

ORGANIZZAZIONE DI DOCUMENTAZIONE STORICA ETEROGENEA PER STUDI SULL'INSTABILITÀ GEO- IDROLOGICA: UN APPROCCIO A BASSO COSTO

Oreste SIGNORE (*), Marta CHIARLE (**), Fabio LUINO (**), Guido NIGRELLI (**),
Simonetta SALIPANTE (***), Valentina SALIPANTE (****)

(*) CNR Istituto di Scienza e Tecnologie dell'Informazione, Via G. Moruzzi,1 – 56124 Pisa, tel 0503152995, fax 0503152040, e-mail oreste.signore@isti.cnr.it

(**) CNR Istituto di Ricerca per la Protezione Idrogeologica, sede di Torino, Strada delle Cacce 73 – 10135 Torino, tel. 0113977257, fax 011343574, e-mail marta.chiarle@irpi.cnr.it; fabio.luino@irpi.cnr.it; guido.nigrelli@irpi.cnr.it

(***) Università degli Studi di Pisa, Facoltà di Scienze M.F.N., Corso di Laurea in Informatica, salipan@cli.di.unipi.it

Riassunto

L'Istituto di Ricerca per la Protezione Idrogeologica del CNR di Torino, dal 1970, anno della sua fondazione, ha raccolto un consistente e diversificato patrimonio di informazioni riguardante i fenomeni d'instabilità naturale (soprattutto frane ed alluvioni) che sono avvenuti nel bacino idrografico del Fiume Po e nei bacini del Triveneto negli ultimi tre secoli. Al fine d'informatizzare la catalogazione delle risorse, rendendo più fruibile il ricco ed eterogeneo patrimonio documentario presente, è stato progettato e realizzato, in collaborazione con l'Istituto di Scienza e Tecnologie dell'Informazione del CNR di Pisa, un sistema di gestione dell'informazione le cui caratteristiche peculiari sono la struttura aperta all'integrazione di altre fonti documentarie, la possibilità di essere utilizzato mediante un qualsiasi *web browser*, un'interfaccia grafica *user friendly* per l'esecuzione di *query*, la coerenza con le linee di sviluppo del Web e le norme di accessibilità dei siti web, e, non ultimo, il basso costo.

Abstract

The Istituto di Ricerca per la Protezione Idrogeologica of the Consiglio Nazionale delle Ricerche (CNR-IRPI) located in Turin, since its foundation in 1970, has gathered a substantial and varied large volume of information dealing with natural instability processes (mainly landslides and floods) occurred in the last three centuries in the hydrographic basins of northern Italy.

In collaboration with the Istituto di Scienza e Tecnologie dell'Informazione of the Consiglio Nazionale delle Ricerche (CNR-ISTI) located in Pisa, a system of information management has been devised and realized, with the aim to computerize all available resources. This system is characterized by an open structure to the integration by other source information, the possibility to be used by any web browser, to query by an user friendly graphic interface, the coherence with web development trends and rules of website accessibilities and, last but not the least, the low cost.

Problematica/obiettivo

L'Istituto di Ricerca per la Protezione Idrogeologica (IRPI) del CNR di Torino ha come compito principale lo svolgimento di attività di ricerca sulla dinamica evolutiva di bacini imbriferi, pianure e litorali, volta alla mitigazione dei rischi geo-idrologici (<http://www.irpi.to.cnr.it>). Nell'adempimento dei propri compiti istituzionali, l'IRPI, nel corso dei suoi 37 anni di attività, ha raccolto un consistente e diversificato patrimonio d'informazioni riguardante i fenomeni d'instabilità naturale (in particolare frane ed alluvioni) che sono avvenuti nel bacino idrografico del Fiume Po e nei bacini del Triveneto negli ultimi tre secoli. Questo patrimonio d'informazioni è raccolto in archivi diversificati a seconda della tipologia dei documenti: aerofototeca, archivio storico, biblioteca,

