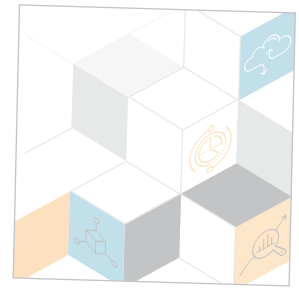
Consortium GARR

Conferenza GARR_17
Selected papers

THE DATA WAY TO SCIENCE

Venezia, 15-17 novembre 2017



Consortium GARR

Conferenza GARR 2017
Selected papers

THE DATA WAY TO SCIENCE

Venezia, 15-17 novembre 2017

ISBN 978-88-905077-7-9
DOI 10.26314/GARR-Conf17-proceedings

Editore: Associazione Consortium GARR
Via dei Tizii, 6, 00185 Roma, Italia
http://www.garr.it
Tutti i diritti riservati.

Curatori editoriali: Marta Mieli, Federica Tanlongo, Carlo Volpe
Progetto grafico: Carlo Volpe
Impaginazione: Marta Mieli, Carlo Volpe

Prima stampa: Febbraio 2018
Numero di copie: 600
Stampa: Tipografia Graffietti Stampati snc
S.S. Umbro Casentinese Km 4.500, 00127 Montefiascone (Viterbo)

Tutti i materiali relativi alla Conferenza GARR 2017 sono disponibili all'indirizzo:
http://www.garr.it/conf17

Indice

Introduzione

Angelo Scribano
Chair del Comitato di Programma della Conferenza GARR 2017

La conferenza GARR 2017 è stata l'occasione per mettere al centro del dibattito scientifico il valore dei dati della ricerca che assumono giorno dopo giorno un'importanza crescente. Il titolo scelto per questa edizione, The data way to Science, ben rappresenta l'attenzione necessaria a tutti gli aspetti relativi ai dati e a come essi stiano sempre più guidando e permeando il lavoro scientifico. Si tratta di un percorso che, nelle varie sfaccettature e declinazioni, è comune a tutte le discipline con requisiti condivisi, caratteristiche specifiche e necessità emergenti.

Durante le giornate della conferenza sono stati presi in considerazione tanti elementi che caratterizzano la vita del dato: produzione, archiviazione, gestione, diffusione, riuso, sicurezza, interoperabilità. Lo abbiamo fatto attraverso la presentazione di casi di successo e buone pratiche e con la discussione di strategie comuni nell'ambito delle infrastrutture di ricerca nazionali e internazionali. Sono stati tanti i temi in programma con speaker di rilievo internazionale che hanno affrontato il tema dei Big Data e di come questi abbiano un impatto trasversale tra i vari settori. In particolare, è stato dato spazio alla riflessione sulle prospettive dell'analisi di grandi quantità di informazioni in ambito biomedico, di come attraverso la circolazione e la condivisione dei dati si possa arrivare ad una diagnosi e ad una cura sempre più precisa e personalizzata, ma soprattutto tempestiva e predittiva. È stato affrontato in modo dettagliato il nodo importante della definizione di una policy per la gestione dei dati scientifici e di quanto sia importante poter accedere ai dati in modo semplice e sicuro da ogni parte del mondo, vista la natura sempre più globale della nostra ricerca.

Il tema degli open data è stato trattato attraverso interessanti esperienze che hanno posto attenzione anche al riuso intelligente dei dati e delle risorse. Particolarmente d'attualità nel panorama scientifico sono stati gli interventi che hanno evidenziato il ruolo che i principi metodologici e le nuove tecniche della Data Science stanno avendo nelle attività di ricerca. Si tratta di nuove frontiere che richiedono competenze ancora scarsamente presenti ma che sono destinate ad avere un impatto sempre più determinante. Ancora una volta emerge in modo chiaro il valore della multidisciplinarità. Man mano che la complessità dello studio dei fenomeni aumenta, si avverte l'esigenza di avere una rete che metta insieme le diverse conoscenze e permetta uno scambio concreto e in tempo reale.

La conferenza ha dimostrato il ruolo di GARR in questo senso, ovvero nella capacità di aggregare e mettere in comune risorse e infrastrutture ma soprattutto persone e ricercatori, da sempre il valore maggiore, e permettere quel proficuo interscambio di conoscenze oggi

più che mai indispensabile. L'evento ha visto una grande partecipazione della comunità accademica e scientifica. I lavori presentati hanno rappresentato 40 diverse organizzazioni (università, enti di ricerca, enti di ricerca biomedica come IRCCS, IZS e Istituto Superiore di Sanità, enti della PA e startup innovative che collaborano con il mondo della ricerca). In queste pagine troverete alcuni dei contributi presentati, mentre tutti i materiali della conferenza sono pubblicati sul sito dell'evento.

Vorrei infine rendere noto che i risultati del questionario di gradimento proposto al termine dei lavori testimoniano apprezzamento per la formula organizzativa e per le scelte sul programma.

Nel ringraziare tutti i partecipanti e tutti coloro che hanno permesso la realizzazione della conferenza, a partire dal comitato di programma che ha svolto con estrema professionalità il suo compito, in particolare quello non facile di selezionare i contributi tra i tanti proposti, vi auguro una buona lettura.

Chair del Workshop

Angelo Scribano, INFN e GARR

Comitato di programma

Giuseppe Attardi - GARR e Unversità di Pisa
Claudia Battista - GARR
Massimo Carboni - GARR
Agostino Cortesi - Università Ca' Foscari Venezia
Emiliano Degli Innocenti - CNR-OVI, DARIAH-IT
Paolo Favali - INGV
Claudio Grandi - INFN
Luciano Milanesi - CNR-ITB
Paola Mello - Università di Bologna
Marisol Occioni - Università Ca' Foscari Venezia
Gabriella Paolini - GARR
Fabio Pasian - INAF
Federico Ruggieri - GARR
Enzo Valente - GARR
Carlo Volpe - GARR

Tutte le presentazioni e maggiori informazioni sono disponibili sul sito dell'evento:
www.garr.it/conf17

DOI: 10.26314/GARR-Conf17-proceedings-01

Una "Big Data Open Platform" italiana per la ricerca e l'innovazione

Roberta Turra

Cineca

Abstract. Cineca, l'infrastruttura di supercalcolo più importante d'Italia, ha avviato un processo di sviluppo per la realizzazione di una piattaforma per la gestione ed elaborazione di grandi moli di dati per la ricerca scientifica e l'innovazione industriale. La nuova piattaforma conta sulle più avanzate risorse di calcolo e archiviazione, e sulla collaborazione delle comunità scientifiche che già raccolgono grandi quantità di dati da sensori e dispositivi e ne producono di nuovi attraverso le simulazioni computazionali.
La piattaforma supporta gli scienziati nella gestione dei dati durante tutto il ciclo di vita del progetto e mette insieme diversi modelli di utilizzo. Un team di esperti aiuta i ricercatori a ottimizzare l'uso delle risorse attraverso lo sviluppo e la selezione di componenti hardware e software appropriate. La piattaforma del Cineca, utilizzata da un numero sempre crescente di progetti nel settore Big Data, è stata di recente riconosciuta come innovation space (i-Space) da parte della BDVA (Big Data Value Association).

Keywords. Big Data, Data Life Cycle, HPC, Simulazioni, Deep Learning

Introduzione

La crescita esponenziale di dati generati e raccolti in quasi tutti i campi di attività apre la strada a processi di innovazione e a scoperte scientifiche "data driven" che necessitano di supporto in termini di competenze, potenza di calcolo, strumenti e servizi.
In questo contesto il Consorzio interuniversitario Cineca ha avviato lo sviluppo di una piattaforma abilitante, mettendo a frutto sia una lunga tradizione in abito di gestione dati, ontologie, data mining e business intelligence, sia le riconosciute competenze in ambito di calcolo ad alte prestazioni. Queste ultime costituiscono, in effetti, un elemento caratterizzante e distintivo rispetto ad altre analoghe piattaforme per i "big data".

La strategia di sviluppo si basa sulla partecipazione a progetti finanziati e sull'attivazione di accordi di collaborazione e di ricerca congiunta con centri di rilevanza nazionale e consolidate comunità scientifiche per la raccolta dei requisiti, lo sviluppo e la validazione di strumenti, servizi e risorse ad hoc. Vista la sua natura trasversale e intrinsecamente complessa, questa attività è svolta coniugando diverse competenze e analizzando i singoli problemi con un approccio end-to-end in stretta collaborazione con gli utenti e/o i clienti finali. Consapevoli del fatto che non esistono soluzioni universali, l'approccio seguito è quello di analizzare i casi singolarmente cercando di individuare classi di soluzioni e intrecciando competenze orizzontali, di dominio e tecnologiche.

Per dare un quadro generale dell'approccio utilizzato, di seguito viene presentato il ciclo di vita del dato declinando la presentazione nel contesto della ricerca scientifica. Ven-

gono inoltre descritte le caratteristiche della piattaforma Big Data allo stato attuale di sviluppo e gli obiettivi verso cui è indirizzata la sua evoluzione.

1. Il ciclo di vita del dato

La gestione efficiente dei dati scientifici costituisce un aspetto cruciale per consentire la ricerca, l'analisi e l'utilizzo dei dati raccolti e/o generati in quanto ne assicura l'organizzazione, l'identificazione e la descrizione. Inoltre, una buona gestione del dato ne garantisce la qualità, la protezione, la condivisione, la riproducibilità, la conservazione per un uso nel lungo periodo e il riuso. Questo ruolo fondamentale è stato riconosciuto dalla Commissione Europea che richiede, per ogni progetto finanziato, un documento descrittivo dei dati generati e della loro gestione (data management plan).

Il ciclo di vita del dato è una rappresentazione ad alto livello dei passaggi e dei processi che sono coinvolti nella gestione del dato. È utile per identificare e pianificare tutte le operazioni che devono essere implementate. Ne esistono diverse versioni a seconda delle prassi in vigore in ciascun dominio e comunità scientifica. Un esempio di riferimento è quello fornito dal Data Observation Network for Earth (https://www.dataone.org/data-life-cycle) che definisce otto componenti: la pianificazione della gestione dati, la raccolta dati, la valutazione della qualità, la descrizione mediante metadati, l'archiviazione, l'identificazione, l'integrazione e l'analisi. In questo schema l'analisi è il fine ultimo della raccolta e gestione dati e i suoi risultati possono dare luogo a nuovi progetti e nuove raccolte dati.

All'interno del progetto EUDAT (European Collaborative Data Infrastructure – www.eudat.eu), la definizione in vigore è quella del ciclo di vita del dato della ricerca scientifica dell'UK Data Service (http://www.data-archive.ac.uk/create-manage/life-cycle) che vede nell'ordine: 1) la pianificazione e raccolta, 2) il trattamento (che comprende inserimento, controllo qualità, pulizia, anonimizzazione, descrizione e archiviazione), 3) l'analisi e la produzione di risultati, 4) la conservazione, 5) la condivisione, 6) il riuso (Figura 1). In questo schema l'analisi dati è una fase intermedia e l'accento è posto sulla condivisione e il riuso. Cineca, come membro di EUDAT, e grazie anche agli altri progetti e collaborazioni, mette a disposizione strumenti e servizi che coprono tutte le fasi della gestione e analisi dati.

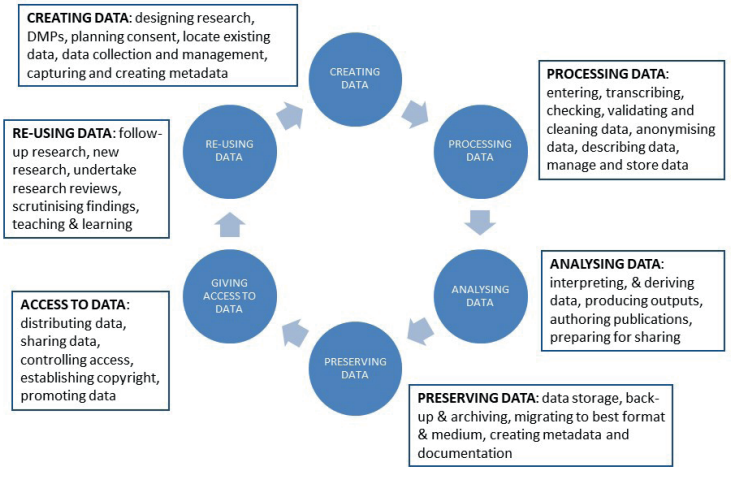


Figura 1
Ciclo di vita del dato della ricerca scientifica in uso in EUDAT
(fonte: UK Data Service)

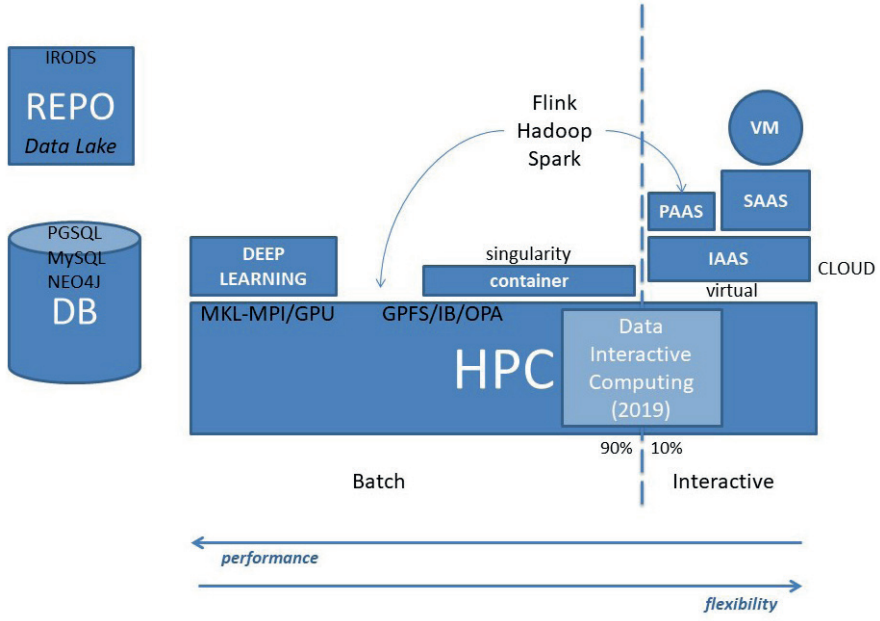
2. La piattaforma Big Data

Sviluppata a partire dal 2014, la piattaforma per i Big Data è utilizzata principalmente dalle comunità di bioinformatica e di calcolo industriale (per quanto riguarda la predictive maintenance e IND4.0), ma ospita anche numerosi progetti in altri ambiti: beni culturali e digital humanities, assicurazioni, media, energy e e-Government.

Attualmente, la piattaforma mette a disposizione risorse di calcolo che privilegiano la flessibilità e abilitano diversi modelli di utilizzo: da carichi computazionali intensi che fanno uso di GPU sull'infrastruttura HPC, in modalità batch, per gli utenti più esperti, all'uso più interattivo di risorse in modalità "container", e ottimizzate per obiettivi di analisi specifici, fino alla PaaS disponibile su risorse cloud in modalità interattiva e liberamente e autonomamente configurabile secondo le esigenze dell'utente (Figura 2).

Anche lo spazio di archiviazione offre diverse opzioni, dal data lake, uno spazio di storage per dati non strutturati, al database, uno spazio con gestione dei metadati, fino al servizio di repository, per dati strutturati.

Figura 2
Disegno logico
della piattaforma Big Data



Oltre agli strumenti e ai servizi di gestione, annotazione e analisi dati, la piattaforma mette a disposizione anche servizi di consulenza (indipendente dai fornitori di tecnologie), training e supporto utenti e competenze che vanno dall'ottimiz-zazione di codice alla data science e alla visualizzazione.

Eccellenza dell'infrastruttura, qualità dei servizi e trend crescente nel numero di progetti big data supportati, hanno consentito a questa piattaforma di ottenere la label di innovation space, i-Space, da parte della BDVA (Big Data Value Association). (http://www.bdva.eu/?q=node/790).

3. Le linee di sviluppo

L'evoluzione dei sistemi tende verso l'erogazione delle risorse, anche di supercalcolo, attraverso il paradigma del cloud computing, per garantire maggiore flessibilità agli utenti senza ridurre le prestazioni, per supportare diversi carichi computazionali e fornire ambienti isolati sicuri e interattivi. Tende inoltre verso il potenziamento dell'infrastruttura dedicata e ottimizzata per processi di deep learning per ridurre i tempi di addestramento delle reti neurali.

Il deep learning è infatti l'ambito dove maggiormente si sposano le necessità di dati e calcolo. Grandi quantità di dati devono essere disponibili per l'addestramento e grande potenza di calcolo deve essere disponibile per valorizzare l'enorme quantità di parametri (pesi) che il modello richiede. Il fatto di poter procedere in parallelo consente inoltre di esplorare diverse architetture di reti neurali simultaneamente e giungere in maniera tempestiva a identificare i modelli più efficaci.

Questa direzione di sviluppo, volta a rendere più efficienti (e quindi anche più efficaci) i processi di deep learning è originato dall'ambito big data classico, che fondamentalmente tenta di modellare il comportamento umano. Per quanto riguarda l'ambito del calcolo scientifico e la modellazione del mondo fisico, si possono porre due obiettivi:

- individuare sinergie tra l'approccio computazionale (simulazioni) e l'approccio data driven,
- sviluppare simulazioni che incorporano il deep learning / machine learning per aumentarne efficienza (accorciare il time-to-science) ed efficacia.

Nel primo caso, si tratta di accoppiare il dato simulato con quello reale, proveniente da sensori, per correggere il modello e migliorare i risultati della simulazione. Questo approccio trova applicazione anche in ambito industriale nella realizzazione di un digital twin sempre più fedele all'oggetto reale con la conseguente possibilità di modificarne il disegno e la produzione. In questo contesto è indicato l'uso del machine learning (non necessariamente del deep learning) per generare modelli empirici del funzionamento degli oggetti reali e prevederne i guasti.

Nel secondo caso si tratta invece di sostituire la parte di simulazione che assorbe maggiore potenza di calcolo con un modello di machine learning addestrato a riprodurre gli stessi risultati, date le condizioni di partenza, della simulazione. In fase di applicazione, un modello di machine learning non richiede infatti grandi potenze di calcolo e può accorciare i tempi e ridurre il consumo energetico.

4. Conclusioni

La necessità di gestire enormi e sempre crescenti quantità di dati, di diversa tipologia e in maniera tempestiva e l'opportunità di derivarne nuove chiavi di lettura della realtà sono aspetti trasversali che permeano sia le discipline scientifiche che il mondo produttivo. La condivisione dei dati, degli strumenti e delle best practices è fondamentale per l'innovazione, per nuove scoperte scientifiche e per affrontare le grandi sfide economico sociali. Per rispondere a questa esigenza è nata la piattaforma Big Data del Cineca, un ambiente che si arricchisce del contributo delle diverse comunità scientifiche cui dà supporto e che,

mettendo a disposizione gli strumenti più opportuni, consente una buona gestione di tutto il ciclo di vita del dato e favorisce la condivisione e il riuso dei dati stessi.
L'elemento distintivo rispetto ad altre piattaforme risiede nella potenza di calcolo e nelle competenze che consentono di sfruttarla al meglio, per questo motivo lo sviluppo strategico va nella direzione di ottimizzare i processi di deep learning, sia a beneficio delle applicazioni big data (data-driven), sia per inglobarli nei processi di simulazione.

Autori

Roberta Turra - r.turra@cineca.it

Roberta Turra coordina il team di Big Data Analytics del dipartimento HPC al Cineca. Si è laureata in Scienze Statistiche ed Economiche all'Università di Bologna nel 1991 e lavora al Cineca dal 1994 dove sviluppa applicazioni di data mining e text mining. Ha partecipato a numerosi progetti di ricerca finanziati a livello nazionale ed europeo e rappresenta Cineca presso la PPP BDVA (Big Data Value Association).

DOI: 10.26314/GARR-Conf17-proceedings-02

Piattaforme per l'analisi di Big Data istanziate on-demand tramite la PaaS di INDIGO-DataCloud

Marica Antonacci[1], Alberto Brigandì[2], Miguel Caballer[3], Giacinto Donvito[1], Germán Moltó[3], Davide Salomoni[4]

[1]Istituto Nazionale di Fisica Nucleare (Sezione di Bari),
[2]Concept Reply, [3]Universitat Politècnica de València,
[4]Istituto Nazionale di Fisica Nucleare (Sezione CNAF)

Abstract. Nell'ambito del progetto europeo H2020 "INDIGO-DataCloud" [1] è stata progettata e sviluppata una soluzione avanzata per il deployment di piattaforme complesse di data analytics su infrastrutture digitali distribuite ed eterogenee. L'obiettivo primario è garantire agli utenti delle comunità scientifiche un accesso immediato e trasparente alle risorse di calcolo e storage, nascondendo loro le complessità operazionali. In questo contributo vengono forniti dettagli sulle tecnologie adottate e le soluzioni implementate.

Keywords. INDIGO-DataCloud, PaaS, Mesos, Spark, TOSCA

Introduzione

La memorizzazione e l'analisi di Big Data sono oggi tra i più importanti trend nel panorama della ricerca e dell'industria, dalla medicina alla sicurezza informatica, dalla fisica delle alte energie alle scienze sociali. Ma l'analisi dei big data si configura come un'operazione tutt'altro che semplice e richiede tecniche e tecnologie diverse da quelle tradizionali.

Apache Spark [2] si è rapidamente affermato come la piattaforma di riferimento e una valida alternativa al MapReduce di Hadoop [3]. Spark è un framework open-source per calcolo distribuito, nato per essere veloce e flessibile: è, infatti, caratterizzato dalla capacità di memorizzare i risultati parziali in memoria. Può essere configurato per utilizzare Apache Mesos [4], un gestore di cluster di nuova generazione, che fornisce un efficiente isolamento delle risorse e la loro condivisione tra le applicazioni distribuite. Inoltre, Spark può essere configurato per leggere e scrivere dati su tipi di storage come, per esempio, HDFS (il filesystem distribuito di Hadoop) e Openstack Swift [5].

L'alto livello di flessibilità fornito da questi tool corrisponde ad un alto grado di complessità nell'installazione e configurazione dei vari componenti che richiedono conoscenze e competenze che spesso esulano dall'attività di ricerca degli utenti finali.

Al fine di superare queste difficoltà, abbiamo sviluppato una serie di tool basati su tecnologie open-source per semplificare e automatizzare la creazione on-demand di piattaforme per l'analisi di Big Data basate su Spark e Mesos e la gestione delle risorse allocate ai vari cluster.

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

1. INDIGO-DataCloud PaaS

La PaaS di INDIGO consiste in un set di micro-servizi che interagiscono tra loro tramite meccanismi snelli, come chiamate ad HTTP API. Dettagli sull'architettura della PaaS si possono trovare in [6].

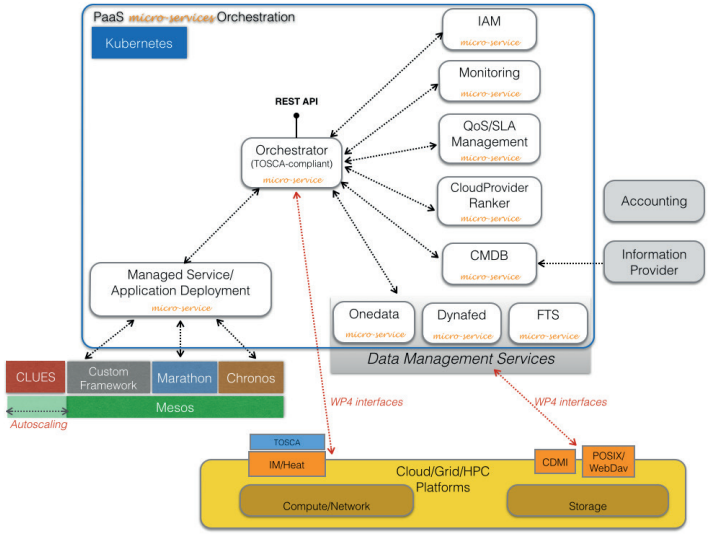


Figura 1
Architettura della PaaS di INDIGO-DataCloud

Il componente centrale della PaaS (Figura 1) è l'Orchestrator che espone l'endpoint REST per la sottomissione delle richieste da parte degli utenti e coordina le operazioni necessarie per effettuare il deployment interagendo con gli altri servizi della PaaS.
La richiesta di deployment è espressa nel linguaggio TOSCA [7] che è lo standard di riferimento per la descrizione della topologia e dell'orchestrazione dei servizi cloud. L'adozione di standard è uno degli elementi chiave del progetto INDIGO per garantire l'interoperabilità in ambienti cloud differenti.

La creazione automatica delle risorse è delegata all'orchestratore di livello IaaS: INDIGO IM (Infrastructure Manager) è capace di orchestrare in maniera trasparente diverse IaaS come Openstack, OpenNebula, Amazon, Azure, etc. Una volta che le risorse sono state allocate, esse vengono auto-configurate tramite ruoli Ansible pubblicati su Ansible-Galaxy sotto il namespace indigo-dc [8].

2. Big Data Analytics as a Service

Attraverso la PaaS di INDIGO è possibile istanziare un cluster Spark su Mesos utilizzando un semplice template TOSCA.
L'architettura del cluster Mesos di INDIGO è mostrata in Figura 2.
Esso presenta le seguenti principali caratteristiche:

- alta affidabilità: non ci sono "single point of failure" nel cluster;
- elasticità: il plugin INDIGO CLUES [9] gestisce la scalabilità automatica del cluster incrementando o riducendo le risorse in base allo stato della coda dei task da eseguire;

- persistenza dello storage: il plugin rex-ray [10] consente di fornire volumi persistenti ai docker container.

