolgere operazioni diverse mediante ricorso a strumenti e tecnologie, a loro volta, differenti. Questa sezione illustra le principali tecnologie e strumenti oggi disponibili, suddividendoli in macro-categorie corrispondenti alle fasi proposte nell'approccio metodologico della sezione precedente.

7.2.1. Tecnologie per la bonifica dei dati

Come riportato nella Sezione 6.2 occorre fare in modo che i dati siano conformi a dei criteri e che rispettino alcuni dei vincoli di qualità e integrità. Questa fase di pulizia ("data cleansing") può essere affrontata con varie tecnologie e strumenti, alcuni dei quali riportati nel seguito.

ETL

Questi strumenti permettono di trattare fonti dati eterogenee, fare operazioni di estrazione, normalizzazione, denormalizzazione, riconciliazione e "data cleansing" in maniera semplice, intuitiva e scalabile.

Un processo di ETL, una volta sviluppato, può diventare di fatto il componente che si occupa dell'aggiornamento del dataset pubblicato. Mentre in passato queste operazioni e questi strumenti erano impiegati prevalentemente per la popolazione di data warehouse, oggi, vista la loro flessibilità d'uso, si

qualificano anche come strumenti per la messa a sistema di operazioni semplici e ripetitive (nella fattispecie l'aggiornamento di dataset) in un ambiente facilmente mantenibile.

Esempi di strumenti ETL Open Source sono Kettle e Talend Open Studio. Entrambi molto utilizzati anche in ambiti enterprise, sono ricchi di molte componenti "già pronte".

Kettle è molto intuitivo e facile da usare. La conferma della facilità di utilizzo si nota immediatamente nel modo in cui vengono realizzati i processi di ETL e di trasformazione. Rispetto a Talend, riduce i passaggi in cui è necessaria la generazione di codice Java per effettuare la trasformazione. La piattaforma Talend, di contro, risulta più aperta e orientata al mondo Java.

SISTEMI DI GESTIONE DEI DATI

Anche i Sistemi di gestione dei dati (DBMS), possono essere utilizzati per compiere operazioni di pulizia di dati, specialmente quando i dati da aprire risiedono proprio all'interno di un DBMS. I database relazionali open source più noti sono PostgreSQL e MySQL. Ultimamente, per contrastare la crescita esponenziale della quantità di dati e la loro multidimensionalità e incompletezza, si stanno sempre più affermando i DBMS non-relazionali (NoSQL).

GOOGLE REFINE è uno strumento open source per la pulizia, l'analisi e l'elaborazione di dati testuali. Esso può operare su diversi tipi di dati in ingresso e offre un piccolo supporto alla funzione di "record linkage" grazie a funzionalità di espansione semantica.

7.2.2. Tecnologie per l'analisi e la modellazione dei dati

In questa fase, è utile individuare strumenti che facilitino la reingegnerizzazione del dataset secondo il nuovo modello logico e concettuale di output.

PROTEGÉ[14] è una soluzione open-source scritta in Java sviluppata dall'Università di Stanford (e in seguito con l'aiuto dell'Università di Manchester) per la definizione e manutenzione di ontologie (schemi, vocabolari) in RDF/OWL.

È un editor di ontologie la cui architettura è basata su plug-in estendibili. Esso ha un'interfaccia grafica intuitiva ed è supportato da una nutrita comunità. L'attenzione è posta su OWL, quindi i Linked Data in RDF puro hanno spesso qualche problema a essere gestiti in maniera adeguata. Il limite principale di Protégé è la mancanza di un plugin per utilizzare SPARQL, dovuta appunto all'approccio "OWL-centric". Protégé implementa un buon debugger logico per le ontologie.

[14] http://protege.stanford.edu

NeOn Toolkit[15] è un'altra soluzione open-source in Java, sviluppata dal progetto europeo NeOn, per la definizione e manutenzione di ontologie (schemi, vocabolari) in RDF/OWL. Anche questo editor di ontologie è basato su plug-in. Diversamente da Protégé, è pienamente compatibile con RDF e gestisce SPARQL, associandolo ai ragionatori automatici per OWL. Essendo molto più recente, la comunità di supporto è limitata. NeOn Toolkit implementa un buon debugger logico, ma anche un debugger basato su "best practices" di modellazione. Attraverso i suoi plug-in, ci sono anche funzionalità di progettazione avanzata che utilizzano i cosiddetti "ontology design patterns", e funzionalità semplici e intuitive di visualizzazione.

TOPBRAID COMPOSER[16] è una soluzione proprietaria per la modellazione di ontologie e dati, centrata su RDF. OWL è supportato, ma l'applicazione è poco interoperabile con i ragionatori automatici più evoluti. Il supporto per SPARQL è probabilmente il migliore disponibile attualmente, grazie all'estensione SPIN che fa di SPARQL un vero e proprio linguaggio per esprimere regole. Contiene anche vari strumenti per l'importazione di dati da formati diversi. Come anche gli altri editor di ontologie, non è molto scalabile, in quanto pensato per la gestione di schemi e non di grandi dataset. Diversamente da altri editor, è però possibile associare "triple store". Il debugging di ontologie e dati è molto limitato e bisogna realizzarlo inserendo "unit test" personalizzati in SPARQL.

D2R[17] (Database to RDF) è un framework che permette l'accesso al contenuto di un database relazionale come se si trattasse di un dataset RDF.

Tramite un linguaggio dichiarativo è possibile definire clausole che, in modo flessibile descrivono le relazioni tra le ontologie di riferimento e la struttura tabellare della base di dati. L'applicazione delle clausole consente, quindi, la produzione di un dataset RDF (anche chiamato "RDF dump") secondo le specifiche desiderate. Il suo limite principale è la mancanza di una semplice e intuitiva interfaccia utente.

STANBOL RULES[18] è un framework open-source, realizzato nell'ambito del progetto europeo IKS, che permette la definizione di regole e la rifattorizzazione su grafi RDF, associato a ragionatori automatici. Fa parte della piattaforma Stanbol[19], un insieme di componenti per l'applicazione di tecnologie semantiche ai CMS. Il linguaggio di RULES è pienamente compatibile con SPARQL e offre soluzioni compatte per la manipolazione di grafi RDF distribuiti.

7.2.3. Tecnologie e linguaggi per l'arricchimento dei dati

Le tecnologie che rientrano in questa categoria permettono di "verificare", "inferire", o "popolare"

15 http://www.neon-toolkit.org
16 http://www.topbraidcomposer.com
17 http://d2rq.org/d2r-server
18 http://incubator.apache.org/stanbol/docs/trunk/rules.html
19 http://incubator.apache.org/stanbol/

automaticamente i dati.

La verifica automatica di dati e ontologie serve a controllare che non siano presenti incoerenze nello schema dei dati (per esempio che qualcosa possa essere sia una città sia una persona, se si è indicato esplicitamente nello schema OWL che la classe Città è disgiunta dalla classe Persona), né inconsistenze nei dati stessi (per esempio che i dati contengano triple per cui una certa entità è sia una città sia una persona).

La verifica è importante per l'arricchimento perché serve a evitare che si introducano automaticamente affermazioni non valide, e serve a garantire le prestazioni di un ragionatore automatico, che non funzionerebbe, o sarebbe limitato, in caso di incoerenze o inconsistenze.

Il ragionamento automatico su dati e ontologie permette l'arricchimento mediante inferenze logiche; in pratica, mediante la materializzazione della conoscenza logicamente implicita negli schemi e nei dati esistenti (Sezione 6). La materializzazione permette di costruire il cosiddetto "modello inferito" dei dati, cioè l'insieme di triple che si possono dedurre logicamente dalla struttura dei dati esistenti. Il ragionamento automatico è tecnicamente limitato a dati consistenti e a schemi coerenti: per questo motivo gli strumenti di ragionamento automatico svolgono anche il ruolo di "consistency checker". I migliori ragionatori automatici sono, in teoria, quelli che permettono inferenze complete su dati e schemi (in particolare su ontologie in OWL), ossia quelli che non trascurano nulla. A questo insieme appartengono attualmente HermiT[20] dell'Università di Oxford, dotato di licenza aperta LGPL, FacT++[21] dell'Università di Manchester (anche questo aperto, ma come GPL), Pellet[22] di Clark&Parsia, che ha una licenza aperta accademica, ma commerciale per le aziende, RacerPro[23] di Racer Systems, che ha solo licenze commerciali.

Esistono poi ragionatori automatici che lavorano solo su sottoinsiemi del linguaggio OWL per ottimizzare l'efficienza computazionale. I più interessanti sono: QuOnto[24] dell'Università La Sapienza di Roma, che lavora sulla parte di OWL equivalente all'espressività di un database relazionale, quindi utile per gli approcci che vogliono usare tecnologie semantiche per accedere direttamente a dati relazionali senza tradurli in Linked Data; Oracle 11g[25] (commerciale), che lavora su un altro sottoinsieme di OWL più efficientemente computabile ed è integrato con il DBMS di Oracle.

Ci sono poi ragionatori automatici che si possono definire incompleti, che si preoccupano più dell'efficienza che della completezza logica e della verifica di consistenza e coerenza. Questo tipo di approccio comprende molte diverse soluzioni ed è raccomandabile soprattutto quando non si lavora con OWL, ma solo con dati RDF(S). Tra questi si può menzionare OWLIM[26].

Il popolamento automatico di dati e ontologie è infine un tipo di arricchimento basato su tecnologie di estrazione della conoscenza. In questo caso, la conoscenza implicita che si vuole inferire è quella linguistica e l'accuratezza non è quasi mai del 100%. Questi strumenti sono quindi utilissimi quando i dati contengono molto testo, o quando è importante far emergere la conoscenza "sepolta" dentro

20 http://www.hermit-reasoner.com/
21 http://owl.man.ac.uk/factplusplus/
22 http://clarkparsia.com/pellet/
23 http://www.racer-systems.com/products/racerpro/
24 http://www.dis.uniroma1.it/quonto/
25 http://www.oracle.com/technetwork/database/enterprise-edition/overview/index.html
26 http://www.ontotext.com/owlim

documenti testuali, accettando in compenso un certo margine di imprecisione.

Le tecnologie disponibili per l'arricchimento automatico sono moltissime ed è difficile darne una valutazione obiettiva. Si va dai *riconoscitori di entità* (cioè nomi propri come "Mario Rossi" o "Berlin") ai *risolutori di entità*, che permettono di dare un'identità specifica alle entità riconosciute (per esempio http://dbpedia.org/resource/Berlin), agli *estrattori di relazioni* (per esempio "Berlin isTheCapitalCityOf Germany"). Esistono molti strumenti, sia open source (e.g., GATE[27] dell'Università di Sheffield, TermExtractor[28] della Sapienza Università di Roma, DBpedia Spotlight[29] dell'Università di Lipsia, Text2Onto[30] dell'Università di Mannheim, FRED[31] dell'STLab del CNR) sia commerciali (e.g., Alchemy[32], Zemanta[33]), che implementano queste funzionalità; tipicamente gli strumenti commerciali offrono maggiore precisione e sono più facili da usare.

7.2.4. Tecnologie e linguaggi per l'interlinking dei dati

Come è ovvio immaginare, il linking è una funzionalità molto importante per i Linked Data e di fatto può essere considerata una forma particolare di arricchimento. La particolarità consiste nel fatto che l'arricchimento avviene grazie all'interlinking fra dataset di origine diversa, tipicamente fra amministrazioni o istituzioni diverse, ma anche, al limite, all'interno di una stessa amministrazione.

SILK [89] è uno degli strumenti più utilizzati per fare "record linkage". È in grado di scoprire somiglianze (*similarity*) fra entità appartenenti a diverse sorgenti e materializzare i collegamenti in nuovi dataset RDF.

7.2.5. Tecnologie e strumenti per la pubblicazione dei dati

Probabilmente la più vasta categoria di tecnologie di LOD è quella sul processo di pubblicazione e accesso ai dati.

Come detto nella Sezione 6.7 al fine di permettere il riuso agile dei LOD, è opportuno esporre uno SPARQL endpoint. In questo modo si sfrutta pienamente la strutturazione dei dati e si consente l'accesso puntuale al dato di interesse. Per una conoscenza più dettagliata dei sistemi che espongono SPARQL endpoint si può far riferimento a un famoso benchmark [77] [86]. In generale, è importante evidenziare come tali sistemi, vista anche la loro recente adozione, siano ancora affetti da problematiche relative alla scalabilità in presenza di grandi dataset sui quali eseguire un elevato numero di interrogazioni. Pur se tale circostanza è indice di successo, essa pone evidentemente problemi

27 http://gate.ac.uk/
28 http://lcl.uniroma1.it/termextractor
29 http://dbpedia.org/spotlight
30 http://code.google.com/p/text2onto/
31 http://wit.istc.cnr.it/stlab-tools/fred/
32 http://www.alchemyapi.com/
33 http://www.zemanta.com

prestazionali da governare. A tal riguardo, opportuni meccanismi forniti dai sistemi descritti possono essere sfruttati per contrastare le suddette problematiche. Per esempio, si può efficacemente limitare il numero di risultati da visualizzare in una singola interrogazione attraverso l'uso di clausole del linguaggio di interrogazione, oppure, in maniera più sofisticata, si possono configurare i sistemi in base all'uso specifico che l'amministrazione ne deve fare (e.g., uso di indici, tecniche di caching, ecc.). In ogni caso, qualora i predetti accorgimenti non risultassero sufficienti, è necessario agire a livello di risorse, dispiegando i sistemi in maniera potenziata e più flessibile così da bilanciare il carico di lavoro atteso.

Di seguito si riportano alcuni dei sistemi attualmente utilizzati per il processo di pubblicazione e accesso ai dati.

Openlink Virtuoso Universal Server è un knowledge store in quanto offre funzionalità di RDBMS, ORDBMS, XML Database, RDF Store, Web Service Server. È utilizzato principalmente come RDF Store poiché rappresenta una soluzione completa e offre un supporto unitario in grado di gestire quasi tutte le principali questioni relative alla gestione dei Linked Data. Raggiunge buone performance grazie a una rappresentazione fisica dei dati RDF in un "Quad Store" e consente di esporre e configurare molto semplicemente uno SPARQL endpoint.

