UN PROGETTO SUGLI OPEN DATA AMBIENTALI IN EMILIA-ROMAGNA

I DATI AMBIENTALI SONO TRA QUELLI DI MAGGIORE INTERESSE PER IL PUBBLICO. LA LOGICA OPEN DATA POTREBBE CONTRIBUIRE A MIGLIORARNE LA QUALITÀ, ANCHE A USO INTERNO, GRAZIE ALLE EVOLUZIONI TECNOLOGICHE. RESTANO DA STABILIRE ALCUNE CONDIZIONI DI FORNITURA, IN MODO CHE I DATI SIANO FACILMENTE FRUIBILI DA UTENTI ESTERNI.

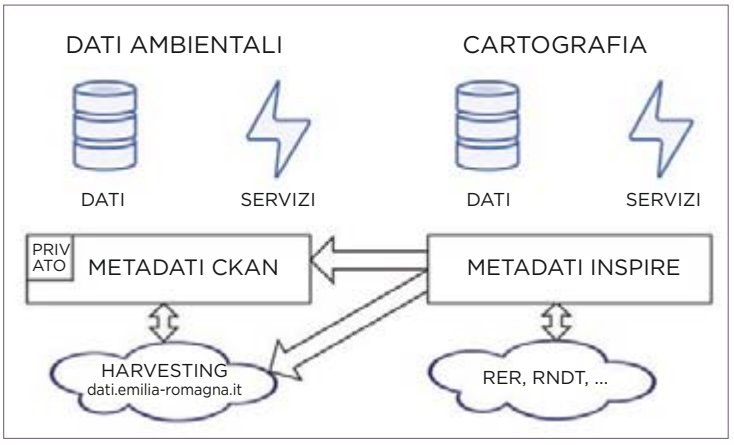
Il percorso avviato dalla pubblica amministrazione verso la liberazione dei dati è tangibile. Oggi esistono numerose esperienze e migliaia di dataset accessibili attraverso sofisticati portali informatici. E scopriamo che i dati che riguardano l'ambiente sarebbero al secondo posto in una ipotetica scala di interesse dei *bloggers*, subito dopo i bilanci pubblici ("Open data nella Pa: verso la trasparenza, tra difficoltà e qualche confusione", La Repubblica, 16 marzo 2015, http://bit.ly/Rep_OD2015). Dati ambientali che, presumibilmente, dovrebbero comprendere quelli riguardanti meteorologia, dissesto idrogeologico e, in generale, previsioni su possibili eventi naturali critici o catastrofici.

Nel 2013 Arpa Emilia-Romagna si è posta il problema di come affrontare il tema: si è partiti dal presupposto che la produzione di dati ambientali è continua e ponderosa, che una parte di questi dati viene pubblicata nel sito ufficiale www.arpa.emr.it e che, nel suo ruolo istituzionale, Arpa dedica cospicue energie per produrre un ulteriore strato conoscitivo, ovvero quello dei servizi di elaborazione e interpretazione dei dati, nonché redazione di rapporti tecnici e studi di settore per gli *stakeholder* istituzionali.
Nel voler affrontare sistematicamente un progetto di apertura dei dati elementari (grezzi) per la società civile, emerge immediatamente un aspetto organizzativo, ben evidenziato nei documenti di Agid (Agenzia per l'Italia Digitale, www.agid.gov.it) e della presidenza del Consiglio (dati.gov.it, vedi sotto). Ovvero, che la pubblicazione di dati non deve essere un'attività "a latere", quindi accessoria, ma compenetrare l'attività istituzionale. Chi produce il dato lo dovrebbe depositare in un contenitore che, in un processo di integrazione e omogeneizzazione, ed entro i limiti della normativa sulla privacy, diventa immediatamente accessibile sia agli *stakeholder* interni (a fini di utilizzo istituzionale), sia contemporaneamente a quelli esterni. L'effetto immediato è che non si generano né sovrapposizioni, né dispersione di risorse.
Un esempio virtuoso è quello che, a fronte di un'eventuale proposta di "correzione del dato" *ex-post*, da qualsiasi parte provenga, si possa bonificare la base dati direttamente alla fonte, mantenendo l'allineamento a caldo con gli open data. Questo aspetto, a prima vista ovvio e banale, in realtà cambierebbe il paradigma di "validazione" che, fino a ora, ha previsto che il dato, una volta validato dal tecnico, non sia più messo in discussione. È un vero e proprio ripensamento del ciclo di vita del dato.