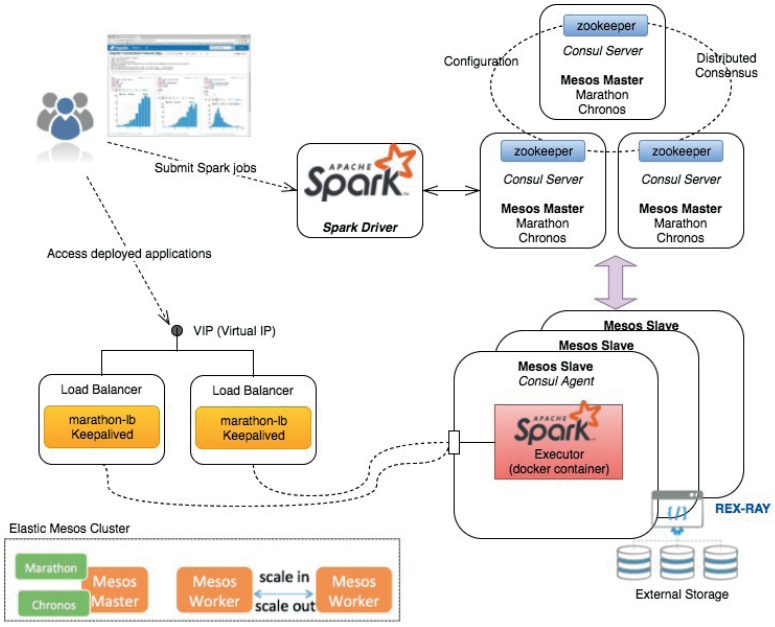


Figura 2
Architettura del cluster Spark/Mesos creato con la PaaS di INDIGO

Sul cluster Mesos viene automaticamente installato e configurato il framework Spark. La dipendenza tra Spark e Mesos è descritta nel TOSCA template come mostrato in Figura 3. L'utente che sottomette la richiesta di creazione del cluster può scegliere se installare solo il Dispatcher, a cui sottomettere i job in modalità batch, oppure installare Apache Zeppelin [11] che consente un uso interattivo di Spark tramite un'interfaccia web a notebook.
Entrambe le applicazioni (il dispatcher e l'applicazione web) sono istanziate su Marathon [12], il framework di Mesos che gestisce i servizi long-running, come applicazioni dockerizzate: l'immagine docker è stata pubblicata su Docker Hub nel namespace indigodatacloud [13]. L'utilizzo dei docker consente di pacchettizzare l'applicazione e le sue dipendenze senza necessità di installare alcun software aggiuntivo sui nodi del cluster.
L'utente ha anche la possibilità di personalizzare il cluster attraverso una serie di parametri

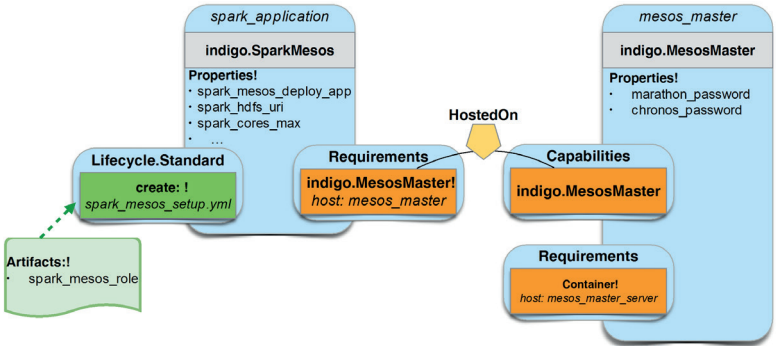


Figura 3
Diagramma logico dei nodi del template TOSCA che mostrano la dipendenza tra Spark e Mesos

di input del TOSCA template che vengono gestiti dai ruoli Ansible per configurare dinamicamente i vari componenti. E' così possibile attivare o meno opzioni specifiche come la configurazione dello storage: sia HDFS che Swift sono supportati nella nostra implementazione.

Infine, con lo stesso meccanismo è possibile richiedere l'istanziazione automatica di un nodo di monitoring per la raccolta dei log e delle metriche dei vari componenti del cluster. Il nodo viene automaticamente installato e configurato con Elasticsearch [14], utilizzato per memorizzare e indicizzare i dati raccolti, e Grafana [15] per la loro visualizzazione. I log generati dai servizi sui vari nodi del cluster vengono inviati al database tramite Fluentd [16] che fa da connettore tra il sistema di logging (rsyslog o journald) ed Elasticsearch. Le metriche, invece, vengono raccolte tramite Metricbeat [17] che le invia direttamente ad Elasticsearch.

3. Conclusioni

Utilizzando gli strumenti sviluppati per la PaaS di INDIGO è possibile istanziare, con un template TOSCA, un'infrastruttura per big data che consiste in un pool di risorse cloud automaticamente allocate e configurate per far eseguire workload Spark su un cluster Mesos. La soluzione implementata consente di eliminare i costi operazionali legati all'installazione e configurazione dell'intero stack software. Con le stesse tecnologie e componenti, abbiamo anche sviluppato una soluzione semplice ed automatizzata per creare e gestire un cluster dinamico istanziato on-demand per l'analisi di dati LHC (CMS) dimostrando la versatilità dell'approccio.

Riferimenti bibliografici

[1] https://www.indigo-datacloud.eu

[2] https://spark.apache.org

[3] http://hadoop.apache.org

[4] http://mesos.apache.org

[5] https://docs.openstack.org/swift

[6] D. Salomoni et al. (2016), INDIGO-Datacloud: foundations and architectural description of a Platform as a Service oriented to scientific computing (arXiv:1603.09536v3)

[7] https://www.oasis-open.org/committees/tosca

[8] https://galaxy.ansible.com/indigo-dc

[9] https://indigo-dc.gitbooks.io/clues-indigo

[10] https://rexray.readthedocs.io

[11] https://zeppelin.apache.org

[12] https://mesosphere.github.io/marathon

[13] https://hub.docker.com/u/indigodatacloud

[14] https://www.elastic.co/products/elasticsearch

[15] https://grafana.com

[16] https://www.fluentd.org

[17] https://www.elastic.co/guide/en/beats/metricbeat/current

Autori

Marica Antonacci - marica.antonacci@ba.infn.it

Lavora come Tecnologo, esperto di Cloud Computing, presso INFN-BARI. Ha partecipato a diversi progetti sperimentando soluzioni open-source per sistemi di calcolo e storage distribuiti.

Ha contribuito alla realizzazione dell'infrastruttura cloud di produzione del sito di Bari (di cui è amministratore di sistema) e alla sua integrazione nella EGI Federated Cloud.

Alberto Brigandì - a.brigandi@reply.it

Lavora dal 2015 presso Santer Reply (Torino), nel dipartimento Ricerca e Sviluppo con il ruolo di Software Architect. Ha preso parte a diversi progetti, sia Italiani che Europei.
I suoi ambiti di ricerca ricoprono l'analisi e il design di infrastrutture cloud private e di piattaforme IoT.

Miguel Caballer - micafer1@upv.es

Dal 2001 è membro del gruppo di ricerca "Grid and High Performance Computing" dell'Institute of Instrumentation for Molecular Imaging (I3M). Ha partecipato a diversi progetti di ricerca europei e nazionali sull'applicazione delle tecnologie di calcolo parallelo, grid e cloud in vari campi dell'ingegneria.

Giacinto Donvito - giacinto.donvito@ba.infn.it

È il responsabile tecnico del data center ReCaS-Bari. Ha partecipato a tutti i progetti EGEE ed EGI e ha lavorato per diversi progetti di bioinformatica acquisendo esperienza nel supportare gli utenti provenienti da diverse comunità scientifiche, sulle più moderne infrastrutture di calcolo distribuito. Nel progetto ReCaS ha avuto un ruolo centrale nella progettazione e nella realizzazione di RecaS-Bari. È stato direttore tecnico e WP leader nel progetto INDIGO-DataCloud.

Germán Moltó - gmolto@dsic.upv.es

Dal 2002 è membro del gruppo di ricerca "Grid and High Performance Computing" (GRyCAP) presso l'I3M. È professore associato nel Dipartimento di Computer Systems and Computation presso UPV. Ha partecipato a diversi progetti europei e condotto progetti nazionali nel campo del cloud computing.

Davide Salomoni - davide.salomoni@cnaf.infn.it

È Dirigente Tecnologo presso l'INFN; attualmente è responsabile del gruppo di Ricerca e Sviluppo del CNAF (Bologna). E' il coordinatore del progetto H2020 INDIGO-DataCloud.

Dirige e partecipa a numerosi progetti nazionali e internazionali come EOSCPilot (call INFRADEV-04-2016).

È il coordinatore dell'INFN Cloud computing workgroup ed è coinvolto in attività che riguardano il trasferimento tecnologico in Università, PA e aziende attraverso seminari e corsi.

DOI: 10.26314/GARR-Conf17-proceedings-03

Big Archaeological Data. The ArchAIDE project approach

Francesca Anichini, Gabriele Gattiglia

Università degli Studi di Pisa

Abstract. Digitisation has changed archaeology deeply and has increased exponentially the amount of data that could be processed, but it does not by itself involve datafication, which is the act of transforming something (objects, processes, etc.) into a quantified format, so they can be tabulated and analysed. Datafication fits a Big Data approach and promises to go significantly beyond digitisation. To datafy archaeology would mean to produce a flow of data starting from the data produced by the archaeological practice, for instance, locations, interactions and relations between finds and sites. The ArchAIDE project goes exactly in this direction. ArchAIDE is a H2020 funded project (2016-2019) that will realise a tool for recognising archaeological potsherds; a web-based real-time data visualization to generate new understanding; an open archive to allow the archival and re-use of ar-chaeological data. This process would move archaeology towards data-driven research and Big Data.

Keywords. Archaeology, Digitisation, Datafication, data-driven research, Big Data

Introduction

Data are what economists call a non-rivalrous good, in other words, they can be processed again and again and their value does not diminish (Samuelson, 1954). On the contrary, their value arises from what they reveal in aggregate. On the one hand, the constant enhancement of digital applications for producing, storing and manipulating data has brought the focus onto data-driven and data-led science even in the Humanities, on the other hand, in recent decades, archaeology has embraced digitisation. In recent years, archaeologists began to ask to themselves if a Big Data approach can be applied to archaeology from both a theoretical and practical point of view (Gattiglia 2015).

For a better understanding of the general concept of Big Data, we adopt the definition proposed by (Boyd et al. 2012): "Big Data is less about data that is big than it is about a capacity to search, aggregate, and crossreference large data sets". In other words, Big Data's high volume, high velocity, and high variety do not have to be considered in an absolute manner, but in a relative way. As suggested by (Mayer-Schönberger et al. 2013), using Big Data means working with the full (or close to the full) set of data, namely with all the data available from different disciplines that can be useful to solve a question (Big Data as All Data). This kind of approach permits to gain more choices for exploring data from diverse angles or for looking closer at certain features of them, and to comprehend aspects that we cannot understand using

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

smaller amounts of data.

1. Datafication

Digitisation has changed archaeology deeply increasing exponentially the amount of data that could be processed, but from a more general point of view the act of digitisation, i.e. turning analogue information into computer readable format, does not by itself involve datafication. Datafication promises to go significantly beyond digitisation, and to have an even more profound impact on archaeology, challenging the foundations of our established methods of measurement and providing new opportunities. To datafy means to transform objects, processes, etc. in a quantified format so they can be tabulated and analysed (Mayer-Schönberger et al. 2013). Moreover, a key differentiating aspect between digitisation and datafication is the one related to data analytics: digitisation uses data analytics based on traditional sampling mechanisms, while datafication fits a Big Data approach and relies on the new forms of quantification and associated data mining techniques, that permit more sophisticated mathematical analyses to identify non-linear relationships among data, allowing us to use the information, for instance, for massive predictive analyses. In other words, to datafy archaeology would mean to produce a flow of data starting from the data produced by the archaeological practice, for instance, locations, interactions and relations between finds and sites. A flow of data that the archaeological community should have available.

2. ArchAIDE project

The ArchAIDE project goes exactly in this direction. ArchAIDE is a three-year (2016-2019) RIA project, approved by EC under call H2020-REFLECTIVE-6-2015. The project consortium is coordinated by the University of Pisa with the MAPPA Lab, and includes a solid set of Human Sciences partners (University of Barcelona, University of Cologne and University of York), some key players in ICT design and development (CNR-ISTI and Tel Aviv University), two archaeological companies (BARAKA and ELEMENTS) and one ICT company.

The work of the project includes the design, development and assessment of a new software platform offering applications, tools and services for digital archaeology. This framework, that will be available through both a mobile application and a desktop version, will be able to support archaeologists in recognising and classifying pottery sherds during excavation and post-excavation analysis. The system will be designed to provide very easy-to-use interfaces (e.g. touch-based definition of the potsherd profile from a photograph acquired with the mobile device) and will support efficient and powerful algorithms for characterisation, search and retrieval of the possible visual/geometrical correspondences over a complex database built from the data provided by classical 2D printed repositories and images. We thus plan to deliver efficient computer-supported tools for draft-ing the profile of each sherd and to automatically match it with the huge archives provided by available classifications (currently encoded only in drawings and written descriptions contained in books and publications). The system will also be able to support the production

of archaeological documentation, including data on localisation provided by the mobile device (GPS). The platform will also allow to access tools and services able to enhance the analysis of archaeological resources, such as the open data publication of the pottery classification, or the data analysis and data visualisation of spatial distribution of a certain pottery typology, leading to a deeper interpretations of the past. Data analysis will be achieved as an ex-ploratory statistical analysis of data related to pottery. It will be mainly concerned with data about size, density, geo-localisation and chronology. The main objective of the exploratory analysis is to disclose statistical relationships (in statistical sense) between the different variables considered. Moreover, it will provide a comprehensive description of the available data, pointing out important features of the datasets, such as: where the information concentrates and where is missing, or where little data more would imply a relevant gain of information. There are different statistical techniques useful for exploratory data analysis, each one concentrating on particular aspects of the description we would like to give for the data. However, it is important to observe that the statistical techniques are not ex-ploratory as such, rather they are used in order to summarize main characteristics of data, identify outliers, trends, or patterns, i.e. they are used as explorative.

Concerning the analysis of pottery datasets, we will concentrate on the following tools:

- classification and clustering techniques, to be used for understanding whether or not some features of the data may possess convenient classifications in a number of categories/groups, subsequently suggesting meaningful interpretation of such categories;
- dimensionality reduction techniques, to be used in order to extract a small number of specific combination of features describing the greatest part of information and variability contained within the data. These specific combinations provide all at once a way to summarize data, and the identification of the major sources of variability;
- spatial statistics, point pattern analysis and Kriging methods will be mainly used in order to highlight the possible patterns within the spatial distribution of data;
- different predictive modelling techniques will be implemented mostly for suggesting where to look for more data in order to get relevant gain of information, or optimal strategies to perform testing.

The results of the data analysis will be made more understandable and easily explicable applying data visualisation techniques. Apart from the quantitative data analysis, data visualization is of extreme importance, in order to: provide an efficient way to understand a vast amount of data; allow non-technical people to do data-driven decision making; communicating the results of the data analysis (Llobera 2011). An important issue is the communicating the visual information about the relationships among different ceramic classes in the same location, the relationships between the location of the finding and the productive centre, and the relationships with pottery found in different locations. A web-based visualisation tools will be built following the principles of data visualization.

Following these guidelines, we will classify the different data into types (categorial, ordinal, interval, ratio types), and will determine which visual attributes (shape, orientation, colors, texture, size, position, length, area, volume) represent data types most effectively,

so giving rise to the visualization, according to the basic principle of assigning most efficient attributes, such as position, length, slope, to the more quantitative data types, and less efficient attributes, like area, volume, or colors to ordinal or categorical types. The process of building the visualisation will be made interactive, letting the user associating the different variables with the different attributes, at the same time explaining the principles above. Moreover, the different relations within pottery production, trade flows, and social interactions, will be visualised applying the same principles, with graphs.

4. Conclusions

The possibilities of such system open to research actors, institutions and general public would be a dramatic change in the archaeological discipline as it is nowadays. Its impact on the field would dramatically change the profile of the professionals involved and will generate new markets.

References

Boyd D., Crawford K. (2012), Critical Questions for Big Data. Information, Communication and Society, (15), pp. 662–679

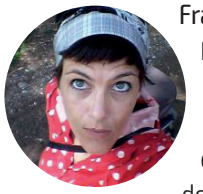
Gattiglia G. (2015), Think big about data: Archaeology and the Big Data challenge, Archäologische Informationen, (38), pp 113-124.

Llobera M. (2011), Archaeological Visualization: Towards an Archaeological Information Science (AISc), Journal of Archaeological Method and Theory, (18), p: 193–223.

Mayer-Schönberger V., Cukier K. (2013), Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, Boston, MA.

Samuelson P.A. (1954), The Pure Theory of Public Expenditure, Review of Economics and Statistics (36,4), pp 387-389.

Authors



Francesca Anichini - francesca.anichini@for.unipi.it

Francesca has worked as professional archaeologist for several years and directed excavations from Roman to Post-medieval ages. Since the 2010 she also works in MAPPALab at the University of Pisa as project and communication manager. She is one of developers of MOD - the Italian repository for Open Archaeological Data, and she deals with methodologic issues related to open data, archaeological potential and communication.

Gabriele Gattiglia - gabriele.gattiglia@for.unipi.it

He is a Researcher at the University of Pisa, coordinator of the ArchAIDE Project and leads the MAPPA Lab, which manages the MOD (Mappa Open Data), the Italian repository for Open Archaeological Data. He is devoted to digital application in archaeology, open data and Big Data issues in archaeology.

DOI: 10.26314/GARR-Conf17-proceedings-04

Digital Biomarkers: why Precision Medicine need them

Enrico Capobianco

Center for Comuptational Science, University of Miami, Fl-USA

Abstract. "The advent of digital tools presents the opportunity to revolutionize the data capture methods — specifically, the ability to collect more specific, relevant data points or digital biomarkers".
According to Rock Health's report, digital biomarkers are "consumer-generated physiological and behavioral measures collected through connected digital tools" which "represent an opportunity to capture clinically meaningful, objective data in a cost-effective manner."
Research can tremendously benefit from longitudinal data capture throughout a subject's daily life, gaining accuracy through disease-specific algorithms designed to process subject data quantify the severity and progression of symptoms or alerting about signal warning suggesting intervention.

Keywords. Precision Medicine; Digital biomarkers

Opinion

Precision Medicine has rapidly become the horizon to look at, and this is a trend which is destined to grow. However, some limitations and open problems remain to be solved.
For example, how effective will be data integration at a global scale? How to harmonize geo-differentiated information and manage ethnic balance? What is the status of consensus over protocols for digital repositories involving transferability of structured and non-structured records? These are examples of "missing links", i.e. gaps that need to be addressed to speed up the optimization of decision making processes governing Precision Medicine. The achievement of actionable steps implies that synergies must be enabled at various levels.

Consider one of the priorities: establishing markers of diseases and associated comorbidities. This no doubt remains an open frontier. What we have now available to diagnose and cure is clearly insufficient, especially with several complex diseases, and despite many clinical progresses, technological advances, and overall changes affecting individuals and societies that have gradually assigned a central role to personalized decisions with regard to prevention, cure and care.

Benefits are expected from the influence of factors with marginal or none role in the past. Interestingly, not only patients but healthy individuals too are targets of next generation studies. Wellbeing, healthy ageing, stress balance have become popular terms. And many interests, including economic ones, are centered on risk profiles derived from personal histories, electronic clinical records, physiological measurements, genetic information, socio-economic determinants, environmental and social influences, nutrition aspects and more.

Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

An effective interoperability between many resources outsourcing all such multi-dimensional information is crucial, and a way to address this need is to provide measures of synthesis defined through Digital Biomarkers. They are defined as indicators of health status obtained from medical devices, mobile technologies, software tools, wearables and apps measuring physio-pathological and behavioral parameters in both active (at the gym) and passive (during sleep) conditions, day and night, everywhere.

This definition can in my view be expanded, considering Digital Biomarkers as extensions of molecular and clinical markers obtained by connecting other layers of complementary information relevant to health status. Not by chance the sector of digital health includes among its categories digital therapeutics, one destined to be highly dynamic in the near future. Therefore, the model which is likely to be sustainable may be based on integrated cross-linked digital records available to doctors for timely and effective patients management.

Ultimately, and most importantly, by looking at digitally organized, standardized and curated data, the quality of care will improve. This would be consequence of ta process of data democratization, an hybrid generated by the fusion of web social influences and data liquidity. An assumption is that a recognized new role in this process involves patients, more actively role controlling factors falling outside the doctor's sphere, say life-style, technology and socials. Social implications, in particular, are key to determine the role of Digital Biomarkers dynamically. This is a natural consequence of the fact that the complex multifaceted underlying data generative process at place is subject to variation at both individual and small-to-large community scales.

Novel uses of social media are de facto inspiring a 360-degree re-assessment of health and disease, in terms of interactions (care and cure), perception of conditions (self-quantified), remote delivery or monitoring (telemedicine), and community medicine models. Collectively, these advances will open up limitless opportunities that may become concrete benefits once cleared from biases and confounders, and instead elucidated by factors significantly tested to be determinants of lifestyle modifications. Let us think for a moment, and objectively: how many years have we listened to possible health risks from multiple forms of addiction, say cellular phones, TV, web, emails, till occasional use of alcohol and drugs? And, what do we really know from the consensus reached so far? Not much, as no consensus is presently there.

Therefore, a sort of new health chapter will be written by Digital Biomarkers, one with a title including the word "epigenetic hallmarks", and with a word that should not be missing, "people".

Author

Enrico Capobianco - ecapobianco@med.miami.edu

Enrico has supported the worldwide growth of Systems Medicine in the last ten years, especially leading Network Science to targeted methodological and algorithmic applications in Precision Medicine. He is currently a scientist at the University of Miami, working at the Center for Computational Science, the Cancer Center and the School of Medicine.

DOI: 10.26314/GARR-Conf17-proceedings-05

Realizzazione di una Infrastruttura di reti di sensori per il monitoraggio dell'ambiente e della persona

Marco Moscatelli[1], Ittalo Pezzotti Escobar[1], Luciano Milanesi[1], Marco Scodeggio[2], Matteo Gnocchi[1]

[1] CNR-ITB Istituto di Tecnologie Biomediche, [2] CNR-IBFM Istituto di Bioimmagini e Fisiologia Molecolare

Abstract. È stata realizzata un'infrastruttura per il monitoraggio remoto di parametri ambientali e fisiologici della persona e dell'ambiente attraverso l'impiego di reti di sensori i cui dati sono trasmessi mediante un apposito modulo Gateway ad un server remoto per la gestione e l'analisi dei dati basato su tecniche di Big Data. La visualizzazione dei risultati ottenuti è successivamente affidata ad una applicazione web dedicata. Tale infrastruttura consente il monitoraggio in tempo reale non intrusivo di alcuni parametri funzionali delle persone che operano all'interno di ambienti specifici.

Keywords. Internet of Things, Interoperabilità dei dati, Big Data

Introduzione

In un contesto dove l'evoluzione della tecnologia internet e delle reti è orientato alla fornitura di servizi specifici per il miglioramento della qualità della vita, la necessità di trovare soluzioni adeguate in grado di adattarsi progressivamente all'ambiente e alla persona sta diventando uno degli obiettivi cruciali nel settore IT. In quest'ambito l'Internet of Things (IoT) (Internet of Things Consortium 2017), si sta affermando come paradigma tecnologico di riferimento poiché ogni dispositivo elettronico è in grado di interconnettersi e comunicare con altri "oggetti" aventi differenti scopi portando così alla creazione e scambio d'informazioni eterogenee. Queste tecniche applicate su numero elevato di apparecchiature e soggetti necessità la raccolta di una grande quantità di dati non gestibile con le tradizionali tecniche informatiche; per questo motivo sono state introdotte tecnologie del tipo "Big Data" (Big Data Definition 2017). In quest'ambito, è in fase di sviluppo un'infrastruttura basata sull'utilizzo di una rete di sensori (Rahmani et al 2015) e metodiche di analisi per il monitoraggio dell'ambiente e di alcuni parametri funzionali per le persone (Dimitrov 2016) che operano all'interno dello stesso (Figura 1).

1. Metodi

L'infrastruttura implementata è suddivisa in 4 componenti (Figura 2):

- rete di sensori: apparati in grado di effettuare differenti tipologie di misurazioni ed eseguire azioni definite;
- modulo gateway: si occupa del dialogo tra la rete di sensori e il server di gestione dei dati;

- server gestione dati: gestione di Big Data che archivia e analizza tutte le misurazioni effettuate dalla rete di sensori e dialoga con l'applicazione utente;
- applicazione utente: software per la gestione e visualizzazione dei dati rilevati;

Figura 1
Sensori ambientali per la persona

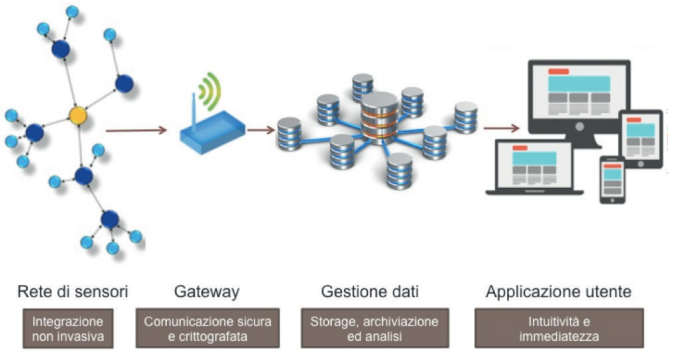


Figura 2
Componenti dell'infrastruttura implementata

La rete di sensori implementata utilizza il protocollo ZigBee (ZIGBEE Alliance 2017) per la comunicazione interna alla rete di sensori e il protocollo ModBus TCP-IP (MODBUS organization 2017) per la comunicazione esterna tra i sensori e il modulo gateway. La struttura della rete è di tipo mesh (Sun and Zhang 2009) dove è presente un coordinatore (che agisce come master ed è in grado di comunicare con tutti gli end device), dei router (che trasmettono i dati da e verso altri dispositivi) e gli end device (che acquisiscono i dati). Per rendere il sistema versatile e adattabile alle diverse tipologie di sensori è stato sviluppato un modulo Gateway con il compito di interfacciarsi alla rete di sensori e dialogare con il server di gestione dei dati tramite delle RESTful API (Rodriguez A 2015) che permettono di inviare tutti i dati campionati e ottenere la lista delle operazioni sulla rete richieste dall'utente.

Il server di gestione dati si occupa di strutturare, archiviare e analizzare tutte le misurazioni effettuate dalla rete di sensori ed è composto da due moduli principali:

- modulo database: responsabile dell'archiviazione dei dati;
- modulo RESTful API: responsabile dell'interazione con il modulo gateway e l' applica-

zione utente.

Il sistema di sensori esegue costantemente un elevato numero di campionamenti da fonti eterogenee producendo una considerevole quantità di dati gestiti mediante MongoDB (MongoDB 2017). Particolare attenzione è stata attribuita alla gestione dei dati prodotti per ciascun sensore in modo da ottimizzare i tempi di lettura e scrittura. Per l'inserimento e l'utilizzo dei dati sono state sviluppate delle apposite API che consentono al modulo gateway e all'applicazione utente di comunicare con il database.

Al fine di poter gestire la rete di sensori e accedere ai dati memorizzati dal server di gestione dei dati è stata sviluppata una specifica applicazione web mediante il framework Angular (Angular 2017) versione 4 (Figura 3). La progettazione e lo sviluppo dell'applicazione e delle relative API sono basati sui principi di semplicità, adattabilità, multilingua, performance, funzionalità ed estensibilità.