Il sistema è disponibile in diverse versioni: commerciale, "cloud" (modalità PaaS) e ridotta rilasciata sotto licenza Open Source.

D2RQ SERVER la piattaforma D2RQ già sopra introdotta offre, oltre alla capacità di trasformazione "bulk" del contenuto di una base di dati relazionale in RDF, anche la capacità di gestire varie modalità di accesso ai dati, fornendone la navigazione Web (servizio di deferenziazione) e mediante hyper-data browser. In particolare, offre anche un punto di accesso SPARQL. Tale risultato è possibile attraverso uno strato intermedio di interpretazione delle richieste in ingresso, che sfrutta un "mapping" definito nel linguaggio D2RQ. Esso è quindi in grado di trasformare un'interrogazione SPARQL in un'interrogazione SQL tramite un confezionamento dell'informazione che maschera la natura relazionale dello strato fisico sottostante.

Sia che la pubblicazione dei dati avvenga mettendo a disposizione i file contenenti le triple del dataset sia che essa sia realizzata scegliendo di non mettere a disposizione dei dati interoperabili, occorre predisporre un portale o sito Web.

Si può far riferimento a una serie di opzioni tecnologiche che facilitano questo tipo di pubblicazione.

PORTALE CKAN

Il sistema CKAN, sviluppato da Open Knowledge Foundation, è un prodotto nato per la catalogazione di risorse aventi la natura di file accessibili tramite URL.

Si tratta di una piattaforma onnicomprensiva, ben integrata e altamente personalizzabile, con cui si possono realizzare tutti gli elementi di un sistema di gestione di Open Data, dalla loro memorizzazione fisica, organizzazione logica, metadatazione e, infine, esposizione su un sito Web. Essa copre i seguenti aspetti del processo di pubblicazione di un dataset:

- sistema di redazione delle schede dei metadati;
- storicizzazione automatica;

- sistema di memorizzazione dei file;
- pubblicazione delle schede all'interno di un portale Web personalizzabile;
- funzioni di ricerca per chiave;
- sistema di esposizione dei dati come servizio mediante API di tipo REST;
- funzioni di anteprima per i tipi di formati più comuni;
- funzioni basilari di analisi statistica degli accessi;
- sistema minimale di quality assurance e feedback dagli utenti.

Il linguaggio con cui è stato programmato CKAN è Python, che per la realizzazione dei servizi sopra elencati si integra con un database PostgreSQL e un motore di ricerca Solr su application server Jetty (un Apache Web Server).

Grazie alla potenza di queste API, quindi, è possibile integrare le funzionalità del sistema di catalogazione all'interno di altri software. Ad esempio, è possibile inserire qualunque funzione di CKAN all'interno di un CMS moderno, svincolando in tal modo la fase di pubblicazione da quella di redazione.

RACCOMANDAZIONI

R21: È preferibile registrare il proprio dataset al portale CKAN.

CMS

Molti dei portali esistenti sono basati, per la realizzazione delle funzioni di gestione di contenuti, su CMS Open Source. Questi sistemi sono utili qualora i fruitori dei dati siano solamente utenti umani.

I tre principali e più utilizzati prodotti CMS sono Wordpress, Drupal e Joomla Essi sono sufficientemente flessibili e personalizzabili da adeguarsi allo scopo di pubblicazione di (Linked) Open Data. In particolare in Drupal sono presenti in maniera nativa le tecnologie legate al Web Semantico, come RDF e RDFa; inoltre esso possiede un modulo per CKAN che integra le funzionalità del catalogo, di un motore di ricerca e di wiki.

Come detto per Stanbol, il progetto europeo IKS[34] ha sviluppato un insieme di componenti per passare agevolmente da un CMS tradizionale a uno dotato di tecnologie semantiche.

Geoportali

Con il termine geoportale si intende un sito Web realizzato in modo tale da costituire un punto di accesso unico (gateway) ai servizi relativi a dati e risorse spaziali, che non devono necessariamente risiedere all'interno del sito stesso ma che possono invece essere distribuiti. Soggetti, siano essi pubblici,

[34] http://www.iks-project.eu

privati o generiche comunità di utenti che realizzano un geoportale, permettono l'accesso a informazioni territoriali per mezzo di un'interfaccia Web e attraverso l'utilizzo di web services. Dall'altro lato, gli utenti Web che hanno la necessità di utilizzare informazioni territoriali, siano essi utenti generici, professionisti o pubbliche amministrazioni, utilizzano i geoportali per ricercare, accedere e utilizzare i dati.

I geoportali possono rappresentare inoltre un elemento fondamentale delle infrastrutture di dati spaziali (SDI - Spatial Data Infrastructure), di cui l'esempio più significativo a livello nazionale e europeo è la SDI definita e regolata dalla direttiva INSPIRE (Direttiva 2007/2/CE).

Per garantire la ricerca e l'accesso alle informazioni territoriali, un geoportale offre all'utente una serie di servizi che, pur essendo spesso fruibili tramite interfaccia grafica per mezzo di un comune browser Web, devono essere implementati, all'interno di un'architettura di tipo Service Oriented (SOA), come veri e propri web-services realizzati in modo conforme a una serie di standard che ne garantiscano l'interoperabilità.

Per quanto riguarda il livello nazionale ed europeo la direttiva INSPIRE prevede che i servizi implementati siano resi disponibili attraverso il geoportale comunitario ed eventualmente attraverso punti di accesso nazionali. I servizi previsti dalla Direttiva sono:

1. discovery services, che permettono la ricerca dei dati territoriali e relativi servizi attraverso i metadati e di visualizzare i metadati stessi;
2. view services, che consentono la visualizzazione dei dati territoriali;
3. download services, che consentono lo scaricamento dei dati;
4. transformation services, che consentono la trasformazione dei dati allo scopo di conseguire l'interoperabilità;
5. invoke services, che consentono di richiamare altri servizi sui dati territoriali.

Per l'implementazione di tali servizi, sono state predisposte apposite linee guida tecniche INSPIRE che fanno riferimento ad una serie di standard internazionali già esistenti; in particolare, per quanto riguarda il contenuto dei metadati, agli standard ISO della serie 19100 (si veda la parte di dati territoriali della Sezione 4) e a quelli OGC (Open Geospatial Consortium), per quanto riguarda i dettagli tecnici e architetturali.

Allo stato attuale, INSPIRE ha predisposto e reso disponibili le linee guida relative ai primi due servizi; per gli altri, invece, le linee guida sono ancora in versione di bozza[35]. Per quanto riguarda i servizi di discovery (1), i servizi di view (2) e i servizi di download (3), gli standard cui fare riferimento sono quelli del OGC, relativamente ai CS-W (Catalogue Services for the Web)[36], ai WMS (Web Map Services)[37] e ai WFS (Web Feature Services)[38].

Attualmente, sono disponibili diverse soluzioni software applicative sia open-source che proprietarie,

[35] Le linee guida tecniche sono disponibili sul sito di INSPIRE al link http://inspire.jrc.ec.europa.eu/index.cfm/pageid/5

[36] La specifica "*OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile*" è disponibile sul sito di OGC al link http://www.opengeospatial.org/standards/specifications/catalog.

[37] Standard ISO 19128:2005 "*Geographic Information -. Web Map Service Interface*"; OpenGIS Web Map Service (WMS) Implementation Specification disponibili al link http://www.opengeospatial.org/standards/wms

[38] Standard ISO 19142 "*Geographic Information - Web Feature Service*"

che permettono di pubblicare all'interno di un geoportale servizi web conformi agli standard OGC (CS-W, WMS e WFS).

Per quanto riguarda il panorama open-source, le soluzioni software più utilizzate per la realizzazione di geoportali, sia in ambito pubblico che privato, possono essere suddivise in tre categorie principali:

- applicazioni server che implementano e consentono la pubblicazione di web services per la visualizzazione e il download di dati territoriali;
- applicazioni server che implementano e consentono la pubblicazione di web services per la gestione dei metadati relativi ai dati territoriali;
- librerie software API per la fruizione tramite browser web di dati territoriali.

Tra la applicazioni open-source per la pubblicazione e il download di dati territoriali (a) si citano Geoserver[39], UMN Mapserver[40] e Degree[41], tra quelli per la gestione dei metadati (b) Geonetwork[42], Degree e Geoportal[43], mentre per quanto riguarda le API per la fruizione di dati territoriali (c) OpenLayers[44].

Un'ulteriore modalità di pubblicazione dei dati è quella dell'"opendata-as-a-service" dove anziché mettere a punto una piattaforma in-house si sfruttano piattaforme di cloud (pubblico o privato).

SOCRATA: è una soluzione proprietaria Open Data as a service composta da: un sistema semantico di archiviazione dell'informazione; un'interfaccia web dinamica, per l'accesso ai dati mediante maschere parametriche; una Socrata Open data API, per esporre interfacce applicative; una serie di strumenti di indagine statistica e di visualizzazione grafica con cui realizzare semplici attività di data mining; un sistema di social networking integrato, con cui gestire il feedback degli utenti; un sistema integrato di metadatazione e classificazione dell'informazione.

Offre la possibilità di caricare dataset su di un sistema esterno, e di utilizzare una serie di funzionalità avanzate che permettono ad un ente anche sprovvisto di un proprio asset IT interno di avere esposti i propri Open Data. La piattaforma SOCRATA è quella utilizzata data.gov.

OpenLab - già OGDI (Open Government Data Initiative) [82]: è un progetto di Microsoft Corp. Nato per fornire ad enti governativi e pubbliche amministrazioni un servizio "opendata-as-a-service". I dati sono messi a disposizione mediante interfacce standard (API), pensate espressamente per la realizzazione di applicazioni Web.

La tecnologia di memorizzazione dei dati è la piattaforma Microsoft Azure, ovvero un sistema cloud orientato alle applicazioni, già integrato con il framework di sviluppo *.NET*. Su questa base si innestano inoltre una serie di strumenti che realizzano un sistema integrato di interfacce per l'accesso e la gestione

39 http://geoserver.org
40 http://mapserver.org
41 http://www.deegree.org
42 http://geonetwork-opensource.org
43 http://www.esri.com/software/arcgis/geoportal
44 http://openlayers.org/

dei dati presenti in Azure. OpenLab è gratuito e contiene componenti Open Source.

Come tutte le soluzioni di hosting, i vantaggi di OpenLab stanno soprattutto nella quantità di strumenti e di interfacce già a disposizione, e nella totale mancanza di carico per quanto riguarda la parte di archiviazione e gestione dei dataset.

Analogamente alla piattaforma Socrata, con OpenLab è necessario comunque aggiungere l'impegno di risorse umane interne per l'associazione della componente semantica ai dataset caricati.

Esistono dei casi in cui i dati sono forniti attraverso API, servizi Web oppure attraverso sistemi di ricerca e di navigazione non convenzionali.

WEB SERVICES: nel caso in cui si volesse sviluppare un sistema informativo interno che si basi sugli Open Data si può pensare anche alle classiche tecnologie di "remote call". In questo caso il consumatore degli Open Data è rappresentato da un programma o da un sistema informativo, anziché una persona, e, l'aspetto cruciale passa quindi dalla presentazione dei dati alle funzionalità che possono essere richiamate e ai loro metodi di invocazione. In questo campo la tecnologia più diffusa è rappresentata dai web service, che si dividono sostanzialmente in due tipologie:

- WS SOAP: usa il protocollo HTTP solo per il trasporto dell'informazione e ricostruisce completamente lo strato di servizio dedicato alla comunicazione e l'interoperabilità tra soggetti;
- WS REST: poggia esclusivamente sui servizi del protocollo http per eseguire operazioni su un insieme di dati che viene opportunamente elaborato e restituito a seconda della URL chiamata, come se già esistesse in quella forma a quell'indirizzo. Nel caso in cui i dati sono fruiti come servizi, è importante che i primi siano restituiti in un formato facilmente elaborabile da un'applicazione software.

API (Application Program Interface): una via di interrogazione percorribile è quella resa possibile dalle *API*, interfacce realizzate tramite applicazioni che permettono la comunicazione con altre applicazioni. Di solito invisibili all'utente, consentono l'interazione di applicazioni Web o mobile, con un determinato servizio o con repository di dati.

MOTORI DI RICERCA E NAVIGAZIONE A FACCETTE: la quasi totalità degli utenti che cercano dati non conosce i linguaggi d'interrogazione né la strutturazione dei dataset. Per questo, sono essenziali nuove forme d'interrogazione. Una di queste è quella tramite parole chiave (keyword), la più comune nei moderni motori di ricerca. Per consentire questo tipo di ricerca è opportuno utilizzare strumenti che siano in grado di compiere un'indicizzazione e una ricerca di tipo "full-text" dei dati. I principali prodotti per queste attività sono Lucene e Solr.

Lucene, della Apache Software Foundation, è sviluppato nativamente in Java, è multipiattaforma e distribuito con Apache License 2.0. È rilasciato come API per l'indicizzazione e la ricerca "full text" ma è ormai divenuto uno standard "de facto" come componente per motori di ricerca. Solr, un altro prodotto sviluppato dalla Apache Software Foundation e che si appoggia a Lucene, è molto utilizzato nell'implementazione di motori di ricerca verticali all'interno di portali Web. Esso consente di abbinare una ricerca "full text" con la ricerca per "tag", ricerche geografiche e altro ancora. Fornisce inoltre

interfacce non solo HTTP ma anche JSON e XML.

Strumento proprietario molto diffuso, è la Google Search Appliance, che viene schierata con una componente hardware da inserire nel proprio centro elaborazione dati, e con un'assistenza da remoto assicurata da Google o da suoi partner.

Qualora i dati siano stati metadatati semanticamente è possibile fornire agli utenti finali una navigazione basata su *facets* (faccette), cioè filtri dinamici che individuano i dati attraverso le informazioni che li descrivono. Questo tipo di ricerca consente d'impostare la ricerca dell'utente utilizzando categorie concettuali e intuitive, e di visualizzare i passi successivi di espansione e raffinamento della ricerca. Questa tecnica è particolarmente utile quando l'utente ha difficoltà a esprimere a parole, o non ha ben chiaro, l'obiettivo della propria ricerca.

