La gestione del dato in BRISEIDE: analisi, acquisizione e pubblicazione

Simone Caldon, Luca Maroni

Reggiani S.p.A., Via Rovera, 40, 21026 Gavirate (VA)
simone.caldon@reggiani.it, luca.maroni@reggiani.it

BRISEIDE è un progetto co-finanziato dall'Unione Europea attraverso il programma ICT-PSP nel settore del Sistema dell'Informazione Geografica (GIS).
Obiettivo di BRISEIDE è di estendere i modelli di dati geografici sviluppati nell'ambito di altri progetti finanziati e nel contesto della direttiva comunitaria INSPIRE, includendo gli aspetti spazio-temporali che, con poche eccezioni, non sono contemplati dalle linee guida sui dati geografici oggi disponibili, o dalle norme in essere. Il progetto sta sviluppando diverse applicazioni basate sull'integrazione delle banche dati e dei servizi esistenti e di servizi aggiuntivi per la gestione, il processamento, l'analisi e la visualizzazione interattiva spazio-temporale dei dati.
BRISEIDE sarà applicato, testato e validato nel contesto di diversi casi d'uso in scenari di protezione civile e gestione del territorio. La fase pilota operativa coinvolgerà molti fornitori di dati a livello europeo, partner tecnologici, e utilizzatori finali con una base dati molto importante.
Tra le attività già svolte nell'ambito del progetto vi è la fase di raccolta, analisi e pubblicazione dei dati e dei servizi esistenti messi a disposizione dai vari partner allo scopo di ottenere un'immediata visione operativa sui dati in oggetto. Formati esistenti, volume complessivo dei dati, criteri di protezione dei dati sensibili e ed armonizzazione del dato sono alcuni dei fattori da considerare e pianificare attentamente.
Lo scopo di questo paper è descrivere l'esperienza maturata nella gestione dei dati in BRISEIDE per fornire linee guida e considerazione utili nella gestione di una base dati importante ed eterogenea.

Gli aspetti trattati saranno:

- Indagine dei dati e servizi esistenti.
- Linee guida per la raccolta dati.
- Architettura della piattaforma per la gestione dati.
- Verifica dei dati raccolti.
- Analisi e acquisizione del dato grezzo.
- Processo di armonizzazione.
- Pubblicazione attraverso servizi standard.
- Verifica dei dati esposti.
- Ottimizzazione delle prestazioni.

Abstract in English
BRISEIDE is an ICT Policy Support Programme project within EU. It involves 15 EU partners on the development of spatio-temporal web processes for geospatial application.
BRISEIDE aims at delivering:
1.time-aware extension of data models developed in the context of previous/on-going EU INSPIRE related projects (e.g. in the context of GMES, eContentPlus);

2.application (e.g.Civil Protection) based on the integration of existing, user operational information;
3.value added services for spatio-temporal data management, authoring, processing, analysis and interactive visualization.
BRISEIDE will be applied, tested and validated within a Civil Protection application context, using the INSPIRE relevant themes, via a chain of stakeholders, data providers, technology partners, and downstream users. The Pilot operational phase will last 12 months and will consider real life events, with extensions in additional domains, being considered and assessed.
One of the first activities carried out within the project was the data collection. To have an operative and concrete view on available data, a key task is the analysis and exposure of available data and existing services through OGC standard services. Data format and size, harmonization of data and authorization to protect sensible data are all factors to be analyzed and planned in a proper way.

The topics covered in this paper will be:

- data and existing services survey
- guideline for data collection
- infrastructure for data management
- data analysis
- data harmonization
- OGC standard services exposure
- data testing
- optimization.

Indagine dei dati e servizi esistenti
Briseide è un progetto Europeo gestito da un consorzio costituito da 14 partner, validato in 10 scenari di protezione civile e gestione del territorio reali. La base dati e servizi esistenti messi a disposizione devono essere attentamente analizzati al fine di predisporre un preciso ed efficace ciclo di raccolta, controllo e pubblicazione del dato. I dati vengono forniti da molteplici attori, con formati ed esigenze diverse.