IL PROGETTO OPEN DATA EMILIA-ROMAGNA

Il progetto *Open data Emilia-Romagna* riveste una importanza strategica nell'avvio di un processo di diffusione e riuso delle informazioni in possesso della pubblica amministrazione del territorio regionale da parte di privati utilizzatori, professionisti e imprese. La capacità di far emergere e moltiplicare il valore dei dati posseduti dalle pubbliche amministrazioni è direttamente proporzionale alla possibilità di renderli pienamente accessibili e riutilizzabili. Il progetto offre supporto agli enti locali o alle strutture regionali interessate a realizzare azioni di open data. Tra gli altri obiettivi, il progetto si propone di pubblicare i dati in formato *open* sul portale regionale http://dati.emilia-romagna.it secondo opportune licenze aperte. La piattaforma alla base del portale implementa un indice di dati basato su metadati quali Dublin Core ed Eurovoc e funzioni di *harvesting* basate su servizi CKAN *compliant*; la piattaforma inoltre implementa un *repository* per dataset "grezzi" e moduli per la gestione ed esposizione di *linked open data* (*triplestore*, *end point Sparql* ecc.).

Prendendo spunto da varie occasioni di confronto tra Arpa e Regione Emilia-Romagna, sta maturando un modello operativo che prefigura la seguente evoluzione: Arpa definisce un proprio *repository* di metadati ambientali unico, a uso interno ed esterno; la Regione attraverso il proprio sistema raccoglie e propaga (con il processo di *harvesting*) i metadati Arpa; come sistema di gestione di metadati cartografici, Arpa può utilizzare il geoportale regionale Inspire o il proprio portale dei metadati ambientali, attivando i relativi canali di comunicazione/allineamento per evitare duplicazioni e garantire l'integrità dei dati. Il sistema può essere rappresentato come nella figura sottostante.



Anzitutto è necessario identificare l'organizzazione necessaria a compiere questo processo virtuoso. Agid, nelle *"Linee guida nazionali per la valorizzazione del patrimonio informativo pubblico (anno 2014)"* al paragrafo 5.1, cita *"un possibile gruppo di lavoro orizzontale e inter-settoriale che un'amministrazione può costituire per avviare e gestire a regime il processo di gestione dei dati in generale e, nello specifico, di apertura dei dati"*, la cui declinazione viene dettagliata nel seguito del documento. È importante notare che al processo di gestione dell'open data viene dato un peso soprattutto organizzativo, ovvero trasversale all'interno dell'amministrazione.
Mentre, riguardo al "cosa" pubblicare per documentare i dati, viene in aiuto la *"Guida sintetica per l'interoperabilità con il catalogo nazionale open data dati.gov.it Ver_1.3_giugno_2014"*, dove si individuano i metadati e le modalità di pubblicazione, tali da renderli fruibili e accessibili anche senza intervento umano (*machine readable*). Nel nostro specifico campo di interesse, ovvero i dati ambientali, possiamo dire qualcosa in più riguardo alla classificazione, grazie al thesaurus Gemet dell'Agenzia europea dell'ambiente, da preferire al più "consigliato" Eurovoc, precisando che le ontologie di Gemet e di altri thesauri a valenza europea sono, in ogni caso, mappati in termini di interoperabilità. Per esser chiari, definendo nei metadati il "tema" e/o il "microtema" codificato secondo Gemet, potremo avere automaticamente la corrispondente catalogazione Inspire (o Eurovoc, o Agrovoc ecc.). Questo imponente incrocio di terminologie multilingua è predisposto per favorire il web semantico, arrivando agli auspicabili dataset/servizi *five stars* (cfr. classificazione open data).

A fronte di un modello concettuale più solido di una semplice pubblicazione di dataset, restano da stabilire alcune condizioni di fornitura, in modo che i dati siano facilmente fruibili da utenti esterni all'amministrazione. Anche se il punto di arrivo potrebbe essere un sistema basato su *linked open data*, dobbiamo prendere atto che sarebbe già un risultato eccellente riuscire a dare coerenza alla disponibilità dei dati ambientali in formati che ne favoriscano il riutilizzo (ritenendo più che soddisfacenti i livelli 3/4 *stars*, ovvero dataset in formati aperti e standard e servizi accessibili via web).
La prima tipologia riguarda dataset a periodicità annuale (o semiannuale) con i relativi metadati, impacchettati in file scaricabili in vari formati *machine readable* (es. Csv, Json). In questo contesto ricadono quelli di un certa consistenza e con lunghe serie storiche, come rilevazioni orarie delle reti di monitoraggio e/o dati fortemente disaggregati sul territorio.
La seconda tipologia di dataset è quella dei dati cartografici, pubblicabili sia come singoli files in formato aperto (es. GeoJSON o KML), ma anche, e soprattutto, resi disponibili come servizi web (OGC, Inspire), che garantirebbero maggiore integrità e garanzia di aggiornamento.
La terza tipologia è quella dei dati *real time* (tramite *open services web based*), come, ad esempio, dati di monitoraggio meteo o qualità dell'aria e relative previsioni a breve.