Per realizzare motori a faccette sono disponibili diverse soluzioni Open Source. Una di esse è Protegé, che consente di associare alla base di conoscenza di un certo dominio le faccette, disegnate alla luce dell'esperienza e dei processi di navigazione che si è scelto di modellare.

Altre soluzioni open-source utili alla creazione di un motore a faccette sono sempre Solr e Sparallax; quest'ultima permette di navigare il contenuto esposto da uno SPARQL endpoint.

7.2.6. Altre tecnologie

Sono da segnalare altri strumenti che non rientrano direttamente nelle categorie definite in base alla metodologia proposta in questo documento, ma che in modo trasversale possono essere utilizzati nel processo di gestione, consumo, interazione e scoperta di LOD.

LIBRERIE PER RDF/OWL

Per l'elaborazione e la trasformazione di dati già presenti in formato Linked Data si può far uso di librerie per RDF/OWL. Ad oggi esistono librerie per ogni linguaggio di programmazione. Tra le più diffuse si possono citare Jena, Sesame e JRDF per Java, RDFlib per Python. Queste librerie sono anche utilizzate per la produzione di applicazioni che usano LOD.

HYPER-DATA BROWSERS

I Linked Data per via della loro proprietà di collegarsi tra loro sono anche chiamati Hyper-data. In tempi recenti sono stati sviluppati browser che permettono di esplorare la basi dati esposte seguendo le norme dei Linked Data. Queste applicazioni possono essere installate sui terminali dagli utenti (i consumatori dei dati) e consentono di navigare fra le informazioni disponibili, seguendo i link che collegano i dati di interesse (e.g., DISCO[45]). Inoltre, questi sistemi forniscono il supporto, lato client, per effettuare complesse interrogazioni che possono richiedere l'accesso a una molteplicità di basi di dati. Alcuni di questi sono già configurati per operare delle ricerche ed esplorazioni sulla nuvola LOD come SWOOGLE [87]e Sindice [88].

[45] http://www4.wiwiss.fu-berlin.de/rdf_browser/

STRUMENTI PER L'INTERAZIONE DEI LOD

Infine, nei casi in cui la produzione di dati si è ispirata o si è evoluta in base a requisiti interni all'organizzazione, può essere importante creare degli strumenti di interazione adeguati. Infatti, nel caso i requisiti manchino, è molto difficile andare oltre gli schemi di interazione generici per i Linked Data, tipicamente gli SPARQL endpoint, eventualmente arricchiti con supporto per i namespace o l'auto-completamento. E' anche possibile visualizzare i dati in HTML mediante dei trasformatori XSLT appropriati; ottimi esempi sono OWLDoc[46] dell'Università di Manchester e LODE[47] dell'Università di Bologna. Uno degli esempi più interessanti di visualizzazione "generica" è RelFinder[48] dell'Università di Stoccarda, che permette di stabilire alcune entità di partenza e di costruire poi i percorsi ottimali per collegare quelle entità nel grafo RDF che ha a disposizione.

Nel caso in cui i requisiti forniscano una motivazione per la costruzione di interazioni motivate e verticalizzate, l'interazione può diventare più sofisticata. Esempi classici di mash-up semantici pensati per scopi specifici sono gli esempi applicativi di SIMILE[49] del MIT di Boston e le applicazioni dimostrative di data.gov sviluppate dal Rensselaer Institute[50]. In Italia, l'esempio a tal riguardo è tra data.cnr.it e il Semantic Scout [57].

RACCOMANDAZIONI

R22: Considerare la piena conformità delle tecnologie agli standard del Web Semantico per la messa in produzione della propria soluzione LOD.

R23: Considerare la maturità delle tecnologie in termini di varietà dell'offerta, ampia diffusione, ampia documentazione, vivacità e supporto della comunità di riferimento.

R24: Adottare meccanismi, nel dispiegamento delle tecnologie, che consentano di alleggerire e bilanciare il carico di lavoro laddove presenti dataset di grandi dimensioni e un elevato numero di potenziali interrogazioni.

46 http://www.co-ode.org/downloads/owldoc/
47 http://www.essepuntato.it/lode
48 http://www.visualdataweb.org/relfinder.php
49 http://simile.mit.edu
50 http://tw.rpi.edu/web/Projects

8. ASPETTI LEGALI E MODELLI DI BUSINESS DEI LINKED OPEN DATA

Questa sezione introduce gli aspetti legali e i modelli di business legati ai (Linked) Open Data. Si è optato di trattare entrambi gli argomenti in un'unica sezione perché, anche se non così evidente, questi due aspetti sono in realtà molto legati tra loro. In particolare, la licenza d'uso è uno strumento per dare attuazione a un preciso modello di business. Ad esempio, la filosofia open source si basa sua una pluralità di licenze che a diverso titolo cercano di favorire lo sviluppo di codice riutilizzabile; in questo contesto i detentori dei diritti permettono lo studio e l'aggiornamento del codice prodotto da parte di altri sviluppatori. La licenza diventa quindi uno strumento per garantire l'accesso al codice di un'opera; questo è esattamente il caso duale rispetto all'esempio tipico del modello "closed source" che non contempla l'accesso al codice sorgente. Nel momento in cui si sceglie il tipo di licenza da applicare al contenuto informativo che si vuole liberare, è molto importante avere ben chiaro e presente gli obiettivi che si vogliono raggiungere e il modello di business cui ci si ispira.

8.1. Licenze d'uso per i dati

L'aspetto relativo alle licenze dei dati è cruciale per quanto riguarda l'uso che gli utenti possono effettivamente fare dei dati, i vincoli di copyright da applicare a lavori derivati da quei dati, il mantenimento della paternità sui dati, ecc. Il mondo delle licenze è vasto e intricato. Questa sezione analizza e confronta le licenze maggiormente utilizzate nel contesto Open Data.

Le licenze più diffuse sono quelle basate su Creative Commons [69] (CC). CC è un'organizzazione non a fini di lucro che nasce con l'intenzione di armonizzare l'articolato mondo del diritto d'autore (in Italia regolato dalla legge n. 633 del 22 aprile 1941) e del copyright. Nel 2002, CC ha pubblicato un primo insieme di licenze che si sono affermate *come standard de facto* a livello internazionale.

In linea generale, la scelta di adottare un modello di licenze basate su CC deriva principalmente dall'esigenza di armonizzare il rilascio di dati aperti con analoghe iniziative di carattere internazionale, semplificando e promuovendo il riuso dei dati stessi. Le licenze CC, infatti, facilitano la comprensione dei dati e consentono un loro ampio riuso grazie a un buon grado di permessi.

In funzione delle specificità dei diversi insiemi di dati la scelta può ricadere su diversi tipi di licenze, anche non necessariamente CC. Nella maggior parte dei casi, è opportuno seguire i criteri derivanti dalla definizione di Open Data della OKF *"[…] dati che possono essere liberamente utilizzati, riutilizzati e redistribuiti, con la sola limitazione – al massimo – della richiesta di attribuzione dell'autore e della redistribuzione allo stesso modo (ossia senza che vengano effettuate modifiche)*". Di seguito si analizzano alcune licenze utilizzate dalle amministrazioni rappresentate all'interno del gruppo di lavoro incaricato della produzione delle presenti Linee guida; per ogni licenza, se ne evidenziano le caratteristiche e si forniscono alcune indicazioni pratiche per il loro corretto uso. In Italia, licenze CC sono state utilizzate e promosse da alcune regioni quali, ad esempio, Emilia-Romagna e Piemonte, e da comuni come quello di Genova; a livello di PAC

lo scenario è, invece, più variegato con amministrazioni che adottano licenze CC (e.g., MIUR) e altre (e.g., INPS) che si orientano per l'adozione di licenze italiane.

LICENZA CREATIVE COMMONS ZERO (CC0)

La Creative Commons Zero [70] esprime "la più ampia e libera utilizzazione gratuita, anche per fini commerciali e con finalità di lucro". Apponendo su un documento la dichiarazione CC0 si rinuncia a tutti i diritti sul documento e sui suoi contenuti, dati inclusi, nella misura massima possibile prevista dalla legge. La Creative Commons Zero deve essere preceduta da una dichiarazione relativa alla provenienza del documento. A titolo di esempio si riporta un passaggio che riguarda l'applicazione della suddetta licenza per il riutilizzo delle Banca dati della rilevazione scolastica della regione Piemonte.

> *Il riutilizzo della "Banca dati della rilevazione scolastica" è stato concesso da Regione Piemonte ai sensi della Legge regionale n. 24/2011 e s.m.i..*
> *Regione Piemonte autorizza, pertanto, la libera e gratuita consultazione, estrazione, riproduzione, modifica e riutilizzo del documento e dei dati in esso contenuti da parte di chiunque vi abbia interesse per qualunque fine secondo i termini della Dichiarazione Creative Commons - CC0 1.0 Universal.*

LICENZA CREATIVE COMMONS ATTRIBUZIONE (CC-BY)

Un'alternativa alla licenza CC0 è la licenza Creative Commons "Attribuzione o equivalente" [71]. Questa permette al soggetto utilizzatore di riprodurre, distribuire, comunicare, esporre, rappresentare, nonché di modificare e usare un insieme di dati anche a fini commerciali con il solo obbligo di attribuire la paternità dell'opera. Anche questa licenza risulta essere, al pari della standard, espressione del principio della "più ampia e libera utilizzazione gratuita anche per fini commerciali e con finalità di lucro". In caso di uso della CC-BY però, l'unico obbligo imposto al licenziatario è quello di citare l'autore della base dati o del documento, oggetto di riutilizzo, nel rispetto delle modalità indicate dall'autore stesso nella, o a corredo, della licenza, come di seguito meglio specificato ("Attribuzione"). In generale, questa licenza è adottabile per le banche dati che risultano chiaramente tutelate dal diritto d'autore e/o dal diritto sui generis[51].

Al fine di prevenire qualsiasi incertezza interpretativa da parte del licenziatario e incoraggiare il riutilizzo dei dati, è opportuno chiarire che la licenza stessa si applica sia agli eventuali diritti d'autore relativi alla banca dati licenziata, sia ai diritti cosiddetti sui generis a tutela dei contenuti della banca dati stessa. Va cioè chiarito che la licenza disciplina tutti i diritti di cui alla L. 633/41 e s.m.i., con esplicita inclusione dei Diritti del costitutore di una banca di dati, di cui al Titolo II-bis della legge stessa.

Un esempio di applicazione di questa licenza a una generica banca dati è riportata qui di seguito:

[51] La tutela giuridica del diritto sui generis è stata introdotta dalla Direttiva 96/9/CE. La tutela sui generis si riferisce alla protezione garantita all'insieme delle informazioni contenute all'interno di una raccolta di dati per distinguerla dalla tutela, riconosciuta dal diritto d'autore, che può interessare invece la struttura o architettura della banca dati.

La titolarità piena ed esclusiva del documento "[DENOMINAZIONE E DESCRIZIONE SINTETICA DEL DOCUMENTO]" è di Regione Piemonte, ai sensi della L. 633/41 e s.m.i. (Licenziante). Regione Piemonte autorizza la libera e gratuita consultazione, estrazione, riproduzione e modifica dei dati in essa contenuti da parte di chiunque (Licenziatario) vi abbia interesse per qualunque fine, purché nel rispetto dei termini della licenza Creative Commons – Attribuzione 2.5 Italia.

Si precisa esplicitamente che con la presente licenza il Licenziante intende autorizzare il Licenziatario ad esercitare, ferme restando le restrizioni della licenza di cui sopra, anche i diritti disciplinati dall'art. 102-bis e ss., L. 633/41 e s.m.i. (c.d. diritto sui generis del costitutore di una banca di dati).

L'attribuzione prevista dalla licenza dovrà avvenire nella seguente forma: [INSERIRE NOTA PER L'ATTRIBUZIONE]

Per quanto concerne l'attribuzione, caratteristica propria delle licenze CC-BY, il licenziatario dovrà provvedere alla menzione, rispetto al mezzo di comunicazione o supporto utilizzato, di:

1. l'autore originale e/o titolare dei diritti;
2. le terze parti designate, se esistenti;
3. la descrizione/titolo del documento;
4. nella misura in cui ciò sia ragionevolmente possibile, l'Uniform Resource Identifier (URI) che il Licenziante specifichi dover essere associato con il documento oggetto di riutilizzo;
5. in caso di documenti rielaborati o opere derivate di vario genere, l'attribuzione dovrà essere effettuata in modo tale da non ingenerare confusione rispetto all'origine del documento stesso, ad esempio: "carta topografica basata su …".

Alle licenze suddette potranno essere altresì allegati l'invito a segnalare errori o imprecisioni, l'invito a inviare alla Direzione competente per materia eventuali versioni aggiornate/rielaborate del documento reso disponibile al riuso. A differenza della licenza CC-BY sopra descritta, versione 2.5, successive versioni della CC-BY Italia producono sul diritto sui generis della banca dati i medesimi effetti della Dichiarazione CC0, ovvero una rinuncia totale e incondizionata ai diritti.

ALTRE LICENZE CREATIVE COMMONS

La licenza CC-BY può essere estesa mediante alcuni attributi, quali:

1. **Share Alike (SA):** obbliga i lavori derivati a essere licenziati con la stessa licenza del lavoro originale
2. **Non Commercial (NC):** consente la copia, la distribuzione e l'uso del lavoro (o dati) solo per scopi non commerciali
3. **No Derivative Works (ND):** consente la copia, distribuzione e l'uso del lavoro, impedendo la creazione di lavori derivati

La combinazione di questi attributi genera di fatto altre licenze. Naturalmente, considerato che alcuni attributi sono mutuamente escludentesi, non tutte le combinazioni hanno senso. Nella pratica si considerano altre cinque licenze oltre alla CC0 e alla CC-BY, che sono:

1. *CC-BY-SA:* Attribution Share Alike;
2. *CC-BY-ND:* Attribution No Derivatives;
3. *CC-BY-NC:* Attribution Non-Commercial;
4. *CC-BY-NC-SA:* Attribution Non-Commercial Share Alike;
5. *CC-BY-NC-ND:* Attribution Non-Commercial No Derivatives.