È stato distribuito un questionario a tutti i partner del progetto per catalogare i dati e servizi esistenti:

1. Vector data
2. Raster data
3. Alphanumeri data
4. Services

Per ogni dataset disponibile le informazioni richieste sono state:

- Dataset name
- Description
- Format (VECTOR: shp, Arcinfo Coverage, dgn, Mapinfo, gml, Kml, PostGis, ArcSDE, Oracle, SQLserver, other; RASTER: jpg, ECW, tiff, Png, grid; ALPHANUMERIC: XML, Excel (xls)m Formatted text (Txt, csv), PostGres, Oracle, SQL server, DB Access, Other)
- Storage (file system, DB, Web serices, other)
- Protocol Access (http, ftp, odbc, WMS, WFS, WCS, SOS)
- Data URL
- Use (Public, Private, Private for partners)
- CRS
- Scale

- Attributes
- Temporal Properties (Geometry change during time, Attirbutes change during time, Data continuosly update (new features added), Result from spatio-temporal analisys, Time series, Geometry and attributes change during time, other)
- Language
- Metadata
- Metadata Standard
- Metadata URL
- Owner

Per i servizi le informazioni richieste sono state:

- Service name
- Description
- OGC-ws
- Use
- XML messaging
- Programming language
- Url
- Metadata
- Metadata Standard
- Metadata URL

Le informazioni raccolte sono state aggregate in un documento riassuntivo condiviso con i vari partner. È stato così possibile avere un quadro preciso della quantità di dati da trattare al fine di pianificare le risorse necessarie per la loro gestione.

Linee guida per la raccolta dati

Con un quadro chiaro dei dati e servizi disponibili all'interno del progetto è stato possibile definire una procedura condivisa per ottimizzare la raccolta e la gestione dei dati /metadati.

E' stata sviluppata una guida per illustrare ai partner del consorzio le modalità di raccolta dati e gli strumenti disponibili per la validazione/creazione di metadati compatibili, in un primo momento, con il modello INSPIRE e successivamente con il modello proposta da BRISEIDE.
Nel dettaglio:

- come caricare i dati in un area FTP condivisa e protetta da password
- formati compatibili che possono essere utilizzati per la condivisione dei dati
- caratteristiche dei metadati da produrre
- tool disponibili per la validazione e la creazione dei metadati
    - MDweb
    - CatMDEdit
    - INSPIRE
    - GeoNetwork
    - MIG
    - Preludio

Grazie a questo strumento è stato possibile standardizzare la modalità di raccolta dei dati e semplificare il lavoro di analisi, correzione e pubblicazione dati.

VECTOR	NOTE
ArcView SHP	.prj file is mandatory
Access GeoMedia Warehouse (.mdb)	no ESRI geodatabase
RASTER	**NOTE**
TIFF + TFW	zipped if heavy. Coordinate Reference System is mandatory
GeoTIFF	zipped if heavy. Coordinate Reference System is mandatory
ECW	zipped if heavy. Coordinate Reference System is mandatory
ALPHANUMERIC	**NOTE**
CSV (comma or character separated values)	PLAIN TXT format is preferred. Check carefully lat/long column format (decimal position). Coordinate Reference System is mandatory
GRID	**NOTE**
ASCII .ASC (ArcGRID)	zipped if heavy. Coordinate Reference System is mandatory
XYZ	zipped if heavy. Coordinate Reference System is mandatory
METADATA	**NOTE**
XML file / URL ref	Metadata of the real DATA. NO Services metadata. ISO 19115 INSPIRE compliant
OTHER INFO	**NOTE**
Thematic information for maps	For each layer please provide info like: Proper scale visualization value Map styling (e.g colour based on attribute values) Additional "Wished" Layer (e.g Filtered by, Geometric Intersection)
Time parameter specification	In case of separate datasets (e.g 3 datasets which 1 shp per year) give guidelines to identify: **common data schema** (if attribute schema is different between the shapes) **timestamp** (format e.g yyyy-mm-dd, how to build it if not exist)

Tabella 1 – Esempio di formati compatibili per la raccolta dati.

Architettura della piattaforma per la gestione dati
Al fine di supportare le fasi iniziali degli sviluppi software e per mettere i partner in condizione di verificare la correttezza dei dati raccolti è stata configurata una piattaforma SDI (Service Data Infrastructure) così composta:

