

Archiviazione di dati digitali nell'infrastruttura cloud ENEA

Una soluzione ENEA in tecnologia cloud per l'archiviazione e la salvaguardia a lungo termine di dati e risultati scientifici, sviluppata per il settore dei Beni Culturali e trasferibile a molteplici contesti applicativi

DOI 10.12910/EAI2016-052

di **Beatrice Calosso, Gabriele Giovanetti, Francesco Iannone, Samuele Pierattini e Andrea Quintiliani, ENEA**

Negli ultimi anni la rapidissima evoluzione delle ICT ha avuto un impatto estremamente rilevante su molti settori scientifici e produttivi. Il settore dei Beni Culturali non fa eccezione a questa regola: le *Information Technologies* giocano un ruolo sempre più di primo piano nella diagnostica, nei sistemi di analisi, nelle tecnologie per la conservazione ed il restauro, e nell'ampia varietà di applicazioni per lo sviluppo di nuove modalità di fruizione dei beni artistici e culturali. Questo stato di cose si riflette tanto negli orientamenti nazionali che in quelli comunitari, nei quali si evidenzia una crescita della domanda di nuovi *digital service* sempre più orientati

alle esigenze della comunità scientifica, della Pubblica Amministrazione e del mondo dell'impresa.

Fra gli argomenti di maggiore interesse resi accessibili dall'evoluzione tecnologica è senz'altro compreso il tema della preservazione digitale su larga scala di dati e documenti. Nella moderna Società della conoscenza è possibile l'accesso ad una grande quantità di informazioni (intese come dati strutturati, ossia forniti di significato) e avere a disposizione tecnologie sempre più sofisticate per consultare, condividere ed utilizzare tali informazioni è di fondamentale importanza. In ambito internazionale sono molteplici le iniziative volte alla preservazione digitale su larga scala (*Large-Scale digital preserva-*

tion initiatives, LSDI). I principali protagonisti di questa rivoluzione digitale sono istituzioni culturali, grandi imprese come Google e Microsoft, e gruppi no-profit, tra cui la *Open Content Alliance* (OCA).

Lo scopo principale è di espandere le possibilità di accesso alle risorse documentali in una visione a lungo termine. Molti enti culturali partecipano alla OCA così come al *Million Book Project* (MBP). I principali obiettivi che spingono le biblioteche alla partecipazione a programmi LSDI sono: **accesso, preservazione, ricerca e sviluppo**. Oltre che per garantire l'accessibilità ai testi per il futuro, le biblioteche prevedono di utilizzare copie digitalizzate come backup per le opere fuori stampa,



Cattedrale di Santa Maria Assunta a Rieti e Torre campanaria

deteriorate o danneggiate in maniera irreparabile.

In ambito nazionale si registrano importanti interventi di adeguamento e rafforzamento strutturale, quali ad esempio quelli intrapresi dal Gruppo Armonizzazione Reti di Ricerca (GARR) con il progetto GARR-X Progress. Va tuttavia sottolineato come su scala europea, il 57% del patrimonio culturale debba essere ancora digitalizzato e reso accessibile alla comunità della ricerca e, più in generale, al cittadino. [1]

La medesima esigenza è infatti condivisa anche a livello europeo. Ad esempio, l'iniziativa *Digital Research Infrastructure for the Arts and Humanities European Research Infrastructure Consortium* (DARIAH

ERIC), di cui l'Italia è capofila, è finalizzata all'allestimento di una rete di strumenti, informazioni, esperti e metodologie per la ricerca nel settore *Digital Humanities*. L'infrastruttura metterà a disposizione della comunità di ricercatori che lavorano per la fruizione digitale del patrimonio culturale, testi, ricerche, *best practices* e specialmente standard metodologici e tecnici. L'iniziativa vede la partecipazione, fra gli altri del MiBACT, del CNR e dell'ENEA, attraverso la *Divisione per lo Sviluppo sistemi per l'informatica e l'ICT del Dipartimento Tecnologie Energetiche* (DTE-ICT) di ENEA.