È evidente che le licenze che non consentono il riutilizzo dei dati non sono compatibili con lo spirito e gli scopi degli Open Data.

LICENZA IODL 2.0

La licenza IODL 2.0 (Italian Open Data License) [72] prevede che l'utente possa liberamente (i) consultare, estrarre, copiare e pubblicare i dati; e (ii) creare un lavoro derivato integrando diversi dataset.

LICENZA IODL 1.0

Questa prima versione della IODL è simile alla IODL 2.0 ma con l'obbligo dell'utente di pubblicare o condividere i lavori derivati con la stessa licenza.

LICENZA ISA OPEN METADATA 1.1

Questa licenza [99] è stata creata nell'ambito del programma ISA della Commissione Europea. Essa è licenza aperta che può essere utilizzata in particolare nei casi in cui governi decidano di condividere i propri metadati e i propri "asset" semantici. La licenza può essere utilizzata sia per lavori commerciali che non, consente la modifica del lavoro pur preservando la paternità sul lavoro originale e obbliga, dove possibile e praticabile, l'indicazione del link alla piattaforma Joinup [91]. La licenza non è interlingua ma è fornita in inglese, notoriamente considerata lingua ufficiale tecnica a livello internazionale. L'applicazione della licenza è soggetta alla legislazione del paese del primo licenziatario.

8.1.1. Analisi critica delle licenze

La Tabella 1 analizza e confronta le diverse licenze prima descritte sulla base di alcune caratteristiche identificate, in funzione dei livelli di interoperabilità e riusabilità dei dati che possono essere garantiti.
Alcune di queste caratteristiche come ad esempio "Uso per fini commerciali", "Uso gratuito del lavoro soggetto alla licenza", ecc, sono ben note e auto-esplicative. Altre, invece, nascono proprio dall'analisi prodotta dal gruppo di lavoro; per esempio, si è voluto analizzare le licenze sotto l'aspetto della "Portabilità inter-lingua", ossia la possibilità che essa sia supportata da una traduzione in diverse lingue. Tale peculiarità è particolarmente importante in un'ottica di riuso transfrontaliero dei dati.
Anche la "Riconoscibilità internazionale del logo", che consente di capire se la licenza gode di sufficiente fama internazionale riconducibile al logo ad essa associato, è una peculiarità individuata importante ancora una volta nell'ottica di riuso dei dati per un'utenza non necessariamente confinata all'Italia. L'analisi delle licenze qui proposta si sofferma anche sulla "Compatibilità inter-licenza per lavori derivati" intesa come la possibilità di applicare licenze differenti a lavori derivati dall'uso dei dati.

	CC0	CC-BY	CC-BY-SA	CC-BY-SA-NC	IODL 1.0	IODL 2.0	ISA Open Metadata 1.1
Portabilità inter-lingua	X	X	X	X			
Riconoscibilità internazionale del logo	X	X	X	X			
Uso per fini commerciali	X	X	X			X	X
Uso gratuito del lavoro soggetto alla licenza	X	X	X	X	X	X	X
Possibilità di modificare i dati	X	X	X	X	X	X	X
Possibilità di preservare la paternità		X	X	X	X	X	X
Compatibilità inter-licenza per lavori derivati	X	X				X	

Tabella 1: Confronto tra le licenze CC, le licenze italiane IODL 1.0 e IODL 2.0 e licenze europee

Dalla Tabella 1 si può notare come la maggior parte delle licenze CC soddisfino la quasi totalità delle caratteristiche considerate, risultando quindi quelle che meglio si prestano a rispondere ai requisiti d'interoperabilità e massimo riuso dei dati.

Ai fini della scelta del tipo di licenza da associare ai dati da pubblicare, infine, è importante sottolineare anche la relazione bidirezionale che intercorre tra le licenze e i modelli di business che possono essere abilitati dagli Open Data. Se da un lato è possibile sostenere che le licenze con caratteristiche di maggior apertura consentono lo sviluppo di maggiori opportunità di business (Sezione 8.2.3) per il mondo delle imprese, è anche vero, dall'altro, che l'aumento delle opportunità di business facilitano la diffusione e l'affermazione delle stesse licenze, con innegabili benefici per l'espansione del movimento Open Data nel mondo.

RACCOMANDAZIONI

R25: Associare una tipologia di licenza che lascia il massimo spazio d'azione ai riutilizzatori dei dati, se non vi sono problematiche specifiche che richiedano l'adozione di licenze meno aperte.

R26: Utilizzare licenze che siano leggibili e riconoscibili non solo a utenti nazionali ma anche a utenti internazionali in un'ottica di interoperabilità transfrontaliera.

8.2. Impatto socio-economico

La presente sezione introduce alcuni aspetti utili nell'analisi, nella comprensione e nel governo dell'impatto sociale ed economico derivante dall'introduzione e dalla diffusione degli Open Data.

Va subito detto che, nonostante le rosee aspettative ed il fervente impulso proveniente dalla comunità scientifica, il processo di diffusione e riuso dell'informazione pubblica stenta ancora a decollare, specialmente nel panorama italiano. Tale situazione è ascrivibile, oltre che a una resistenza organizzativa e culturale al cambiamento ancora presente in numerose PA, a una scarsa chiarezza in capo agli operatori economici in merito ai meccanismi di creazione del valore associabile agli Open Data.

Obiettivo di questa sezione è quindi quello di introdurre un modello generale del sistema delle dinamiche di mercato abilitate dagli Open Data e caratterizzarlo rispetto a tre categorie principali di aspetti da considerare nel processo decisionale: aspetti legati alla domanda di dati; aspetti legati ai modelli di business sottesi; aspetti legati a indicatori territoriali che concorrano al governo delle iniziative di apertura dei dati e a misurarne le prestazioni.

8.2.1. Open Data: attori e ruoli

I perimetri delle organizzazioni pubbliche diventano sempre più sfumati, aprendosi in maniera selettiva a contributi e conoscenze provenienti dalla società civile e dal mondo delle imprese. A causa della crescente complessità dei sistemi socio-economici e della rapidità con cui l'innovazione tecnologica procede in molti settori, le organizzazioni pubbliche, infatti, faticano sempre più a svolgere l'intero ventaglio dei compiti loro assegnati. Le crescenti restrizioni sui bilanci preventivi della PA e gli imperativi imposti dagli attualissimi criteri di revisione della spesa, dal canto loro, contribuiscono ad acuire il fenomeno. Una progressiva apertura della galassia pubblica diventa pertanto necessaria non solo per incrementare la trasparenza della "macchina governativa" ma, soprattutto, per spianare la strada a nuovi modelli di gestione capaci di combinare in maniera efficace ed efficiente *asset* pubblici e risorse messe a disposizione dalla società. Il rilascio del patrimonio informativo della Pubblica Amministrazione (PSI) in modalità "aperta" da parte delle organizzazioni governative rappresenta un elemento abilitante per l'implementazione di siffatti modelli in cui una molteplicità di attori può

realizzare nuovi prodotti e servizi a partire da un profluvio di dati liberati dai "forzieri" della Pubblica Amministrazione: il dato pubblico, libero di circolare, diventa una linfa vitale che alimenta una miriade di attività destinate a creare valore.

Il paradigma degli Open Data rende quindi permeabile il confine che separa il comparto pubblico da quello privato e genera un ecosistema in cui i vari attori, per loro convenienze specifiche, condividono in modalità aperta parte dei loro dati e riutilizzano dati di altri attori (Figura 4). Nel paradigma degli Open Data la gestione del dato tradizionale è superata a favore di un modello in cui l'informazione rappresenta un elemento di valore condiviso su cui sviluppare servizi a valore aggiunto per cittadini ed imprese.

Ogni attore ha aspettative diverse sui ritorni dell'investimento derivanti dall'ingresso nell'ecosistema: la PA, oltre ad offrire uno strumento di trasparenza dei processi, può promuovere l'innovazione negli ambiti nei quali opera creando le condizioni per generare ricchezza e nuovi posti di lavoro; le aziende possono cogliere nuove opportunità di business; le comunità, infine, sfruttano la possibilità di migliorare i propri prodotti e servizi con l'aiuto di tutti.

Dalla molteplicità di soggetti presenti nell'ecosistema scaturisce uno scenario piuttosto eterogeneo: molte entità si trovano infatti ad occupare posizioni differenti all'interno dell'articolata rete del valore giacché assolvono compiti specifici destinati, in forme diverse, ad arricchire il dato grezzo rilasciato a monte dagli organismi pubblici. Il ricorso al concetto di "catena del valore" [73] consente di fare chiarezza rispetto all'ecosistema risultante, schematizzando una serie di stadi disposti in cascata secondo un approccio di interazione sequenziale (Figura 4).

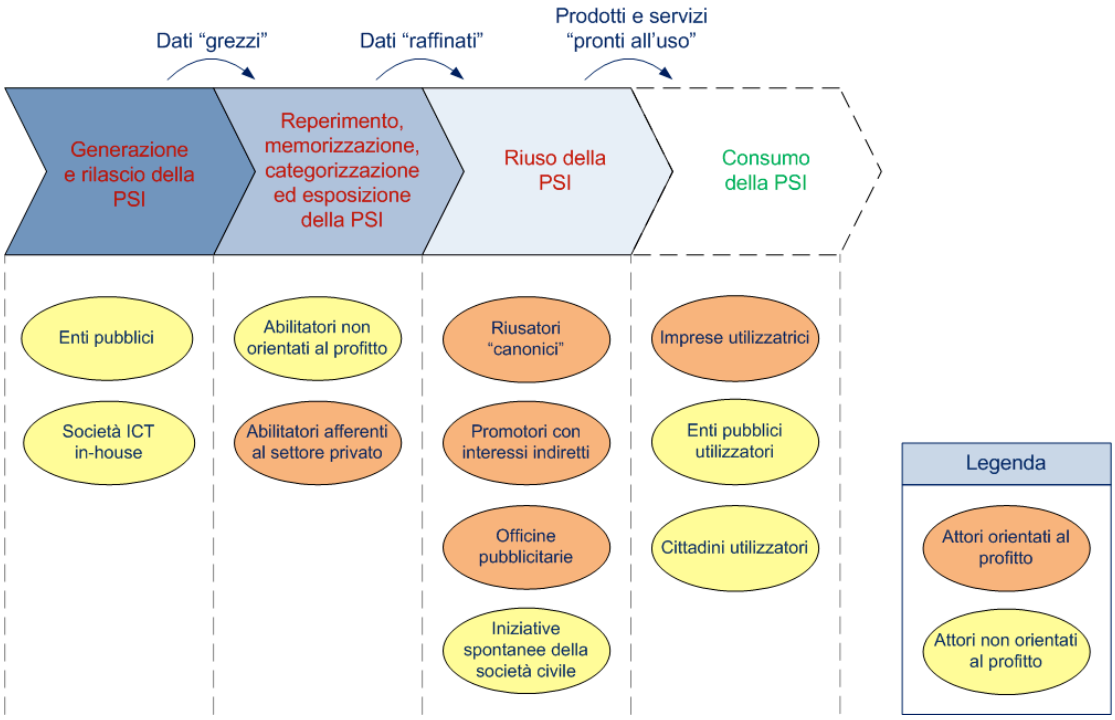

Figura 4: Catena del valore legata alla PSI

Le fasi sequenziali individuate sono di seguito presentate sinteticamente.

A monte della filiera si collocano le attività di generazione e la pubblicazione del dato degli attori pubblici detentori di PSI (i cosiddetti PSI "holder") che possono essere o direttamente le PA oppure apposite società ICT interne alle PA.

Proseguendo verso valle, il dato rilasciato viene progressivamente arricchito. Il primo stadio comprende, a titolo esemplificativo, la raccolta dei dati provenienti da fonti diversificate, la loro memorizzazione su infrastrutture virtualizzate, la classificazione dei dataset attraverso metadati e l'esposizione mediante interfacce applicative. Queste operazioni, sebbene non intervengano sensibilmente sulla natura del dato e non aggiungano valore in termini di logica applicativa, rendono più agevole il riuso agendo sulla cercabilità, sull'accessibilità e sulla leggibilità automatica dei dati. Le posizioni ricomprese in questo stadio possono essere ricoperte sia da imprese sia da attori non orientati al profitto, guidati da intenti profondamente dissimili. Le stesse PA o le società ICT interne, nell'ottica di investire nell'interoperabilità semantica, possono giocare un ruolo centrale in questa fase.

La fase successiva, come mostrato in Figura 4, è rappresentata dal riuso della PSI, che assume forme estremamente variegate. Anche in questa fase è possibile rinvenire la presenza di attori orientati al profitto, e attori invece provenienti dalla società civile che, spinti dalla ricerca di trasparenza governativa, mettono a frutto le proprie capacità tecniche e la propria creatività per originare servizi di pubblica utilità fondati sul dato pubblico.

Infine, i risultati del riuso, che si declinano in prodotti o servizi, giungono nelle mani degli utilizzatori finali che, alla luce dell'ampio spettro di funzionalità (ed eventualmente di prezzi), possono essere imprese, consumatori ed enti governativi.

Non è stato invece evidenziato in Figura 4 un possibile feedback sulla PA della catena del valore degli Open Data, legato da un lato alla possibilità di avere dati aggiornati da cittadini e imprese e dall'altro alle opportunità per l'ente pubblico stesso di attingere ai nuovi bacini informativi pubblicati da altri soggetti (anche in questo caso vi potrebbero essere potenzialmente interessanti riduzioni di costi e senz'altro un miglioramento della base informativa).

Di fatto l'ecosistema Open Data diventa una forte leva d'innovazione, dove i diversi attori competono per erogare servizi efficaci ed efficienti a cittadini ed imprese. I dati rappresentano la materia necessaria per creare l'ecosistema ma, per ottenere l'attivazione di questi processi, occorrono dei campioni (intesi come attori del quadro economico, istituzionale o sociale che riescono a diventare esempi da seguire ed emulare) che interpretino al meglio quanto oggi è presente in termini di tecnologia, norme e comunicazione. In questo momento la PA è il principale attore che può interpretare un ruolo guida nell'ambito dell'Open Data in grado di far partire lo sviluppo dell'ecosistema.