Si noti che questa articolata modalità di fornitura non deve essere orientata solo a un ipotetico utente esterno. Anzi, proprio l'utilizzo contestuale dall'interno potrebbe contribuire a migliorare il servizio in termini di qualità e documentazione del dataset, di disponibilità di vari formati e di eventuali servizi accessori.
Il ripensamento concettuale coinvolge, com'è ovvio, i sistemi informativi. Si passa da una visione come in *figura 1* a un modello dove gli open data fanno parte di un servizio completamente integrato nei sistemi informatici che, in un certo senso, viene alimentato dal *data warehouse* interno (che supporta la reportistica direzionale), attraverso la pubblicazione dei *data mart* (ovvero estrazioni di dati mirate in formato dataset), come in

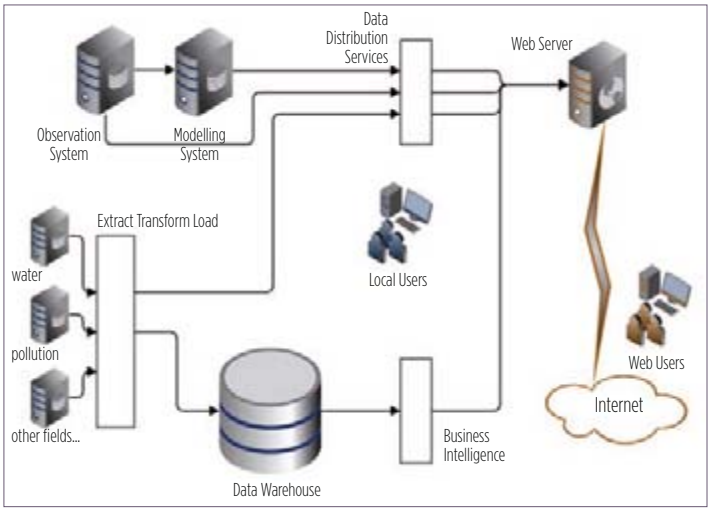


FIG. 1
SITUAZIONE ATTUALE

Schema della situazione attuale del flusso dei dati.

GLOSSARIO

CKAN: prodotto software *open source* realizzato da Open Knowledge Foundation (okfn.org) per la gestione dei metadati e la loro diffusione in rete (v. *repository* dei metadati).

Gemet (www.eionet.europa.eu/gemet) e Eurovoc (http://eurovoc.europa.eu/): sono dizionari multilingua dell'Unione europea utilizzati anche in ambito web semantico; servono a classificare univocamente e in modo interoperabile i temi e gli argomenti, secondo tassonomie e terminologie univoche e interoperabili. Il lavoro continuo di interoperabilità semantica comprende anche altri dizionari, tra i quali quello consolidato in ambito cartografico, ovvero Inspire (http://inspire.ec.europa.eu/). Si precisa che i thesauri citati, compreso Gemet, rientrano nel framework SKOS (www.w3.org/2001/sw/wiki/SKOS e www.w3.org/2004/02/skos) su cui si basa il web semantico.

Harvest: processo di raccolta e propagazione dei metadati dai *repository* disponibili in rete, che non ne modifica in alcun modo né il contenuto, né i riferimenti. Consente di accedere ai dati (fornendo l'*end point*, ovvero l'Url di rete univoco), anche accedendo a cataloghi replicati.

Linked open data: i dati aperti collegati sono una modalità di pubblicazione di dati aperti strutturati e collegati fra loro. Si basa su tecnologie e standard del web semantico e ne estende l'applicazione per fornire informazioni che possano essere lette e comprese da computer. Questo rende possibile collegare e utilizzare dati provenienti da diverse sorgenti.

Metadati: informazioni descrittive su dataset e risorse collegate (es. riferimenti, catalogazione, unità di misura, data di aggiornamento).

Repository dei metadati: database strutturato di metadati, generalmente accessibile in rete.

figura 2. Un altro modo di visualizzare la fornitura dai sistemi informativi secondo una visione basata sul catalogo di metadati (nella fattispecie CKAN), è quello esposto in *figura 3*.