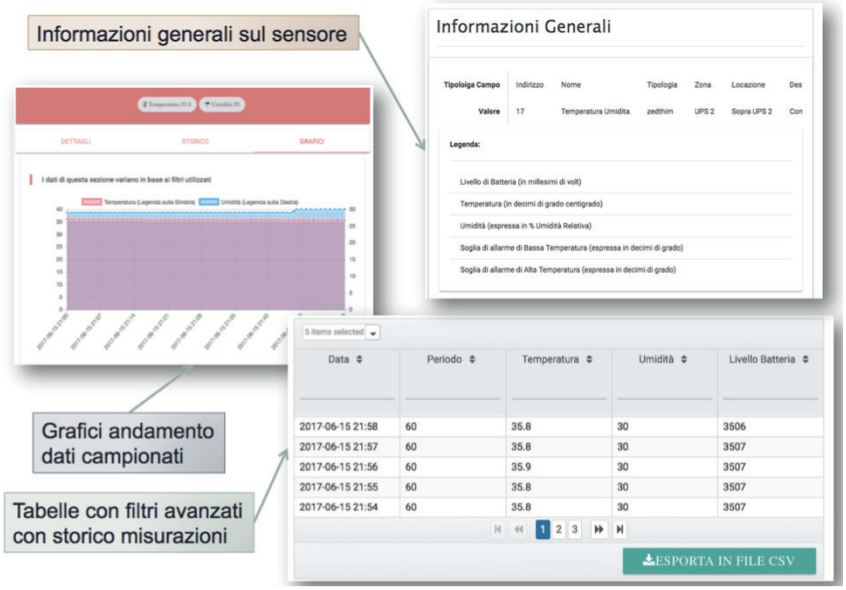


Figura 3
Interfacce per la gestione del sensore in Angular

2. Applicazioni

La piattaforma presentata può essere impiegata in un ampio numero di applicazioni progettuali. In particolare; nel contesto del progetto promosso dal MIUR PON "OPLON Care & Cure" (Oplon 2017), il quale si propone di studiare e monitorare la fragilità dell'anziano, che si manifesta prima di una non-autosufficienza conclamata con elevati costi economici e sociali; sono in fase di realizzazione le seguenti azioni:

- archiviazione e analisi dei dati per lo sviluppo di uno strumento predittivo del rischio di declino funzionale dell'anziano;
- standardizzare delle procedure di raccolta dati;
- utilizzo di API e tecnologie legate ai Big Data;
- servizio di Monitoraggio;
- accesso ai dati mediante GUI;

• sviluppo di un sistema capace di intervenire autonomamente in specifici casi.

Infine, per questo ambito particolare è in fase di sviluppo una applicazione per Android con interfaccia semplificata per il monitoraggio della rete così da garantire una consultazione immediata dei dati (Figura 4).

Particolare attenzione è stata data alla gestione della sicurezza e comunicazione tra i componenti dell'infrastruttura mediante connessione protetta e crittografata all'interno della quale ogni utente è tracciato tramite un identificativo univoco.

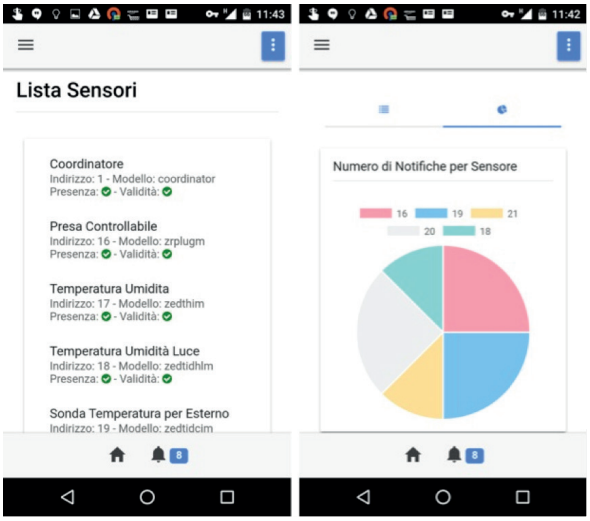


Figura 4
Schermate applicazione Android relative al monitoraggio di un locale

3. Conclusioni

L'infrastruttura descritta consente il monitoraggio in tempo reale di uno o più ambienti e delle persone che operano all'interno degli stessi; l'utilizzo delle tecnologie descritte consente di individuare eventuali parametri fuori norma permettendo di intervenire immediatamente qualora venga riscontrato un problema.

La suddivisione in moduli specifici per ogni servizio ne permette l'installazione e utilizzo in ambienti eterogenei richiedendo interventi minimi di adeguamento infrastrutturale. E' in fase di sviluppo l'integrazione di uno spazio private-cloud dedicato e l'integrazione di nuovi sensori legati all'uomo (biosensori) come ad esempio la misurazione del PH della saliva o della frequenza cardiaca.

Riferimenti bibliografici

Angular 2017: https://angular.io

Big Data Definition 2017: https://www.oracle.com/big-data

Dimitrov, D. V. 2016. Medical Internet of Things and Big Data in Healthca-re. Healthcare Informatics Research, 22(3), 156–163.

Internet of Things Consortium 2017: http://iofthings.org

MODBUS organization 2017: http://www.modbus.org

MongoDB 2017: https://www.mongodb.com

Oplon 2017, MIUR PON "OPLON Care & Cure": http://www.oplon.eu

Rahmani AM, Thanigaivelan NK, Gia TN, Granados J, Negash B, Liljeberg P and Tenhunen H. 2015, "Smart e-Health Gateway: Bringing intelligence to Internet-of-Things based ubiquitous healthcare systems," 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), pp. 826-834. Las Vegas, NV, 2015

Rodriguez A 2015, "RESTful Web services: The basics": https://www.ibm.com/developerworks/library/ws-restful

Sun J. and Zhang X. 2009, "Study of ZigBee Wireless Mesh Networks," 2009 Ninth International Conference on Hybrid Intelligent Systems, pp. 264-267, Shenyang

ZIGBEE Alliance 2017: http://www.zigbee.org

Autori

Marco Moscatelli - marco.moscatelli@itb.cnr.it

Laureato in Bioinformatica, lavoro presso l'istituto di tecnologie biomediche del CNR sia come ricercatore nell'ambito dello studio e analisi dei Big Data sia come sistemista in quanto gestisce l'infrastruttura del gruppo di Bioinformatica prestando attenzione all'efficienza e dinamicità dell'infrastruttura.

Ittalo Pezzotti Escobar - ittalo.pezzotti@itb.cnr.it

Nato in Colombia; con una laurea in ingegneria, ha la passione per la tecnologia dei sistemi hardware e software. Negli ultimi 10 anni ha completato i suoi studi con un dottorato in ingegneria dei sistemi sensoriali partecipando a diversi progetti europei nell'ambito di applicazioni agro-alimentari e ambientale pubblicando lavori su sensori e biosensori.

Luciano Milanesi - luciano.milanesi@itb.cnr.it

Laureato in Fisica. Direttore di ricerca presso il CNR Istituto di Tecnologie Biomediche. Ha partecipato a diversi progetti nazionali ed internazionali. E' autore di più di 350 pubblicazioni nel campo della bioinformatica, Systems Biology e Informatica medica.

Marco Scodeggio - marco.scodeggio@ibfm.cnr.it

Responsabile dell'Area di Ricerca Milano 4 del CNR e della FabLab nata per supportare i gruppi di ricerca degli Istituti del CNR insediati presso l'Area stessa.

Matteo Gnocchi - matteo.gnocchi@itb.cnr.it

Laureato in Tecnologie dell'Informazione e della Comunicazione; dal 2012 lavora come tecnologo presso l'Istituto di Tecnologie Biomediche del CNR in qualità di responsabile delle infrastrutture Web utilizzate in molteplici progetti. Dal 2014 collabora con il nodo italiano (BBMRI.it) dell'infrastruttura europea BBMRI-ERIC. Dal 2016 si occupa della fruizione di tecnologie legate al mondo IoT e Big Data.

DOI: 10.26314/GARR-Conf17-proceedings-06

Piattaforma per l'epidemiosorveglianza e per la categorizzazione del rischio nelle filiere produttive

Giorgio Bontempi, Federico Scali, Giovanni Loris Alborali

Istituto Zooprofilattico Sperimentale Lombardia Emilia Romagna "Bruno Ubertini"

Abstract. Il sistema sviluppato per l'epidemiosorveglianza e classificazione del rischio ha come obiettivi la stima dell'esposizione degli animali agli antimicrobici e la verifica e valutazione dell'efficacia dell'applicazione dei mezzi per ridurre e razionalizzare l'utilizzo degli antimicrobici in azienda: biosicurezza, benessere; e si fonda su quattro vincoli architetturali ben definiti e precisi:

- l'utilizzo di data analytics;
- una stretta integrazione con i sistemi informatici esistenti;
- l'implementazione di algoritmi di calcolo a valenza scientifica riconosciuti a livello internazionale;
- il divieto di duplicazione di funzionalità di elaborazione tra i sistemi informatici esistenti.

Nel corso del contributo sarà illustrato il sistema, le sue caratteristiche, ma soprattutto la sua capacità di estendere gli ambiti di applicazione senza dover subire modifiche.

L'impiego di data analytics e la stretta interoperabilità con i sistemi informatici esistenti permette al sistema di acquisire sempre un maggior numero informazioni attraverso le quali sarà possibile riconoscere e definire le relazioni tra ambito umano e animale nel campo dell'antibiotico resistenza.

Keywords. Big Data, Data analytics, interoperabilità, antibiotico resistenza

Introduzione

La piattaforma per l'epidemiosorveglianza e per la categorizzazione del rischio nelle filiere produttive (monitoraggio di benessere animale, biosicurezza, consumo antimicrobici, antibiotico resistenza, patologie, macello) è parte integrante dei Servizi Veterinari Nazionali e si pone come strumento di elaborazione di informazioni raccolte da sistemi informativi esistenti o da indagini direttamente sul campo finalizzato al monitoraggio di eventi e delle azioni intraprese per valutare i rischi e porre in essere le conseguenti azioni correttive.

Al momento la piattaforma monitora:

- le condizioni di benessere e biosicurezza negli allevamenti;
- il consumo di antimicrobici misurati utilizzando standard basati sulle DDD (defined daily dose) e sulle DCD (defined course dose);
- le lesioni e le patologie degli animali rilevate al macello;
- l'efficacia dei mezzi o delle misure adottate per ridurre e razionalizzare l'utilizzo degli antimicrobici in azienda: biosicurezza, benessere, etc.

Il sistema è stato presentato il 5 ottobre 2017 a Roma presso il Ministero della Salute durante il G7 Chief Veterinary Officers Meeting riscuotendo unanimi apprezzamenti dai partecipanti.

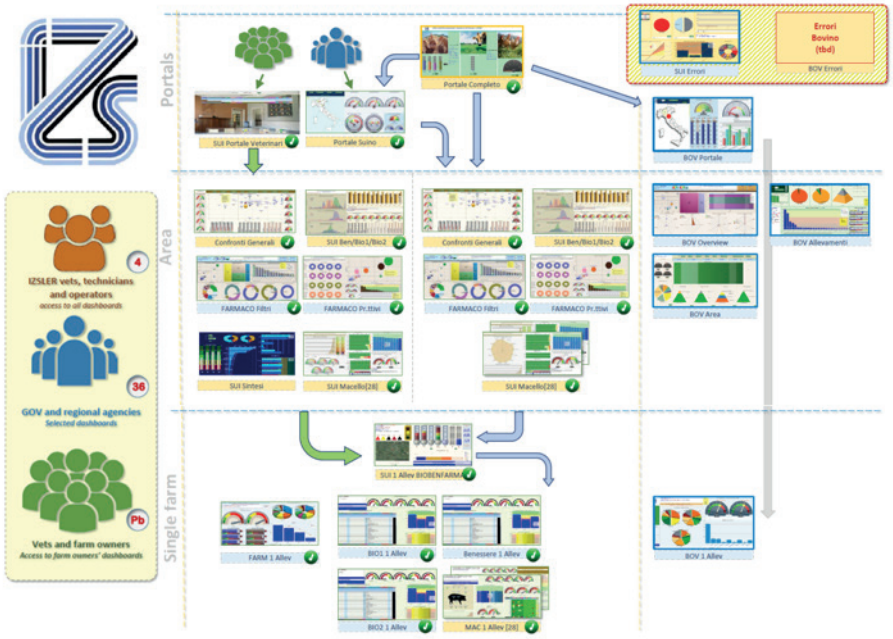
Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

1. Architettura del sistema

Il sistema (Figura 1) è composto da:

1. uno strumento web (iDashboards) per la generazione e consultazione delle dashboards con cui si monitora la categorizzazione del rischio;
2. un contenitore di dati (data lake) provenienti dai sistemi ministeriali esistenti (Banca Dati Nazionale, prontuario farmaceutico, centri di referenza internazionali/nazionali,...) arricchiti con le informazioni raccolte direttamente sul campo (biosicurezza, benessere ...) o da latri sistemi informatici esistenti nelle varie realtà territoriali;
3. un insieme di modelli di machine learning in grado di armonizzare i dati raccolti nel data lake, per analizzare la qualità del dato raccolto, per individuare relazioni tra gli stessi al fine di avere una comprensione territoriale e temporale del rischio nella filiera suinicola;
4. un insieme di applicativi sviluppati in ambiente mobile per la raccolta dei dati destinati a supplire ad eventuali assenze di software per l'acquisizione dei dati presso allevamenti.

Figura 1
Modello del sistema



Il Sistema opera acquisendo i dati presenti nelle Banche Dati Nazionali (anagrafe aziende zootecniche, prontuario farmaceutico, centri di referenza internazionali/nazionali) e arricchendoli con informazioni:

- raccolte direttamente sul campo attraverso l'utilizzo di dispositivi mobili:
  - benessere animale (compilazione check list);
  - biosicurezza (compilazione di check list). Utilizzo dello strumento Biocheck.UGent® secondo l'algoritmo di ponderazione definito dall'Università di Gent (Belgio) e riconosciuto a livello internazionale;

- stima dell'esposizione degli animali agli antimicrobici secondo degli standard di misurazione ponderati sui singoli principi attivi (DDD/DCD);
- gli esiti degli esami di laboratorio sui campioni prelevati presso le aziende zootecniche (PRRS, antibiotico resistenza, ...);
- rilevazione delle lesioni e delle alterazioni ante- e post-mortem nei macelli;

• acquisite mediante meccanismi di interoperabilità con i sistemi informatici esistenti e in uso presso ASL/ATS, ...; tra queste sinergie con sistemi informatici esistenti si sottolinea l'integrazione con la ricetta elettronica veterinaria che nel corso del 2018 sarà utilizzata sull'intero territorio nazionale sostituendo l'attuale gestione cartacea;

Il Sistema consente agli operatori abilitati non solo di monitorare la situazione in tempo reale o di verificare l'efficacia delle azioni applicate per la riduzione del consumo degli antibiotici veterinari, ma anche d'identificare il migliore mix di azioni atte a contrastare il fenomeno dell'antibiotico resistenza in ambito animale.

Le informazioni raccolte sul campo e nei database dei Servizi Veterinari vengono elaborate e visualizzate, secondo processi ben strutturati (Figura 2), attraverso dei cruscotti interattivi ("dashboard") intuitivi che consentono di navigare facilmente e velocemente dall'elaborazione di sintesi al dettaglio.

Dettaglio che attraverso questo processo è stato validato non solo da un punto di vista formale, ma anche in relazione al contesto in cui è stato raccolto e in cui l'azienda opera. La qualità del dato acquisito, frutto di processi automatici di validazione, è uno dei principali pilastri su cui si poggia l'intero Sistema. Questa attenzione sistematica alla correttezza di quanto acquisito dalle base dati esistenti è resa possibile dall'utilizzo di strumenti di data analitycs e di dashboards interattive.

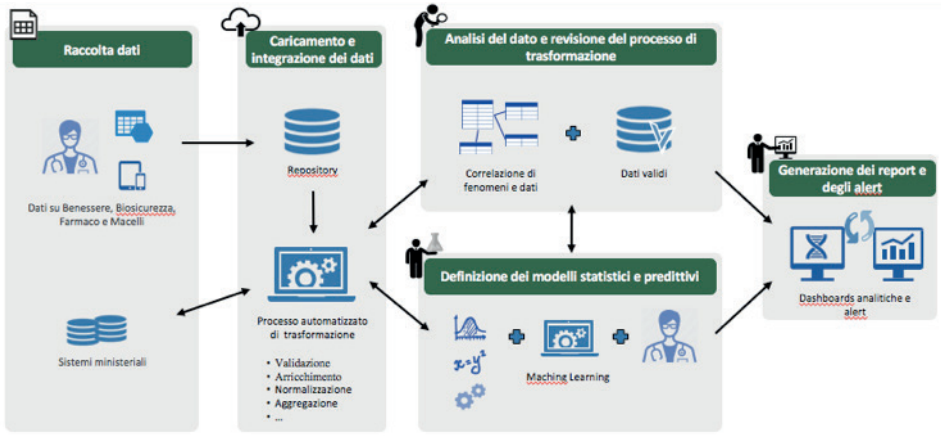


Figura 2
Macro-processi del sistema

Ai dati acquisiti vengono applicati algoritmi di "machine learning" per:

• verificare la qualità del dato raccolto, potendo in automatico introdurre e gestire livelli di confidenza di affidabilità;
• analizzare i dati per arricchirli di significato utile: l'obiettivo non è generare report su ciò che è accaduto individuare dalle informazioni raccolte le azioni più efficaci per ridurre l'utilizzo del farmaco in ambito animale.

Le elaborazioni prodotte dal sistema di data analytics sono consultabili dall'utente attraverso appositi cruscotti che permettendo di partire da un'aggregazione a livello nazionale o territoriale (Figura 3) per potere poi consultare, con pochi clic del mouse, i dati di sintesi di una singola azienda (Figura 4).

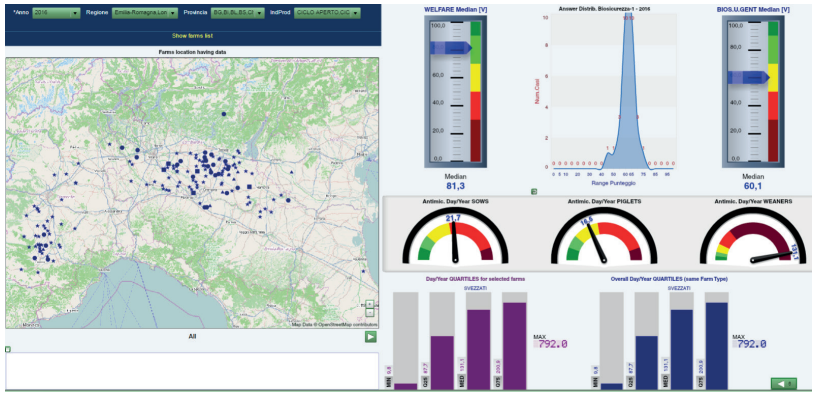


Figura 3
Situazione territoriale

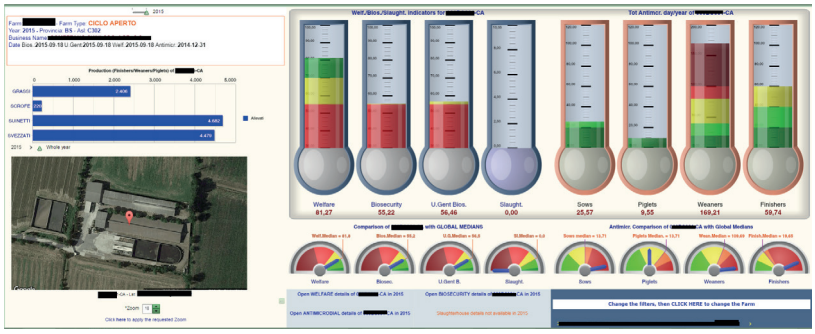


Figura 4
Dettaglio azienda

Sulla medesima azienda è possibile inoltre andare nel dettaglio del consumo del farmaco (Figura 5), verificando la posizione dell'azienda a livello nazionale, regionale e territoriale, con la sua valutazione in quartili.

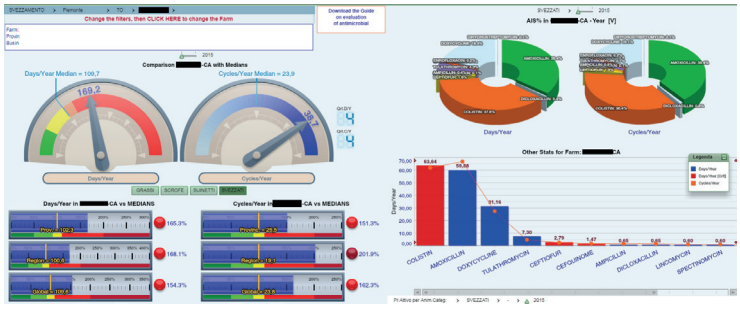


Figura 5
Dettaglio consumo del farmaco

Si sottolinea la disponibilità d'informazioni che permettono di analizzare sia il consumo globale di antimicrobici sia quello dei singoli principi attivi impiegati; informazione, quest'ultima, fondamentale nella lotta all'antibiotico resistenza. Analogo dettaglio sulla singola azienda è disponibile per benessere e per biosicurezza (Figura 6) arrivando fino

alla visualizzazione delle singole risposte alle check list.

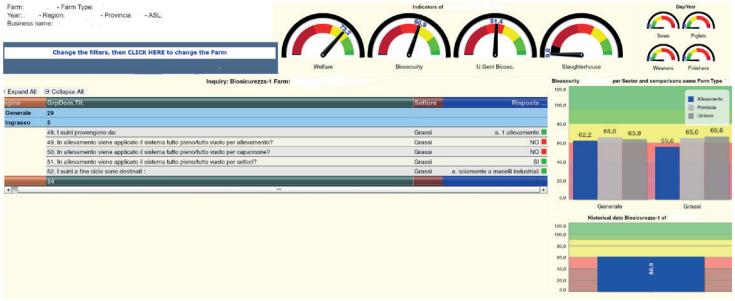


Figura 6
Dettaglio biosicurezza

2. Conclusioni

L'applicazione dei data analytics in ambito sanitario veterinario e l'implementazione di meccanismi di interoperabilità con i sistemi informatici esistenti ha consentito lo sviluppo e messa in produzione di un sistema di monitoraggio dell'esposizione degli animali agli antimicrobici in tempi rapidi. Il sistema attraverso l'utilizzo di algoritmi predittivi è in grado di analizzare e valutare l'efficacia delle azioni messe in campo per la riduzione dell'utilizzo degli antibiotici presso le aziende zootecniche.

L'utilizzo di tecnologie data analytics ha permesso di rendere disponibili e accessibili i dati memorizzati nei silos dei vari attori coinvolti a tutti i portatori di interesse. Si è pertanto realizzato quanto indicato nel piano triennale pere l'informatica nella pubblica amministrazione. Inoltre il sistema sta ampliando la rete dei dati acquisiti al fine di consentire l'individuazione e valutazione delle relazioni tra l'esposizione degli animali agli antimicrobici e i fenomeni di antibiotico resistenza che sempre di più si stanno manifestando in ambito umano.

Autori

Giorgio Bontempi - giorgio.bontempi@izsler.it
Giorgio Bontempi è ingegnere informatico: si è laureato all'Università degli Studi di Milano in Scienze dell'Informazione. È dirigente responsabile dei Sistemi Informativi dell'Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini", con il quale ha realizzato progetti a carattere nazionale e internazionale.

Federico Scali - federico.scali@izsler.it
Laureato in medicina veterinaria presso l'Università degli Studi di Milano. Dottorato di ricerca in Igiene Veterinaria e Patologia Animale. Dal 2015 collabora con l'Istituto Zooprofilattico Sperimentale Lombardia Emilia-Romagna "Bruno Ubertini".

Giovanni Loris Alborali - giovanni.alborali@izsler.it
Laureato in Medicina Veterinaria presso l'Università degli Studi Milano. È dirigente responsabile della Sezione Diagnostica di Brescia dell'Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna "Bruno Ubertini". È diplomato al College europeo ECPHM e Presidente della (SIPAS) Società di Patologia ed Allevamento suino.

DOI: 10.26314/GARR-Conf17-proceedings-07

Analisi, valutazione del rischio e sicurezza informatica di dati e informazioni dei dispositivi medici connessi alle reti IT-medicali

Catello Chierchia[1], Enrico Guerra[2], Martina Balloccu[2], Lorenzo Monasta[3], Francesca Deluca[1], Michele Bava[1,2]

[1]Ufficio Sistema Informativo – SC Ingegneria Clinica, Informatica e Approvvigionamenti – IRCCS Burlo Garofolo di Trieste, [2]DIA – Università degli Studi di Trieste, [3]SSD Epidemiologia e Statistica – IRCCS Burlo Garofolo di Trieste

Abstract. Questo lavoro ha come scopo la realizzazione di un modello per l'analisi e la gestione del rischio dei dispositivi medici (DM) collegati a una rete IT-medicale. Il modello determina un Indice di Valutazione del Rischio (IVR) che è funzione di alcuni fattori di rischio preventivamente selezionati, calcolando i relativi pesi ottenuti da due modelli statistici: la regressione lineare e il modello logistico.
I fattori di rischio si concentrano principalmente sulla parte informatica che negli ultimi anni sta ricoprendo un ruolo fondamentale nell'ambito della sanità, dove occorrono misure adeguate di sicurezza per proteggere reti ospedaliere, sistemi clinici, dati e informazioni.

Keywords. Rete IT-Medicale, Sicurezza informatica, Analisi del rischio, Dispositivi Medici, Metodi statistici

Introduzione

In un mondo che si avvia verso la completa informatizzazione e digitalizzazione di dati, informazioni e conoscenza, anche il settore ospedaliero e il mondo sanitario in generale si trovano ad affrontare nuove opportunità e sfide che hanno a che fare con la gestione dei dati e la sicurezza informatica. Nella società attuale infatti possiamo considerare la protezione dei dati e le informazioni che essi generano (*data and system security*), come "il nuovo petrolio" (Simi 2016), cambiando e innovando il modo di produrre ricchezza. La mancanza d'idonei strumenti di protezione di una rete dati ospedaliera quindi può portare non solo al danneggiamento del dispositivo che ha generato le informazioni, ma anche rendere la rete stessa facile preda di *hackers* malintenzionati, che realizzano gran parte del loro fatturato con la sottrazione e la violazione di dati di aziende impreparate ad affrontare efficacemente la minaccia (CINI 2017).

È quindi fondamentale, e auspicabile, che gli enti si tutelino con misure di sicurezza adeguate per proteggere capitale, tecnologia e conoscenza, in particolar modo i sistemi e i servizi che trattano dati sensibili, attraverso investimenti sulla messa in sicurezza della rete informatica (CLUSIT 2017). Lo scopo è di ottenere un elevato grado di protezione da attacchi esterni e garantire così la continuità operativa della struttura, ma non basta: è necessario infatti che questo grado di protezione sia periodicamente verificato attraverso

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

un'analisi del rischio che produca parametri oggettivi per la valutazione di tutti i sistemi, servizi e le apparecchiature collegate alla rete IT-medicale (Cacciari et al. 2015).

1. Metodo

Nel progetto proposto, partendo da quanto acquisito in ambito legislativo (D.Lgs. 196/03, nuovo Regolamento Privacy 679/2016), e normativo (ISO 27001, ISO 80001 e ISO 30001) si attribuisce, a ogni apparecchiatura o dispositivo medico (DM) collegato a una rete ospedaliera, un indice, l'Indice di Valutazione del Rischio (IVR), che valuti il relativo livello di sicurezza nelle condizioni d'uso tipiche. L'IVR è ottenuto attraverso l'implementazione di metodi statistici e una stima dei pesi oggettiva che renda il modello ripetibile e quindi convalidabile.