8.2.2. La domanda che caratterizza il mercato degli Open Data

Per lo sviluppo dell'ecosistema prima descritto è necessario che il flusso informativo che lo alimenta, proveniente dalla PA, comunità e aziende, sia continuo e che le informazioni e i dati pubblicati siano aggiornati in modo costante. Per assicurare l'impegno dagli attori dell'ecosistema su questo fronte è necessario che si affermi una domanda qualificata e forte che incentivi gli investimenti necessari per

mantenere attivi nel medio/lungo termine i servizi che via via sono attivati. Ecco perché quindi è importante individuare gli elementi che caratterizzano la struttura della domanda di informazioni e dati aperti a livello territoriale.

La domanda di OD da parte dei singoli cittadini, pur essendo potenzialmente interessante si crede sia difficile da stimolare. Questo dipende sia dal fatto che i formati con cui sono resi disponibili gli OD (non facilmente gestibili dai singoli privati), sia dalla natura della transazione col cittadino, che dovrebbe essere per definizione a titolo non oneroso (e.g., OD per ricerche scolastiche), sia perché molti dei dati necessari al privato cittadino sono già disponibili gratuitamente e sono fruibili con facilità su internet (per esempio, indirizzi, localizzazioni di punti di interesse, ecc.).

Una componente della domanda di dati aperti forte e pressante nei confronti della PA arriva dalle comunità. Nelle comunità il limite tecnologico viene superato agevolmente dal fatto che molto spesso di queste organizzazioni orizzontali di cittadini fanno parte anche tecnici molto preparati. È del tutto evidente che in questo ambito il ritorno dell'investimento non può essere misurato in termini di solo impatto economico, si tratta infatti di considerare nel calcolo complessivo anche l'insieme di servizi, anche di valore pubblico, che le comunità mettono a disposizione del contesto locale.

In ogni caso, ciò che può effettivamente fare la differenza sono i servizi a valore aggiunto sviluppati dalle imprese sulla base delle informazioni e dei dati messi a disposizione in modalità aperta. Per questo diventa centrale analizzare la domanda delle imprese che, in prima battuta, può essere caratterizzata in funzione della dimensione geografica dei dati richiesti e delle opportunità che le imprese possono cogliere da essi. I dati da un punto di vista geografico possono avere carattere locale, regionale, sovraregionale, nazionale o sovranazionale. Le opportunità che l'impresa può cogliere vertono su: lo sviluppo del core business (ad esempio, sviluppo di servizi a valore aggiunto basati sui dati aperti, identificazione di mercati potenziali, posizionamento di punti di vendita o di assistenza, quantificazione e distribuzione di potenziali clienti profilati); la promozione dell'innovazione di prodotto o di processo (ad esempio, programmi di ricerca e sviluppo, open innovation); lo sviluppo e la gestione delle risorse umane (ad esempio, reclutamento del personale, valutazione di possibili benefit da assegnare ai dipendenti); la soluzione di problematiche legate alla logistica (ad esempio, localizzazione magazzini o centri distributivi, ottimizzazione dei mezzi di trasporto e delle tratte, sicurezza dei trasporti) o sulla gestione della supply chain (ad esempio, identificazione di possibili fornitori alternativi di beni commodity).

	Dati Locali	Dati Regionali	Dati Sovra-regionali	Dati Nazionali	Dati Sovra-Nazionali
Business Model					
Promuovere Innovazione					
Gestione Risorse Umane					
Gestione					

Logistica					
Supply Chain					

Figura 5: Matrice per valutare la domanda potenziale di Open Data

La matrice proposta in Figura 5 costituisce uno strumento preliminare per valutare la "forza" della domanda delle imprese del contesto locale. La dimensione orizzontale individua i dataset potenzialmente disponibili, mentre quella verticale evidenzia il livello di interesse di un'impresa in un determinato dataset. Tanto minore è l'interesse in uno specifico dataset "aperto" tanto minore è l'impatto economico potenziale che lo stesso ha per l'azienda.

Valutare la domanda territoriale di dati aperti è un modo oggettivo per orientare investimenti ed impegni della PA in progetti e azioni di apertura dei dataset interni. Per questo è importante valutare, nel caso in cui dall'analisi preliminare la domanda risulti debole, l'opportunità di promuovere assieme alle iniziative Open Data anche dei programmi di riposizionamento e riqualificazione della domanda.

8.2.3. Modelli di business abilitati dagli Open Data

La presente sezione si pone l'obiettivo di esaminare come le imprese possano sfruttare le potenzialità di business scaturenti dall'inclusione della PSI nei prodotti e nei processi aziendali. I risultati riportati si basano sullo studio *Modelli di Business nel Riuso dell'Informazione Pubblica* [24] pubblicato e messo a disposizione dall'Osservatorio sulle ICT di Regione Piemonte.

Lo scacchiere strategico. L'analisi delle tipologie di imprese afferenti all'ecosistema porta all'individuazione di alcuni archetipi significativi organizzati entro due dimensioni fondamentali.

In primis, si può definire un asse relativo al posizionamento dell'impresa nel processo di creazione del valore. In questo ambito si vogliono differenziare gli attori collocati a valle, che fronteggiano il mercato finale offrendo agli utilizzatori soluzioni frutto del riuso della PSI, da coloro che assumono le fattezze di intermediari, presidiando una posizione più a monte che non prevede punti di contatto con gli utilizzatori finali.

Il secondo asse attiene invece alla visione strategica che l'impresa possiede rispetto al ruolo della PSI agli occhi del cliente finale. In questa dimensione si vogliono differenziare gli attori che considerano la PSI come una "fonte di sostentamento", capace di garantire alle imprese che la riusano un'adeguata redditività, da coloro che associano la PSI ad uno "strumento di attrazione" su cui le imprese riusatrici finali possono far leva per raggiungere in realtà finalità diverse dal profitto per se (e.g., potenziare la visibilità dell'impresa agli occhi dei clienti finali, incrementale la reputazione del sodalizio, intessere nuove *partnership* promettenti, spianare la strada per altre linee di business complementari dissociate dal dato pubblico).

L'incrocio delle due dimensioni suddette origina una matrice, riportata in Figura 6, in cui si può osservare il posizionamento degli attori archetipali in funzione del loro collocamento rispetto alle

dimensioni delineate.

Figura 6: Classificazione degli attori archetipali

Partendo dal quadrante in alto a sinistra di Figura 6, vi sono i riusatori "canonici", ovvero imprese che offrono agli utilizzatori finali prodotti o servizi basati su PSI con l'intento cristallino di generare flussi di ricavi dalla vendita dei beni stessi o da altre linee di business intimamente legate alla PSI. Nel quadrante in basso a sinistra di Figura 6, invece, vi sono gli attori abilitatori che, operando a monte dei riusatori "canonici", offrono a questi ultimi dati "arricchiti" (e.g., set informativi in formati maggiormente fruibili, set informativi frutto di integrazioni di fonti diverse) o servizi ad essi legati (e.g., offerta di capacità computazionale, interrogazioni dinamiche attraverso chiamate ad API). Spostandosi in alto a destra in Figura 6, vi sono i promotori con interessi indiretti, ossia imprese che offrono agli utilizzatori finali prodotti o servizi basati su PSI senza realizzare ricavo a partire da essi; l'attività promozionale posta in essere da questi attori economici si candida a generare ricadute positive in senso lato, andando potenzialmente a favorire i risultati economici registrabili su linee di business estranee alla PSI. I promotori con interessi indiretti possono internalizzare lo sviluppo di iniziative legate al comparto della PSI o, alternativamente, possono affidarsi a professionisti del settore, capaci di "confezionare" per conto terzi soluzioni chiavi in mano incentrate sulla PSI: emergono così le officine pubblicitarie, aziende che operano "dietro le quinte" al servizio di operatori desiderosi di promuovere la propria immagine attraverso la realizzazione di servizi di utilità collettiva legati alla PSI.

Panoramica sui modelli di business emergenti. Nell'ottica della comprensione delle modalità attraverso cui la creazione e l'appropriazione di valore possono aver luogo, è necessario effettuare una riflessione sulle differenti logiche di business che gli attori possono adottare. Una panoramica dei modelli di business archetipali (Figura 7) delineabili è di seguito enucleata.

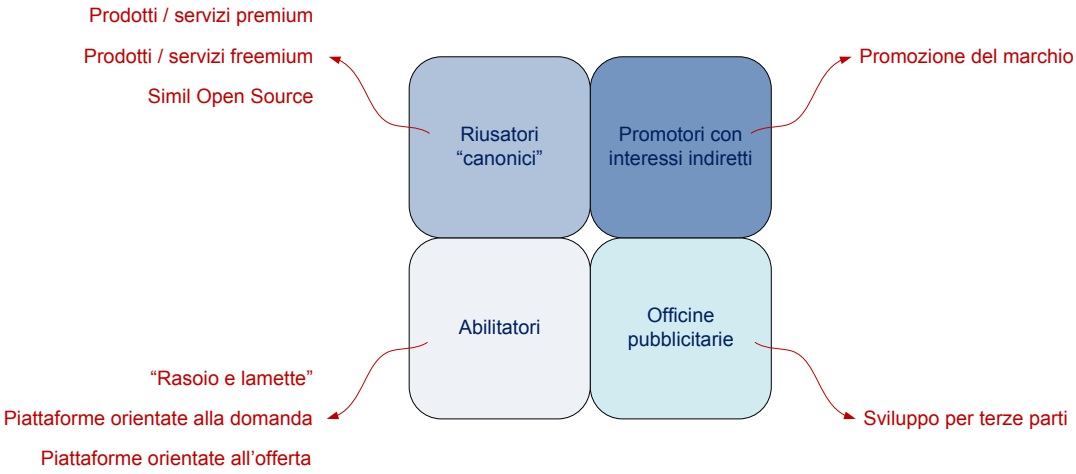

Figura 7: Panoramica dei modelli di business archetipali

Prodotti/servizi "premium". In questo modello di business il riusatore "canonico" offre al mercato finale un prodotto o un servizio, presumibilmente caratterizzato da elevato valore intrinseco, in cambio di un pagamento effettuato in contropartita al bene scambiato. Le modalità di pagamento generalmente sono *à la carte* o previa sottoscrizione[52]: mentre la prima, assimilabile ad un *pay-per-use*, prevede la corresponsione di un importo per ciascuna unità di prodotto acquistata, la seconda contempla una tariffa "tutto compreso" che, per un arco temporale prestabilito, consente l'utilizzo di determinate funzionalità in conformità alle pattuizioni contrattuali. Alla luce del meccanismo di *pricing* e dell'elevato valore intrinseco, la clientela target è perlopiù di tipo B2B e viene gestita mediante logiche di medio-lungo termine che vanno al di là del mero approccio transattivo.

Prodotti/servizi "freemium". Nel suddetto modello di business il riusatore "canonico" propone al mercato un prodotto o un servizio attraverso un'offerta apparentemente conveniente agli occhi del cliente finale. Il meccanismo di fissazione del prezzo, infatti, si articola secondo l'approccio *freemium*, il quale prevede la fruizione gratuita delle funzionalità di base mentre, per funzionalità più avanzate, è richiesta la corresponsione di una tariffa. Nel mondo della PSI, questo paradigma si declina attraverso limitazioni applicate ai prodotti e servizi offerti: pagamenti *ad-hoc* possono essere infatti necessari per accedere a funzionalità supplementari o a set di dati aggiuntivi. In questo caso, la clientela appartiene tendenzialmente al segmento B2C; la gestione del cliente segue logiche di breve-medio termine e non prevede, se non in rari casi, una vera personalizzazione del servizio. L'erogazione della proposizione di valore si avvale generalmente dei canali Web e *mobile* (attraverso *app*), dalla cui sinergia è possibile raggiungere un numero cospicuo di basi installate.

Simil Open Source. Il modello di business "*open*" prevede l'offerta di prodotti, servizi o set di dati non "confezionati", senza il pagamento di alcuna tariffa. Le voci di ricavo per il riusatore "canonico" che adotta questa logica di business possono essere legate a pagamenti per servizi a valore aggiunto[53] o per

[52] In realtà, nelle fattispecie esaminate esiste una pluralità di forme intermedie di pagamento (cosiddette tariffe a due parti) che prevedono sia una componente fissa che una variabile.

[53] Esempi collocabili in questa fattispecie annoverano elaborazioni analitiche effettuate su richiesta, l'implementazione di specifiche *query* o l'incrocio dei dati originari con altre fonti informative, piuttosto che un ampio spettro di attività di consulenza legate all'armonizzazione dei dati con determinate *suite* applicative ed all'impiego fruttuoso dei dati all'interno dei processi di business precipui del cliente.

modifiche alla licenza originaria[54]. Si attua così una cosiddetta sovvenzione incrociata, in cui le entrate da ascrivere a queste linee di business supplementari, comunque legate alla PSI, vengono impiegate per coprire i costi ingenerati dalle linee di business offerte a titolo gratuito: in altre parole, agli occhi del cliente finale il dato è offerto gratuitamente ed in formati aperti e rielaborabili senza barriere di natura tecnica (da qui il riferimento al mondo dell'Open Source). La flessibilità che contraddistingue il modello di business rende questa logica adatta ad una variegata platea di clienti gestiti attraverso relazioni di volta in volta definite a seconda dell'interlocutore.

"Rasoio e lamette". Gli attori abilitatori fautori di questa logica agiscono come intermediari che facilitano l'accesso a risorse PSI da parte di sviluppatori orientati al profitto o di scienziati non guidati da intenti commerciali. Il nome conferito a questo modello deriva dalla strategia comunemente definita come "razor &blades"[55]: inizialmente un prodotto viene venduto ad un prezzo molto basso[56] al fine di alimentare il successivo acquisto ripetuto di un bene complementare (generalmente un consumabile), avente una domanda piuttosto inelastica, sul quale è invece possibile realizzare marginalità significative. Nel caso della PSI, il predetto modello viene implementato posizionando set di dati pubblici su piattaforme di cloud computing e rendendoli accessibili via API da chiunque, tariffando poi il solo utilizzo della capacità computazionale richiesta per il processamento dei dati stessi. Anche in questo caso è ravvisabile una cosiddetta sovvenzione incrociata, poiché i profitti realizzati dalla messa a disposizione di capacità elaborativa "on-demand" vanno a coprire i costi ascrivibili allo storage e all'organizzazione del dato. In ambito PSI, è superfluo specificare che le applicazioni di questo modello si limitano a contesti in cui gli oneri computazionali sono consistenti.