Si noti che il *repository* di metadati ambientali, a regime, potrebbe diventare l'evoluzione naturale del portale Infoambiente (http://infoambiente.arpa.emr.it), nato e gestito da Arpa insieme alla Regione Emilia-Romagna per ottemperare alla legislazione vigente sulla pubblicazione delle informazioni ambientali. Il sito Infoambiente, il cui progetto risale all'epoca del Dlgs 195/05, è stato organizzato in schede inserite in un albero concettuale.
Oggi, grazie ai notevoli sviluppi in termini di software di supporto e di motori di ricerca evoluti, è possibile unificare catalogazione, fornitura di informazioni e di dati grezzi, nel rispetto delle normative vigenti e in evoluzione, attraverso i paradigmi dell'open data.
Il confine tra "informazione" e "dato" non è (ancora) sufficientemente dettagliato nella normativa vigente. Ma sappiamo che il paradigma open data presuppone dati grezzi alla massima granularità possibile. Quindi potremmo spingerci a definire come "obbligatoria" la fornitura dei dati che, ove non siano disponibili elaborazioni di secondo livello, ricomprende il livello informativo. Ma non si può affermare il viceversa: il dato genera informazione elaborata, ma dal dato elaborato è impossibile risalire al dato grezzo.

La pubblicazione degli open data, secondo la logica attuale, è favorita dalla disponibilità di prodotti informatici come CKAN (http://ckan.org/) che, oltre a fornire un eccellente supporto per la pubblicazione di metadati e dati ed essere disponibile come prodotto *open source*, è in grado di interfacciarsi con tutte le fonti e secondo tutti i protocolli standard in ambito web e cartografico. In questo senso la scelta tecnologica di Arpa e della Regione Emilia-Romagna risulta naturale e convergente. Scelta che sarà presto concretizzata nell'ambito di uno specifico progetto.

Stefano Cattani[1], Massimo Fustini[2]

1. Arpa Emilia-Romagna
2. Regione Emilia-Romagna

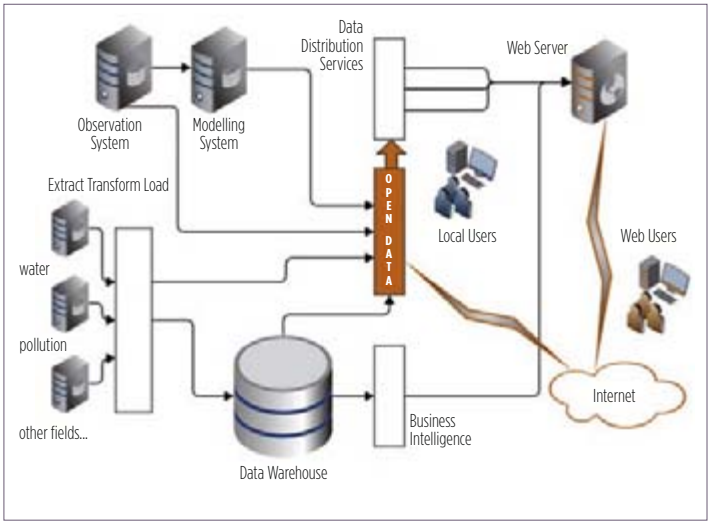


FIG. 2
SITUAZIONE DI PROGETTO

Schema del possibile flusso dei dati in logica open data.

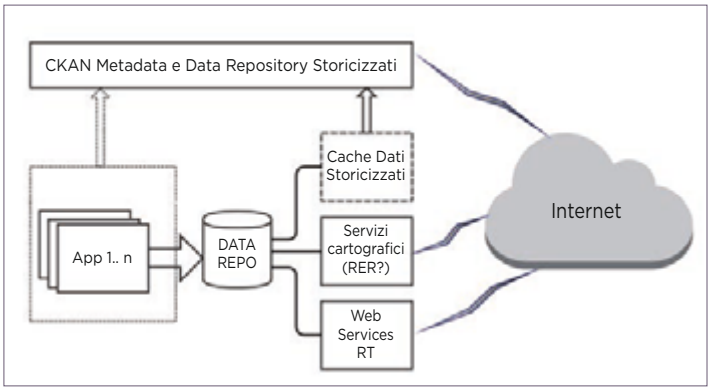


FIG. 3
SITUAZIONE DI PROGETTO

Schema della fornitura di dati secondo una visione basata sul catalogo di metadati.

OPEN DATA SU ECOSCIENZA

Alla trattazione del paradigma open data, *Ecoscienza* aveva dedicato il servizio "Open data, nuova vita per i dati pubblici" sul n. 3/2013 (http://bit.ly/ES3_2013), disponibile anche in versione ebook (.epub e .mobi) con il titolo "Ambiente open data" all'indirizzo www.arpa.emr.it/ebook.