L'IVR è distribuito in un range da 1 (basso rischio) a 10 (alto rischio) e suddiviso in macrocategorie che tengano conto sia delle tematiche tipiche dell'ingegneria clinica (la documentazione e la manutenzione delle apparecchiature, i rischi collegati al paziente) sia di aspetti ICT solitamente trascurati nell'analisi del rischio delle tecnologie biomediche. Attraverso l'assegnazione di una serie di regressori è realizzata la formula per il calcolo dell'IVR relativo al rischio rilevabile sulla singola apparecchiatura nelle condizioni di esercizio.

In particolare per la parte relativa alla sicurezza informatica si è scelto di considerare come regressori la presenza/assenza di credenziali di accesso per accedere al sistema, se è presente ed è aggiornato l'antivirus, se è stato effettuato il backup dei dati, se è avvenuta perdita dei dati, se è attivo il firewall, se il dispositivo è sotto gruppo di continuità e se risulta positivo ai test di vulnerabilità; tutti argomenti che riguardano la privacy, l'information security e la cybersecurity (Bava et al. 2009).

Per ricavare i pesi di ciascuna categoria, nel nostro studio sono stati utilizzati due modelli allo scopo di confrontarne i risultati: la regressione lineare multipla e il modello logi-

DOCUMENTAZIONE E MANUTENZIONE				
X1 Documentazione completa	**X2 Controlli e verifiche effettuati periodicamente**	**X3 Disponibilità ditta**	**X4 Costo di manutenzione**	**X5 Disponibilità muletti**
Si=0	Si=0	Si=0	Sotto contratto=0	Si=0
No=1	No=1	No=1	Nessun contratto=1	No=1

RISCHIO PER IL PAZIENTE			
Y1 Funzione apparecchiatura	**Y2 Conseguenze per il paziente**	**Y3 Età del dispositivo**	**Y4 Frequenza d'utilizzo**
Altro=2	Nessun rischio =1	Minore di 8 anni=0	Annuale=1
Analisi=3	Terapia inappropriata=2	Maggiore di 8 anni=1	Mensile=2
Diagnostica=4	Danno=3		Settimanale=3
Terapeutica=5	Morte=4		Giornalireo=4

SICUREZZA INFORMATICA						
Z1 Credenziali di accesso al sistema	**Z2 Antivirus**	**Z3 Backup**	**Z4 Perdita dei dati**	**Z5 Test di vulnerabilità**	**Z6 Firewall**	**Z7 UPS**
Si=0	Installato e aggiornato=0	Effettuato=0	No=0	Negativo=0	Attivo=0	Si=0
No=1	Installato e non aggiornato=1	Non effettuato=1	Si=1	Positivo=1	Non attivo=1	No=1
	Non presente=2					

Figura 1
Fattori di rischio relativi alla documentazione, rischio paziente e IT security

stico. Nel primo caso si studia la dipendenza di una variabile quantitativa Y dall'insieme dei regressori $X_1,\ldots,X_m$, $Y = f(X_1, \ldots, X_m) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$ attraverso un modello lineare (Montanari 2015), e una tripartizione del livello di rischio in alto, medio e basso; nel secondo caso invece il modello di regressione è applicato nei casi in cui la variabile dipendente Y può assumere esclusivamente valori dicotomici, in questo caso alto e medio/basso rischio $Y=\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$.

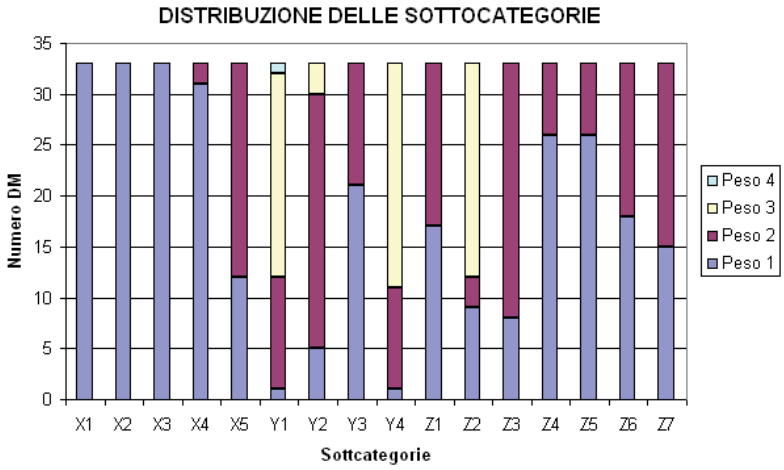


Figura 2
Distribuzione dei DM rispetto le varie sottocategorie/fattori di rischio

2. Risultati

I risultati finora ottenuti evidenziano che entrambi i modelli possono essere presi in considerazione ed essere valutati per la stima dei pesi per le singole categorie e quindi trovare l'IVR dell'apparecchiatura; il modello logistico (procedura Firth) però riesce a simulare me-

MODELLO LOGISTICO PROCEDURA FIRTH			
Predizione Lineare	Medio/Basso Rischio	Alto Rischio	Totale
-10,51835	3	0	3
-5,56893	1	0	1
-4,652917	1	0	1
-4,243236	1	0	1
-2,740673	2	0	2
-2,619888	5	0	5
-2,505077	2	0	2
-1,294194	1	0	1
-1,149973	1	0	1
-1,002514	1	0	1
-0,5046952	2	1	3
0,2083686	2	3	5
1,118652	0	1	1
1,200816	0	1	1
2,411699	0	2	2
3,914261	0	1	1
3,946909	0	1	1
6,863303	0	1	1
Totale	**22**	**11**	**33**

Figura 3
Valore di predizione lineare con procedura Firth e cut-off

glio l'andamento desiderato con una sensibilità del 100% e una specificità dell'82%, identificando tutti i dispositivi ad alto rischio, mentre per il medio/basso rischio non individua quattro macchine, ponendole a un livello di allerta superiore al necessario.

3. Conclusioni

Avendo un modello ripetibile e convalidabile la prospettiva futura è di impiegarlo nelle strutture ospedaliere per fornire una valutazione del rischio realistica e affidabile, con una formula predittiva che permetta l'intervento tempestivo sul DM e sulla rete dati, riducendo così i possibili rischi legati ad attacchi informatici o relativi allo stato delle apparecchiature. Non solo: l'espansione del suo utilizzo a livello territoriale permetterebbe di centralizzare l'archiviazione dei dati relativi ai DM delle strutture ospedaliere di tutta una regione.

L'aggregazione di una grande quantità di questi dati consentirebbe di applicare il modello descritto al dominio dei big data e ottenere così sia risultati dell'IVR sempre più attendibili, ma anche un sensibile miglioramento nell'attività decisionale. Attraverso lo studio di una tale mole di dati si potrebbero analizzare variazioni dell'IVR e trovare i trend che permettano di prevenire i guasti, garantendo un ciclo di vita più lungo delle macchine e/o inviare un alert alle ASL interessate prima che la criticità diventi troppo rilevante.
Gli stessi big data potrebbero produrre differenti IVR relativi a strutture analoghe e questo essere sintomo di una non adeguata manutenzione e controllo delle apparecchiature stesse.

L'utilizzo di reti neurali o altri sistemi d'intelligenza artificiale o machine learning inoltre, permetteranno di valorizzare ulteriormente la bontà del modello, con un supporto immediato e ancora più efficace alle decisioni in fase di analisi e valutazione del rischio sia locale che globale.

Riferimenti bibliografici

Bava M.et al. (Indore, India 23-25 July 2009), "Information Security Risk Assesment in Healthcare: the experience of an Italian Pediatric Hospital", CICSYN 2009, pp. 321-326, IEEE Computer Society

Cacciari D., Zotti D., Sossa E., Bava M. (2015), "Infrastructure Security in Pediatric Hospital: Architectural Evolution, Virtualization and Network Management Systems", Special Issue in Advances in Networks- Secure Network Communications, volume 3, Issue 3-1, pp 23-26, Science Publishing Group 2015

CINI Laboratorio Nazionale di Cybersecurity (2017), "Italian Cybersecurity Report", pag. 3

CLUSIT – Associazione italiana per la sicurezza informatica (2017), "Rapporto CLUSIT 2017 sulla sicurezza ICT in Italia"

Montanari A (2015) "La regressione lineare multipla", pag. 1 http://www2.stat.unibo.it/montanari/Didattica/dispensa2.pdf

Simi L. (2016), "I dati sono il nuovo petrolio"

http://www.pagina99.it/2017/03/16/industria-dei-dati-italia-nuovo-petrolio

Autori

Catello Chierchia - catello.chierchia@burlo.trieste.it

Laureato in Ingegneria Clinica presso l'Università di Trieste, da marzo 2017 è ricercatore presso l'IRCCS Burlo Garofalo per la sicurezza informatica dei dispositivi medici in una rete IT-medicale.

Enrico Guerra - enricoguerra@live.com

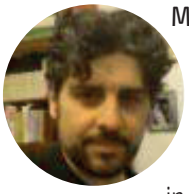
Studente di Ingegneria Clinica presso l'Università di Trieste dove ha coltivato l'interesse per l'informatica sanitaria, sta conducendo la tesi sull'implementazione di un sistema di supporto alla decisione clinica utilizzando IBM Watson.

Martina Balloccu - martina.balloccu@gmail.com

Laureata in Ingegneria Clinica presso l'Università di Trieste ha conseguito precedentemente la laurea triennare in Ingegneria Biomedica presso l'Università degli Studi di Cagliari.

Francesca Deluca - francesca.deluca@burlo.trieste.it

Laureata in ingegneria clinica presso l'Università di Trieste, dal 2016 è impiegata presso l'IRCCS Burlo Garofolo di Trieste dove si occupa di sistemi informativi sanitari all'interno della S.C. Ingegneria Clinica, Informatica e Approvvigionamenti.



Michele Bava - michele.bava@burlo.trieste.it

Laureato in Ingegneria Elettronica, Specialista in Ingegneria Clinica e Informatica Medica, PhD in Ingegneria dell'Informazione lavora dal 2003 presso l'Ufficio Sistema Informativo dell'IRCCS Burlo Garofolo di Trieste e dal 2009 in qualità di Amministratore di Sistema. Negli anni titolare di diversi progetti di ricerca svolge attualmente ricerche nel campo dell'ICT in Sanità, della Telemedicina e della Sicurezza Informatica.

DOI: 10.26314/GARR-Conf17-proceedings-08

Microbial Resource Research Infrastructure: stato e prospettive sull'integrazione dei dati

Paolo Romano[1], Giovanna Cristina Varese[2]

[1]Ospedale Policlinico San Martino, Genoa,
[2]Mycotheca Universitatis Turinensis, Università di Torino

Abstract. In questo manoscritto, viene presentato lo stato dell'infrastruttura Microbial Resource Research Infrastructure (MIRRI), con particolare attenzione alla situazione nazionale e alla realizzazione di un nodo italiano. Sono inoltre presentati lo stato attuale e le prospettive sullo scambio di dati tra collezioni di servizio per i microorganismi (microbial domain Biological Resource Centers, mBRC) e sulla realizzazione di un sistema informativo condiviso, MIRRI-IS, che sia al contempo in grado di offrire un accesso integrato ai cataloghi degli mBRC e di interoperare con sistemi specializzati sull'analisi dei microorganismi e, in generale, con le più rilevanti banche dati di biologia molecolare, nell'ottica di inserire i dati relativi ai microrganismi nel contesto di un ambiente bioinformatico realizzato con un approccio FAIR (Findable, Accessible, Interoperable, Reusable). Infine, viene ipotizzato un coinvolgimento di GARR nella realizzazione di un prototipo a livello nazionale.

Keywords. Microorganismi, infrastruttura di ricerca, integrazione dati, interoperabilità di sistemi

Introduzione

Nell'ambito dell'iniziativa ESFRI è compresa la Microbial Resource Research Infrastructure (MIRRI) (http://www.mirri.org/) che, nella fase preparatoria, ha incluso 16 partner e 28 istituti collaboranti da 19 stati europei. La missione di MIRRI consiste nel superamento della frammentazione esistente nell'offerta di risorse e servizi in ambito microbiologico. La sua azione è focalizzata sulle esigenze, opportunità e sfide poste ai microbial domain Biological Resource Centres (mBRCs) e agli utilizzatori di microorganismi, sia industriali sia del mondo accademico. MIRRI intende offrire un punto di accesso unico ai servizi e alle risorse offerte dai mBRC.

Il raggiungimento degli obiettivi di MIRRI richiede anche una maggiore interoperabilità tra i sistemi informativi dei mBRC e un'accresciuta offerta di dati. Con l'avvento delle tecnologie high-throughput e il conseguente spostamento della ricerca dai dati a livello cellulare a quelli molecolari, è diventato necessario includere nei cataloghi dei mBRC informazioni di sequenze e d'interazione tra molecole. È necessario implementare un'architettura informatica in grado di gestire dati di sequenza, fenotipici e immagini, che sono intrinsecamente "big data".

1. I sistemi informativi degli mBRC

La maggior parte dei mBRC europei propone il proprio catalogo on-line. Esistono pochi esempi di accesso integrato a più cataloghi. Inoltre, i sistemi informativi dei mBRC sono

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

disomogenei per modalità di accesso ed eterogenei nei contenuti e nel formato dei dati, sostanzialmente non in grado di interoperare e lontani dal un approccio FAIR (Findable, Accessible, Interoperable, Reusable).

L'accesso integrato a più cataloghi è possibile tramite Common Access to Biological Information and Resources (CABRI, http://www.cabri.org/) (Romano P et al. 2005), StrainInfo (http://www.straininfo.net/) (Verslyppe B et al. 2014) e il Global Catalogue of Microorganisms (GCMs, http://gcm.wfcc.info/) (Wu L et al. 2013). CABRI consente l'accesso integrato a 25 cataloghi che includono più di 130.000 risorse microbiologiche. La sua implementazione si basa sull'adozione di dataset e formato dati condivisi (http://www.cabri.org/guidelines/catalogue/CPdata.html). L'indicizzazione dei cataloghi e la ricerca dei loro contenuti è effettuata tramite Sequence Retrieval System (SRS), un motore di ricerca per database di biologia molecolare. Gli utenti possono eseguire ricerche diversificate utilizzando l'"Extended Query Form" dell'interfaccia SRS o l'apposito modulo di ricerca "Simple Search".

Il database StrainInfo comprende alcuni metadati estratti dai cataloghi dei mBRC. L'analisi e integrazione di questi dati consente l'identificazione dei ceppi presenti nelle diverse collezioni, ma originate dallo stesso ceppo iniziale, e permette così di creare una sintesi del contenuto dei cataloghi centrata su ceppi identici. Da questa sintesi ("strain passport") sono resi disponibili link alle collezioni e a database esterni in grado di fornire informazioni estese su tassonomia, sequenze, riferimenti bibliografici e altro. StrainInfo comprende dati su ca. 300.000 ceppi presenti in più di 60 cataloghi con ca. 700.000 numeri di collezione diversi (http://www.straininfo.net/stats).

GCM è il database che include il maggior numero di cataloghi di mBRC. Alcune delle sue caratteristiche, però, non vanno incontro alle esigenze degli utilizzatori. Ad esempio, i dati dei cataloghi devono essere trasferiti manualmente in GCM e questo comporta che molti cataloghi non siano aggiornati. Inoltre, le possibilità di ricerca nel database sono limitate. È necessario quindi realizzare un sistema informativo che superi le limitazioni di CABRI, StrainInfo e GCM. Il nuovo sistema informativo deve allo stesso tempo includere dataset più estesi, ad esempio alle sequenze, e nuove funzionalità che siano in grado di eseguire alcune analisi di tipo bioinformatico.

2. L'architettura proposta e alcuni risultati preliminari

Nel corso della fase preparatoria di MIRRI, è stata definita un'architettura informatica per il MIRRI Information System (Smith D et al. 2017) (figura 1). L'architettura prevede le seguenti componenti:

- un formato standard per lo scambio di dati tra mBRC, sviluppato a partire dal Microbiological Common Language (MCL) di StrainInfo (Verslyppe B et al. 2010),
- un dataset minimo per i dati essenziali di ogni ceppo disponibile, destinato ad evolvere nel tempo sino a comprendere ogni informazione potenzialmente interessante per valutare le applicazioni dei singoli ceppi, da definire come Minimum Information about Biological Resources (MIaBRe),
- un'interfaccia "user-friendly" per le ricerche da inglobare all'interno di un Collaborative Working Environment (CWE),

- opportune Application Programming Interfaces (APIs) e Web Services / automatic workflow per i più diffusi e adottati software di integrazione (come Galaxy (Goecks J et al. 2010) e Taverna (Wolstencroft K et al. 2013).

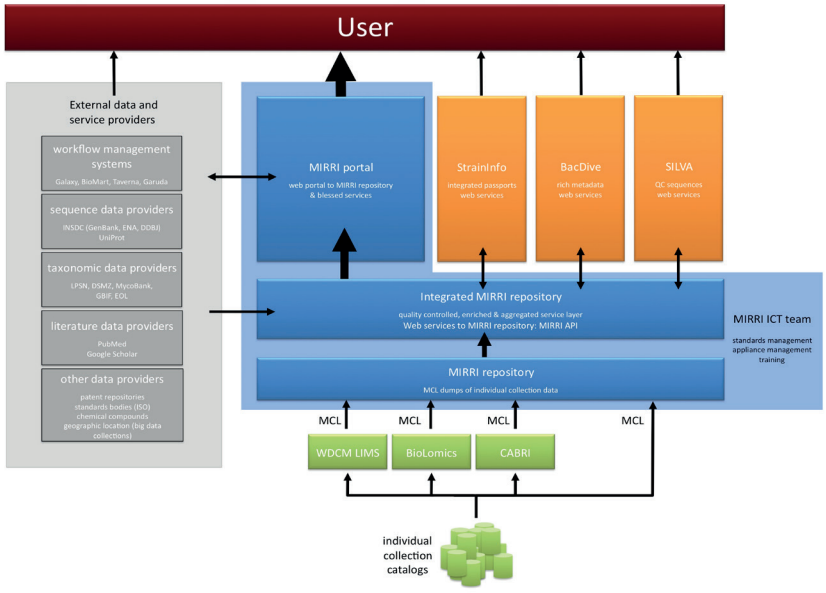


Figura 1
Possibile architettura per il MIRRI Information System

Per valutare la fattibilità delle scelte architetturali, tre diversi prototipi, chiamati "MIRRI demonstrators", sono stati sviluppati durante la fase preparatoria di MIRRI. I tre prototipi affrontano aspetti diversi, ma ugualmente rilevanti, del futuro sistema informativo.

Il prototipo legato al Bacterial Diversity Metadatabase (BacDive, https://bacdive.dsmz.de/) (Söhngen C et al. 2015) mira a estendere i contenuti dei cataloghi e a migliorarne la qualità dei dati. Si tratta di un lavoro impegnativo che ha portato a descrivere in maniera qualitativamente elevata un numero limitato di ceppi. Lo sforzo necessario a descrivere con precisione i ceppi è però superiore a quello che un collezione può normalmente permettersi. Il prototipo ha comunque consentito di identificare e caratterizzare molte informazioni utili e dimostra come l'estensione dei contenuti dei cataloghi sia possibile e si possa raggiungere progressivamente, selezionando ambiti d'interesse specifico e concentrandosi sui ceppi di maggior interesse, portando a collezioni più specializzate, con un numero minore di ceppi, ma appropriatamente caratterizzati e descritti.

Il prototipo di StrainInfo mira a ottenere una migliore integrazione dei contenuti delle collezioni tramite una precisa identificazione dei ceppi comuni, in possesso di più collezioni. Rende così possibile sia la riorganizzazione e focalizzazione delle collezioni, sia lo scambio di dati e l'arricchimento della descrizione dei ceppi comuni.

L'USMI Galaxy Demonstrator (UGD, http://bioinformatics.hsanmartino.it:8080/), è un server Galaxy per i curatori e utenti dei cataloghi dei mBRC ed è mirato a facilitare la gestione dei dati di catalogo ("data curation") e l'integrazione nei cataloghi di dati estratti da database esterni (Colobraro DP and Romano P. 2015). Questo server consente anche di accedere e

utilizzare terminologie e formati standard, migliorando così la qualità dei dati contenuti nel catalogo. UGD consente quindi l'integrazione dei dati dei cataloghi con informazioni di altri database sfruttando un software molto diffuso e senza la necessità di sviluppi informatici.

3. Esigenze e prospettive a livello nazionale

L'effettiva realizzazione di MIRRI dipende, come per ogni infrastruttura ESFRI, dall'attivo coinvolgimento degli stati interessati a sostenerla, che sono chiamati a sostenere sia il nodo centrale, con funzioni di coordinamento, sia le esigenze della comunità scientifica e industriale nazionale. Mentre ogni accordo per il nodo centrale viene stabilito nel contesto europeo, le forme e i modi del sostegno nazionale possono variare a seconda della specifica situazione nazionale.

In Italia esistono molte collezioni di ceppi microbici, ma poche che abbiano una chiara e definita missione di servizio. Inoltre, il coordinamento tra collezioni è molto limitato. La creazione di una rete tra le collezioni, di servizio e non, ha comunque una grossa ricaduta potenziale in vari settori, quali le diverse declinazioni delle biotecnologie, sia nell'ambito della ricerca sia in quello industriale, la salute e l'ambiente. Siamo quindi convinti che il governo italiano dovrebbe sostenere la creazione di una rete nazionale di mBRC.

Recentemente, è stata costituita una Joint Research Unit per la creazione di un nodo italiano di MIRRI (MIRRI-IT). Ad essa hanno già aderito gli enti che hanno partecipato alla fase preparatoria di MIRRI: le Università di Torino, Perugia e Modena e Reggio Emilia, l'Ospedale Policlinico San Martino e il CNR. Lo sviluppo di una stretta connessione tra i mBRC italiani è uno degli obiettivi importanti della JRU, insieme all'implementazione di un sistema informativo per l'accesso integrato ai dati relativi ai microorganismi conservati presso le collezioni, da sviluppare secondo l'architettura di MIRRI-IS. Per questo obiettivo sono richieste notevoli risorse IT. Data la natura pubblica della JRU e dei suoi aderenti, tutti membri del Consortium GARR, il partner ideale è la rete GARR.

4. Conclusioni

Abbiamo presentato in sintesi gli obiettivi e lo stato attuale di MIRRI, un'infrastruttura di ricerca sulle risorse microbiologiche in via di sviluppo nel contesto di ESFRI. Uno degli obiettivi fondamentali di MIRRI è la realizzazione di un sistema informativo flessibile in grado di integrare i dati delle collezioni partecipanti e di numerosi altri database e software bioinformatici per offrire agli utilizzatori di ceppi microbici un ambiente informativo ricco di dati e applicazioni e di facile utilizzo. Abbiamo quindi presentato l'architettura informatica ipotizzata per la realizzazione di MIRRI-IS e alcuni prototipi software che hanno dimostrato la fattibilità delle sue principali componenti. Infine, abbiamo presentato la situazione nazionale, evidenziando come sia ipotizzabile la realizzazione di una piattaforma simile a MIRRI-IS per le esigenze nazionali. Tale piattaforma potrebbe logicamente e utilmente essere implementata nel contesto della rete GARR.

Riferimenti bibliografici

Colobraro DP, Romano P. (2015) A Galaxy approach to microbial data integration: the

USMI Galaxy Demonstrator. Conference Proceedings 12th BITS Annual Meeting, June 3-5, 2015, Milano, Italy. Milanesi L, Mauri G, Masseroli M (Eds). BUP - Bononia University Press SpA, Bologna Italy, 2016, pp. 43-45.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible and transparent computational research in life sciences. Genome Biology, 11:R86

Romano P, Kracht M, Manniello MA, Stegehuis G, Fritze D. (2005) The role of informatics in the coordinated management of biological resources collections. Applied bioinformatics 4 (3), 175-186.

Smith D, Stackebrandt E, Casaregola S, Romano P, Glöckner FO. (2017) MIRRI Recommendations for Exploiting the Full Potential of Micro-Organism Data. Ann Biom Biostat 4(1):1027.

Söhngen C, Bunk B, Podstawka A, Gleim D, Vetcininova A, Reimer LC, Overmann J. (2015) BacDive - the Bacterial Diversity Metadatabase. Nucleic Acids Res. 2015.

Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P. (2010) Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Research in microbiology 161.6:439-445

Verslyppe B, De Smet W, De Baets B, De Vos P, Dawyndt P. (2014) StrainInfo introduces electronic passports for microorganisms. Syst Appl Microbiol. Feb;37(1):42-50.

Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C. (2013) The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 41(Web Server issue):W557-561

Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, Vasilenko A, Suzuki K, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Philippe D, Ma J. (2013) Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. BMC Genomics, 14:933.

Autori

Paolo Romano - paolo.romano@hsanmartino.it

Bioingegnere, lavora come bioinformatico dal 1993. I suoi interessi di ricerca sono legati all'integrazione di dati biomedici e allo sviluppo di procedure automatiche per l'analisi dei dati. Si è occupato a lungo di database per le risorse biologiche. Dal 2012 lavora nel laboratorio di proteomica. Organizza i workshop NETTAB sulle tecnologie ICT emergenti per la ricerca biomedica.

Giovanna Cristina Varese - cristina.varese@unito.it

Professore Associato in Botanica Sistematica presso l'Università di Torino. Responsabile Scientifico della Mycotheca Universitatis Taurinensis (MUT), una delle più grandi collezioni ex situ di microrganismi in Italia e in Europa. Partecipa attivamente alla creazione dell'Infrastruttura Europea MIRRI e sta coordinando la creazione del network italiano delle collezioni di microorganismi MIRRI-IT.

DOI: 10.26314/GARR-Conf17-proceedings-09

Il nuovo Regolamento Privacy, cloud computing e big data

Nadina Foggetti

Università degli Studi di Bari, Dipartimento di Giurisprudenza

Abstract. Il primo aspetto che esamineremo riguarda le problematiche connesse alla privacy. Il trattamento dei dati personali assume particolari peculiarità a seconda della tipologia di cloud computing che prendiamo in considerazione. In particolare si prenderanno in considerazione le problematiche connesse al maggiore livello di responsabilizzazione dei responsabili del trattamento, mediante l'introduzione di un sistema innovativo di valutazione dei rischi connessi alla tutela dei dati personali. Nel cloud computing si può qualificare l'attività svolta rispetto al trattamento di dati come "data processor", tuttavia nella prassi, le attività svolte sono maggiormente inquadrabili nell'ambito di quelle garantite da un "responsabile del trattamento". Il secondo obiettivo riguarda la disciplina giuridica del trattamento dei Big Data nel diritto internazionale e dell'UE. Gli aspetti di rilievo che occorre analizzare riguardano la Data discovery, la raccolta e la profilazione alla luce della disciplina vigente a livello europeo ed internazionale. I Big Data introducono una rivoluzione significativa in materia di privacy inserendosi in un settore caratterizzato già da un'elevata complessità e pongono la questione relativa alla definizione degli scopi perseguiti attraverso la raccolta e il trattamento dei dati stessi, ovvero quale sia la tipologia dei dati trattati e quindi la relativa disciplina. L'analisi verrà condotta alla luce del Reg. 2016/679 del 27 aprile 2016.

Keywords. Big Data, Privacy, RGPD, Cloud Computing.