cartoteca, emeroteca, fototeca e videoteca. Complessivamente sono conservati circa 300.000 fotogrammi stereoscopici, 250.000 documenti inediti, 15.000 testi scientifici, 9.000 carte topografiche o tematiche, 100.000 immagini, 500 videocassette e diverse migliaia di articoli tratti da quotidiani (Fig. 1). L'attività di archiviazione del materiale è sempre stata accompagnata da catalogazione, avvenuta con criteri differenti in funzione della tipologia della risorsa. Tale schedatura non si è mai avvalsa finora di strumenti informatici. Tuttavia, al crescere della mole di documentazione archiviata, la mancanza di una catalogazione informatizzata ha rappresentato un ostacolo via via crescente ad un reperimento rapido e completo delle risorse documentarie disponibili. Al fine di sopperire a tale carenza, rendendo più fruibile il ricco ed eterogeneo patrimonio documentario presente, in collaborazione con l'Istituto di Scienza e Tecnologie dell'Informazione (ISTI) del CNR di Pisa, è stato progettato e realizzato un sistema di gestione dell'informazione, il cui nucleo è costituito da un database relazionale. Questo database è stato progettato per essere uno strumento flessibile ed aperto all'integrazione di fonti documentarie diversificate, per essere utilizzato anche da utenti non esperti sotto l'aspetto informatico e per essere consultato ed interrogato via *web*, mediante rete *intranet* ed *internet*.

Inoltre, considerato lo stretto legame con il territorio che caratterizza gli studi realizzati dall'IRPI di Torino, il DB deve altresì essere predisposto per un'integrazione con sistemi informativi geografici (GIS). Infine, poiché le ricerche dell'IRPI sono indirizzate fundamentalmente ai processi d'instabilità geo-idrologica, le risorse documentarie devono essere relazionate, quando possibile, con i fenomeni naturali ai quali si riferiscono. L'interrogazione del DB deve consentire l'individuazione delle risorse disponibili per uno specifico ambito territoriale e/o per un dato fenomeno d'instabilità.

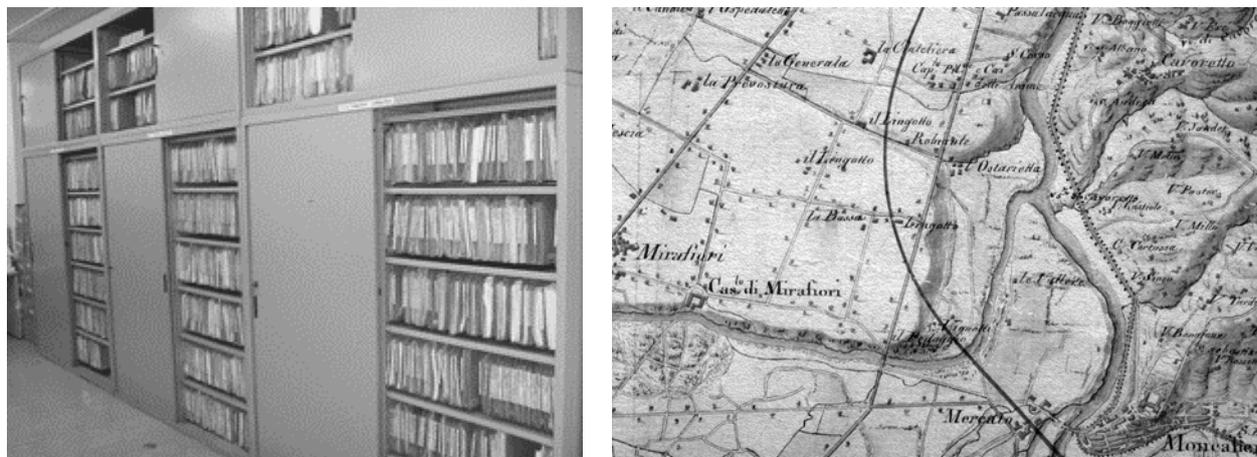


Figura 1 – A sinistra, l'archivio dell'IRPI di Torino, a destra, particolare di cartografia storica (Gran Carta degli Stati Sardi in Terraferma, anno 1819, F. Torino, zona Lingotto-Moncalieri)

Metodologia

Il ricco patrimonio informativo presente all'IRPI di Torino, per le sue caratteristiche e per le modalità con cui si è costituito nel tempo, è stato organizzato e catalogato con criteri diversi. Di conseguenza, pur trattandosi delle risorse informative di proprietà di un unico soggetto, ci si trovava di fronte al problema tipico d'integrazione d'informazioni eterogenee, che potevano essere associate in base a vari criteri (geografico, amministrativo, idrografico, temporale, causale) di volta in volta più adeguati per le varie esigenze, non esistendo un motivo per privilegiare un criterio rispetto ad un altro. Inoltre, i vari criteri non sempre sono indipendenti (per esempio, un luogo ha una sua caratterizzazione in termini amministrativi, geografici ed idrografici), e potrebbero essere fornite informazioni incoerenti, in quanto alcune informazioni, come ad esempio la localizzazione amministrativa, possono variare nel tempo.