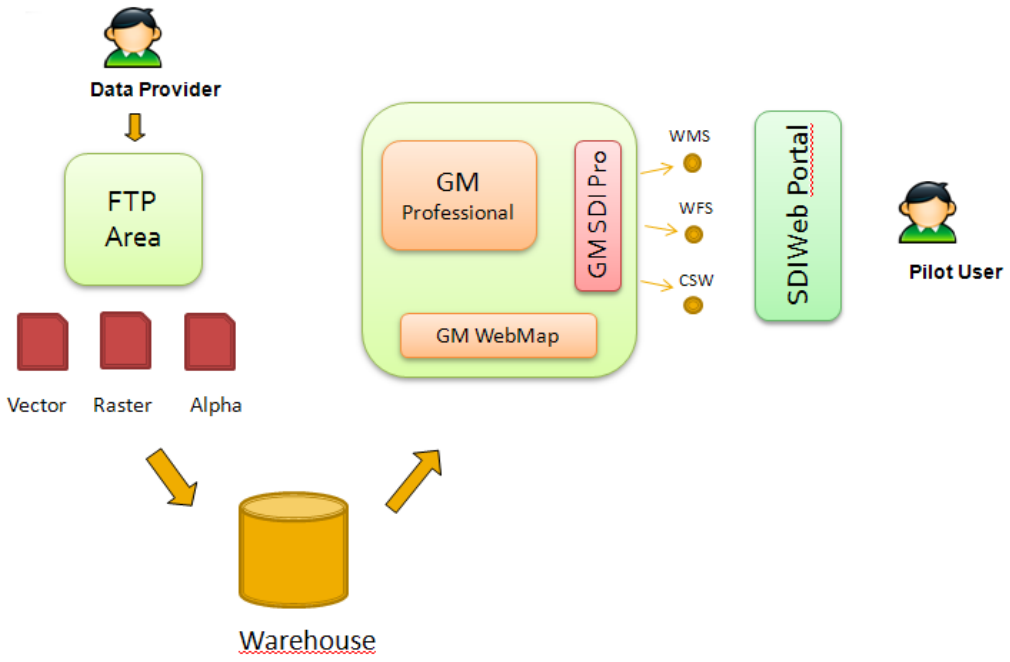


Figura 1 – Architettura piattaforma per la gestione dei dati.

L'architettura prevede le seguenti componenti:

- Area FTP: E' stato creato un ambiente con autenticazione dedicata per ciascun pilot, definendo una struttura directory base in grado di mantenere una raccolta ordinata dei file:

  PILOT-NAME/
  /ALPHANUMERIC-DATA
  /DB
  /XML
  /RASTER-DATA
  /DataSet1
  /VECTOR-DATA
  /DataSet1

- Warehouse: I dati nei formati originali sono stati importati in un database con estensioni geo-spaziali (Oracle). La conversione è stata realizzata con Intergraph Geomedia Fusion, definendo mapping tra gli attributi dei dati originali e i campi delle tabelle per i dati spaziali. Sono state, inoltre, applicate trasformazioni come ad esempio la creazione di nuovi campi per definire il timestamp del dato e l'accorpamento di più versioni temporali in un'unica struttura

- Componenti SDI Server-side: I dati importati sono stati verificati con un desktop GIS (GeoMedia Professional) e pubblicati come servizi OGC standard (WMS/WFS/WCS e CSW) tramite GeoMedia WebMap con estensioni SDI Pro

- SDI Client: E' stato configurato un portale web con profilo di autenticazione dedicato per ogni pilot al fine di facilitare la visualizzazione e verifica dei dati.

I riferimenti per poter accedere ai servizi esposti sono stati diffusi attraverso uno documento condiviso avente il seguente schema:

- Dataset ID: codice univoco del dataset

- Service: modalità di esposizione del dato

- End Point: URL di accesso al servizio

- EPSG: sistema di coordinate utilizzato

PILOT			
Dataset Id	**Service**	**End Point**	**EPSG**
VD1	WMS	http://host/serviceEndpoint1	EPSG:4326, EPSG:2100, EPSG:3785
VD2	WFS	http://host/serviceEndpoint2	EPSG:4326, EPSG:2100, EPSG:3785

Tabella 2 – Esempio elenco servizi esposti.

Sono stati inoltre configurati strumenti per il monitoring delle performance e log di accessi che hanno fornito utili indicazioni sull'utilizzo dei servizi per l'ottimizzazione delle prestazioni.

Verifica dei dati raccolti

Per la raccolta dati sono state definite delle scadenze diverse per ogni categoria di dato:

1. Vector data
2. Raster data
3. Alphanumeri data

Il lavoro di raccolta è stato monitorato mettendo a disposizione di tutti i partner un documento condiviso e accessibile tramite web. Il documento contiene l'elenco dei dataset disponibili divisi per pilot. Per ogni dataset è stato indicato:

- Nome dataset
- Descrizione del dataset
- Se la richiesta del dataset è stata inoltrata al fornitore del dato
- Se il dato è stato fornito
- Se il metadato associato è stato fornito
- Note generiche

DATASETS	Usage	Data Collection			
		Data requested	Data available	Metadata available	Note
Dataset 1 Dataset description	Public	Yes	Yes	Yes	EPSG:32633
Dataset 2 Dataset description	Public for partners	Yes	Yes	No	EPSG:32632
Dataset 3 Dataset description	Private	Yes	No	No	Not yet available

Tabella 3 – Esempio datasheet raccolta dati.