Introduzione

La libera circolazione dei dati è al centro del nuovo Regolamento privacy dell'UE (RGPD). La nuova disciplina uniforma di fatto la normativa in materia di privacy a livello europeo ed è orientata ad accogliere le nuove sfide della società dell'informazione, tra cui in particolare la profilazione, i Big Data ed il diritto all'oblio.

1.Big data e profilazione

La tecnologia cloud ha contribuito a determinare la diffusione dei Big Data. Il WP29 definisce i Big Data come "gigantesche banche dati digitali ... analizzate in modo estensivo attraverso algoritmi elettronici". I Big Data introducono una rivoluzione in materia di privacy inserendosi in un settore caratterizzato già da un'elevata complessità. Spesso, come avviene ad esempio nell'ambito della bioinformatica, i dati prodotti a seguito del trattamento, possono avere una natura giuridica differente rispetto a quella attribuibile ai dati inizialmente raccolti o forniti dall'interessato. Il RGPD introduce il diritto dell'interessato a non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici nei suoi confronti,

a meno che non sia necessaria per la conclusione o l'esecuzione di un contratto tra lo stesso e un titolare del trattamento; non sia autorizzata dal diritto; o si basi sul consenso dell'interessato. Il Regolamento introduce un obbligo specifico per il titolare del trattamento di effettuare una valutazione di impatto (DPIA), qualora vi sia un trattamento sistematico e globale di aspetti personali relativi a persone fisiche, basato su un sistema automatizzato, compresa la profilazione.

Il RGPD impone ai titolari di adottare misure atte a dimostrare la compliance normativa, tenendo conto dei "rischi aventi probabilità e gravità diverse per i diritti e le libertà delle persone fisiche". L'obbligo di condurre una DPIA è collocato nel contesto della responsabilità di gestire correttamente i rischi connessi al trattamento di dati personali. Per "rischio" si intende uno scenario descrittivo di un evento e delle conseguenze che sono stimate in termini di gravità e probabilità. La "gestione del rischio" è definibile come l'insieme coordinato delle attività finalizzate a guidare e monitorare un ente o organismo nei riguardi di tale rischio.

Il riferimento ai "diritti e le libertà" degli interessati afferisce innanzitutto al diritto alla privacy, ma può riguardare anche altri diritti fondamentali quali la libertà di espressione e di pensiero. Coerentemente con l'approccio basato sul rischio che informa il Regolamento, la DPIA è obbligatoria solo se una determinata tipologia di trattamenti "può presentare un rischio elevato per i diritti e le libertà delle persone fisiche" (art. 35, paragrafo 1). Anche in assenza delle circostanze che impongono l'adozione di una DPIA, i titolari devono valutare in modo continuativo i rischi creati dai propri trattamenti così da individuare quelle situazioni in cui una determinata tipologia di trattamenti "può presentare un rischio elevato".

L'elenco disposto dall'art. 35 non è esaustivo, possono esservi trattamenti "a rischio elevato" non ricompresi nell'elenco, per questo il WP29 ha specificato dei sub criteri da considerare nella valutazione del rischio quali l'esistenza di trattamenti di scoring, compresa la profilazione e attività predittive, in particolare a partire da "aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze o gli interessi personali", come ad esempio una società biotecnologica che somministri test genetici gratuiti ai consumatori per finalità predittive del rischio di determinate patologie. Un ulteriore criterio è rappresentato dal monitoraggio sistematico quali "la sorveglianza metodica di un'area accessibile al pubblico". Un altro criterio riguarda i trattamenti di dati su larga scala. Il WP29 ha stabilito che è opportuno tenere conto di alcuni indicatori quali il numero di soggetti interessati; il volume dei dati e/o ambito delle diverse tipologie; la durata; l'ambito geografico.

Nella valutazione del rischio occorre considerare se vi sia una combinazione o raffronto di insiemi di dati, per esempio derivanti da due o più trattamenti svolti per diverse finalità e/o da titolari distinti, secondo modalità che esulano dalle ragionevoli aspettative dell'interessato. Infine occorre considerare l'esistenza di dati relativi a interessati vulnerabili poiché in questa situazione risulta accentuato lo squilibrio di poteri fra interessato e titolare. Il Regolamento prevede il "rischio del trattamento", inteso come l'impatto negativo sulle libertà e i diritti degli interessati. Si tratta di un approccio

risk based, che ha il vantaggio di pretendere degli obblighi che possono andare oltre la compliance, più adattabile al mutare degli strumenti tecnologici, ma che delega al titolare del trattamento la valutazione del rischio, rendendo più difficili le contestazioni in caso di violazioni.

2. I nuovi diritti introdotti

L'articolo 20 del RGPD introduce il diritto alla portabilità dei dati, che permette agli interessati di ricevere i dati personali, in un formato strutturato, di uso comune e leggibile meccanicamente e di trasmetterli a un diverso titolare. L'obiettivo è accrescere il controllo degli interessati sui propri dati personali. Consentendo la trasmissione diretta dei dati personali da un titolare all'altro, attua il principio della libera circolazione dei dati ed è teso a garantire trasparenza nella disciplina dei contratti facilitando il passaggio da un fornitore di servizi all'altro.

Ai fini dell'applicazione del diritto è necessario il consenso dell'interessato, la presenza di un trattamento effettuato con mezzi automatizzati, non applicandosi, invece, ai registri cartacei. Il WP29 precisa che il Regolamento non prevede un diritto generale alla portabilità dei dati il cui trattamento non si fondi sul consenso o su un contratto. Il WP29 ha chiarito che il diritto alla portabilità dei dati si configura rispetto ai dati forniti consapevolmente e in modo attivo dall'interessato, nonché rispetto ai dati generati dalle attività svolte dall'interessato, diversamente la natura dello stesso risulterebbe svuotata del suo contenuto.

Il principio richiede l'adozione da parte dei titolari di dispositivi atti a facilitare l'esercizio del diritto, quali strumenti per il download dei dati e API. Vi è, infatti, l'obbligo di garantire che i dati personali siano trasmessi in un formato strutturato, accessibile, in capo ai titolari. Sarebbe pertanto necessario garantire l'adozione di un formato standard che assicuri l'interoperabilità dei formati con cui i dati sono messi a disposizione. L'innovazione apportata è funzionale a garantire un maggiore controllo dei propri dati personali, effettuando, di fatto un "bilanciamento" nel rapporto tra interessati e titolari. È possibile individuare alcuni elementi di cui il diritto in parola si compone. Include il diritto dell'interessato a ricevere un insieme di dati da un titolare al fine di conservarli in vista di un utilizzo futuro per scopi personali. La conservazione può avvenire tramite un cloud privato, senza richiedere la trasmissione dei dati ad un altro titolare, rappresentando un'estensione del diritto di accesso. Un secondo elemento è il diritto di trasmettere dati personali ad un altro titolare senza impedimenti al fine di garantire all'interessato un margine di controllo sui dati, impedendo forme di "lock-in" tecnologico. Il diritto in parola non è un diritto assoluto, poiché il suo esercizio non deve pregiudicare nessuno degli altri diritti. Qualora l'interessato intenda esercitare il diritto all'oblio, il titolare non può procrastinare o negare l'applicazione dello stesso in virtù dell'applicazione del diritto di cui all'articolo 20 del RGPD.

Un aspetto importante riguarda l'applicazione del principio ai dati generati dal titolare, per esempio mediante l'analisi di dati grezzi originati da un contatore intelligente. Occorre effettuare una distinzione tra i dati forniti dall'interessato, anche nella fruizio-

ne di un servizio e quelli forniti consapevolmente dall'interessato. Il WP29 ritiene che la nozione di dati "forniti da" un interessato debba riferirsi anche ai dati personali osservati sulla base delle attività svolte dagli utenti, come per esempio i dati grezzi generati da un contatore intelligente o altri oggetti connessi, le registrazioni delle attività svolte, la cronologia della navigazione su un sito web. L'espressione "forniti da" si riferisce ai dati personali relativi ad attività compiute dall'interessato o derivanti dall'osservazione del comportamento, con esclusione dei dati derivanti dalla successiva analisi di tale comportamento. Viceversa, tutti i dati personali che siano creati dal titolare nell'ambito di un trattamento, per esempio attraverso procedure di personalizzazione o finalizzate alla formulazione di raccomandazioni, o attraverso la categorizzazione o profilazione degli utenti, sono dati derivati o dedotti dai dati personali forniti dall'interessato e non ricadono nell'ambito del diritto alla portabilità. Il GRPD introduce una tempistica entro la quale è necessario garantire l'esercizio del diritto alla portabilità dei dati, stabilendo all'art. 12, paragrafo 3, che il titolare fornisce "informazioni relative all'azione intrapresa" all'interessato "senza ingiustificato ritardo" e comunque "entro un mese dal ricevimento dalla richiesta" che si estende previa motivazione a tre mesi in casi di particolare complessità.

3. Conclusioni

Il GRPD, in definitiva introduce una serie di diritti innovativi, quali quello alla portabilità dei dati, il diritto all'oblio ed alcuni obblighi quali quello di adozione di DPIA. Tuttavia occorre rilevare che si basa su un'ottica fondata sul consenso, apparendo di fatto maggiormente orientato alla libera circolazione dei dati e quindi alle esigenze connesse alla società dell'informazione.

Riferimenti bibliografici

E. Pelino, L. Bolognini, C. Bistolfi (2016), Il regolamento privacy europeo. Commentario alla nuova disciplina sulla protezione dei dati personali Copertina flessibile, Milano.

R.H. JR. Carpenter (2010), Walking From Cloud To Cloud: The Portability Issue In Cloud Computing, in Washington Journal of Law, Tech. & Arts, (6), pp 1-25.

B. Custers, H. Ursic (2016), Big data and data reuse: a taxonomy of data reuse for balancing data benefits and personal data protection, in International Data Privacy Law, (6), p. 15-23.

A. Diker Vanberg, M.B. Ünver (2017), The right to data portability in the GDPR and EU competition law: odd couple or dynamic duo?, in European Journal of Law and Technology, (8), pp. 26-33.

P. Lee, K. Pickering (2016), The general data protection regulation: a myth-buster, in Journal of Data Protection & Privacy, pp-1-5.

H.T. Tavani, J.H. Moor (2000), Privacy Protection, Control of Information, and Privacy -Enhancing Technologies, in ACM SIGCAS Computers & Society, pp. 24-31.

Article 29 Data Protection Working Party (2017), Guidelines on Data Protection Impact Assessment (DPIA) and determining whther processing is "likely to result in a hight risk"

for the purposes of Regulation 2016/679.

Autori

Nadina Foggetti nadinafoggetti@gmail.com

Avvocato del Foro di Bari, mediatrice familiare e dottore di ricerca in Diritto Internazionale e dell'UE, ha conseguito un Master in Diritto Europeo e Transnazionale presso l'Università di Trento. Docente in corsi di perfezionamento e post-lauream, cultore della materia in Diritto Internazionale, Diritto dell'Unione Europea, del Commercio e Informatica Giuridica, attualmente ha un contratto per lo svolgimento dell'attività di ricerca presso l'Università degli studi di Bari "Aldo Moro" e ha partecipato a vari progetti nazionali e internazionali sui temi di cybercrime e cloud computing, nonchè sul diritto all'istruzione. Autrice di diversi articoli scientifici di carattere internazionale su varie tematiche tra cui gli aspetti giuridici collegati alle nuove tecnologie ICT. Fa parte di gruppi di lavoro internazionali sul tema della tutela delle persone con disabilità.

DOI: 10.26314/GARR-Conf17-proceedings-10

Applicazione di strumenti di business intelligence agli studi epidemiologici in sanità pubblica veterinaria

Laura D'Este, Elena Mazzolini, Andrea Ponzoni, Giuseppe Arcangeli, Antonio Barberio, Lisa Barco, Monia Cocchi, Gabriella Conedera, Michela Corrò, Debora Dellamaria, Ilenia Drigo, Nicola Pozzato, Karin Trevisiol, Fabrizio Agnoletti

Istituto Zooprofilattico Sperimentale delle Venezie

Abstract L'ottimizzazione degli strumenti di campionamento di convenienza per le ricerche in sanità animale rientra nell'applicazione delle strategie aziendali di process management. Allo scopo, l'IZSVe ha introdotto QlikView, un prodotto di Business Intelligence (BI), nella pianificazione, gestione e monitoraggio dei sistemi di sorveglianza veterinaria. QlikView è stato affiancato al LIMS IZSVe per individuare i campioni analitici, inviati ad IZSVe per altre finalità, utilizzabili nei disegni di studio predisposti per tre diverse indagini epidemiologiche, tese a verificare la presenza o la prevalenza di microrganismi di interesse in sanità pubblica. Dopo 20 mesi di applicazione QlikView ha permesso di identificare e raccogliere il 98% dei campioni previsti dagli studi, offrendo in tal modo interessanti prospettive di supporto da parte degli strumenti di BI all'attività di ricerca e sorveglianza in veterinaria.

Keywords. Ricerca, Business Intelligence, Epidemiologia, QlikView

Introduzione

Negli ultimi anni negli Istituti Zooprofilattici Sperimentali si è osservata una contrazione dei budget dedicati alla ricerca, a fronte di una crescente necessità di sorvegliare la presenza, o talvolta la prevalenza, di agenti di interesse in sanità pubblica, adottando numerosità campionarie sufficientemente rappresentative della popolazione di riferimento. L'IZSVe ha una popolazione di riferimento estremamente ampia, sia in senso geografico (area di provenienza) che per l'origine dei campioni analitici, rappresentati da animali, alimenti o ambiente. Ogni anno vengono conferiti ad IZSVe un milione e mezzo di campioni originati dall'attività di sorveglianza veterinaria, dal servizio di diagnostica delle malattie infettive e dal controllo degli alimenti. Questi campioni possono essere riutilizzati per altri studi; a tale proposito dal 2015, per due progetti di ricerca e tre diversi studi epidemiologici, è stato utilizzato un campionamento di convenienza impiegando strumenti di Business Intelligence (QlikView) applicati alla base di dati di supporto del LIMS per la selezione, la verifica e l'arruolamento dei campioni per indagare: a) la presenza di Clostridium difficile in molluschi eduli bivalvi; b) la diffusione di Staphylococcus aureus meticillino resistente (MRSA) in latte bovino; c) la diffusione di Escherichia coli produttori di beta-lattamasi a spettro esteso (E. coli ESBL+) in bovini, suini, cani e pollame. QlikView era utilizzato in IZSVe dal 2008 per finalità di controllo di gestione e questo lavoro descrive la prima applicazione epidemiologica di tale stru-

Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

mento in IZSVe a supporto del campionamento di convenienza.

1. Materiali e Metodi

Per i tre studi epidemiologici è stato predisposto un disegno che identificava la popolazione di riferimento, l'unità epidemiologica (l'animale, l'allevamento oppure l'area geografica), la definizione del campione arruolabile (ovvero la matrice biologica idonea all'analisi), la numerosità campionaria rispetto all'obiettivo di studio ed infine la distribuzione temporale del campionamento di convenienza.

L'analisi dei dati storici ha permesso di individuare le numerosità campionarie tenendo in considerazione il bacino di utenza dei laboratori coinvolti nella ricerca e l'origine geografica dei campioni, rendendo così possibile l'applicazione di un campionamento di convenienza proporzionale alla popolazione di riferimento (quest'ultima individuata dai dati reperibili in BDN) per la ricerca di MRSA nel latte, oppure l'area geografica di origine dei campioni di molluschi per la ricerca di C. difficile.

La ricerca di E. coli ESBL+ in suini, bovini e pollame è avvenuta in tutti i gruppi/allevamenti che conferivano campioni ad IZSVe utilizzando l'ID dell'allevamento (codice 317) per evitare l'overclustering, oppure arruolando tutti i soggetti disponibili nel caso della ricerca di E. coli ESBL+ nel cane. I campioni da arruolare sono stati quindi distribuiti in un periodo compreso tra i 16 e i 20 mesi, proporzionale all'attività del laboratorio e alla popolazione zootecnica di riferimento (solo per i bovini), a partire dal primo giorno di ogni mese. Per il campionamento dei molluschi è stato stabilito che ogni area geografica fosse ricandidabile all'arruolamento di nuovi campioni solo dopo un periodo proporzionale al numero di prelievi storici. Lo studio intendeva così arruolare in 20 mesi 600 campioni di molluschi, 1000 campioni di latte bovino, e 900 campioni di feci di bovino, suino e cane, e, in 16 mesi, ulteriori 500 campioni di feci da pollame.

Il disegno di studio e le sue modalità, le definizioni e gli assunti, sono stati quindi descritti in un modello decisionale (Figura 1) e in una matrice numerica che sono stati poi tradotti nel software di Business Intelligence QlikView. In Figura 2, viene descritto un esempio applicato di QlikView allo studio di C. difficile in due specie di molluschi. Tutti i campioni conferiti in Istituto e registrati nel LIMS aziendale (Izilab) sono stati dunque quotidianamente processati da QlikView per identificare attraverso una serie di flag la presenza di campioni consistenti con il modello decisionale e la matrice numerica dei diversi studi epidemiologici.

Questa elaborazione veniva svolta in due step: nel primo step QlikView selezionava tutti quei campioni che, all'atto della registrazione, presentavano caratteristiche tali da poter rientrare in uno degli studi effettuati. Nel secondo step venivano verificate le con-

	A	B	C	D	E	F
1	**MATERIALE**	**SPECIE**	**LABORATORIO**	**ANALISI**	**MOTIVO DEL PRELIEVO**	**SUBACCERTAMENTO**
2	PRODOTTI DELLA PESCA FRESCHI	MOLLUSCO VONGOLA	PN Alimenti Ufficiali	CONTA ESCHERICHIA COLI ß-GLUCORONIDASI POSITIVI (MPN)	PIANI LOCALI (REG-PROV)	PIANO MONITORAGGIO REGIONE FVG
3		MOLLUSCO VONGOLA (C. gallina)	PD Microbiologia Alimentare	RICERCA SALMONELLA SPP (KIT) IN 25G		MOLLUSCHI-MONITORAGGIO AMBITO
4		MOLLUSCO VONGOLA (R. decussatus)	AD Alimenti Ufficiali	SALMONELLA SPP. - IN 25G		
5		MOLLUSCO VONGOLA (R. philippinarum)	PD Biofood			
6		MOLLUSCO MITILO (M. galloprovincialis)				

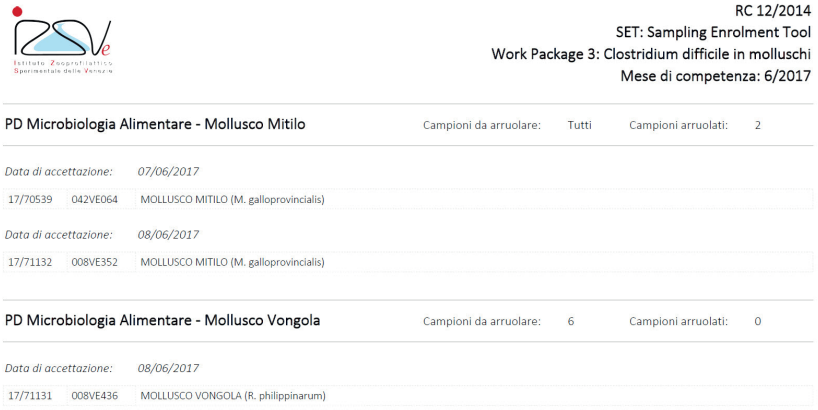
Fig. 1

Esempio di modello decisionale per la ricerca di per la ricerca di C. difficile in molluschi.

	A	B	C	D	E	F
1	UO	TIPO_MOLLUSCO	MESE	ANNO	NUMERO_CAMPIONI	MESI_ESCLUSIONE
2	AD Alimenti Ufficiali	M	11	2015	9999	0
3	PD Microbiologia Alimentare	M	11	2015	9999	2
4	PN Alimenti Ufficiali	M	11	2015	9999	0
5	AD Alimenti Ufficiali	V	11	2015	5	0
6	PD Microbiologia Alimentare	V	11	2015	9	9999
7	PN Alimenti Ufficiali	V	11	2015	4	0
8	AD Alimenti Ufficiali	M	12	2015	9999	0
9	PD Microbiologia Alimentare	M	12	2015	9999	2
10	PN Alimenti Ufficiali	M	12	2015	9999	0
11	AD Alimenti Ufficiali	V	12	2015	5	0
12	PD Microbiologia Alimentare	V	12	2015	8	9999
13	PN Alimenti Ufficiali	V	12	2015	4	0

Fig. 2
Esempio di traduzione in Excel del campionamento (modello decisionale e matrice numerica) dello studio epidemiologico per la ricerca di C. difficile in molluschi funzionale all'utilizzo in QlikView.

dizioni più complesse, come ad esempio il rispetto della rotazione dei luoghi di prelievo, il raggiungimento del limite di arruolamento per quello specifico laboratorio oppure l'avvenuto arruolamento di quel campione. Una volta identificati tramite delle clausole sui flag creati nello step precedente tutti i record pronti per essere segnalati, BIREPORT, un secondo prodotto specializzato in reportistica pixel perfect capace di estrarre i dati da QlikView, costruiva un report personalizzato per ogni laboratorio (Figura 3) segnalando a tutto il personale coinvolto nella ricerca tramite e-mail, la presenza di un campione candidabile allo studio. Il laboratorio verificava quindi la possibilità di effettuare l'arruolamento e, in caso positivo, procedeva inserendo le ricerche microbiologiche previste dallo studio per il campione selezionato.


RC 12/2014
SET: Sampling Enrolment Tool
Work Package 3: Clostridium difficile in molluschi
Mese di competenza: 6/2017

PD Microbiologia Alimentare - Mollusco Mitilo — Campioni da arruolare: Tutti — Campioni arruolati: 2

Data di accettazione: *07/06/2017*

17/70539 042VE064 MOLLUSCO MITILO (M. galloprovincialis)

Data di accettazione: *08/06/2017*

17/71132 008VE352 MOLLUSCO MITILO (M. galloprovincialis)

PD Microbiologia Alimentare - Mollusco Vongola — Campioni da arruolare: 6 — Campioni arruolati: 0

Data di accettazione: *08/06/2017*

17/71131 008VE436 MOLLUSCO VONGOLA (R. philippinarum)

Fig. 3
Esempio di report giornaliero di SET prodotto da BIREPORT indirizzato al Laboratorio per l'arruolamento di campioni idonei alla ricerca di C. difficile in mitili

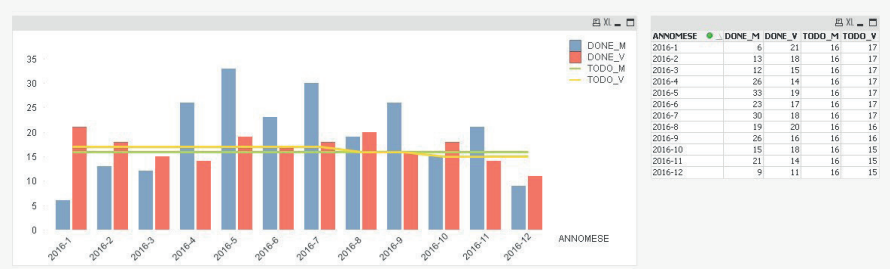
L'intero processo, rappresentato dal modello decisionale e dalla matrice numerica, dalle loro revisioni, dalla traduzione in QlikView e dalla produzione di avvisi e report tramite BIREPORT, è stato definito "Sampling Enrollment Tool" (SET). SET rappresenta quindi uno strumento informatico principalmente progettato per la conduzione di indagini epidemiologiche, a supporto dell'attività di campionamento e process management.

2. Risultati

Nel corso dei due progetti di ricerca SET ha permesso di monitorare l'arruolamento dei campioni per i tre diversi studi epidemiologici ed il calcolo delle performance del progetto:

campioni ottenuti rispetto ai previsti nel mese corrente e alla fine del progetto (Figura 4).

Con l'eccezione dello studio per la ricerca di E. coli ESBL+ in suini, e di E. coli ESBL+ in pollame (quest'ultimo tutt'ora in corso con performance corrente dell'89%) tutti gli altri studi (presenza di Clostridium difficile in molluschi, MRSA in latte bovino e E. coli ESBL+ nel bovino e nel cane) sono risultati in linea con quanto programmato, ottenendo a fine campionamento una performance compresa tra 93% e 117% rispetto al numero di campioni previsto. Per lo studio E. coli ESBL+ in suini, SET ha permesso di individuare già dalle prime fasi del progetto delle performance insoddisfacenti; gli interventi di mitigazione adottati sono risultati anch'essi insufficienti comportando la necessità di allungare i tempi di campionamento; questi problemi, tuttavia, sono attribuibili ad una sovrastima dei conferimenti di campioni suini a livello di disegno dello studio e non ad un difetto delle tecniche di arruolamento.



ANNOMESE	DONE_M	DONE_V	TODO_M	TODO_V
2016-1	6	21	16	17
2016-2	13	18	16	17
2016-3	12	15	16	17
2016-4	26	14	16	17
2016-5	33	19	16	17
2016-6	23	17	16	17
2016-7	30	18	16	17
2016-8	19	20	16	16
2016-9	26	16	16	16
2016-10	15	18	16	15
2016-11	21	14	16	15
2016-12	9	11	16	15

Fig. 4
Esempio di reportistica QlikView (numero campioni raccolti per mese per specie, rispetto al numero pianificato) funzionale al project management

3. Conclusioni

SET, strumento informatico-decisionale che vede applicate tecniche di BI agli studi epidemiologici, può rappresentare un valido supporto al processo di campionamento e all'attività di process management, consentendo di ottimizzare l'impiego delle risorse disponibili per la sorveglianza epidemiologica veterinaria.

4. Fonti di finanziamento

Questo lavoro fa riferimento ad attività svolte nell'ambito del progetto di ricerca IZSVE 12/14 RC "Migliorare l'efficacia della sorveglianza IZSVE verso le resistenze batteriche agli antimicrobici ed altri rischi emergenti in sanità pubblica" finanziato dal Ministero della Salute, e del progetto CCM 2015 "Il modello One-Health per il contenimento delle resistenze microbiche di possibile origine zoonosica in sanità pubblica: sviluppo di un network medico-veterinario applicato alla prevenzione e controllo della circolazione di Escherichia coli produttore di ESBL" finanziato dal Ministero della Salute.

Autori

Laura D'Este ldeste@izsvenezie.it
LIMS manager e BI Developer presso il Laboratorio Gestione Risorse Informatiche e Innovazione Tecnologica

Elena Mazzolini emazzolini@izsvenezie.it
Epidemiologo e microbiologo veterinario responsabile dell'Osservatorio epidemiologico veterinario e sicurezza alimentare della Regione Autonoma Friuli Venezia Giulia, Istituto Zooprofilattico Sperimentale delle Venezie (SCS4 Epidemiologia veterinaria) e Direzione Centrale Salute, Regione Autonoma Friuli Venezia Giulia

Andrea Ponzoni aponzoni@izsvenezie.it
Responsabile del Laboratorio Gestione Risorse Informatiche e Innovazione Tecnologica

Giuseppe Arcangeli garcangeli@izsvenezie.it
Direttore Centro specialistico di Ittiopatologia.
Direttore del Centro di Referenza Nazionale per le malattie dei pesci, molluschi e crostacei.