DARIAH gioca un ruolo importante e complementare a quello della

European Research Infrastructure for Heritage Science (E-RIHS): infrastruttura fisica distribuita per la scienza e le tecnologie applicate ai beni culturali. DARIAH ed E-RIHS lavoreranno nei prossimi anni affiancate a supporto della ricerca sul tema della conservazione e gestione dei dati e della documentazione, condividendo l'obiettivo primario di *standardizzare le modalità di storage dei dati*, oltre che di armonizzare la loro fruizione da parte delle due comunità di riferimento del settore: scientifica e culturale. L'operazione coinvolgerà un gran numero di infrastrutture, strumenti, metodologie e *repository* in tutta Europa.

Coerentemente, DARIAH IT, in cui

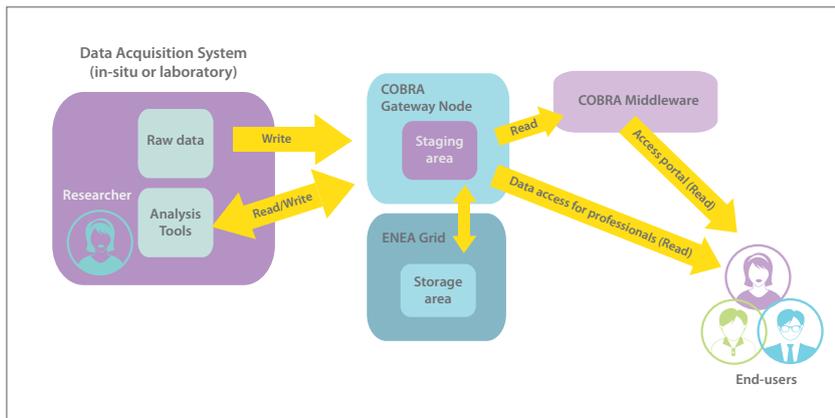


Fig. 1 Schema logico del Sistema Staging Storage Sharing (E3S)
Fonte: elaborazione ENEA

la Divisione DTE-ICT dell'ENEA è partner, mira a sviluppare una rete di *e-infrastructure* digitali di ricerca, per lo *storage* e la condivisione dei dati scientifici derivanti da applicazioni per il patrimonio culturale. [2] La Divisione DTE-ICT è in grado di rispondere alle esigenze finora descritte impiegando la propria *infrastruttura di supercalcolo scientifico*, fortemente integrata e supportata dall'insieme di risorse di rete, servizi di base, servizi *cloud*, strumenti di comunicazione e collaborazione a distanza: ENEA-GRID. È proprio questa interoperabilità di molte risorse e strumenti software e hardware a garantire non solo la qualità dei servizi, ma anche un'economia di scala del sistema. Il *cloud-storage* ENEA si appoggia sull'infrastruttura di calcolo ENEA-GRID che gestisce un sistema di *storage* distribuito e una specifica strategia di backup che ottimizza la conservazione e gestione dei dati [3].

Una significativa esperienza di *cloud-storage* è stata sviluppata nell'ambito del progetto COBRA - Sviluppo e diffusione di metodi, tecnologie e strumenti avanzati per la COnservazione dei Beni culturali, basati sull'applicazione di Ra-

diazioni e di tecnologie Abilitanti - finanziato dalla Regione Lazio (a valere sulla Legge Regionale 13 del 2008). Anche il progetto COBRA si avvale dell'infrastruttura ENEA-GRID per la creazione di un archivio digitale condiviso, denominato cloud-storage ENEA Staging Storage Sharing (E3S).

In ENEA-GRID i dati sono organizzati in due file system:

- AFS di ~ 40 TB
- IBM GPFS di ~1,3 PB.

AFS fornisce un ambiente di lavoro comune per l'accesso ai servizi e risorse informatiche, indipendentemente dalla posizione fisica degli utenti. Al contempo, GPFS è usato principalmente per applicazioni parallele e per sfruttare al meglio la potenzialità del cluster HPC CRESCO, la principale struttura di calcolo di ENEA-GRID.

AFS è un file system geograficamente distribuito che offre un'architettura client-server federata per il file-sharing; ciò significa che è possibile usufruire delle risorse presenti in diversi domini e gestite da più server. I principali vantaggi derivanti dall'utilizzo di AFS sono, oltre all'indi-

pendenza dalla posizione fisica e geografica dell'utente, la sicurezza e la capacità di migrazione trasparente. La sicurezza dei file condivisi in AFS è salvaguardata dal meccanismo di autenticazione e autorizzazione Kerberos 5, ed il meccanismo di *access control list* (ACL) è molto più potente e flessibile rispetto ai permessi standard POSIX. Un'importante caratteristica di AFS è la sua capacità di mantenere un'istantanea (*snapshot*) di ciascun volume (unità logica), memorizzato in un apposito spazio di backup.