In una tabella riassuntiva è stata evidenziata la percentuale di dati forniti per poter monitorare costantemente l'andamento della raccolta dati.

Data provider	Vector data	Raster data	Alphanumeric data	Metadata	Note
DP 1	5/9	0/1	4/5	8/14	
DP 2	10/11	38/38	7/7	50/56	new alphanumeric data has been provided but not yet checked.
DP 3	6/6	¾	0/5	4/15	
DP 4	1/10	1/37	0/6	1/53	
DP 5	10/10	½	½	11/14	
DP 6	20/21	3/3	1/3	0/22	
DP 7	14/14	6/6	6/6	20/26	2 metadata are not ISO valid.
DP 8	1/2	½	0/0	3/5	3 metadata provided are not ISO valid
DP 9	5/6	0/2	0/2	0/5	
ALL	**72/89**	**53/95**	**19/36**	**95/210**	
	82%	**56%**	**53%**	**45,23%**	

Tabella 4 – Report raccolta dati.

Analisi e acquisizione del dato grezzo

Per ogni dataset messo a disposizione è stato applicato il seguente flusso di acquisizione:

1. Accesso del dato su Area FTP
2. Controllo delle caratteristiche del dato con quanto dichiarato nell'indagine iniziale
3. Controllo preliminare sulla corretta visualizzazione/sovrapposizione del dataset mediante applicativo GIS
4. Acquisizione e trasformazione del dato nel database geospaziale

5. Creazione di una tematizzazione base per la pubblicazione del dato
6. Pubblicazione del dato attraverso servizi OGC standard e profilazione delle autorizzazioni
7. Acquisizione dei metadati associato al dato nel database di catalogo
8. Configurazione del workspace dedicato sul portale web

Partendo da una situazione eterogenea e dovendo gestire una quantità considerevole di dati provenienti da fornitori di dati diversi sono emerse problematiche che possono essere riassunte nelle seguenti categorie:

- Layer "corposi" (>10000 feature)
- Attributi delle feature class equivalenti a parole riservate nel database
- Anomalie nelle geometrie (Kickbacks,Duplicate points, Loops)
- Attributi con descrizioni UNICODE
- Data temporale e timestamp da ricostruire in quanto espressi su più campi
- Caratteri di separatore decimale e campo non eterogenei (nel caso di tabelle CSV)
- Identificatori di metadato non univoci

Processo di armonizzazione
Conclusa la fase di raccolta e analisi dei dati grezzi si è reso necessario iniziare una fase di armonizzazione del dato. Tutti i dataset forniti sono stati adattati per aderire ai requisiti del progetto:

- Formato del dato
- Correzione delle geometrie e ottimizzazioni dei dati
- Estensione del dato con la variabile temporale dove necessario
- Estensione del metadato per aderire al nuovo modello proposto da Briseide

Per ogni dataset è stata compilata una scheda di dettaglio al fine di tenere traccia delle operazioni eseguite per l'armonizzazione del dato, così composta:

- Descrizione delle caratteristiche del dataset
- Operazioni eseguite per eseguire l'armonizzazione del dato e del metadato
- Tools usati durante l'armonizzazione
- Feedback/commenti relativi alle operazioni di armonizzazione

I dati armonizzati sono stati poi collezionati seguendo la stessa procedura usata per la raccolta del dato grezzo:

- Upload dei dati sull'area FTP

- Monitoraggio della fase di raccolta utilizzando un documento condiviso e accessibile dai vari fornitori di dati

Pubblicazione attraverso servizi standard
La pubblicazione è stata realizzata tramite la configurazione di servizi GeoMedia SDI Pro abilitando dove necessario un profilo di autenticazione dedicato. La pubblicazione ha seguito in generale il seguente approccio:

Tipo di dato	Servizio	CRS
Vector	WMS, WFS	EPSG:4326 e CRS nativo
Raster (GeoTIFF)	WMS, WCS	EPSG:4326 e CRS nativo
Raster (ASCII)	WMS,WCS	EPSG:4326 e CRS nativo
Alphanumeric (CSV,TXT)	WMS,WFS	EPSG:4326 e CRS nativo

Tabella 5 – Esposizione dati.