Antonio Barberio abarberio@izsvenezie.it
Dirigente veterinario presso SCT1 – Verona e Vicenza

Lisa Barco lbarco@izsvenezie.it
Dirigente veterinario presso SCT4 – Friuli Venezia Giulia

Monia Cocchi mcocchi@izsvenezie.it
Dirigente veterinario presso SCT4 – Friuli Venezia Giulia, sezione territoriale di Udine. Laurea in Medicina veterinaria. Si occupa di microbiologia diagnostica e ricerca (biofilm e resistenza agli antimicrobici)

Gabriella Conedera gconedera@izsvenezie.it
Direttore struttura complessa territoriale SCT4 - Friuli VG (Udine, Pordenone)

Michela Corrò mcorro@izsvenezie.it
Dirigente veterinario presso SCT3 – Padova e Adria – Diagnostica in sanità animale

Debora Dellamaria ddellamaria@izsvenezie.it
Dirigente veterinario presso SCT5 – Trento

Ilenia Drigo idrigo@izsvenezie.it
Dirigente biologo presso SCT2 – Treviso, Belluno e San Donà di Piave.

Nicola Pozzato npozzato@izsvenezie.it
Dirigente veterinario presso SCT1 – Verona e Vicenza

Karin Trevisiol ktrevisiol@izsvenezie.it
Dirigente veterinario presso SCT6 – Bolzano

Fabrizio Agnoletti fagnoletti@izsvenezie.it
Direttore struttura complessa territoriale SCT2 (Treviso, Belluno, Venezia)

DOI: 10.26314/GARR-Conf17-proceedings-11

Documenti delle pubbliche amministrazioni, un patrimonio da preservare nel tempo: regole e prospettive per la realizzazione di una rete di poli di conservazione

Patrizia Gentili, Raffaele Montanaro, Cristina Valiante

AGiD, Agenzia per l'Italia Digitale

Abstract. Il tema della conservazione è importante, in particolare per le P.A. Senza la conservazione non è concepibile un processo di dematerializzazione, poiché non vi è garanzia che le informazioni in formato digitale siano preservate nel tempo, in modo autentico e accessibile, come per i documenti analogici. Inoltre le P.A. hanno, istituzionalmente, il compito di conservare i propri documenti e archivi, sia come testimonianza diretta della loro azione amministrativa, sia come memoria storica. Il contributo, esponendo alcuni concetti essenziali come: documento informatico, ciclo di vita dei documenti, sistema di conservazione e modelli organizzativi per l'attività di conservazione, poli di conservazione, offre una panoramica delle regole tecniche in materia e coerentemente con quanto indicato nel Piano triennale per l'informatica nella P.A. 2017-19, preannuncia l'avvio di una sperimentazione volta ad identificare un modello di riferimento di polo di conservazione dei documenti informatici ed a promuovere la realizzazione di una rete logica dei Poli di conservazione, nonché a definire la relativa infrastruttura tecnica funzionale all'interconnessione tra i Poli stessi.

Keywords. Documento informatico, conservazione, poli di conservazione, dematerializzazione.

Introduzione

La definizione di documento informatico fornita dal Codice dell'amministrazione digitale, di seguito CAD, ossia "il documento elettronico che contiene la rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti", evidenzia la centralità delle informazioni contenute nel documento stesso a prescindere dalla forma o dal supporto su cui sono salvate. La definizione di documento elettronico di cui al Regolamento UE n.910/2014, è ancor più ampia, poiché il termine indica "qualsiasi contenuto conservato in forma elettronica, in particolare testo o registrazione sonora, visiva o audiovisiva".

L'ampiezza di tali definizioni è il riflesso della capacità delle tecnologie ICT di accrescere le possibili modalità di formazione dei documenti informatici, come indicato dall'articolo 3 delle regole tecniche in materia di formazione, trasmissione, copia, duplicazione, riproduzione e validazione temporale dei documenti informatici nonché di formazione e conservazione dei documenti informatici delle pubbliche amministrazioni.

Al crescere dei mezzi con cui formare documenti corrisponde l'aumento della mole di documenti informatici e dati prodotti e gestiti. La quantità di informazioni è composta da: documenti informatici, singoli o variamente aggregati, dai relativi metadati e dagli

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

archivi digitali, e costituisce, anche legalmente, un patrimonio informativo da preservare nel tempo.

1. La conservazione

Il tema della conservazione è essenziale in particolare per le pubbliche amministrazioni. Senza la conservazione non si può concepire un processo di dematerializzazione, poiché non vi è garanzia che documenti e informazioni in formato digitale siano preservati nel lungo periodo, in modo autentico e accessibile, come avviene per i documenti analogici. Inoltre le P.A. hanno, istituzionalmente, il compito di conservare i propri documenti e archivi, sia come testimonianza diretta della loro azione amministrativa, sia come memoria storica. I documenti e gli archivi degli enti pubblici sono beni culturali, tutelati anche a livello costituzionale. La realizzazione di archivi accessibili e strutturati, che rendano disponibile l'enorme patrimonio informativo della P.A., è, quindi, un elemento indispensabile. La conservazione digitale, intesa come attività complessa volta a mantenere inalterate nel tempo le sequenze binarie degli oggetti trattati, risponde all'esigenza di assicurarne la possibilità di accesso e fruizione.

Grazie alla conservazione si dovrà sempre più offrire funzionalità idonee a soddisfare le richieste di consultazione e di esibizione formulate sia dal titolare dei documenti, cioè, in termini archivistici, il soggetto che ha prodotto l'archivio, costituito dai documenti ricevuti o prodotti nel corso della sua attività, sia dai cittadini utenti, cioè coloro che richiedono di fruire delle informazioni di interesse conservate e, in futuro da studiosi e altri portatori d'interesse. La memoria digitale pubblica dovrà essere consultabile nel rispetto delle norme sulla tutela dei dati personali e della riservatezza. In particolare dovrà essere garantita la confrontabilità dei documenti e degli archivi secondo la disciplina del Codice dei Beni Culturali. La conservazione digitale dovrà costruire per le P.A. gli archivi storici del futuro.

A tal fine, occorre collocare la conservazione nell'ambito del ciclo di vita del documento informatico. Infatti per poter far sì che la memoria digitale sia conservata è necessario che già in fase di produzione e di gestione documentale corrente siano rispettati alcuni requisiti fondamentali, quali l'utilizzo di formati idonei e corrette azioni di gestione documentale, quali identificazione e registrazione dei documenti e loro organizzazione in base al piano di classificazione e fascicolazione.

2. I documenti informatici

Mantenere nel tempo i documenti informatici significa preservarne specifiche peculiarità, che derivano dalle caratteristiche oggettive del documento informatico stesso: di qualità, sicurezza, integrità e immodificabilità.

I documenti informatici per loro natura sono e devono essere soggetti a un continuo processo di mantenimento e trasformazione che ne consenta l'accesso nel tempo. L'attenzione dell'attività di conservazione deve spostarsi dal supporto al contenuto, riducendo i rischi di perdite e alterazioni provocati dalla rapida obsolescenza delle tecnologie informatiche e operare per la salvaguardia nel lungo termine del valore giuridico e delle caratteri-

stiche oggettive predette.

La conservazione digitale è un'attività permanente e deve realizzarsi con tempi e modi adeguati, prevedendo sia la cosiddetta conservazione dei bit (Bit preservation), cioè la capacità di accedere ai bit come erano stati originariamente registrati, anche in caso di degrado del supporto, di obsolescenza dell'hardware e/o disastri di sistema sia, soprattutto, la conservazione logica (Logical preservation) cioè la capacità di comprendere e usare l'informazione in futuro, conservando il contenuto intellettuale anche in presenza di futuri cambiamenti tecnologici e di conoscenza. Occorre bilanciare gli indubbi vantaggi offerti dalle tecnologie ICT in termini di capacità di ricerca e di riproduzione e gli svantaggi derivanti da maggiori rischi, rispetto al passato, di perdita della possibilità di recuperare, restituire o interpretare informazioni.

In quest'ottica, il concetto di sistema di conservazione è inteso come un insieme di persone, apparecchiature, applicazioni e procedure volte ad assicurare la conservazione a lungo termine dei documenti e delle aggregazioni documentali informatiche e dei rispettivi metadati, per garantire il mantenimento delle caratteristiche sopra citate.

Tale concetto è stato poi sviluppato in apposite regole tecniche, che hanno: esplicitato la funzione del sistema di conservazione, contestualmente individuato gli oggetti trattati (come pacchetti informativi, in conformità allo standard OAIS) per i quali il sistema stesso deve garantire fin dalla presa in carico dal produttore le già ricordate caratteristiche oggettive, definito i modelli organizzativi della conservazione, nonché stabilito ruoli e responsabilità coinvolti. Inoltre lo stesso CAD ha affidato ad AgID il compito di accreditare i conservatori.

La scelta del modello organizzativo più indicato deve essere compiuta in base ad efficienza e sostenibilità e attuata valutando le risorse finanziarie, tecnologiche e professionali disponibili nell'ente.

3. Il piano triennale della P.A. e i Poli di conservazione

Nell'ambito del Piano triennale per l'informatica nella Pubblica Amministrazione 2017-2019 è stato definito un percorso, da compiersi in più tappe nel lasso temporale di riferimento, che persegue l'obiettivo finale dell'interoperabilità tra sistemi di conservazione, al fine di permettere l'accesso unico ai documenti della P.A.

Attualmente nell'ambito della conservazione sono state rilevate sia problematiche funzionali, i poli di conservazione oggi esistenti non sono interoperabili tra di loro poiché manca un linguaggio comune ed ogni polo utilizza software e sistemi di archiviazione differenti dagli altri, sia problematiche organizzativo-economiche, le P.A. non hanno sufficienti competenze e disponibilità economiche per sviluppare dei sistemi di conservazione propri così da dover esternalizzare il servizio.

L'AgID, per realizzare tale percorso, intende identificare un modello di riferimento di polo di conservazione dei documenti informatici e promuovere la realizzazione di una rete logica dei Poli di conservazione e la definizione della relativa infrastruttura tecnica funzionale all'interconnessione tra i Poli stessi. A tal fine si sta avviando un percorso di sperimentazione per individuare le misure necessarie e sufficienti a garantire l'interopera-

bilità fra i Poli.

Il progetto si basa sull'interesse a delineare un processo di conservazione che agevoli l'accesso e il controllo da parte dei soggetti cui spetta la funzione di vigilanza ed ottimizzi il flusso dei pacchetti di conservazione mantenendo le menzionate caratteristiche oggettive del documento informatico.

I punti qualificanti del progetto si riassumono in:

- Accesso: costruire un punto unico di accesso ai documenti informatici della PA per i Cittadini, le Imprese e le P.A.;
- Economicità: l'accesso ai poli pubblici, in particolare per la documentazione dei cittadini e degli archivi della PA da conservare a lungo termine, consente di fare rilevanti economie di scala;
- Uniformità: occorre garantire l'uniformità a livello nazionale in tema di accesso e sicurezza dei documenti amministrativi;
- Localizzazione: la distribuzione dei poli deve poter coprire geograficamente tutto il territorio nazionale;
- Numerosità della documentazione: il numero di documenti sta crescendo e crescerà sempre più, gli attuali poli non riusciranno a gestire tutto.

4. Conclusioni

L'obiettivo primario del progetto è di impedire la perdita o la distruzione non autorizzata dei documenti e di mantenere nel tempo le loro caratteristiche di autenticità, integrità, affidabilità, leggibilità, reperibilità, offrendo alle amministrazioni un servizio di conservazione che garantisca la presenza sul territorio nazionale di almeno una copia operativa per ciascun documento informatico conservato.

Riferimenti bibliografici

Decreto legislativo 7 marzo 2005, n. 82 e s.m.i. Art. 1, comma 1, lett. p).

2 Regolamento (UE) n. 910/2014 del Parlamento europeo e del consiglio del 23 luglio 2014 in materia di identificazione elettronica e servizi fiduciari per le transazioni elettroniche nel mercato interno e che abroga la direttiva 1999/93/CE. Art. 3, definizione n. 35.

Decreto del presidente del consiglio dei ministri 13 novembre 2014.

DPCM 3 dicembre 2013 "Regole tecniche in materia di sistema di conservazione"

L'art.3, comma 1 del citato DPCM stabilisce che gli oggetti della conservazione, per i quali il sistema di conservazione deve garantire, dalla presa in carico dal produttore, le caratteristiche di qualità, sicurezza, integrità e immodificabilità, sono: i documenti informatici e i documenti amministrativi informatici con i metadati ad essi associati; i fascicoli informatici ovvero le aggregazioni documentali informatiche con i metadati ad essi associati.

Tali pacchetti sono distinti in: pacchetti di versamento (Submission Information Package, SIP), pacchetti di archiviazione (Archival Information Package, AIP), pacchetti di distribuzione (Dissemination Information Package, DIP).

L'art. 5 del DPCM 3 dicembre 2013, prevede che la conservazione può essere svolta: all'in-

terno della struttura organizzativa del soggetto produttore dei documenti informatici da conservare, oppure affidandola, in modo totale o parziale, ad altri soggetti, pubblici o privati che offrono idonee garanzie organizzative e tecnologiche, accreditati come conservatori presso l'Agenzia per l'Italia digitale. Viene altresì specificato che Le pubbliche amministrazioni realizzano i processi di conservazione all'interno della propria struttura organizzativa o affidandoli a conservatori accreditati, pubblici o privati, di cui all'art. 44 bis, comma 1, del CAD.

L'art. 44bis del CAD stabilisce che "i soggetti pubblici e privati che svolgono attività di conservazione dei documenti informatici e di certificazione dei relativi processi anche per conto di terzi e intendono conseguire il riconoscimento dei requisiti del livello più elevato, in termini di qualità e sicurezza".

Autori

Patrizia Gentili gentili@agid.gov.it

Funzionario presso l'Agenzia per l'Italia digitale, esperto di progetti di dematerializzazione della P.A., è responsabile del servizio "Documentali" dell'Agenzia. Fin dal 2004 ha curato per l'allora CNIPA il monitoraggio della diffusione dei sistemi di gestione documentale presso la P.A. Dal 2010 al 2016 ha lavorato presso INPS come responsabile di pianificazione strategica e controllo del budget ICT.

Raffaele Montanaro montanaro@agid.gov.it

Funzionario presso l'Agenzia per l'Italia digitale, esperto di informatica giuridica e di semplificazione di procedimenti amministrativi pubblici, con riferimento alle tematiche della gestione documentale ed accessibilità. Da dieci anni collabora alle attività nell'ambito dell'azione di informatizzazione della normativa.

Cristina Valiante cristina.valiante@agid.gov.it

Collaboratrice presso l'Agenzia per l'Italia digitale, esperta di gestione e conservazione dei documenti informatici, con particolare riferimento agli aspetti giuridico-archivistici ed alla sicurezza informatica. Ha seguito il processo di accreditamento dei conservatori e coordina le attività del Forum della Conservazione dei documenti informatici.

DOI: 10.26314/GARR-Conf17-proceedings-12

A View on the Implementation of the European Open Science Cloud

Elena Bianchi[1], Paolo Budroni[2], Augusto Celentano[3], Marisol Occioni[4], Sandra Toniolo[4], Maurizio Vedaldi[1], Antonella Zane[1]

[1]Università degli Studi di Padova, Centro di Ateneo per le Biblioteche; [2]University of Vienna, Dpt. e-Infrastructures, Library and Archive Services; [3]Università Ca' Foscari Venezia, Dipartimento di Scienze Ambientali, Informatica e Statistica; [4]Università Ca' Foscari Venezia, Sistema Bibliotecario di Ateneo

Abstract. This paper outlines the discussion and presents the main outcomes of two workshops held in Padova and Venice in August–September 2017 on Open Science and the European Open Science Cloud (EOSC) initiative. The paper describes the three layers on which EOSC is grounded: governance, service and data layer, and discusses the new emerging roles for researchers and research support services. Suggestions about the EOSC implementation at local level are given.

Keywords. European Open Science Cloud, Data Management Plan, FAIR Principles, Research Data Management, Research Policies

Introduction

The European Commission is promoting the European Open Science Cloud (EOSC), a supporting environment for Open Science whose ultimate goal is the building of a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes under well-defined and trusted conditions.

According to the resolutions adopted so far, the EOSC is not an actual cloud service, but is based on the reengineering of existing e-infrastructures based on scientific data. As such, it is a bottom-up process based on existing and emerging elements in the Member States, with lightweight international guidance and governance and a large degree of freedom regarding practical implementation. As research is not an individual task, but the result of a joint effort between research and research support, the EOSC implementation impacts the whole organization of any research institutions; in particular it requires a shift from vertical to horizontal thinking by integrating several skills and knowledge into a coordinate set of services.

The goal of this paper is to outline the discussion and to present the main outcomes of the workshop "The European Open Science Cloud (EOSC) versus the single Research Institution. Drawing the scenario at local level" held in Padova and the "Workshop on Open Science" held in Venice (August-September 2017) organized by the University of Padova,

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

the Ca' Foscari University of Venice and the University of Vienna.
The results of the workshops are discussed here together with the final recommendations about the implementation of an EOSC strategy at local level.

1. The European Open Science Cloud - Three Layers

The implementation foresees the development of three layers: a governance layer, a service layer and a data layer. The governance layer addresses the issues of policies, good governance, trust, legacy and sustainability; as noted above, it suggests a bottom-up strategy based on federation of existing infrastructures. The service layer supports the governance strategy in several directions: (1) research support, legal and ethical issues, exploitation rights, statistics and analytics; (2) IPR protection, privacy and personal data protection; (3) big data processing and high-performance computing;(4) data storage, access and re-use; (5) data management plans; (6) terminology; (7) data exchange, integration and fusion across different disciplines. The data layer provides technical support in terms of data storage, manipulation, conversion, export and re-use, discovery strategies and cataloguing functions.

Different research areas have different demands about the amount of data; physics, life sciences and Earth sciences are the leading users with the most data intensive environments, while humanities and citizen science, at the other side of the range, are less demanding; such distribution justifies a high degree of flexibility in the organization of the research support systems, coherent with the approach suggested by the European Commission.

2. New Roles for Researchers and Research Support

To pave the way to the realization of EOSC at local level the Research Institution must provide a digital workflow to manage the research process and assure the convergence of knowledge into shared transversal services to support research. This organization is motivated by the idea that excellent research is possible only if accompanied by optimal research support.

A goal of this process is the offering of advice and the concrete monitoring on cost generation and development along the entire chain concerning data production, storage and reuse, e-infrastructures, human resources development, funding, services, timing. The key elements are:

- The digital workflow of research processes, to assure the compliance with the FAIR principles required by the EOSC,
- Research data (RDM) management policies, regarding roles and responsibilities of researchers, research support entities and the institution, as well as good governance models,
- Data management plans (DMP) defining data and all processes concerning their production, use and final reuse . DMPs are structured guidelines (documents or online tools) that depict the entire lifeline of data. DMPs must assure that research data are traceable, available, authentic, citable, properly stored, and that they adhere to clearly defined legal parameters and appropriate safety measures governing subsequent use.
- A single reference point gathering transversal knowledge for research support involving a set of competences and skills (internal or/and outsourced). The reference point is inten-

ded to be a service (internal or external). In any case the reference point will be able to solve questions referring to shared services, central services and cross-disciplines services. The implementation of the EOSC at a local level will improve the visibility and the attraction of the Research Institution, contributing to improve also its ranking. Research support entities will be called to play a major and strategic role in this process. The improved quality of training will attract further resources and more qualified personnel and students.

3. Conception and adoption of RDM Policies

RDM policies are key issues for the implementation of the EOSC Governance Layer. In this context we refer to the outputs and findings of the project LEARN and the results of the Italian working group GDL-Dati della ricerca.Policies concern: jurisdiction, intellectual property rights (IPR), handling of research data, responsibilities, rights, duties (e.g., "Researchers are responsible for…", "the Research institution is responsible for…"), validity.

Data management plans are the key elements of policies. They refer to description and management of information, the content acquired and generated by the projects and the context in which they are used. DMPs must be generated at the start of the project and may evolve into versions during the project development. They must address the following issues: making data findable, making data accessible, making data interoperable, increase data re-use, allocation of resources and data security.

The FAIR principles are the guiding elements in the conception of the DMPs . Particular attention should be paid to the following issues: the management of resources (especially time), reproducibility and reusability of the produced data, the assignment of proper licenses, security (infrastructure and processes), compliance of legal and ethical issues. At present, there are several models of DMPs available[1], offered by the issuing institutions according to the domain research processes of the related disciplines. Therefore, the research workflow may have different expressions, e.g., depending on data formats, size, objectives, etc.[2]

[1]See DCC-DMP online, the template of the EC, national versions of DMP online, locally tailored versions, etc.

[2]Future developments foresee the creation of machine actionable plans, process management plans and data stewardship plans.

4. A set of recommendations for successful implementation of EOSC at the local level

A bottom-up process grows and extends according to the dynamics of the organization's components, which are most often different in size, data usage, temporal scale, requirements and practices. Partial failures (or, worse, a complete failure) cannot be avoided, in principle, if the dynamics are left free to evolve independently. Hence, the importance of the governance layer is evident, and the Policies and DMPs are instruments effective only if properly driven by a common superior view, leading to an integration of bottom-up and top-down approaches.

Precise recommendations, even if plausible in principle, should be considered more than practical guidelines, and a control must be exercised over them to ensure the successful implementation of a common EOSC strategy. From the workshops in Padova and Venice the following issues were emerging as primaries:

- Enhance the shift of mentality from vertical based thinking to horizontal based thin-

king; create and offer new horizontal cross-disciplines services; make convergence of knowledge possible and gather efforts into Reference Points for research support.
- Start policy development and alignment at all levels, and introduce especially RDM policies. As a further step generate and adopt Data Management Plans, supporting data stewardship .
- Acknowledge the increasing relevance of the roles of research support units versus the researcher community.
- Get involved into the bottom-up processes of EOSC and participate to the networks and initiatives concerning the EOSC. Activate all stakeholders in your Research Institution for the realization of the EOSC.

5. Conclusions

The realization of the European Open Science Cloud will generate changes. The implementation of the EOSC at a local level will improve the visibility and the attraction of the Research Institution, contributing to improve also its ranking. Research Support entities will be called to play a major and strategic role in this process. It will be not easy but the sooner a Research Institution starts to adapt its organization, the sooner it will achieve the goal: EOSC is planned to be a tangible reality in 2018.

References

http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

https://hdl.handle.net/11168/11.334202

https://www.dtls.nl/fair-data/go-fair/

www.learn-rdm.eu

http://wikimedia.sp.unipi.it/index.php/OA_Italia/Risorse_sugli_open_research_data

http://learn-rdm.eu/wp-content/uploads/RDMToolkit.pdf

https://www.dtls.nl/fair-data/go-fair/

http://libereurope.eu/wp-content/uploads/2015/11/Open-Cloud-Principles.pdf

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

Authors

Elena Bianchi elena.bianchi@unipd.it

Elena Bianchi graduated in Foreign Languages at University of Padova in 1989. Since 2012 she is responsible for the University of Padova Library System internal and external Communication. She coordinates the Library System support initiatives on Digital library services, Open Access publication and Research Data Management. She is member of the Italian Open Science Support Group (IOSSG).

Paolo Budroni paolo.budroni@univie.ac.at
Paolo Budroni is Head of the Dpt. E-Infrastructures of Vienna University Library and Austrian National Delegate in the e-Infrastructures Reflection Group. He developed the first CRIS System of the University of Vienna (1991), its Research Data Management Repository Phaidra (2007) and led the Project e-Infrastructures Austria. He is also TAIEX Expert for ICT, and specialized in DMPs, RDM-Policies, Open Data, Open Science, EOSC and development of sustainable Infrastructures.

Augusto Celentano auce@unive.it
Professor Emeritus at Ca' Foscari University of Venice, he received a Master Degree in Electronic Engineering from Politecnico di Milano. At Ca' Foscari he has been Head of the Computer Science Department, Deputy Rector for the University Information Systems and Head of the Data Monitoring Board. His research interests are focused on advanced human-computer interaction, multimedia and information technology applications to cultural heritage.

Marisol Occioni occioni@unive.it
Currently Director of the Digital Library at the Ca' Foscari University of Venice and member of the Data Monitoring Board of the University. She is member of AISA (Associazione Italiana per la promozione della Scienza Aperta), of IOSSG (Italian Open Science Support Group) and of the CTS of Consortium IDEM-GARR.

Sandra Toniolo sandra.toniolo@unive.it
Director of Libraries at Ca' Foscari University of Venice, Head of Staff and Financial Resources. She supports internal and external projects on the development of information services and promotes the growth of the digital environment for researchers and e-learning. She has been a consultant at the Polytechnic University of Bari on the re-engineering of the library processes, then member of the Board and consultant of the CIPE consortium.

Maurizio Vedaldi maurizio.vedaldi@unipd.it
Maurizio Vedaldi is the Director of the University of Padova Library Centre. He coordinates, manages and fosters the development of library services for the benefit of scientific research and teaching, by means of the latest technologies, in accordance with international standards. The Library Centre also promotes national and international cooperation through agreements with other Institutions and Universities. From 2014 to 2016 he was the Director of the University of Padova IT Services.

Antonella Zane antonella.zane@unipd.it
Antonella Zane has a doctor's degree on Earth Sciences and 10 years of research background in Petrology and Archaeometry. She is the Head of the Sector "Digital Library and International projects" for the University of Padova Library System since 2011 and coordinates the activities of the Library System institutional repositories for graduated and PhD thesis, scholarly papers and cultural heritage. She is member of the Italian Open Science Support Group (IOSSG).

DOI: 10.26314/GARR-Conf17-proceedings-13

I dati della ricerca biomedica in Italia: verso la definizione di una policy nazionale?

Moreno Curti[1], Paola De Castro[2], Corrado Di Benedetto[3], Rosalia Ferrara[4], Pietro La Placa[5], Cristina Mancini[4], Luisa Minghetti[6], Elisabetta Poltronieri[2], Filippo Santoro[6], Franco Toni[4], Angela Vullo[7]

[1]IRCCS, Fondazione Policlinico San Matteo di Pavia, Servizio documentazione e biblioteca scientifica, [2]Istituto Superiore di Sanità, Servizio conoscenza e comunicazione scientifica, Attività editoriali, [3]Istituto Superiore di Sanità, Servizio controllo di gestione e informatica, [4]Istituto Superiore di Sanità, Servizio conoscenza e comunicazione scientifica, Biblioteca, [5]Istituto Zooprofilattico Sperimentale della Sicilia, Servizio editoria e biblioteca, [6]Istituto Superiore di Sanità, Servizio tecnico scientifico di coordinamento e supporto alla ricerca, [7]Istituto Zooprofilattico Sperimentale della Sicilia, Dipartimento sanità territoriale

Abstract. Il sistema Bibliosan opera per la condivisione e la diffusione delle risorse informative in campo biomedico, tramite la rete delle biblioteche degli Istituti di ricerca afferenti al Ministero della Salute. Nel 2016, si è costituito il Gruppo di lavoro inter-istituzionale BISA (Bibliosan per la Scienza Aperta), promosso dall'Istituto Superiore di Sanità (ISS) a sostegno dei principi dell'Open Science. Il Gruppo ha lanciato nel 2017 un'indagine sul trattamento dei dati della ricerca nei suoi molteplici risvolti (utilizzo, diffusione, accesso, conservazione). In esito a tale indagine, diffusa a tutti gli enti Bibliosan, che ha rilevato alcune criticità nella gestione dei dati, si è attivata l'elaborazione di una policy per regolare organicamente la gestione dei dati della ricerca. Con il supporto delle componenti istituzionali ISS, il documento è in corso di discussione ai fini di definire un documento che, con l'approvazione del Ministero della Salute possa costituire un riferimento normativo comune agli enti Bibliosan.