AFS è usato in ENEA-GRID perché costituisce uno spazio scalabile, distribuito e sicuro contenente i dati dei singoli utenti, e perché permette di condividere applicazioni e librerie ma anche aree per il lavoro collaborativo con i dati prodotti nei vari progetti.

La capacità di memorizzare e conservare è praticamente illimitata: i dati di AFS sono conservati in 10 file server per un totale di circa 10 Terabyte (TB) su 30 TB di capacità, strutturato in circa 3.000 volumi disponibili [4].

Il sistema AFS risponde in modo diretto e semplice alle esigenze di *data preservation* riguardanti il supporto fisico dei dati (backup, *disaster recovery* ecc.), e mette a disposizione dei laboratori gli strumenti necessari per rispondere facilmente anche a tutte le esigenze relative all'elaborazione dei dati e alla produzione dei metadati.

Mediante questi strumenti è stato progettato e realizzato da esperti ENEA l'archivio digitale COBRA, in grado di risolvere il problema della continuità della fruibilità dei dati archiviati nel tempo (*data preservation*). Questo è necessario a maggior ragione nel caso dei dati relativi ai beni culturali, siano essi provenienti



da laboratori o da campagne di misure in situ. In questo caso infatti l'analisi dei dati su tempi lunghi può essere cruciale per la valutazione degli interventi sui beni da pianificare (restauro, messa in sicurezza ecc.) e per il monitoraggio dello stato di conservazione delle opere.

Le soluzioni adottate comprendono, infatti, strategie di backup dei supporti fisici ed una corretta progettazione di procedure per il *disaster recovery*, ma non si limitano solo a questo. L'effettiva utilità nel tempo dei dati può essere messa in discussione anche dalla rapida obsolescenza delle tecnologie, che rende di fatto inintelligibili i formati, così come dalla carenza di informazioni circa i dati stessi. Per ovviare a questa problematica – che si manifesta a lungo e lunghissimo termine – i dati devono sempre essere accompagnati da meta-dati che comprendono la dettagliata specifica del formato, le informazioni relative alla strumentazione di acquisizione e all'eventuale processamento che i dati stessi hanno subito. Inoltre, è possibile conservare anche le informazioni sui software necessari ad elaborare e post-processare i dati.

Il progetto COBRA ha, inoltre, affrontato un'altra situazione peculiare del contesto ENEA. Nei decenni, molte delle tecnologie sviluppate all'interno dell'Agenzia hanno trovato applicazione nell'ambito dei beni culturali. Ciò ha dato origine a varie linee di ricerca di grande interesse raggruppate più in base ai dipartimenti di provenienza che non all'area di applicazione. Per questo motivo la grande mole di dati accumulata in decenni di lavoro nei diversi laboratori è affetta da una grande disomogeneità. Molto spesso i singoli ricercatori si sono dovuti affidare a soluzioni "fatte in casa" per



Fig. 2 Esempio di dato archiviato: modello 3D submillimetrico, ottenuto con scanner a luce strutturata, di un reperto archeologico rinvenuto negli scavi presso il sito di Santa Maria delle Mole (Parco Regionale dell'Appia Antica)

Fonte: elaborazione ENEA

l'archiviazione dei dati.

Il sistema di *cloud-storage* ENEA *Staging Storage Sharing* (E3S) vuole porre rimedio a tale situazione, offrendo un servizio centralizzato ma al tempo stesso personalizzabile e facile da usare per l'archiviazione dei dati prodotti dai laboratori coinvolti. Laddove necessario il sistema è stato integrato con soluzioni ad hoc per il trasferimento, l'elaborazione ed anche la presentazione dei dati all'esterno. Uno schema generale di E3S è riportato in Figura 1.