Nei casi in cui il dato presentava caratteristiche temporali è stato configurato un servizio WMS in grado di servire una versione del dato corrispondente al parametro temporale specificato (parametro TIME).

Verifica dei dati esposti
Nella fase iniziale del progetto è stata utilizzata una soluzione commerciale (Intergraph SDI Suite), che ha permesso una rapida configurazione di un'infrastruttura SDI completa per la raccolta e verifica dei dati:

- Tutti i dati sono stati esposti con servizi standard OGC (WMS, WFS, WCS)
- È stato implementato un servizio di catalogo (CSW)
- I servizi esposti sono protetti e accessibili tramite autenticazione
- È stato implementato un portale web per permettere ai partner di accedere ai dati e verificarli

Con questi strumenti ogni data provider ha potuto accedere ai propri dati per verificarne la correttezza. Per gestire al meglio la fase di test è stata definita una procedura condivisa da tutti i partner con sessioni di test registrate su formato multimediale.

E' stato definito il seguente modello per il piano di test:

No: codice univoco per identificare il test.

REQ. ID: codici univoci dei requisiti funzionali identificati nel fase di analisi del progetto

TEST DESCRIPTION: descrizione del test da eseguire

INPUT DATA: dati richiesti per l'esecuzione del test

TEST STEPS: descrizione delle azioni da eseguire per completare il test (lista analitica)

PANEL: immagine relative al test

EXPECTED RESULT: descrizione del risultato atteso

TEST RESULT: risultato del test (OK/KO)

COMMENTS: commenti utili da inoltrare al team di sviluppo riguardo a problemi emersi durante il test

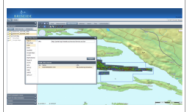
No	Pilot FUN. REQ. ID	REQ. ID	TEST DESCRIPTION	INPUT DATA	TIME RESPONSE	TEST STEPS	PANEL	EXPECTED RESULT	TEST RESULT OK/KO	COMMENTS
CAT.TEST.01	NaN	CAT.GEN.01	Implement Catalogue Services.	None	1sec	• Open Geoportal Geofoto workspace • Click on "Data Sources" • Click on "New Data Source" • To the left side of the Panel that is appeared select "CSW" • To the bottom side on the Panel select the appropriate Public Data Source "INGR BRISEIDE CSW" • Click on "Register" • Click on "Show data sources" and check if CSW is activated		Implement correctly a Catalog Service	OK	NaN
CAT.TEST.02	NaN	CAT.FUN.01	Return a list of identifiers for corresponding features for a request expressed in an OGC query language (CSW Discovery.GetRecords).	RD15-IGN	5sec	• Open Geoportal Geofoto workspace • Click on "Search for data (metadata)" • On "Search criteria" request for "IGN" Index of ignition • Click "Search" • On "Metadata Results", click on "IGN" • On "Metadata Information", view "File Identifier"		View the File Identifier of IGN	OK	NaN

Figura 2 – Esempio di test plan.

I test sono stati eseguiti in connessione remota. I partner coinvolti , insieme ai team di sviluppo, hanno partecipato in teleconferenza. Il desktop della persona che ha eseguito il test è stato condiviso per permettere a tutti di assistere allo svolgimento in tempo reale. La persona che ha eseguito il test ha descritto e illustrato ai partecipanti le azioni eseguite e le eventuali impressioni positive/negative emerse.

Tutti i test sono stati condivisi attraverso il canale YouTube del progetto:
http://www.youtube.com/user/BriseideEU#p/c/5EC22571286749DB

Alla fine di ogni test il documento contenente il test plan è stato aggiornato con i risultati ottenuti ed è stato inoltrato ai team di sviluppo per la correzione dei problemi riscontrati.

Ottimizzazione delle prestazioni

Nella prima fase del progetto il sistema è stato utilizzato in configurazione standard, dove l'accesso ai dati era effettuato mediante una connessione diretta al database geospaziale Oracle.

A seguito dei primi accessi da parte degli utenti, analizzando strumenti di log ed indicatori di performance è stata rilevata la necessità di ottimizzare il caricamento dei servizi. Si è quindi scelto di introdurre una base dati ottimizzata per la pubblicazione in grado di garantire tempi di risposta rapidi a fronte della crescita della dimensione dati complessiva.

E' stato inoltre configurato uno script di "polling" sui servizi per mantenere le risorse di sistema allocate ed in grado di servire tempestivamente le richieste.