Keywords. Ricerca biomedica, Accesso all'informazione, Diffusione dell'informazione, Politica pubblica, Italia.

Introduzione

La ricerca biomedica in Italia può contare dal 2003 sulla condivisione di risorse informative e sullo sviluppo di temi legati alla diffusione della documentazione scientifica assicurati dal sistema Bibliosan tramite la sua rete di biblioteche biomediche afferenti agli istituti di ricerca del Servizio Sanitario Nazionale (SSN).

Gli enti della rete Bibliosan sono 60 e includono l'Istituto Superiore di Sanità (ISS), 46 Istituti di ricovero e cura a carattere scientifico (IRCCS), 10 Istituti zooprofilattici (IZS), l'Agenzia nazionale per i servizi sanitari regionali (AGENAS), l'Agenzia italiana del farmaco (AIFA) e l'Istituto nazionale assicurazione contro gli infortuni sul lavoro (INAIL). Tali istituzioni sono tradizionalmente produttori e utilizzatori di dati della ricerca, acquisiti in forza del proprio mandato istituzionale e organizzati con diverse modalità per attuare

gli obiettivi di tutela della salute e di analisi dello stato sanitario del Paese. Gli strumenti destinati a tali finalità sono principalmente:

- i registri regionali e nazionali. Sistemi di censimento sistematico di dati per monitorare e studiare fenomeni specifici, ad es. i casi di rischio per la salute legati al manifestarsi di una data patologia o alla tossicità chimica di una sostanza. Consistono in "statistiche per dati aggregati", cioè dati elaborati (non grezzi) condensati in report annuali. Alcuni registri sono istituiti con provvedimenti legislativi o ministeriali mentre altri sono registri "spontanei" sorti, ad es., su iniziativa di un'istituzione per contribuire a chiarire quali fattori ambientali, genetici e stili di vita influenzino la salute psicofisica della popolazione;
- le basi di dati. Aggregazioni strutturate di dati di vario genere prodotte, ad es., nell'ambito di studi epidemiologici e di monitoraggio ambientale, condotti da un'istituzione e consultabili, a seconda dei casi, liberamente o con accesso riservato.

1. Accesso aperto ai risultati della ricerca

Tra gli enti del sistema Bibliosan, l'ISS ha coltivato storicamente una missione di apripista in adesione ai principi del movimento Open Access (OA), confluiti e ampliati nella galassia dell'open science, favorendo la comunicazione e lo scambio dei risultati della ricerca. Nel 2008 si è dotato di una "Politica istituzionale per il libero accesso alle pubblicazioni scientifiche" http://www.iss.it/binary/sae4/cont/policy_ISS.jpg con finalità di certificazione, libera disponibilità online e archiviazione digitale permanente delle pubblicazioni scientifiche prodotte dall'Ente e da istituzioni partner del SSN. Per l'attuazione di tali finalità, l'ISS ha istituito nel 2006 un archivio digitale in rete, DSpace ISS http://dspace.iss.it/dspace.

In adesione alle Raccomandazioni europee, l'ISS si è attivato con iniziative e progetti nazionali e internazionali di promozione della cultura dell'accesso aperto: nel 2013 è tra i primi firmatari del position paper sull'accesso aperto ai risultati e ai dati della ricerca scientifica in Italia http://www.cnr.it/sitocnr/Iservizi/Biblioteche/PositionAccessoAperto.html, firmato dai principali enti di ricerca italiani (CNR, ENEA, INFN, ecc. e dalla CRUI) su iniziativa del progetto MedOAnet e partecipa in qualità di stakeholder della ricerca ad una consultazione della Commissione europea sugli open data; nel 2014 firma la "Dichiarazione di Messina 2.0: la via italiana all'accesso aperto" http://decennale.unime.it/?page_id=2039 per confermare insieme agli atenei e agli enti di ricerca italiani il sostegno all'attuazione di politiche istituzionali per lo sviluppo dell'accesso aperto.

2. Verso una policy di gestione dei dati della ricerca

A giugno 2016, sotto il coordinamento dell'ISS è stato istituito, in seno al sistema Bibliosan, il Gruppo di lavoro BISA "Bibliosan per la scienza aperta", con l'obiettivo, in primis, di elaborare un questionario sulle pratiche di gestione dei dati della ricerca in ambito biomedico in Italia. I risultati del Questionario, diffuso a inizio 2017 agli enti Bibliosan, sono stati illustrati nel corso di un Convegno organizzato presso l'Istituto Superiore di Sanità il 15 maggio 2017 (Gruppo di lavoro Bibliosan per la Scienza Aperta, BISA, 2017), Fig. 1. Dall'indagine BISA è emerso il profilo di una comunità di ricerca consapevole delle pro-

blematiche e delle aspettative legate alla gestione dei dati aperti della ricerca, in merito, soprattutto, alla fruizione di linee guida o politiche di indirizzo e all'individuazione, al contempo, di professionalità ad hoc (Tab. 1).
Fattori strutturali (strumentazioni e tecnologie) e gestionali (linee di indirizzo e coordina-

Fig. 1
Aula Pocchiari, Istituto Superiore di Sanità, sede del Convegno "I dati aperti, cemento della scienza: risultati dell'indagine Bibliosan per la Scienza Aperta (BISA)", Roma 15 maggio 2017

Tab. 1
Politiche istituzionali sulla gestione dei dati e iniziative attese (Indagine BISA)

Azione istituzionale attesa	**Risposte**	**%**
Linee guida/politiche di indirizzo	1.537	48,75
Previsione di una professionalità *ad hoc* per la gestione dei dati della ricerca	1.494	47,38
Altro	122	3,87
Totale	**3.153**	**100**

Fonte: dati ricavati dal Questionario BISA e in corso di pubblicazione nella serie "Rapporti ISTISAN" editi dall'Istituto Superiore di Sanità.

mento centrale) sono risultati i motori per attivare la cultura della condivisione, al riparo da pratiche burocratiche. L'orientamento è verso la richiesta di attuazione di una legislazione nazionale competitiva, che superi i legalismi e promuova concretamente il trattamento dei dati, soprattutto nella casistica di studi di coorte, caso controllo, multicentrici e di sorveglianza epidemiologica che rappresentano il nerbo della ricerca biomedica. Una volta analizzate, le criticità di gestione dei dati della ricerca andrebbero eliminate attuando una regolamentazione che ne disciplini il trattamento e il riutilizzo. Un'ipotesi potrebbe essere quella di istituire un database dei dati generati dalle ricerche, ripartiti in settori tematici. Si attuerebbe così la catalogazione del patrimonio di dati raccolto onde favorirne il riutilizzo e l'adeguata valorizzazione. Attuando un'infrastruttura informatica deputata a ospitare i dati a livello nazionale, si offrirebbe la possibilità di massimizzare la circolazione dell'informazione scientifica, liberando così il potenziale della ricerca biomedica a beneficio della più vasta platea di ricercatori e scienziati.

Attualmente l'ISS sta elaborando una bozza di policy sulla gestione dei dati della ricerca, con l'obiettivo di recepire le aspettative dei ricercatori dell'Ente, di convalidare le buone pratiche in uso e di costituire un'infrastruttura organica interna per la registrazione, la protezione e la condivisione dei dati. Si tratta di agevolare i responsabili di progetto aderenti ai bandi di ricerca Horizon 2020 che sono tenuti a depositare ad accesso aperto, tra-

mite un Data Management Plan, i cosiddetti "underlying data", raccolti o generati durante la ricerca e necessari per validare i risultati presentati negli articoli.
Gli argomenti attualmente in discussione sono:

- la natura dei dati (dati già rilasciati al pubblico o non ancora pubblicati);
- l'accesso ai dati, nei casi di ricerca finanziata con fondi pubblici o non pubblici;
- la distinzione tra dati generati internamente e dati acquisiti;
- la questione dei vincoli all'accesso (protocolli d'intesa, brevetti, consenso del paziente per i dati clinici;
- il funzionamento di una infrastruttura istituzionale con riferimento all'integrazione di tipi e formati di dati, all'uso di descrittori comuni, ai livelli di autorizzazione di chi vi opera.

L'iniziativa di delineare i contenuti di una policy sui dati della ricerca si è avvalsa della stesura approntata in sede CRUI da un Gruppo di lavoro ad hoc (GdL Dati della ricerca, 2017) ed è stata integrata da elementi che riflettono la specificità della mission istituzionale dell'ISS. Nell'iter di definizione di tale documento si prevede di raccogliere le indicazioni dei ricercatori ISS, per poi pervenire all'approvazione dei vertici dell'Ente, anche in raccordo con gli orientamenti e le aspettative del Ministero della Salute. Esaurita questa fase, si potranno accogliere, su una base di valutazione condivisa, le integrazioni al documento da parte delle comunità di ricerca degli enti Bibliosan. Sarà molto proficuo istituire una base di confronto con il mondo accademico e la comunità di ricerca internazionale per approdare a un documento inclusivo di tutti gli aspetti e le implicazioni del trattamento dei dati della ricerca.

3. Conclusioni

L'attività del Gruppo di lavoro BISA, istituito nel 2016, testimonia la vitalità dei tanti colleghi che nelle strutture informative degli enti di ricerca del sistema Bibliosan si adoperano per attuare concretamente l'accesso e la condivisione ai risultati della ricerca (pubblicazioni e dati).

Tra le prime iniziative del Gruppo c'è stata l'esigenza di coinvolgere attivamente la comunità dei ricercatori e le varie componenti istituzionali dei singoli enti. Si tratta infatti di avviare un processo comune di ricognizione e potenziamento delle risorse interne, rispetto alla sfida dell'e-science, un concetto associato sempre più ad una conoscenza scientifica ad alta densità di dati (Napolitani, 2017; Poltronieri et al.).

Con la diffusione dell'indagine BISA presso gli enti Bibliosan, nei primi mesi del 2017, si è compiuta una prima fase di lavoro centrata sulla rilevazione delle pratiche legate al trattamento dei dati della ricerca nei suoi vari aspetti. Le criticità rilevate dai rispondenti al Questionario BISA, relativamente a responsabilità manageriali, formazione e infrastrutture informatiche, hanno delineato un quadro di vaste aspettative in termini di servizi e linee guida sulla gestione dei dati generati e/o acquisiti dagli enti del comparto biomedico. Ne è derivato l'auspicio di una gestione organica dei dati della ricerca, terreno sul quale si è innestato un ulteriore piano d'azione del Gruppo di lavoro BISA, tuttora in atto: la definizione di un documento di policy circostanziato quanto a tipologia dei dati, ambiti di applicazione, responsabilità dei ricercatori, aspetti bioetici e legali associati all'accesso, al

deposito e al riuso dei dati della ricerca.

Riferimenti bibliografici

Gruppo di lavoro Bibliosan per la Scienza Aperta (BISA). Convegno I dati aperti cemento della scienza: risultati dell'indagine Bibliosan per la Scienza Aperta (BISA). Notiziario dell'Istituto Superiore di Sanità 2017;30(5):19-21.

http://www.iss.it/binary/publ/cont/ONLINE_MAGGIO.pdf Le relazioni presentate al convegno sono disponibili all'indirizzo: http://www.bibliosan.it/ftp/bisa_atti_15052017/bisa_15_05_2017.html

GdL Dati della ricerca. Modello di Policy sulla gestione dei dati della ricerca. Documento elaborato dal gruppo di lavoro informale sui dati della ricerca costituito da Politecnico di Milano, Università di Milano, Università di Torino, Università di Trento, Università di Venezia Ca' Foscari, Marzo 2017.

http://wikimedia.sp.unipi.it/images/RDMpolicyresearchdata27_03_2017.pdf

Napolitani F. Open data: another step towards open science. Journal of EAHIL 2017;13(2):2.

Poltronieri E, Cognetti G, La Placa P, Vullo A, De Castro P, Ferrara R, Mancini C, Toni F. Let your open data smoothly settle in your publishing habits. In: Diversity in Practice: integrating, inspiring and innovative. Book of abstracts of the 12th International Congress on Medical Librarianship and the 2017 EAHILWorkshop. Dublin (Ireland), June 12-16, 2017. pp. 20-21.

Autori

Moreno Curti curtim@smatteo.pv.it
Coordinatore Nazionale di Bibliosan, la rete delle Biblioteche Biomediche vigilate dal Ministero della Salute. Dirigente Medico presso la Direzione Scientifica della Fondazione IRCCS Policlinico San Matteo di Pavia e responsabile della struttura semplice Grant Office e Documentazione Scientifica

Paola De Castro paola.decastro@iss.it
Responsabile delle attività editoriali dell'Istituto Superiore di Sanità e di numerosi progetti di ricerca multidisciplinari volti a favorire la disseminazione delle conoscenze scientifiche a target diversi. Sviluppa e promuove strategie di comunicazione con lo scopo di utilizzare le evidenze scientifiche per favorire processi di conoscenza e contribuire ai percorsi di salute.

Corrado Di Benedetto
corrado.dibenedetto@iss.it
Laureato in Fisica presso l'Università degli Studi di Roma "La Sapienza", responsabile del Servizio di Informatica dell'Istituto Superiore di Sanità.

Rosalia Ferrara rosalia.ferrara@iss.it
Responsabile dei servizi all'utenza presso la Biblioteca dell'Istituto Superiore di Sanità

(ISS). Coordina il Gruppo di lavoro per la valorizzazione e la conservazione del Fondo Rari della Biblioteca ISS.

Pietro La Placa pietrolaplaca@izssicilia.it

Laureato in Scienze dell'Informazione presso l'Università degli Studi di Catania, responsabile del Servizio Editoria e Biblioteca dell'Istituto Zooprofilattico Sperimentale della Sicilia.

Cristina Mancini cristina.mancini@iss.it

Coordina presso la Biblioteca dell'Istituto Superiore di Sanità l'attività della Sezione Pubblicazioni in serie. E' membro del gruppo di lavoro Bibliosan per la scienza aperta (BISA). E' referente per le attività di formazione proposte dalla Biblioteca, direttore/docente nei corsi ECM, coordinatore dei Seminari in Biblioteca per l'utenza interna, tutor nei percorsi di Alternanza scuola/lavoro.

Luisa Minghetti luisa.minghetti@iss.it

Neurobiologa, lavora dal 1992 all'Istituto Superiore di Sanità, dove ha svolto ricerche su meccanismi infiammatori alla base delle malattie neurodegenerative e diretto fino al 2016 il Reparto Neurologia Sperimentale. Attualmente, vi dirige il neocostituito Servizio per il Coordinamento e Supporto alla Ricerca, dedicato alla promozione della ricerca biomedica e sanitaria nazionale e internazionale.

Elisabetta Poltronieri elisabetta.poltronieri@iss.it

Coordina presso il Servizio Conoscenza e Comunicazione Scientifica dell'Istituto Superiore di Sanità (ISS), la registrazione delle pubblicazioni scientifiche dell'Ente. In collaborazione con il Settore informatico dell'ISS, cura la gestione dell'archivio digitale aperto DSpace ISS. E' coordinatore del Gruppo di lavoro BISA (Bibliosan per la Scienza Aperta) che promuove i principi dell'accesso, riuso e condivisione dei dati della ricerca.

Filippo Santoro filippo.santoro@iss.it

Laureato in Fisica presso l'Università degli Studi di Roma "La Sapienza", è responsabile, presso il Servizio per il Coordinamento e Supporto alla Ricerca dell'Istituto Superiore di Sanità, di piattaforme informatiche per la collezione di dati clinici nell'ambito di progetti di ricerca.

Franco Toni franco.toni@iss.it

Bibliotecario dal 1985 presso la Biblioteca nazionale centrale di Roma e dal 1994 responsabile dell'Ufficio automazione e di SBN. Ricopre dal 1999 il ruolo di Direttore della Biblioteca dell'Istituto Superiore di Sanità (ISS) e dal 2015 è responsabile del Servizio Conoscenza e Comunicazione Scientifica dell'ISS. E' inoltre membro del Comitato di gestione di Bibliosan dal 2005.

Angela Vullo angela.vullo@izssicilia.it

Statistico dell'Istituto Zooprofilattico Sperimentale della Sicilia, coordina le attività della Direzione generale. Nell'ambito della sua attività professionale ha applicato metodi statistici avanzati e risolto problemi analitici in diversi settori disciplinari: economia, sociologia, medicina, biologia, veterinaria.

DOI: 10.26314/GARR-Conf17-proceedings-14

A DMP template for Digital Humanities: the PARTHENOS model

Sheena Bassett[1], Sara Di Giorgio[2], Franco Niccolucci[1], Paola Ronzino[1]

[1]Pin, Prato, [2]ICCU

Abstract. PARTHENOS, stands for "Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies" and it is a Horizon 2020 project funded by the European Commission. It's an umbrella-project which includes both legal entities of the CLARIN and DARIAH ERICs and also their affiliated projects, including EHRI, ARIADNE, CENDARI, E-RIHS and others in the related fields across Humanities and Cultural Heritage. PARTHENOS is working on the definition and support of common standards, the coordination of joint activities, the harmonization of policy definition and implementation, and the development of pooled services and of shared solutions to the same problems. Within this framework, it has been working on the definition of a Data Management Plan (DMP) template, that will support researchers form Humanities and Cultural domains to effectively share their data according to the FAIR Data Principles.

Keywords. DMP, Research Infrastructure, Digital Humanities, FAIR Principles

Introduction

The PARTHENOS project empowers digital research in the fields of History, Language Studies, Cultural Heritage, Archaeology, and related fields across the (Digital) Humanities. PARTHENOS supports the work of relevant research infrastructures, such as CLARIN (language resources) and DARIAH (digital humanities), ARIADNE (digital archaeological research infrastructure), EHRI (European Holocaust research infrastructure), CENDARI (digital research infrastructure for historical research), CHARISMA and IPERION-CH (EU projects on heritage science), along with other relevant integrating activities projects, by building a cross-disciplinary virtual environment to enable researchers of the humanities to have access to data, tools and services based on common policies, guidelines and standards. As part of this work, one task is the implementation of a series of recommendations and guidelines based on FAIR Data Principles, targeted at researchers and research centres about which policies to apply during and after their research or infrastructure work.

Within this framework, PARTHENOS has been working on a Data Management Plan (DMP) template which is based upon the Horizon 2020 template. Indeed since 2016, the Horizon 2020 Programme has produced Guidelines on FAIR Data Management to support Horizon 2020 beneficiaries in making their research data findable, accessible, interoperable and reusable (FAIR). Funded projects are requested to deliver an implementation of Data Management Plan (DMP) which aims to improve and maximise access to and reuse of research data generated by the projects. This is part of the Commission's policy actions on Open Science to reinforce the EU's political priority of fostering knowledge circulation. Open Science is in practice about "sharing knowledge as early as practically possible in the discovery process' and because DMPs gather information about what data

will be created and how, and outline the plans for sharing and preservation, specifying the nature of the data and any restrictions that may need to be applied, DMPs ensure that research data is secure and well-maintained during a project and beyond, when it might be shared with others.

1. The PARTHENOS DMP

A first draft of DMP has been presented in the "Report on Guidelines for Common Policies Implementation' , that gives an overview of existing policies concerning data management as well as policies regarding quality of data, metadata and repositories, IPR, open data and open access, presented according to FAIR Data Principles. The PARTHENOS DMP and Guidelines will support researchers to effectively share their data according to the FAIR Data Principles through publicly accessible, digital repositories.

The PARTHENOS DMP will offer a clear guidance to its stakeholders, mainly researchers and data repositories, to plan the life cycle of data. It will offer a long-term perspective by outlining how data will be generated, collected, documented, shared and preserved, taking into consideration commonalities and specific requirements of the disciplines involved in the project. The PARTHENOS DMP template comprises sections about data collection and documentation, ethics, legal and security issues, data storage and preservation, and data sharing and reuse.

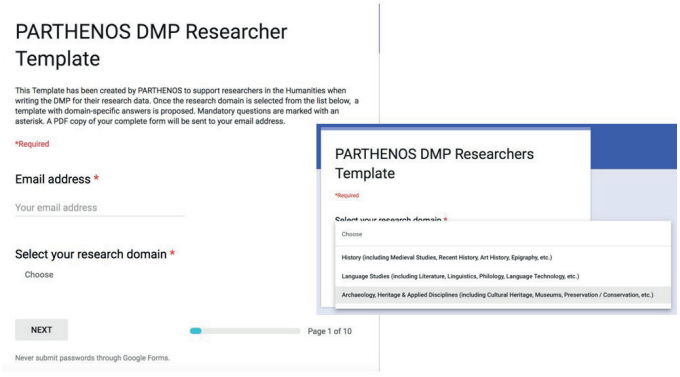


Fig. 1
The PARTHENOS DMP Template on-line

The PARTHENOS DMP template provides a guidance to help researchers/data repositories for completing their data management plan and it is composed of two levels: a basic template that can be applied to all the disciplines in the humanities, as well as the closely related stakeholders: cultural heritage institutions, data archives and research infrastructures, and a domain level that contains more detailed questions where policies, specific standards and practices may vary according to the research area. Moreover, some of the questions will only apply to researchers or to data repositories. The PARTHENOS DMP template will also encourage researchers to follow common policies and use best practice for their data management.

2. Conclusions

The PARTHENOS DMP template will be made available as an online tool to support better researchers to freely access, mine, exploit, reproduce and disseminate their data and identify the tools needed to use the raw data for validating research results, or provide the tools themselves, a significant step towards the realization of Open Science.

References

ARIADNE, www.ariadne-infrastructure.eu

CARISMA, http://www.charismaproject.eu/transnational-access.aspx

CENDARI, www.cendari.eu/

CLARIN, https://www.clarin.eu/

DARIAH, http://www.dariah.eu/

EHRI, https://www.ehri-project.eu/

FAIR principles, https://www.force11.org/group/fairgroup/fairprinciples

H2020 Programme, Guidelines on FAIR Data Management in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oapilot/h2020-hi-oa-pilot-guide_en.pdf

IPERION, www.iperionch.eu/

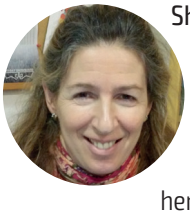
Open Innovation, open science, open to the world, produced by the European Commission's Directorate-General for Research & Innovation, 2016-05-17, https://publications.europa.eu/it/publication-detail/-/publication/3213b335-1cbc-11e6-ba9a-01aa75ed71a1

PARTHENOS, https://www.parthenos-project.eu

Wilkinson, M. D. et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, DOI: 10.1038/sdata.2016.18

PARTHENOS, 2017, Report on Guidelines for Common Policies Implementation: available at https://goo.gl/N6UxZZ

Authors



Sheena Bassett sheena.giess@gmail.com

Sheena Bassett is the Project Manager for the PARTHENOS project and works for PIN Scrl, Prato, Italy. She has been involved in research for over 30 years, having started out a computer programmer and subsequently working as a researcher and consultant in a number of industries such as image analysis, publishing and packaging before moving into cultural heritage. Her wide and varied work experience helps contribute to her understanding of all aspects of research, from technical through to the everyday activities.

Sara Di Giorgio sara.digiorgio@beniculturali.it

Since 2004, she collaborates with ICCU in developing CulturaItalia, the portal of the Italian culture, as well as many European projects for cultural heritage digitalization, on-line access and digital preservation. She participates in PARTHENOS by coordinating the task on definition of policy requirements concerning the data lifecycle.

Franco Niccolucci franco.niccolucci@pin.unifi.it

Franco Niccolucci is the director of the VAST Lab Laboratory at the PIN in Prato, and the Coordinator of the European project PARTHENOS, that is a cluster of research infrastructures in the Digital Humanities and Cultural Heritage fields. He coordinated the ARIADNE infrastructure (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) in the digital archaeology sector, which created a registry of over 2 million archaeological datasets. In the past, it has coordinated various European projects in the cultural heritage domain. As a mathematician, he was a professor at the University of Florence until 2008 and then, until 2013, he was the director of the Science and Technology in Archaeology Research Center at the Cyprus Institute in Nicosia, Cyprus.

Paola Ronzino paola.ronzino@pin.unifi.it

Paola Ronzino is a Post-doc researcher at PIN specialized in ITC for Cultural Heritage. She holds a Master Degree in Archaeology and a PhD in Science & Technology in Cultural Heritage. Her research interests are concerned with the development of ontologies and metadata standards for archaeological documentation and cultural heritage, with interest in the documentation of buildings archaeology. Participating in several of PIN's EU projects on digital cultural heritage, she has been actively involved in the research activities of the ARIADNE project. Currently, she is engaged in the PARTHENOS project activities, particularly in what concerns the definition of the user requirements and the design of the PARTHENOS Data Management Plan for the Humanities.

DOI: 10.26314/GARR-Conf17-proceedings-15

Data Management per la ricerca: un approccio metodologico

Paola Galimberti[1,4], Jordan Piščanc[2,4], Susanna Mornati[3,4]

[1]Università degli Studi di Milano, [2]Università degli Studi di Trieste, [3]4Science, [4]IOSSG

Abstract. Il contesto europeo pone la gestione dei dati della ricerca e la loro corretta archiviazione e conservazione per possibili ed eventuali utilizzi futuri al centro delle politiche sulla research integrity. La possibilità di riprodurre i risultati delle ricerche pubblicate è diventata infatti uno degli elementi essenziali nelle politiche della ricerca della Commissione Europea per poter restituire una credibilità alla scienza, da tempo minata da male pratiche e frodi scientifiche.
Queste tematiche hanno trovato fino ad ora poco riscontro in Italia, sia da parte dell'Ente finanziatore unico della ricerca (il MIUR) sia da parte delle istituzioni.
L'intervento descrive il percorso fatto bottom up da un gruppo di atenei per cercare di fornire ai propri ricercatori politiche e strumenti che avvicinino le nostre pratiche a quelle del resto dei paesi europei.

Keywords. Research Data Management, FAIR data

Introduzione

Il contesto europeo pone la gestione dei dati della ricerca e la loro corretta archiviazione e conservazione per possibili ed eventuali utilizzi futuri al centro delle politiche sulla research integrity. La possibilità di riprodurre i risultati delle ricerche pubblicate è diventata infatti uno degli elementi essenziali nelle politiche della ricerca della Commissione Europea per poter restituire una credibilità alla scienza, da tempo minata da male pratiche e frodi scientifiche.