Piattaforme orientate alla domanda. Un modello siffatto ambisce a fornire agli sviluppatori consistenti facilitazioni nell'accesso a risorse PSI, le quali vengono memorizzate su *server* intermedi proprietari[57] ad elevata affidabilità, catalogate mediante metadati, armonizzate in termini di formati ed esposte attraverso API, rendendo agevole il reperimento dinamico dei dati in maniera "on-demand". Una cospicua gamma di criticità intrinseche al dato grezzo vengono quindi rese ininfluenti grazie al ricorso a piattaforme di accesso ai dati capaci di tradurre dataset in "datastream", contribuendo sensibilmente alla "commoditization" ed alla "democratizzazione" del dato. Le suddette strutture adibite all'intermediazione offrono così agli sviluppatori un ampio ventaglio di dati secondo un approccio "one stop shopping" teso a minimizzare il costo di ricerca da parte del compratore: quest'ultimo, rivolgendosi a un solo fornitore, accede attraverso API standardizzate ad una molteplicità di risorse informative, che esulano anche dai confini nella PSI[58], senza doversi preoccupare dell'interfaccia di accesso a ciascuna delle fonti originali. In termini di costo, poiché un bene che nasce libero e aperto non può essere commercializzato a fini di lucro in assenza di elaborazioni su di esso, gli abilitatori che impiegano questo modello di business sulla PSI ricevono flussi finanziari in entrata in cambio di servizi

[54] Un primo esempio non raro di questa fattispecie si ha quando la licenza di partenza non contempla riusi di tipo commerciale (*e.g.*, CC BY-NC-SA); nel momento in cui un attore è intenzionato ad effettuare un riuso con finalità di vendita, egli deve negoziare con il detentore dei dati un pagamento al fine di ottenere il consenso previo rinuncia formalizzata da parte del detentore dei diritti. Un secondo esempio interessante attiene all'uso della *Open Database License* (ODbL), la quale offre ai riusatori molteplici gradi di libertà rispetto ai dati contenuti in un determinato archivio; qualora il riusatore intenda rilasciare un lavoro derivato senza seguire i vincoli imposti dalla licenza (*e.g.*, *share-alike*, *keep open*), egli si troverà a negoziare con il detentore dei dati forme alternative di *licensing* che siano funzionali allo specifico riuso.

[55] Gli esempi non si limitano a rasoi e lamette, poiché questo modello può essere rinvenuto anche in altri casi quali stampanti e cartucce, macchine automatiche del caffè e cialde, console *videogame* e giochi, talvolta anche telefoni cellulari e contratti per il traffico voce/dati.

[56] Idealmente, si pensi che il prezzo sia pari o inferiore al costo marginale; talvolta il prezzo può anche essere nullo.

[57] L'infrastruttura di calcolo può essere anche di proprietà di terze parti che, attraverso il ricorso alla virtualizzazione, disaccoppiano il livello fisico da quello logico, permettendo agli attori abilitatori di eseguire i propri ambienti applicativi su hardware non proprio.

[58] L'approccio al dato da parte di questi operatori è "olistico", ovvero orientato ad offrire accesso ad una "miniera" di dati prescindendo dall'organizzazione che li ha generati. Il mix informativo messo a disposizione dai suddetti *player* ingloba, oltre alla PSI, dataset (o, più propriamente, flussi di dati) provenienti dai Social Media, da piattaforme Open Data collaborative e da società private.

avanzati o di set di dati raffinati, configurando un modello di costo orientato al paradigma "freemium" in cui l'importo viene modulato in base a limitazioni funzionali.

Piattaforme orientate all'offerta. Questo modello prevede nuovamente la presenza di un attore intermediario fornitore di servizi infrastrutturali; l'impresa qualificata come abilitatore, però, in questo caso non applica alcuna tariffa agli sviluppatori, ricaricando invece le Pubbliche Amministrazioni detentrici di PSI. Osservando lo scenario nel suo complesso, si nota la presenza di un mercato a due versanti [74], costituito dai PSI "holder" da una lato e dagli sviluppatori dall'altro lato. Seguendo le regole auree proprie dei mercati multi-versante, l'abilitatore modula il prezzo in funzione del grado di esternalità indiretta positiva che ciascun versante è in grado di esercitare: vengono così azzerate le barriere per gli sviluppatori, i quali accedono gratuitamente a dati ben strutturati, e vengono applicate delle tariffe agli enti pubblici che diventano conseguentemente titolari delle piattaforme di gestione dei dati. A livello tecnico, infatti, vengono predisposte soluzioni tese a facilitare l'esposizione di dati PSI che siano capaci di adattarsi alle esigenze degli enti pubblici desiderosi di avviare programmi di apertura dei dati. Generalmente le suddette piattaforme garantiscono lo "storage", il rapido aggiornamento di nuovi dataset da parte del personale pubblico, la catalogazione mediante metadati, la standardizzazione dei formati e l'esposizione verso l'esterno dei dati sia attraverso interfacce applicative (API) sia mediante interfacce grafiche (GUI). Le Pubbliche Amministrazioni che aderiscono ai programmi avviano quindi una relazione di lungo periodo con i *provider* e sono tenute alla corresponsione periodica di tariffe in funzione della sofisticazione delle soluzioni acquistate[59] e di alcuni parametri tecnici[60].

Promozione del marchio. I promotori con interessi indiretti offrono agli utilizzatori finali prodotti o servizi basati sulla PSI senza realizzare diretto ricavo a partire da essi; l'attività promozionale condotta da questi attori economici si candida a generare ricadute positive in senso lato, andando potenzialmente a favorire i risultati economici registrabili su linee di business del tutto estranee alla PSI. Gli "economics" sottesi a questa pratica sono connotati da una forte spinta pubblicitaria, che porta l'impresa a considerare i costi non coperti da relativi ricavi alla stregua di investimenti promozionali compresi nel "marketing mix", o dalla presenza di costi marginali nulli, situazione che si verifica allorché i costi di distribuzione e di utilizzo siano non significativi, rendendo l'attività non coperta da ricavi non gravosa a livello di centro di costo. Le linee di business incentrate sulla PSI, dal canto loro, diventano quindi strumenti per porre in essere il cosiddetto "service advertising", modello promozionale emergente che sostituisce l'esposizione di comunicazioni visuali[61] atte a influenzare in maniera intenzionale e sistematica gli atteggiamenti e le scelte degli individui in relazione al consumo di beni e all'utilizzo di servizi, con la fornitura di servizi di pubblica utilità che, in maniera più indiretta ma anche più incisiva, tentano di evidenziare una condotta "illuminata" dell'azienda promotrice. Il pubblico target di riferimento è oltremodo ampio e può essere raggiunto sia attraverso il "tradizionale" canale Web sia mediante il canale mobile. Un'efficace implementazione del modello di promozione del marchio si incentra sul binomio costituito da competenze tecniche e competenze creative: quando esse non albergano entrambe sotto lo stesso tetto, può diventare indispensabile esternalizzare *in toto* o parzialmente la produzione dei servizi promozionali che assurgono a "strumenti di attrazione".

Sviluppo per terze parti. Le officine pubblicitarie entrano in gioco allorché i promotori con interessi

[59] Le funzionalità avanzate che possono determinare il grado di sofisticazione delle soluzioni sono, a titolo meramente esemplificativo, la registrazione di un dominio proprietario, la possibilità per gli utenti di interagire con i PSI *holder*, la georeferenziazione automatica dei dati spaziali, la creazione di una galleria contenente le applicazioni finora sviluppate a partire dai dati e l'integrazione con strumenti analitici di monitoraggio del traffico Web.

[60] I parametri quantitativi generalmente presi in esame sono le dimensioni delle basi di dati, l'ampiezza di banda richiesta ed il numero di chiamate mediante API per unità di tempo.

[61] Il modello promozionale basato sull'esposizione di comunicazioni visuali in spazi a pagamento è spesso detto "*display advertising*".

indiretti decidono di esternalizzare lo sviluppo dei servizi di utilità collettiva forniti con finalità promozionali. Lo sviluppo per conto di terze parti si rivolge quindi ad aziende desiderose di promuoversi attraverso la PSI, con le quali le officine pubblicitarie instaurano relazioni di medio-lungo periodo orientate alla personalizzazione. Alla luce del ruolo vitale ricoperto dal marchio dei promotori con interessi indiretti, le officine pubblicitarie adottano la strategia "white label", in virtù della quale viene eclissato il *brand* dell'intermediario per conferire piena visibilità al marchio dell'attore economico che affronta il mercato finale. Il committente, in cambio della soluzione chiavi in mano così sviluppata, corrisponde pagamenti "lump sum" o periodici, a seconda che il bene confezionato si configuri come prodotto o servizio.

8.2.4. Indicatori territoriali legati allo sviluppo degli Open Data

Gli indicatori d'impatto territoriale forniscono delle indicazioni sugli effetti prodotti dall'attuazione di una politica, di un'azione, di un programma o di un progetto promossi da un'istituzione o da un altro attore del tessuto economico e sociale territoriale. Di fatto un indicatore è uno strumento essenziale per misurare e osservare l'incidenza delle politiche di sviluppo e delle loro realizzazioni e dell'evoluzione e trasformazioni indotte nei territori, nonché per il miglioramento continuo delle politiche e delle azioni avviate. Si tratta quindi di un supporto importante al fine di sostenere la programmazione, la valutazione di una determinata azione ed inoltre fornisce ai decisori elementi oggettivi per monitorare ed eventualmente re-indirizzare gli interventi finanziati.

Com'è già stato introdotto in precedenza nel contesto di questo documento le aspettative di ricaduta dei progetti Open Data avviati dalla PA riguardano essenzialmente l'impatto organizzativo sulla Pubblica Amministrazione (in termini di efficienza, trasparenza, partecipazione e collaborazione) e l'impatto economico sul territorio (in termini di politiche di spesa pubblica, di vantaggio competitivo del sistema territoriale e di aumento della conoscenza collettiva).

Recentemente la commissione europea ha commissionato uno studio, sintetizzato in [78], con l'obiettivo di individuare un insieme di indicatori di misura del riutilizzo dell'informazione pubblica in uno specifico periodo di tempo e dell'impatto economico dello stesso sui costi marginali della PA. Oltre a questo lavoro per meglio dettagliare aspetti di impatto non catturati dallo studio in questione è possibile fare riferimento a strumenti consolidati che si trovano in letteratura. Per quanto riguarda la misura d'impatto economico, ad esempio, è interessante l'insieme di "core indicators" sviluppati dalla commissione europea per misurare i fenomeni associati ai programmi di finanziamento del Fondo Sociale Europeo [79].

Indicatore	**Tipo di indicatore**	**Descrizione**
Dataset OGD	Input	Numero di dataset pubblicati dalle pubbliche amministrazioni locali secondo il paradigma Open Data.
Volume dati della PA	Input	Misura quantitativa dei dati pubblicati dalle

rilasciati in modalità aperta		pubbliche amministrazioni locali secondo il paradigma Open Data.
Numeri di progetti R&S legati all'OD	Input	Numero dei progetti R&S nell'ambito degli Open Data attivati nel contesto locale. Con progetti R&S ci si riferisce a quelle iniziative progettuali che si focalizzano sulla creazione di nuova conoscenza (Ricerca), o adattamento/applicazione di conoscenza esistente (Sviluppo).
Numero di aziende nell'ambito degli OD	Intermedio	Numero di aziende (intermediari informativi) operanti sul territorio nell'ambito Open Data.
Numero degli addetti delle aziende operanti nell'ambito degli OD	Intermedio	Numero degli addetti delle aziende operanti sul territorio nell'ambito Open Data.
Numero di comunità attive nell'ambito degli OD	Intermedio	Numero di comunità sul territorio che generano o utilizzano Open Data.
Numero di progetti finalizzati a favorire la diffusione e l'adozione degli OD	Intermedio	Numero di progetti ed iniziative private o pubbliche volte a promuovere lo sviluppo di linee specifiche di business all'interno delle imprese, di nuova imprenditorialità e l'utilizzo di servizi innovativi legati all'ambito Open Data.
Numero di posti di lavoro creati; di cui posti di lavoro creati per uomini e posti di lavoro creati per donne.	Output	Numero di posti di lavoro lordi creati (full time equivalents - FTE): Una nuova posizione di lavoro creata (prima inesistente) come risultato diretto di progetti completati (i lavoratori impiegati nella realizzazione dei progetti non sono contati). La posizione lavorativa deve essere coperta (i posti vacanti non sono considerati) e deve aumentare il numero totale di posti di lavoro dell'organizzazione.
Indicatori di performance economico-finanziaria delle aziende operanti nell'ambito OD	Output	Dati di performance economico-finanziaria delle aziende che operano nell'ambito dello sviluppo di prodotti o servizi legati ai dataset pubblicati secondo il paradigma Open Data. Con dati di performance economico-finanziaria ci si riferisce a: fatturato, valore aggiunto e margine operativo lordo(ebitda).

Indicatori di mercato - dataset	Output	Indice dei top 5 dataset pubblicati secondo il paradigma Open Data utilizzati da aziende pubbliche o private.
Indicatori di mercato - imprese	Output	Numero delle imprese private o pubbliche utilizzatori di dataset pubblicati secondo il paradigma Open Data.
Indicatori di mercato - cittadini	Output	Numero di cittadini utilizzatori, per il tramite dei servizi sviluppati dalle aziende pubbliche o private, di dataset pubblicati secondo il paradigma Open Data.