È apparso subito evidente che l'ambiente supportato dalle tecnologie del *Semantic Web* sarebbe stato quello più adeguato per rispondere alle esigenze (Sheth, Ramakrishnan, 2003). In questo

contesto ogni risorsa è individuata da un URI (*Uniform Resource Identifier*) e descritta da un insieme di metadati. Il livello ontologico permette di rappresentare ed esportare la conoscenza, consentendo di operare ragionamenti e dedurre ulteriore informazione (Berners-Lee et al., 2001; Antoniou, van Harmelen, 2004). Le ontologie possono gestire le associazioni di tipo geografico, amministrativo e concettuale (relazioni causa-effetto, documentazione relativa a un fenomeno, progetti ed enti finanziatori, etc.). Questa soluzione, aperta anche a successive integrazioni ed allargamento ad altre realtà in possesso di dati d'interesse idrogeografico, permette di mettere in comune la conoscenza senza imporre lo spostamento delle risorse in un unico contenitore, lasciando ai detentori dell'informazione la proprietà dei dati e il controllo dell'accesso ad essi, evitando così i frequenti e ben noti fenomeni di scarsa collaborazione tra istituzioni. L'approccio culturale del *Semantic Web*, infatti, è quello di un sistema completamente decentralizzato, che non ipotizza un punto centrale di raccolta delle informazioni, ma consente l'annotazione semantica dei contenuti e il reperimento dell'informazione pertinente da parte di agenti *software* intelligenti.

Il problema maggiore per l'adozione *tout-court* di un approccio semantico/ontologico è costituito dalla necessità di rappresentare in modo formale la conoscenza, e quindi di risolvere tutte le potenziali ambiguità e incoerenze (Gómez-Pérez et al., 2004). Un approccio così rigoroso avrebbe determinato tempi lunghi e costi elevati. Ricapitolando, i problemi essenziali erano:

- eterogeneità del materiale documentario;
- complessità delle relazioni intercorrenti tra le risorse documentarie;
- apertura ad altre istituzioni per la messa in comune del patrimonio informativo;
- formalizzazione della realtà in esame (individuazione dei metadati e definizione di una ontologia minima di riferimento);
- necessità di contenere i costi;
- opportunità di verificare la bontà della soluzione adottata.

Anche se le tecnologie del *Semantic Web* sono ormai abbastanza mature per realizzare applicazioni innovative (Signore, 2005a), in prima battuta si è preferito optare per una soluzione basata su tecnologie consolidate, per favorire la raccolta delle informazioni e la loro messa a disposizione sul sito *web*, rinviando l'adozione di un approccio più propriamente ontologico a una fase successiva. È stato quindi definito uno schema di metadati, che può fungere da nucleo per la realizzazione di un ambiente completamente interoperabile: saranno necessari aggiustamenti e contributi da parte di tutti, ma lo sforzo necessario dovrebbe essere basso. Nel frattempo, la disponibilità di un sistema leggero e flessibile, ma adeguatamente robusto come tecnologia di base, avrebbe consentito, con costi e tempi contenuti, la catalogazione delle risorse esistenti e l'indicazione di metadati essenziali. I maggiori problemi incontrati nella realizzazione sono stati la modellazione delle classi, con l'identificazione delle associazioni esistenti, e il supporto della componente spaziale. Nella modellazione si è tenuto conto della duplice necessità presente all'IRPI, e cioè quella di poter effettuare, oltre alla semplice catalogazione delle risorse documentarie presenti, anche la schedatura dei fenomeni d'instabilità naturale studiati. L'integrazione di queste due forme di catalogazione risulta di particolare utilità soprattutto quando si devono effettuare *query* per studi aventi come obiettivo l'individuazione di scenari di pericolosità per specifici bacini idrografici, ai fini della mitigazione del rischio idraulico e geologico. Questo ha portato all'identificazione di due insiemi di classi (Fig. 2) che identificano due sottoinsiemi distinti, ma volutamente correlati fra loro, che corrispondono a due criteri di catalogazione diversi da utilizzare su tipologie d'informazioni diverse, ma per un'unica finalità. Tutte le istanze delle classi sono identificabili tramite URI, e quindi possono essere residenti su un qualunque *host*. In estrema sintesi, oltre alla classe *Risorsa*, che raggruppa tutte le risorse documentarie, sono state definite, tra le altre, le classi:

- *Evento*: evento sismico o meteorologico da cui derivano uno o più processi d'instabilità naturale;
- *Fenomeno*: processo d'instabilità (frana, alluvione, etc.), che può essere a sua volta collegato ad un altro Fenomeno;

logico creare un duplicato. Si è quindi preferito adottare una soluzione abbastanza grossolana, ma di minimo impatto e di agevole utilizzo, ritenendo più utile una ricerca d'informazioni che privilegia il richiamo rispetto alla precisione¹. In definitiva, ad ogni Risorsa georeferenziale sono state associate una o più istanze della classe *AreaDiInteresse*, individuata da un nome e un rettangolo, determinato da una coppia di coordinate assolute. Analogamente, alla classe Fenomeno viene associato un rettangolo che identifica l'area in cui si è manifestato, e il bacino idrografico di riferimento.