Il sistema E3S si basa sulla tecnologia open-source ownCloud, modificata per automatizzare lo storage dei dati, prevenendo errori da parte degli utenti. Comprende una *staging area* dove i dati vengono messi a disposizione per elaborazione ed analisi, ed una *long-term storage area* dove i dati sono immagazzinati e conservati. Offre, inoltre, servizi software sia per i ricercatori sia per gli *end-user*, in modo tale che il ricercatore possa avere accesso riservato ai

dati e al tempo stesso scegliere cosa condividere con gli utenti esterni (in particolare, i detentori dei beni e le imprese che vi operano). La *staging area* risiede su una macchina virtuale denominata *Gateway Node*, i cui servizi sono accessibili a tutti gli utenti di ENEA-GRID, dall'interno della rete ENEA. L'accesso mediante il sistema di *cloud-sharing* ai dati da parte di utenti esterni autorizzati è invece garantito dal nodo *middleware*, accessibile da qualunque posizione, sia agli utenti con credenziali per l'Accesso ai Servizi Informatici ENEA (ASIE), sia ad utenti appositamente definiti dai ricercatori responsabili dei dati. L'interfaccia, in tutto e per tutto simile a quella di ownCloud, permette al ricercatore di definire in modo semplice e sicuro le policy di condivisione dei dati. Questa architettura si adatta bene alle filiera dei beni culturali, dove i risultati degli esperimenti devono essere condivisi con soggetti numerosi e diversi, dalle soprintendenze

Trasferimento dei dati per il monitoraggio strutturale della torre campanaria della Cattedrale di Santa Maria Assunta a Rieti

I dati acquisiti dai sensori in fibra ottica posizionati sulla torre sono trasferiti quotidianamente attraverso la rete cellulare GSM mediante un sistema di monitoraggio dinamico e strutturale impiegante sensori in fibra ottica (referente: Michele

Arturo Caponero, email: michele.caponero@enea.it).

I dati così trasferiti sono automaticamente inseriti nel sistema E3S e resi quindi disponibili ai ricercatori attraverso il sistema di *sharing*. Sul *COBRA gateway node* è stata realizzata un'applicazione web riservata agli utenti del laboratorio di riferimento che consente di uniformare il formato dei dati grezzi. Una volta elaborati, i dati potranno essere resi disponibili per le diverse forme di utilizzo previste.

ai restauratori, dagli studenti alle imprese fino al pubblico più vasto, salvaguardandone però la proprietà intellettuale.

Un'altra delle esigenze comuni a molti laboratori ENEA impegnati nel campo dei beni culturali è quella di permettere la *fruizione da remoto degli esperimenti*, visto che non sempre è possibile la presenza fisica sul luogo dell'esperimento di tutti i soggetti interessati. Perciò, nel progetto COBRA è stato realizzato anche un sistema di remotizzazione degli esperimenti che permette il collegamento multimediale e in diretta dal laboratorio durante lo svolgersi dell'esperimento e, in alcuni casi, consente lo streaming in tempo re-

ale, non solo di audio e video, ma anche dei dati prodotti.

L'integrazione tra sistema di remotizzazione e archiviazione consente di rivedere gli esperimenti in tempo differito. Questa possibilità si presta bene anche ad un utilizzo didattico e divulgativo.

Nei prossimi mesi sul nodo *middleware* saranno implementati servizi rivolti non solo allo storage ed archiviazione dei dati, ma anche alla loro presentazione. Sarà, inoltre, possibile integrare nel sistema algoritmi e strumenti per l'analisi e l'elaborazione dei dati. È in programma anche l'integrazione del sistema E3S con dispositivi mobili *low-cost* per il monitoraggio della torre campa-

narica della cattedrale di Rieti (vedi riquadro).

Infine, un valore aggiunto dell'infrastruttura realizzata per il progetto COBRA è la sua trasferibilità ad altri contesti, diversi da quello dei beni culturali. È stata infatti progettata per essere facilmente replicata in qualunque altra disciplina o filiera che condivida le caratteristiche di eterogeneità dei dati prodotti e dei soggetti coinvolti, e la sua architettura flessibile consente l'integrazione con altri componenti e servizi per soddisfare esigenze nuove e in continua evoluzione.

Per saperne di più:
beatrice.calosso@enea.it

BIBLIOGRAFIA

[1] www.garrxprogress.it/utenti/beni-culturali

[2] www.iccrom.org/it/e-rihs-a-new-heritage-research-infrastructure; www.e-rihs.eu

[3] www.eneagrid.enea.it

[4] F. Ambrosino, G. Bracco, A. Colavincenzo, A. Funel, G. Guarnieri, S. Migliori, G. Ponti (2015), “Storage architecture and backup strategy of ENEAGRID/CRESCO systems”, in *High Performance Computing on CRESCO infrastructure: research activities and results 2014*