Tabella 2: Esempi di indicatori di impatto territoriale derivanti dai documenti citati o rielaborazione degli stessi

Un primo insieme di indicatori che possono essere utili per misurare l'impatto a livello territoriale delle politiche, delle azioni e dei progetti legati agli Open Data attivati dalla PA è riportato in Tabella 2. Gli indicatori riportati in tabella sono organizzati rispetto a tre categorie: *indicatori di input*, *indicatori intermedi* ed *indicatori di output*. Gli indicatori di input danno una misura dell'impegno diretto della Pubblica Amministrazione e del territorio in ambito OD, ad esempio il numero di dataset liberati. Gli indicatori intermedi danno l'insieme degli asset in ambito OD su cui il territorio può contare, ad esempio il numero delle comunità attive. Infine gli indicatori di output danno una misura degli impatti prodotti dai diversi investimenti, ad esempio il numero degli occupati o la dimensione del mercato.

A scanso di equivoci è importante chiarire che la definizione e lo sviluppo dell'insieme di indicatori utile per misurare uno specifico fenomeno è un'attività complessa che richiede un'analisi sia delle caratteristiche che si intende valutare, sia delle specificità del sistema locale. In questo senso l'insieme di indicatori presentato in tabella non deve essere considerato esaustivo rispetto alla rappresentazione di tutti gli effetti significativi derivanti dai programmi intrapresi dalle amministrazione nell'ambito dell'Open Data e ad esso andrebbe aggiunto un set di indicatori legato allo specifico territorio.

RACCOMANDAZIONI

R27: Esporre dataset di largo interesse per permettere interrogazioni distribuite su una molteplicità di basi di dati.

R28: Esporre dataset di ampio interesse per gli sviluppatori di applicazioni.

9. SERVIZI LINKED OPEN DATA SPC ABILITANTI L'INTEROPERABILITÀ SEMANTICA

Per garantire l'interoperabilità semantica seguendo l'approccio fin qui definito, il Sistema Pubblico di Connettività e Cooperazione (SPC) gioca un ruolo indispensabile in quanto framework nazionale in grado di fornire agli utenti della PA servizi integrati tramite la condivisione di regole e standard. Le tecnologie semantiche, infatti, permettono di raggiungere i suddetti obiettivi, definendo da una parte le condizioni e gli standard che aiutano a costruire un tale framework, e dall'altra fornendo una soluzione di gestione di grandi e complessi volumi di dati provenienti da fonti differenti. È attraverso l'implementazione di standard condivisi e di tecnologie "user-oriented" che i processi di pubblicazione e di sviluppo dei dati all'interno di tutte le pubbliche amministrazioni possono essere integrati. In questo modo, anche tutte quelle attività correlate e fortemente dipendenti dalle PA stesse possono essere messe nelle condizioni di beneficiare del risparmio in termini di tempo e spesa ottenuto dall'adozione di un framework comune di pratiche e processi condivisi.

Si ritiene, quindi, opportuno prevedere in SPC la definizione di un insieme di servizi abilitanti l'interoperabilità semantica. I servizi agiscono a due diversi livelli di astrazione: un livello "orizzontale" costituito da servizi comuni infrastrutturali, e un livello "verticale" di servizi e-government applicativi per la PA. La presente sezione descrjve entrambe le tipologie di servizi evidenziandone il ruolo nell'attuazione dell'approccio LOD proposto.

9.1. Servizi infrastrutturali: il ruolo delle infrastrutture condivise SPC

Le infrastrutture condivise SPC costituiscono l'architettura di supporto ai nuovi sistemi informativi "semantici" delle PA, destinati a valorizzare la grande mole di dati in loro possesso. Le infrastrutture devono essere progettate per erogare un insieme di servizi comuni "orizzontali", suggeriti anche dall'EIF, attraverso cui raccogliere Linked Data e, in generale, ontologie e schemi comuni di dati ai quali riferirsi nella cooperazione tra PA.

Già nello specifico delle infrastrutture condivise SPC, diverse tipologie di dati pubblici sono disponibili all'apertura. Con particolare riferimento agli ambiti identificati in Sezione 4, possono essere aperti:

- i "dati relativi a spese della PA", grazie alle informazioni di controllo acquisite dagli organi di governance SPC e contenute nei contratti esecutivi pubblici stipulati dalle PA per l'acquisizione di servizi di connettività e applicativi;
- i dati relativi ai riferimenti elettronici (e non) di tutte le pubbliche amministrazioni italiane inclusi nell'IPA, una delle infrastrutture condivise SPC;
- i metadati relativi ai dati territoriali delle PA, anch'essi contenuti in una base di dati di interesse nazionale SPC che è il Repertorio Nazionale dei Dati Territoriali (RNDT).

Una prima applicazione dell'approccio LOD proposto nel presente documento è stata avviata nei mesi scorsi da DigitPA con l'iniziativa Linked Open IPA [61], che prevede l'apertura dei dati dell'IPA attraverso la loro trasformazione in RDF.

La scelta di partire proprio da tale infrastruttura condivisa SPC è motivata dal fatto che l'IPA, contenendo tutti quei dati che identificano in maniera *univoca* le PA italiane, può essere vista come il nucleo di una possibile "nuvola" LOD da cui partire per collegare, tramite tali tecnologie, sia altre tipologie di dati riferiti all'insieme di infrastrutture condivise SPC, sia dati LOD **autentici** gestiti e pubblicati da singole PA centrali e locali. Analogamente al ruolo assunto da DBpedia nel Web dei Dati (Sezione 5.1), l'infrastruttura condivisa IPA, e quindi una parte dello stesso SPC, potrebbe diventare l'"hub" nazionale di LOD della PA a cui altre pubbliche amministrazioni, che decidono di intraprendere un processo di apertura interoperabile dei loro dati come fin qui descritto, possono collegarsi.

I vantaggi di tale approccio sono molteplici. In primo luogo, si può dare un forte impulso allo sviluppo di LOD della PA in Italia, e quindi a una nuova concezione dell'organizzazione dei dati, grazie alla facilità di collegamenti che si possono creare tra i dati della PA tutta. In secondo luogo, l'approccio contribuisce alla formazione del WoD SPC con dati *autentici*, *certificati* dalle PA italiane e di *qualità*, continuamente aggiornati tramite la gestione delle infrastrutture condivise SPC. Infine, l'approccio può favorire la definizione di un'ontologia più ampia che descriva la nuvola LOD SPC e che nasca come integrazione di altre ontologie più specifiche di dominio già rese disponibili da singole amministrazioni. L'obiettivo, in quest'ultimo caso, è quello di facilitare la messa in relazione dei dati tra loro attraverso le informazioni che li descrivono. L'ontologia generale, e quelle specifiche eventualmente riutilizzabili, possono essere raccolte in infrastrutture condivise SPC dedicate per favorire l'interoperabilità semantica tra le diverse PA: questo, infatti, assicura l'uniformità dei tipi di informazioni che descrivono le risorse, permettendo nel medio lungo periodo lo sviluppo di migliori soluzioni di front-end dedicati, e maggiore integrazione nei processi di back-end.

Il processo di espansione della conoscenza, ottenibile grazie alla facilitazione del riuso in seno a SPC, può valorizzare differenti percorsi:

- *riuso nella PA* – permette un maggior controllo sulle attività e sui processi interni, come sulle attività di vigilanza esterne su enti, organizzazione e territorio. Consente un risparmio di risorse in termini di tempo e di spesa nell'individuazione di fenomeni sociali ed economici d'interesse rappresentati nei dati stessi, ma che molto spesso sono in forma d'informazione tacita. Consente inoltre una più facile collaborazione tra PA;
- *riuso privato* – consente di offrire un'informazione più dettagliata e trasparente ai cittadini;
- *riuso commerciale* – consente lo sviluppo di applicazioni in modo più veloce e integrato, favorendo l'interesse del mondo aziendale verso i modelli economici di sviluppo di dati aperti e allargando il bacino aziendale anche al mondo (sempre più in crescita) delle Startup.

Tuttavia, per mettere le PA (e non solo) nelle condizioni di fruire e riutilizzare il più possibile i dati, anche servizi di pubblicazione, interrogazione e ricerca degli stessi devono essere previsti nel contesto SPC. Già nel DPCM 1° Aprile 2008, recante regole tecniche e di sicurezza di SPC, come descritto in Sezione 3.1, è previsto un servizio di interoperabilità semantica che raccoglie l'insieme degli schemi di

dati e ontologie (i.e., il "Catalogo Schemi e Ontologie"). Nel nuovo modello SPC, [14][15] il servizio può essere rivisto in "chiave moderna" trasformandosi in quell'hub nazionale di LOD prima menzionato auspicabilmente applicabile non solo ai meri dati pubblici ma anche a tutti i dati che le pubbliche amministrazioni possono scambiarsi nell'ambito ristretto e riservato SPC per l'esercizio quotidiano delle loro funzioni.

In questo scenario, quello che il DPCM 1° aprile 2008 definisce "Catalogo Schemi e Ontologie" può essere profilato come il servizio LOD SPC, che consente di produrre LOD a partire da dati generati e scambiati in questo contesto, di collegare tali dati ad altri dati delle PA (centrali e locali), e di arricchire i dati con opportuni metadati semantici per stabilire uno standard di qualità a livello di pubblicazione, di utenza e di interoperabilità nella PA. Il servizio può essere erogato adottando un modello cloud ibrido in cui la gestione dei meri dati pubblici può avvenire attraverso l'impiego di una cloud pubblica, e la gestione invece di tutti quei dati relativi alle funzioni di back end delle PA attraverso una cloud di comunità (la cloud SPC).

L'idea generale è quella di raccogliere tali dati e renderli disponibili per successive interrogazioni (anche da macchina a macchina) e per lo sviluppo di applicazioni interoperabili che interrogano i dati. Tali interrogazioni possono essere fatte attraverso l'uso di SPARQL, avendo cura di includere nell'insieme di infrastrutture condivise SPC, per il servizio di catalogo schemi e ontologie, un vero e proprio Platform as a Service (PaaS) rappresentato da un store di triple RDF e da uno SPARQL end-point.

Per incentivare sempre più il riuso e lo sviluppo di applicazioni, le infrastrutture condivise SPC possono giocare un ruolo importante supportando la raccolta e la maggiore fruibilità di servizi e applicazioni progettate a partire dal Web dei Dati SPC. In particolare, il servizio di "catalogo schemi e ontologie" può essere affiancato da un catalogo di servizi e "apps" creati a partire dalla nuvola LOD SPC.

9.2. Servizi LOD per le PA

Parallelamente alla dimensione orizzontale, al fine di aiutare le PA nella produzione di LOD nel rispetto delle presenti linee guida, un'azione sussidiaria, non vincolante e non obbligatoria, è quella di profilare servizi applicativi SPC di LOD, acquisibili da singole PA in base alle proprie esigenze di apertura e di gestione dei dati.

Tali servizi, per essere efficaci, devono essere definiti in modo da soddisfare requisiti funzionali per l'implementazione di tutte le fasi dell'approccio LOD descritto in Sezione 6. Accanto a tali requisiti, devono essere opportunamente specificati aspetti di standardizzazione, diffusione, "openess", sia tecnologica che delle licenze per il riuso, per offrire alle PA soluzioni scevre da problematiche legate al "vendor lock-in" e che vadano nella direzione di erogare servizi capaci di garantire elevati livelli di qualità del dato e di interoperabilità semantica.

Per quel che riguarda la pubblicazione, interrogazione e ricerca dei dati, le presenti linee guida trovano applicazione attraverso la definizione di relativi servizi, le cui modalità di erogazione possono seguire i principi definiti nelle linee guida della Commissione di Coordinamento SPC riportate in [14], in un'ottica di economie di scala che devono essere ricercate. Tali servizi contribuiscono a una maggiore fruibilità e al massimo riutilizzo dei dataset prodotti, non solo in contesti nazionali ma anche internazionali, in un'ottica di trasparenza verso utenti finali e di sviluppo applicativo che può scaturire dalla maggiore accessibilità dei dati.

Le diverse fasi di produzione LOD e pubblicazione, interrogazione e ricerca dati dovrebbero essere tenute separate nella profilazione dei servizi così da consentire ampia flessibilità alla PA nella scelta dei servizi da acquisire mediante apporti esterni, e da gestire internamente in maniera autonoma.

RACCOMANDAZIONI

R29: Acquisire servizi LOD tenendo conto delle diverse specializzazioni e competenze richieste (innovative e tradizionali) nell'approccio metodologico proposto.

R30: Potrebbe essere utile ed efficace avvalersi di strumenti del procurement pre-commerciale in quanto i servizi LOD rivestono ancora in diversi casi carattere di "sperimentazione" e possono quindi prestarsi bene a questa forma di acquisizione.

10. GOVERNANCE E SOSTENIBILITÀ

Aspetti di governance e sostenibilità di iniziative LOD sono già stati affrontati dalle presenti linee guida. Tuttavia, questa sezione ha lo scopo di riassumere alcuni dei principali aspetti al fine di fornire raccomandazioni in tal senso.

LOD governance. Per l'attuazione dell'approccio metodologico proposto, nel caso di amministrazioni di medio-grandi dimensioni è utile affiancare azioni organizzative di sostegno, che consentano di coordinare le iniziative di pubblicazione delle diverse aree dell'amministrazione.

È utile a tale scopo che l'amministrazione individui un'area organizzativa di riferimento per attuare tale governance e sostenga con apposite iniziative le azioni individuate da tale area (comunicazione, linee guida, ecc.)

Come nel caso dell'uso interno del Patrimonio Informativo Pubblico è opportuno attuare azioni coordinate di Data Governance, così anche nel caso di pubblicazione all'esterno del mondo pubblico del patrimonio dati, al fine di massimizzarne i benefici, è opportuno che vengano attuate azioni di governance. Queste azioni devono essere anche finalizzate a garantire la "sostenibilità" della pubblicazione all'esterno, attraverso l'attuazione di sinergie tra i benefici esterni e quelli interni (ad esempio la definizione di una base dati di riferimento unica per l'uso interno all'ente da parte di diverse procedure può portare anche ad un grande beneficio esterno allorché la si pubblica in formato aperto, in quanto elimina la necessità da parte dei fruitori esterni di dover integrare informazioni di fonti diverse).