Per le classi *Evento* e *Danno* viene fornita una localizzazione descrittiva, e per la classe *Danno* la coordinata del centroide, quando questa informazione sia ritenuta utile. La collocazione amministrativa è gestita utilizzando due classi (*Comuni* e *Regioni*) basate sui dati ISTAT, e associando ad ogni comune le sue coordinate e un rettangolo che ne approssima grossolanamente il confine amministrativo. È evidente che questa soluzione è molto meno sofisticata di quelle adottate nei *GIS*, ma è anche chiaro come molte operazioni di tipo spaziale possano essere risolte facilmente e con basso carico computazionale.

Risultati

È stato realizzato un database di risorse documentarie eterogenee, eventualmente residenti su *host* diversi, identificate da un URI e corredate di metadati. Il sistema, leggero e flessibile ma adeguatamente robusto come tecnologia di base, consente la catalogazione delle risorse esistenti, con l'indicazione dei metadati essenziali, con costi e tempi contenuti. Per la realizzazione dell'applicazione *Web* è stato adottato l'ambiente oggi più diffuso². Il *database* è installato sul *server* dell'IRPI di Torino, a sua volta amministrato con *Microsoft Windows Server 2003*, ma la piattaforma utilizzata non impone vincoli specifici sul sistema operativo. La base di dati è attualmente in fase di popolazione, ma manca un sistema di *data entry*, la cui realizzazione è stata rinviata in attesa di verificare la funzionalità del sistema e la validità della formalizzazione adottata, per cui il processo di caricamento avviene mediante le *utility* di sistema³.

Per le esigenze più comuni, oltre ad utilizzare le potenzialità che offre PhpMyAdmin, è stata sviluppata un'apposita interfaccia grafica, con la quale effettuare *query* specifiche già predefinite (Salipante et al., 2007). Le pagine sono conformi alla DTD Strict dell'XHTML, e rispettano i requisiti di legge per l'accessibilità dei siti *web* (legge 4/2004)⁴. Fanno eccezione le interazioni di tipo mappale, utilizzabili a richiesta dell'utente in alternativa ad una più tradizionale presentazione sotto forma di lista. L'interfaccia supporta alcune funzioni di personalizzazione (modifica della lingua e delle modalità di presentazione dei menù a tendina) che possono essere ulteriormente arricchite.

È possibile formulare *query* sul contenuto testuale, utilizzando i normali operatori booleani, grazie alle potenzialità delle *query full text* di MySQL (nel prototipo dell'applicazione *web* è al momento supportato solo l'operatore AND), e quindi si possono individuare le risorse di interesse specificando le parole chiave che le caratterizzano.

Le *query* parametriche predefinite forniscono una lista succinta di risultati, da cui si possono richiedere approfondimenti o accedere ad altre informazioni connesse ontologicamente, secondo la metafora del “*search and link*” (Signore, 2005b).

¹ Si fa qui riferimento ai parametri di valutazione (Precision e Recall) di un sistema di Information Retrieval (Salton, McGill, 1983).

² Gli applicativi utilizzati sono *free open source*; dato che in questo settore sempre più importante dell'informatica vi sono da alcuni anni numerosi *software* che garantiscono elevati livelli di qualità e di affidabilità a costi nulli. Come *web server* è stato scelto Apache HTTP Server (versione 2.2.4) e come linguaggio di *scripting* PHP (versione 5.2.3). Il DBMS relazionale utilizzato è MySQL, il più diffuso *database open source* (versione 5.0.41).

³ Per amministrare MySQL si è scelto il *software* PhpMyAdmin (versione 2.10.2) che può essere utilizzato con qualsiasi *web browser*, e fornisce all'amministratore una semplice interfaccia grafica per gestire gli utenti che utilizzano MySQL (inserimento e cancellazione, modifiche *password*, attribuzione privilegi diversi, ecc.). L'utente, sempre mediante PhpMyAdmin, può effettuare il *data entry*, formulare le *query* e gestire le copie di *backup*.