Le strutture che avranno il compito di attuare la governance interna, dovranno anche rapportarsi con altri livelli della PA al fine di contribuire ad una governance allargata, aggiornando ed allineando anche periodicamente il modello di governance interno.

Sostenibilità. Per rendere l'approccio sostenibile è importante tenere in considerazione una serie di aspetti:

- coinvolgere i diversi attori interessati alle iniziative in tema di Linked Data durante *tutte* le fasi del progetto;
- fare in modo che la produzione di LOD sia parte integrante dei processi di un'organizzazione o di un'amministrazione;
- valutare se la pubblicazione dei dati grezzi (privati della logica applicativa e della relativa semantica che deriva da tale logica), gestiti da una stessa amministrazione, possono sembrare tra loro in conflitto;
- valutare se i dati sono prodotti con una dinamica frequente; nel tal caso i dati devono essere aggiornati costantemente e ciò potrebbe richiede anche risorse da parte di unità che non sono

coinvolte nella manutenzione dei LOD;

- valutare attentamente l'integrazione di dati con licenze che sono potenzialmente in conflitto tra loro;
- predisporre semplici dimostratori in grado di evidenziare in modo chiaro i benefici, per l'amministrazione, derivanti dall'adozione delle tecnologie e dei modelli di Linked Data;
- incentivare la creazione di applicazioni che creino forte consenso con i dati dell'amministrazione.

11. BIBLIOGRAFIA

[1] Agenda Digitale Europea, http://ec.europa.eu/information_society/digital-agenda/index_en.htm, 2012

[2] European Public Sector Information Platform, "Review of Recent PSI Re-Use Studies Published", http://epsiplatform.eu/content/review-recent-psi-re-use-studies-published, 2012

[3] Open definition, "Defining the Open in Open Data, Open Content and Open Services", http://opendefinition.org/okd/, 2012

[4] Open Government Data, "8 Principles of Open Government Data", http://www.opengovdata.org/home/8principles, Sebastopol, California USA, 2007

[5] W3C, "W3C Semantic Web", http://www.w3.org/2001/sw/, 2012

[6] Tim-Berners-Lee, "Linked Data", http://www.w3.org/DesignIssues/LinkedData.html, 2012

[7] HMGovernment, "Data.gov.uk, Opening up Government – Linked Data", http://data.gov.uk/linked-data, 2012

[8] Regione Piemonte, "Bollettino Ufficiale n. 48 del 1 / 12 / 2005", http://www.regione.piemonte.it/governo/bollettino/abbonati/2005/48/siste/00000114.htm, 2012

[9] Regione Piemonte, "Linee guida per il riuso", http://www.dati.piemonte.it/media/files/DGR31-11679 ALLEGATO A Linee guida riuso.pdf, 2012

[10] Regione Piemonte, "Linee guida per le informazioni del settore pubblico", http://www.dati.piemonte.it/media/files/Linee guida PSI.pdf, 2012

[11] Agenda Digitale Italiana, http://www.agenda-digitale.it/agenda_digitale/ 2012.

[12] European Communities, "Draft document as basis for EIF 2.0", 2008

[13] Governo Italiano, "dati.gov.it – i dati aperti della PA", http://www.dati.gov.it/, 2012

[14] GdL 4 – Commissione di Coordinamento SPC, "Contenuti delle future gare S2, S3", Novembre 2011.

[15] GdL 6 – Commissione di Coordinamento SPC, "Definizione dei requisiti tecnici per la transizione, l'evoluzione e il funzionamento delle infrastrutture condivise", Marzo 2012.

[16] DigitPA, Dipartimento per la Funzione Pubblica, Dipartimento per la digitalizzazione e l'innovazione tecnologica, FormezPA, "Linee guida per i siti web delle PA", 2011.

[17] J. Goodwin, C. Dolbear, G. Hart, "Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web", in Transactions in GIS, 12(Suppl. 1): 19 – 30, 2008.

[18] S. Auer, J. Lehmann, S. Hellmann, "LinkedGeoData: Adding a Spatial Dimension to the Web of Data", International Semantic Web Conference (ISWC) 2009.

[19] Garante per la protezione dei dati personali, "La privacy tra i banchi di scuola", http://www.garanteprivacy.it/garante/doc.jsp?ID=1723730, 2012.

[20] Ministero dell'Istruzione, Università, Ricerca, "La scuola in chiaro", http://archivio.pubblica.istruzione.it/scuola_in_chiaro/open_data/index.html, 2012

[21] ISTAT, "Classificazioni", http://www.istat.it/it/archivio/classificazione, 2012.

[22] SDMX - Statistical Data and Metadata Exchange, rif http://sdmx.org/, 2012

[23] JSON-stat, http://json-stat.org/, 2012.

[24] E. Ferro, M. Osella, "Modelli di Business nel Riuso dell'Informazione Pubblica", Osservatorio ICT della Regione Piemonte, 2011, http://www.osservatorioict.piemonte.it/it/images/phocadownload/modelli%20di%20business%20nel%20riuso%20dellinformazione%20pubblica.pdf

[25] T. Heath, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, http://linkeddatabook.com/editions/1.0, 2012.

[26] Linked Data – Connect Distributed Data across the Web, http://linkeddata.org/, 2012.

[27] Freie Universität Berlin, "State of the LOD Cloud", http://www4.wiwiss.fu-berlin.de/lodcloud/state/, Version 0.3, 19 Settembre 2011.

[28] DBpedia, http://dbpedia.org/About, 2012.

[29] Uniprot, http://www.uniprot.org/, 2012.

[30] LinkedGeoData, http://linkedgeodata.org/, 2012.

[31] OpenStreetMap – The free Wiki World Map, http://www.openstreetmap.org/, 2012

[32] GeoNames, http://www.geonames.org/, 2012.

[33] Tetherless World Constallation, http://tw.rpi.edu/, 2012.

[34] Data gov, Open Data Sites, http://www.data.gov/opendatasites, 2012.

[35] Europe's Public Data, http://publicdata.eu/map, 2012.

[36] Open Government Data – Catalogues, http://opengovernmentdata.org/data/catalogues, 2012.

[37] CKAN – The Data Hub, "Linking Open Data Cloud", http://thedatahub.org/group/lodcloud, 2012.

[38] W3C, Linking Open Data, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData, 2012

[39] W3C, DataSetRDFDumps, http://www.w3.org/wiki/DataSetRDFDumps, 2012.

[40] W3C, RDF Working Group, http://www.w3.org/2011/rdf-wg/wiki/Main_Page, 2012.

[41] W3C, Semantic Web Deployment Working Group, http://www.w3.org/2006/07/SWD/, 2012.

[42] W3C, Government Linked Data Working Group, http://www.w3.org/2011/gld/charter, 2012.

[43] W3C Incubator Group, Provenance XG Final Report, http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/, 8 Dicembre 2010.

[44] EU 7FP LOD2 project– Creating Knowledge out of Interlinked Data, http://lod2.eu/Welcome.html, 2012.

[45] W3C, OWL Working Group, http://www.w3.org/2007/OWL/wiki/OWL_Working_Group, 2012.

[46] W3C, RIF Working Group, http://www.w3.org/2005/rules/wiki/RIF_Working_Group, 2012.

[47] W3C, SPARQL Working Group, http://www.w3.org/2009/sparql/wiki/Main_Page, 2012.

[48] W3C, SKOS Simple Knowledge Organization System - Home Page, http://www.w3.org/2004/02/skos/, 2012

[49]W3C, RDF Web Application Working Group, http://www.w3.org/2010/02/rdfa/, 2012.

[50] N. Huijboom, T. Van den Broek, "Open data: an international comparison of strategies", *European Journal of ePractice*, N. 12, March/April 2011, ISSN: 1988-625X.

[51] K.Braunschweig, J.Eberius, M.Thiele, W.Lehner, "The State of the Open Data - Limits of Current Open Data Platforms", In *Proceedings of WWW,* Lyon France, 2012.

[52] CNR, http://data.cnr.it/site/, 2012.

[53] Comune di Firenze, http://opendata.comune.fi.it/linked_data.html. 2012.

[54] DigitPA, SPCData, http://spcdata.digitpa.gov.it, 2012.

[55] Regione Piemonte, http://www.dati.piemonte.it/rdf-data.html, 2012.

[56] Camera dei deputati, http://dati.camera.it/it/, 2012.

[57] CNR, Semantic Scounting, http://webtemp.src.cnr.it/semanticscouting/semanticscouting.php, 2012.

[58] C. Baldassarre, E. Daga, A. Gangemi, A. M. Gliozzo, A. Salvati, G. Troiani, "Semantic Scout: Making Sense of Organizational Knowledge". In *Proceedings of 17th International Conference, EKAW 2010*, Lisbon, Portugal, October 11-15, 2010.

[59] Portale Geocartografico del Trentino, http://www.territorio.provincia.tn.it/, 2012.

[60] P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi. "A semantic geo-catalogue implementation for a regional SDI", *In Proceedings of the INPSIRE Conference*, 2010.

[61] G. Lodi, A. Maccioni, F. Tortorelli, "Linked Open Data In The Italian E-Government Interoperability Framework", *In Proceedings of the 6th International Methodologies, Technologies and Tools enabling e-Government (METTEG)*, luglio 2012.

[62] Linked PA – Portale Semantico della Pubblica Amministrazione, http://www.ontologiapa.it/, 2012.

[63] DBpedia Italian, http://it.dbpedia.org/, 2012.

[64] W3C, Cool URIs for the Semantic Web – 3 Dicembre 2008, http://www.w3.org/TR/cooluris/.

[65] Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmes-xml/, 2012.

[66] Creative Commons, "Describing Copyright in RDF", http://creativecommons.org/ns, 2012.

[67] W3C, RDF Validation Service, http://www.w3.org/RDF/Validator/, 2012.

[68] Ontology design patterns, http://www.ontologydesignpatterns.org, 2012.

[69] Creative Commons Italia, http://www.creativecommons.it/Licenze, 2012.

[70] Creative Commons, CC0 1.0 Universal, http://creativecommons.org/publicdomain/zero/1.0/legalcode, 2012.

[71] Creative Commons, CC-BY Attribution 2.5, http://creativecommons.org/licenses/by/2.5/legalcode, 2012.

[72] Italian Open Data License v.2.0, http://www.dati.gov.it/iodl/2.0/, 2012.

[73] M. E. Porter, V. E. Millar, "How information gives you competitive advantage", *Harvard Business Review*, 63(4), 149-162, 1985.

[74] J. C. Rochet, J. Tirole, "Platform Competition in Two-Sided Markets", *Journal of the European Economic Association,* 1(4), 990-1029, 2003.

[75] European Commission Joinup, National Interoperability Framework Observatory (NIFO) Factsheet, https://joinup.ec.europa.eu/elibrary/factsheet/national-interoperability-framework-observatory-nifo-factsheets, 2012.

[76] R. Cyganiak, D. Reynolds, J. Tennison, "The RDF Datat Cube vocabulary", 14 luglio 2010, http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html.

[77] C. Bizer, A. Schultz, "The Berlin SPARQL Benchmark", in *International Journal Semantic Web Inf. Syst. (IJSWIS),* 5(2):1-24 (2009), http://www4.wiwiss.fu-berlin.de/bizer/berlinsparqlbenchmark/

[78] C. Corbin, "Public Sector Information Economic Indicators & Economic case study on charging models", Agosto 2010, http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/economic_study_report_final.pdf

[79] European Commission, "Report sui core indicators per fondo europeo di sviluppo regionale e fondo di coesione", Luglio 2009, http://www.dps.tesoro.it/documentazione/QSN/docs/indicatori/WD7CE_ITA_core_indicators_lug09.pdf

[80] W3C, http://www.w3c.org, 2012.

[81] Commissione di Coordinamento SPC, http://www.digitpa.gov.it/spc/commissione/attivita, 2012.

[82] OpenLab, Microsoft, https://github.com/openlab, 2012.

[83] European project LAPSI (Legal Aspect of Public Sector Information), http://www.lapsi-project.eu/, 2012.

[84] Progetto Italiano in lingua inglese EVPSI (Extracting value from Public Sector Information: Legal Framework and Regional Policies), http://www.evpsi.org/, 2012.

[85] EVPSI, "Libro bianco per il riutilizzo dell'informazione del settore pubblico", versione 1.0 (beta), http://www.evpsi.org/librobeta, 2012.

[86] The Berlin SPARQL benchmark, http://www4.wiwiss.fu-berlin.de/bizer/berlinsparqlbenchmark/, 2012.

[87] SWOOGLE – Semantic Web Search, http://swoogle.umbc.edu/, 2012.

[88] Sindice – The Semantic Web Index, http://sindice.com/, 2012

[89] Freie Universitat Berlin, SILK – A Link Discovery Framework for the Web of Data, http://www4.wiwiss.fu-berlin.de/bizer/silk/, 2012.

[90] ISA programme, "Towards open government metadata", Settembre 2011, https://joinup.ec.europa.eu/sites/default/files/towards_open_government_metadata_0.pdf

[91] Joinup, http://joinup.ec.europa.eu/, 2012.

[92] Asset Description Metadata Schema (ADMS), https://joinup.ec.europa.eu/asset/adms/description, 2012.

[93] W3C, Asset Description Metadata Schema (ADMS), Namespace Document 25 June 2012, http://www.w3.org/ns/adms

[94] ISA programme, "Report on existing Semantic Asset Repositories", deliverable, https://joinup.ec.europa.eu/elibrary/document/isa-deliverable-report-existing-semantic-asset-repositories, 2012.

[95] ISA programme, "Core Vocabularies", https://joinup.ec.europa.eu/asset/core_business/document/core-vocabularies-working-group-members, 2012.

[96] Provenance Vocabulary, http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Provenance_Vocabulary, 2012.

[97] Open Provenance Model Vocabulary Specification, http://open-biomed.sourceforge.net/opmv/ns.html#sec-intro, 2012.

[98] Provenir Ontology, http://wiki.knoesis.org/index.php/Provenir_Ontology, 2012.

[99] ISA Open Metadata License 1.1, https://joinup.ec.europa.eu/category/licence/isa-open-metadata-licence-v11, 2012.