⁴ <http://www.pubbliaccesso.gov.it/normative/DM080705.htm>

La visualizzazione su base geografico/spaziale è supportata da *Google maps*, che permette anche una metafora d'interazione/interrogazione spaziale limitata e grossolana, ma familiare alla maggior parte degli utenti e a costo nullo.

Conclusioni

Un importante problema di organizzazione ed integrazione d'informazione eterogenea è stato affrontato tenendo conto delle promettenti prospettive offerte dalle tecnologie del *Semantic Web*. Le risorse documentarie sono state classificate in diverse sottoclassi, per ognuna delle quali è stato individuato un insieme appropriato di metadati. È stata elaborata una semplice ontologia per rappresentare le correlazioni tra eventi, fenomeni, effetti e danni. Tutte le istanze delle classi sono individuate in maniera univoca da un URI, e questo consente di annotare semanticamente risorse documentarie e informative residenti su qualunque *host*, e quindi rendere disponibile tutto il patrimonio informativo relativo a fenomeni di instabilità naturale, da chiunque posseduto.

Le associazioni spaziali tra le varie risorse sono gestite mediante un'approssimazione abbastanza grossolana, ma efficace e a costo nullo.

In via preliminare, per verificare la validità dello schema di metadati elaborato, e per valutare l'adeguatezza delle soluzioni proposte, sono state utilizzate tecnologie tradizionali ed è stata realizzata un'applicazione web che consente di interrogare l'archivio mediante *query* parametriche e supporta l'interazione spaziale/geografica.

Ringraziamenti

A Guglielmo Cresci per aver creato l'occasione per svolgere questo lavoro e per le utili discussioni nelle fasi iniziali.

Bibliografia

1. Allegra P, Cristaldi F, Lollino G, Nigrelli G. (2002), "Rischio idrogeologico: applicazione di ArcView GIS per la pianificazione territoriale", 6° Conferenza Nazionale ASITA "Geomatica per l'ambiente, il territorio e il patrimonio culturale", Perugia, 5-8 novembre, 119-124
2. Allegra P, Fadda M, Lollino G, Nigrelli G, Porpiglia M. (2003), "WebGIS application to an informatic territorial database for natural risk mitigation", 4° Congresso Europeo "Cartografia Geoscientifica Regionale e Sistemi Informativi, dati geoscientifici per la pianificazione territoriale", Bologna, 17-20 giugno
3. Antoniou G, van Harmelen F. (2004), *A Semantic Web Primer*, The MIT Press, Cambridge, Massachusetts, April 2004, ISBN 0-262-01210-3.
4. Berners-Lee T, Hendler J, Lassila O. (2001), "The Semantic Web", *Scientific American*, May 2001
5. Gómez-Pérez A, Fernández-López M, Corcho O. (2004), *Ontological Engineering*, Springer-Verlag (2004), ISBN 1-85233-551-3
6. Govi M., Turitto O. (1994) "Ricerche bibliografiche per un catalogo sulle inondazioni, piene torrentizie e frane in Valtellina e Valchiavenna", Supplemento a *GEAM – Geoingegneria Ambientale e Mineraria*, Anno XXXI, n 4, dicembre, pp. 249, allegati 1, 2a, 2b, 3 floppy disk
7. Salipante S, Salipante V. (2007), "Progettazione e realizzazione in MySQL di una banca dati per risorse documentarie eterogenee", Università degli Studi di Pisa, Facoltà di Scienze M.F.N., Corso di Laurea in Informatica, Relazione di Tirocinio, pp. 33
8. Salton G, McGill M J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York
9. Sheth A, Ramakrishnan C. (2003), "Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis", *IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real*, U. Dayal, H. Kuno, and K. Wilkinson, Eds. December 2003.
10. Signore O. (2005a), "Semantic Web: il futuro è già qui?" - 10th Knowledge Management Forum - Siena, 24-25 Novembre 2005
11. Signore O. (2005b), "Ontology Driven Access to Museum Information". CIDOC 2005 Annual Conference of the International Committee for Documentation of the International Council of Museums ICOM-CIDOC. May 24 -27, 2005 Zagreb, Croatia - ISBN 953-6942-15-1, <http://www.w3c.it/papers/cidoc2005.pdf>
12. Turitto O, Di Nunzio F. (1989), "Progetto di archivio automatico per il controllo idrogeologico del territorio", Atti I Workshop Informatica e Scienze della Terra, Sarnano, ottobre, 33.1-33.9
13. Turitto O. (1991), "Banca-dati su eventi alluvionali e frane in Valtellina", Atti III Workshop Informatica e Scienze della Terra, Sarnano, ottobre, 246-252