Le iniziative legate alla gestione dei dati sono numerosissime e sono stati messi a disposizione una serie di toolkit: dalla roadmap sui research data della LERU ai risultati del progetto LEARN, dai Data Management Plan (cartacei o online) alle linee guida FAIR. L'idea è quella di mettere a fattor comune esperienze e strumenti che favoriscano, almeno in Europa, l'adozione di comportamenti coerenti e condivisi.
La necessità di collegare i risultati delle ricerche con i dati che ne stanno alla base è molto sentita nei Paesi "research intensive" dove sono stati organizzati servizi centralizzati sia dal punto di vista delle infrastrutture che da quello delle politiche, linee guida, supporto legale.

In Italia la necessità di indicazioni (tecniche e legali), di formazione (di figure esperte nel Research Data Management) e soprattutto di infrastrutture deputate alla raccolta archiviazione e conservazione dei dati, e' molto sentita dai ricercatori che spesso si trovano ad affrontare le richieste della Commissione Europea senza strumenti, senza indicazioni precise e senza infrastrutture di supporto, mentre il MIUR per ora non ha ancora dato

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

alcun segnale d'interesse rispetto a questo argomento così rilevante.

Nell'attesa che il Ministero elabori politiche e indicazioni sul Research Data Management, un gruppo di lavoro costituito da atenei in cui la sensibilità verso il tema della gestione dei dati della ricerca era maggiormente sentito ha intrapreso un percorso per la definizione di strumenti condivisi (e da condividere) in particolare un modello di policy sul Research Data Management, un modello di Data Management Plan, un'ipotesi di infrastruttura per la raccolta archiviazione e conservazione dei dati.

1. Dalla teoria alla pratica: un percorso di apprendimento

Dalle linee guida di OpenAire si evince che per soddisfare le direttive della Commissione Europea recentemente espresse anche nella EOSC declaration per quel che riguarda la gestione dei dati della ricerca ci si può anche avvalere di Data Repository generici quali Zenodo o FigShare. Numerosi sono anche i repository tematici di dati che possono essere usati per la gestione del ciclo di vita dei dati. Si ripresenta però la stessa situazione di quando, più di 10 anni fa, si iniziò a gestire l'accesso aperto alle pubblicazioni scientifiche e la loro archiviazione negli archivi istituzionali. Per usare archivi generici quali Zenodo o Repository tematici bisogna in ogni caso "accettare" i loro termini d'uso e sottostare ai loro limiti (ad esempio una dimensione massima di spazio per utente usando account gratuiti). Se poi i ricercatori usano archivi diversi per i propri progetti si rischia di avere una frammentazione della produzione dei dati generati.

Si presenta perciò la necessità di istituire e creare anche per le gestione dei dati della ricerca una infrastruttura istituzionale che sia in grado di garantire un'identità e un presidio continuo fornendo altresì servizi a valore aggiunto. In questo caso la messa in produzione di un archivio per i dati della ricerca che gestisca tutto il ciclo di vita (dai dati grezzi, attraverso le diverse versioni, alla descrizione, fino al Data Management Plan) fino ad arrivare al dataset definitivo, risulta ben più onerosa.

Guardando alle esperienze già consolidate in altri paesi europei, dove ci sono buone pratiche e realtà già in atto, ci si rende conto che per gestire efficientemente tutto il ciclo di vita dei dati della ricerca diventa una scelta strategica "fare rete" e instaurare la collaborazione tra più organizzazioni (come ad esempio centri di ricerca e università), come nei Paesi Bassi dove 10 Università e diversi centri di ricerca hanno aderito e usano un Data Research Repository condiviso: DataVerseNL gestito dal DANS.

2. L'esperienza dell'università di Milano e di Trieste

In questa prospettiva le Università di Milano e Trieste insieme ad altri Atenei stanno lavorando a un progetto comune per sperimentare una soluzione di archivio dei dati della ricerca che possa rispondere alle esigenze di comunità scientifiche anche molto diverse fra di loro. Il gruppo di lavoro ha il compito di definire i requisiti di un sistema per la gestione dei dati della ricerca da parte di gruppi disciplinari diversi e di definire, monitorare e validare le attività di test coinvolgendo ricercatori e gruppi di ricerca provenienti da aree diverse. I requisiti saranno analizzati sia dal punto di vista dell'infrastruttura IT e delle risorse umane a supporto, sia dal punto di vista legale che di policy.

L'università di Milano è stata fra le prime in Italia ad adottare un archivio istituzionale per le pubblicazioni e il lavoro sui dati appare il naturale proseguimento di un'attività di supporto all'apertura e alla trasparenza dei risultati della ricerca che si è consolidata nel corso degli anni.

Trieste sarà Capitale Europea della Scienza ESOF 2020 e per l'Università di Trieste l'adozione di una solida soluzione di archiviazione dei dati della ricerca diventa di importanza strategica. L'Università di Trieste è anche coinvolta in UnityFVG, progetto di cooperazione tra gli Atenei del Friuli Venezia Giulia che prevede tra l'altro anche la condivisione delle risorse e delle best practices della ricerca. Sperimentare perciò una soluzione di archivio dei dati della ricerca porterebbe vantaggio anche agli altri Atenei della Regione e potrebbe coinvolgere un maggior numero di Ricercatori.
L'università di Milano ha inoltre coordinato il gruppo di lavoro che ha portato alla definizione di un modello di Policy per il research data management ora a disposizione dell'intera comunità italiana. Sia Trieste che Milano hanno aderito a IOSSG, il gruppo di lavoro che ha il compito di fornire supporto ai ricercatori italiani sulle questioni legate ai dati della ricerca.

Lo strumento scelto da Milano e Trieste per il progetto pilota è Dataverse.
Si è arrivati a questa scelta dopo aver svolto una serie di interviste nei dipartimenti che hanno messo in luce le esigenze principali dei ricercatori. Si sono considerati una serie di strumenti adatti a soddisfare tali esigenze e la scelta è caduta su Dataverse perché è un software open source, che può contare su una comunità di sviluppatori ampia e su un grandissimo numero di utilizzatori in tutto il mondo.

Il gruppo di lavoro si avvale del supporto tecnico di 4Science, azienda specializzata nel fornire soluzioni open source per la ricerca, con cui si cercherà di tradurre in pratica le esigenze espresse nei due atenei.

Dataverse è un'applicazione web open source per condividere, conservare, citare, esplorare e analizzare i dati della ricerca, sviluppato dall'Institute for Quantitative Social Science della Harvard University. Permette di mettere a disposizione della comunità scientifica i propri dati, aumentandone la visibilità. Facilita inoltre il riutilizzo dei dati stessi e, di conseguenza, la replicabilità delle ricerche.

Tramite Dataverse, i ricercatori possono organizzare, condividere e conservare i propri dati, corredati di metadati descrittivi, possono gestirne le diverse versioni e attraverso citazioni formali ricevere credito (citazioni) e adempiere alle richieste dei finanziatori della ricerca. I dati restano sotto il controllo del ricercatore che decide cosa e quando rendere pubblico; la piattaforma è inoltre interoperabile con altre fonti di dati attraverso il protocollo OAI-PMH.

Dataverse mette a disposizione funzionalità avanzate per l'analisi dei dati tabellari, mediante l'integrazione con l'applicazione "TwoRavens". L'interfaccia per le analisi statistiche fornita da TwoRavens è utilizzabile da un'utenza con diversi livelli di competenze statistiche e si configura anche come possibile strumento didattico. Infatti mediante tale interfaccia è possibile visualizzare statistiche di base relative alle variabili che caratterizzano il dataset, effettuare analisi su un sottoinsieme di valori e testare modelli statistici.

TwoRavens è stato concepito proprio come uno strumento per aumentare il numero di utenti in grado di effettuare ragionamenti di tipo quantitativo, mettendo a disposizione funzionalità di analisi che non hanno necessità di grandi infrastrutture e un'interfaccia grafica mediante la quale effettuare le analisi.

Viene supportata anche l'analisi dei dati geospaziali (shapefile) che possono essere esplorati e manipolati attraverso l'integrazione con WorldMap, uno strumento per la visualizzazione e l'analisi di dati geospaziali, sviluppato dal "Center for Geographic Analysis" dell'Università di Harvard.

Attraverso un apposito plug-in, infine, è possibile collegare un repository Dataverse a una specifica rivista gestita attraverso OJS (Open Journal System), in modo da consentire agli autori di sottomettere, oltre all'articolo, anche i dataset ad esso collegati.

3. Conclusioni

Il progetto è appena iniziato ed è presto per poter fare un bilancio. Allo stato attuale sono molto chiare le esigenze di formazione e la richiesta di competenze specifiche all'interno degli atenei che possano supportare i ricercatori nella attività di gestione dei dati.
La sperimentazione che durerà circa sei mesi dovrà aiutare il gruppo di lavoro a capire se Dataverse sia lo strumento in grado di rispondere alle esigenze di trasparenza, ricercabilità, e riusabilità ormai diventate urgenti anche qui da noi.

Autori

Paola Galimberti paola.galimberti@unimi.it
Paola Galimberti si occupa di accesso aperto, qualità dei dati, cura gli strumenti a supporto della valutazione interna ed esterna della ricerca, di etica e integrità della ricerca.

Jordan Piščanc piscanc@units.it
Responsabile IT all'Università di Trieste degli Archivi Istituzionali e sistemi CRIS OpenstarTS, ArTS. Attivo da più di 10 anni nella community DSpace. Il focus principale della sua attività sono gli Open Archive e infrastrutture DSpace-CRIS/GLAM. Segue con molto interesse gli argomenti di Open Science ed è membro di IOSSG.

Susanna Mornati susanna.mornati@4science.it
Direttore Operativo a 4Science S.r.l., ha un'esperienza trentennale in sistemi informativi per la ricerca e la gestione di progetti complessi come l'implementazione di un nuovo Research Information Management System basato su DSpace-CRIS in oltre 60 enti e l'adozione nazionale di ORCiD nel 2015. Susanna ha una reputazione internazionale come sostenitrice dell'Open Access e Open Science ed è membro di IOSSG.

DOI: 10.26314/GARR-Conf17-proceedings-16

Biblioteche accademiche e data literacy: un primo (parziale) rapporto dall'Italia

Anna Maria Tammaro

Università di Parma, International Master DILL

Abstract. Vengono presentati i risultati di una prima indagine sul supporto dato dalle biblioteche accademiche per aumentare la consapevolezza dei ricercatori sui dati della ricerca in Italia. Un questionario in linea è stato inviato ad un gruppo selezionato di biblioteche accademiche che sono state pioniere nell'estendere i servizi tradizionali a supporto dell'intero ciclo della ricerca. Ci sono numerose varianti di corsi e la presentazione descrive le prime esperienze di "Data Literacy" realizzate in Italia.

Keywords. Biblioteche accademiche, Research Data Management, Data Literacy.

Introduzione

Dalla fine degli anni '90, le biblioteche universitarie in Italia hanno realizzato un processo continuo di cambiamento, anche se non guidato da una chiara strategia. Si possono così indicare alcune innovazioni importanti, come il sostegno dato alle politiche dell'Accesso Aperto dopo la Dichiarazione di Messina (2004) a cui hanno aderito circa 70 Atenei italiani. Sono anche numerosi i depositi istituzionali che sono stati aperti soprattutto dalle biblioteche universitarie (100 circa sono elencati nel Registro Open Doar, anche se molti sembrano inattivi). Nel 2006 la CRUI ha aperto un sottogruppo sull'Accesso Aperto che fino ad oggi ha predisposto linee guida sui periodici, le tesi ed i depositi istituzionali, ma non sui dati di ricerca. Dal 2013, la maggioranza delle Università italiane utilizza IRIS (Institutional Research Information System), usato in collegamento ai depositi istituzionali da circa 60 Atenei. Tuttavia, le Università che si sono dotate in Italia di una "Policy" per l'Accesso Aperto registrate nel sito Wiki OA Italia sono appena 25. Ancora poche sono le biblioteche accademiche che offrono un servizio di supporto per i dati di ricerca, ed a riprova di ciò solo 4 istituzioni sono elencate in Open Doar per i dati di ricerca accessibili per il ri-uso.

Tra le biblioteche accademiche in Italia, alcune sono state pioniere per la gestione dei risultati della ricerca e si sono assunte la responsabilità di un'estensione dei servizi alla gestione dei dati di ricerca (Research Data management - RDM). Un primo Gruppo di lavoro informale sui dati di ricerca è stato costituito dalle Biblioteche delle Università di Milano e del Politecnico di Milano, delle Università di Venezia, Padova, Torino, Trento. Altre iniziative sui dati di ricerca sono state avviate dalle Biblioteche dell'Università Bologna e da Biblioteche speciali come la Biblioteca dell'Area di ricerca CNR di Bologna e la Biblioteca dell'Istituto Superiore di Sanità. Altre iniziative sono in corso di sviluppo e la fonte più aggiornata per monitorare il ruolo delle biblioteche accademiche per i dati di ricerca è il sito Wiki OA/Italia.

1. Ciclo della ricerca e Data literacy

La ricerca scientifica in ambito digitale è per sua natura un processo sociale e si basa sul cambiamento di comportamento dei ricercatori, che sempre più hanno bisogno della condivisione dei risultati e della collaborazione spesso interdisciplinare per creare una base di conoscenza comune. Penso ad esempio alla ricerca sull'energia che si fa nel CERN o alla ricerca sulla terra di DataOne, ma anche a ricerche in campo umanistico su corpora di testi, per edizioni critiche e così via. Come conseguenza di questo cambiamento di comportamento dei ricercatori, i modelli di servizio delle biblioteche accademiche devono essere centrati sul ciclo della ricerca; inoltre i servizi delle biblioteche innovative sono fruiti spesso "fuori" dalle mura delle biblioteche, con nuove partnership con i ricercatori e in collaborazione con altri settori delle istituzioni universitarie, come centri di calcolo ed uffici della ricerca.

Le attività previste dai servizi di supporto al ciclo della ricerca (come pianificate dal JISC) si dividono in tre blocchi: Data Governance, Data Management, Data Guidance.
Data Governance include le attività di politiche, strategie e linee guida che devono regolare e gestire il servizio di supporto.

Data Management comprende tutte le attività a supporto della cura digitale e della gestione dei dati. In particolare i servizi previsti includono: Data Management Plan, Manging active data, Selection and appraisal, Data repositories, Data catalogues.
Data Guidance implica un servizio di orientamento personalizzato, con attività continue di informazione, e comprende attività di formazione identificate come Data literacy.

1.1 Data literacy

La gestione dei dati di ricerca è un metodo che consente l'integrazione, la cura e l'interoperabilità dei dati, vale a dire la produzione, l'accesso, la verifica, l'archiviazione persistente e il riutilizzo di questi dati con l'aiuto di strumenti adeguati e facili da usare, come piattaforme per la ricerca virtuale. Tutti i dati dovrebbero essere mantenuti disponibili nei tre diversi settori che uno scienziato deve avere disponibile per fare efficacemente la sua ricerca: un dominio privato, uno collaborativo ed uno pubblico.
Molti docenti hanno sentito parlare della gestione dei dati di ricerca solo recentemente, perché un Data Management Plan è stato richiesto da molti progetti finanziati dalla Commissione Europea. Tuttavia i ricercatori non sono generalmente consapevoli della necessità di gestire i dati lungo l'intero ciclo della ricerca.

1.2 Competenze e Obiettivi formativi

Nella letteratura professionale degli ultimi 5-7 anni, gli articoli sui corsi di Data Literacy sono in costante crescita. Sono numerosi gli esempi di "pedagogia creativa" realizzati dalle biblioteche accademiche, che affrontano la formazione dei ricercatori in modi interessanti ed innovativi. Sono state realizzate inoltre molte risorse educative aperte (OER) (come tutorial online, corsi di formazione MOOC, pagine web di guida) che consentono di parlare di un "curriculum" condiviso sulla Data literacy, anche se non ce n'è uno ancora "ufficiale".
La Data literacy è stata ritenuta molto vicina all'Information literacy: i corsi di alfabetizza-

zione informativa possono facilmente costituire la base per la creazione di un curriculum per la Data literacy. L'alfabetizzazione sulla Data literacy ha lo scopo principale di diffondere la consapevolezza dell'importanza, e in alcuni casi, della necessità per i ricercatori ed anche per gli studenti di saper gestire i propri risultati di ricerca, in collaborazione con servizi di supporto dell'istituzione.

Un primo obiettivo formativo fondamentale è quello di motivare i ricercatori a rendere i dati di ricerca disponibili ed a documentare il contesto dei dati per un futuro utilizzo, sensibilizzandoli all'interesse generale (data commons) di poter riusare i dati, per replicare e costruire ulteriore ricerca sui risultati pubblicati.

L'alfabetizzazione sulla Data literacy ha come ulteriore obiettivo formativo quello di rendere consapevoli i ricercatori delle politiche istituzionali e dei requisiti richiesti dalle fonti di finanziamento. Il Data Management Plan, che viene richiesto come obbligatorio da alcune fonti di finanziamento è uno degli esempi più diffusi su cui si concentrano i corsi disponibili.

Nel complesso, esiste un forte consenso nella letteratura professionale sulle competenze riguardo ai principali temi che dovrebbero essere affrontati nell'alfabetizzazione sulla Data literacy. Alcuni autori si sono concentrati sulle competenze di Data literacy (Qin e D'Ignazio 2010b, Carlson et al. 2011, Piorun et al. 2012, Calzada Prado e Marzal Miguel 2013, Schneider 2013); anche alcuni Progetti si sono concentrati nell'identificare le competenze di vari attori interessati, come DigCurV e l'International Digital Curation Education Action (IDEA) (Hank & Davidson, 2009).

Sono state quindi identificate dodici competenze che qui vengono associate ai servizi di supporto elencati sopra:

Tab. 1
Competenze associate al Ciclo della ricerca

	RDM policies	**Data catalogues**	**Data repositories**	**Selection**
Data Governance	Qualità dei dati e documentazione	Etica Citazione dei dati	Formati dei dati e banche dati	Preservazione dei dati
Data Management	Gestione e organizzazione dei dati	Metadati	Cura dei dati Interoperabilità Conversione dei dati	Ricerca e recupero dei dati Riuso
Data Guidance	Cultura specifica della comunità disciplinare	Descrizione contesto dei dati		Visualizzazione dei dati Analisi dei dati

Fonte: ispirato ai servizi di supporto JISC. http://www.dcc.ac.uk/resources/developing-rdm-services

2. Esperienze realizzate dalle biblioteche accademiche in Italia

Per una prima indagine sulle buone pratiche delle biblioteche accademiche in Italia per servizi di supporto sui dati di ricerca, un questionario è stato inviato via mail ai responsabili dei servizi bibliotecari. Gli obiettivi erano di capire i servizi di Data Guidance disponibili e i contenuti dei curriculum formativi. Alcuni esempi delle esperienze realizzate in Italia sono brevemente descritti di seguito.

Università di Venezia

L'Università di Venezia ha avviato un servizio di supporto sui dati di ricerca, composto da

un'infrastruttura tecnologica basata sui depositi istituzionali (ARCA e PHAEDRA) ed ha avviato una serie di attività formative e di consulenza.

Sono organizzati incontri, in collaborazione con il settore ricerca, per i nuovi ricercatori e docenti con frequenza periodica, per varie tematiche legate ai dati. Sono state anche create delle pagine Web di Guida che integrano il servizio in presenza di supporto e sono integrate da un'attività di consulenza personalizzata. In particolare, l'attività di orientamento e formazione si è concentrata sul bisogno di visibilità dei ricercatori, con strumenti come: Vademecum per la pubblicazione e la corretta citazione dei lavori, guide varie sul funzionamento di ORCID, SCOPUS, RESEARCHID, Google Scholar Citation. Ci si è inoltre concentrati sul bisogno di consapevolezza sulle politiche sull'Open Access e informazione sulle Linee guida esistenti. Una linea di supporto è sui temi legati al copyright con temi come: Copyright, schemi contrattuali depositati presso le segreterie dipartimentali, informazioni sulle licenze, brevetti ed utilizzo di banche dati brevettuali.

Istituto Superiore di Sanità

L'Istituto Superiore di Sanità coordina dal 2016 il Gruppo di lavoro BISA (Bibliosan per la scienza aperta). Ci si è preoccupati di capire i bisogni della comunità di utenti e la Biblioteca ha avuto come primo obiettivo la realizzazione di un'indagine per sondare le pratiche di archiviazione dei dati, gli aspetti legali di copyright e privacy, l'attitudine alla condivisione e le aspettative circa le politiche di gestione dei dati della ricerca nel comparto degli enti biomedici di ricerca affiliati al sistema Bibliosan.

La formazione ed i servizi di orientamento per i ricercatori sono affrontati sia con tutorial online che con attività di formazione in presenza. I temi affrontati nei corsi sono: come recuperare l'informazione, l'accesso e l'interrogazione delle banche dati, la valutazione della ricerca ed i relativi indicatori di analisi delle citazioni, il document delivery, come scrivere un articolo scientifico, come usare i servizi e le risorse di Bibliosan.

Collaborazione tra CNR e Università di Bologna, Parma, Torino e Trento

Una prima iniziativa di collaborazione tra università e CNR per la Data Literacy è stata svolta nel 2015 presso la Biblioteca dell'Area di ricerca del CNR di Bologna. Gli obiettivi formativi del Corso sono stati da un lato quello di far acquisire il valore dei dati aperti per una scienza più collaborativa, dall'altro quello di fornire gli strumenti per il Data Management Plan. Questo è infatti il servizio di base per tutti coloro che sono tenuti a offrire servizi di supporto alla gestione dei dati di ricerca.

3. Conclusioni

La barriera più importante alla gestione dei dati di ricerca che abbiamo identificato nelle risposte al questionario è la mancanza di competenze del personale, includendo sia i ricercatori che il personale di supporto. Ci sono numerose varianti di corsi sui dati di ricerca che sono stati resi disponibili ed in questo articolo abbiamo descritto solo alcune prime esperienze in Italia. C'è tuttavia la necessità di offrire a docenti e studenti una formazione più collegata al ciclo di ricerca, anche specifica per singole discipline. La Data literacy deve essere quindi più concentrata sui bisogni delle comunità di docenti e studenti, mentre questo è stato identificato come una lacuna nella formazione esistente (Swan e Brown 2008; Goldstein, 2010).

Riferimenti bibliografici

Calzada Prado, J. and Marzal Miguel, A. (2013), Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents, Libri 63(2): pp.123-134

Carlson J. R., Fosmire M., Miller C., Sapp N., Megan R. (2011) Determining Data Information Literacy Needs: A Study of Students and Research Faculty. Libraries Faculty and Scholarship and Research. Paper 23.

http://docs.lib.purdue.edu/lib_fsdocs/23

Goldstein S. (2010) Data management, information literacy and DaMSSI.

http://www.rin.ac.uk/our-work/researcher-development-and-skills/data-management-and-information-literacy

Davidson J. and Hank C. (2009) International data curation education action (IDEA) working group: a report from the second workshop of IDEA. D-Lib Magazine, 15 (3/4)

http://eprints.gla.ac.uk/80151/1/80151.pdf

Piorun M. E., D. Kafel, T. Leger-Hornby, S. Najafi, E. R. Martin, P. Colombo and N. R. LaPelle (2012) Teaching Research Data Management: An Undergraduate/Graduate Curriculum, Journal of eScience Librarianship 1(1): e1003.

http://dx.doi.org/10.7191/jeslib.2012.1003

Qin J. and J. D'Ignazio (2010), The Central Role of Metadata in a Science Data Literacy Course, Journal of Library Metadata, 10(2/3): pp.188-204

Schneider R. (2013) Research data literacy. In Worldwide Commonalities and Challenges in Information Literacy Research and Practice, Springer International Publishing, pp. 134-140.

Swan A. and Brown S. (2008) The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the JISC. Truro: Key Perspectives

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf

http://wikimedia.sp.unipi.it/index.php/OA_Italia/Regolamenti_e_Policy_sull%27Open_Access

http://wikimedia.sp.unipi.it/index.php/OA_Italia/Risorse_sugli_open_research_data

http://www.dcc.ac.uk/resources/developing-rdm-services

http://www.bibliosan.it/ftp/bisa_atti_15052017/bisa_15_05_2017.html

Autori

Anna Maria Tammaro annamaria.tammaro@unipr.it

Anna Maria Tammaro, PhD Information Science, è attualmente membro della Commissione IFLA Library Theory and Research e Segretario di ASIS&T European Chapter. Docente di "DILL International Master Digital Library Learning" Master congiunto con Tallinn University e l'Università di Parma, e del MOOC "Digital Library in Principle and Practice" nella piattaforma EMMA. Partecipa al Progetto ROMOR finanziato dalla Commissione Europea per costruire l'infrastruttura Open Access per gestire i risultati di ricerca in Palestina. Come Chair della Sezione IFLA Library Theory and Research ha realizzato un'indagine sul profilo a livello mondiale del Data curator.

DOI: 10.26314/GARR-Conf17-proceedings-17

Politiche e linee guida per la gestione dei dati della ricerca: l'esperienza di IOSSG

Paola Gargiulo

Cineca

Abstract. La relazione descrive brevemente il contesto in cui il gruppo di lavoro informale IOSSG – Italian Open Science Support Group si è costituito, gli obiettivi del gruppo e le attività che svolge. Il gruppo IOSSG è nato in ambito accademico, nell'ambito della promozione, dello sviluppo e la diffusione della scienza aperta in Italia, per fornire supporto e strumenti che rispondono alle sfide che la scienza aperta pone al mondo della ricerca, con particolare attenzione, ai processi della ricerca, al ciclo di vita e gestione dei dati, ai servizi e alle infrastrutture di supporto. Le attività del gruppo consistono nell'elaborare linee guida, toolkit, raccogliere buone pratiche, condurre analisi, studi, e inoltre produrre materiali e corsi di formazione sulle tematiche della scienza aperta.
Il programma di lavoro suddiviso in 4 sottogruppi contempla, per l'anno 2018, quattro linee di attività: la prima si concentra sulla produzione di linee-guida, raccomandazioni per l'implementazione di politiche istituzionali sulla gestione dei dati; la seconda sullo studio di modelli per la creazione di unico punto di accesso alle diverse competenze e risorse presenti nell'istituzione, al fine di costruire un efficiente servizio di sostegno trasversale alla ricerca; la terza sulla formazione, al fine di diffondere la conoscenza e promuovere l'acquisizione di competenze su tutti gli aspetti relativi alla gestione dei dati della ricerca, una quarta sullo studio e analisi di soluzioni per la creazione di infrastrutture a livello locale per la gestione dei dati della ricerca in un'ottica aperta e federata, basata su tecnologia open source.

Keywords. Open Science, research data management, politiche e raccomandazioni, dati aperti della ricerca

Introduzione

Quando si parla di scienza aperta non si fa riferimento solo ai dati della ricerca (open research data), alle pubblicazioni (open access publications) risultanti dai progetti finanziati, prevalentemente con fondi pubblici, ma anche alle metodologie usate nel corso della ricerca (open methodology), ai software utilizzati o sviluppati (open source), alle attività di revisione da parte dei pari alle pubblicazioni in modalità aperta (open peer reviewing) e anche ai materiali per scopo didattico (open educational resources). Si tratta dell'intero processo della ricerca, dei suoi risultati, che viene reso accessibile e fruibile nelle modalità più aperte, della disseminazione dei risultati, della loro conservazione, ma anche della loro divulgazione con mezzi e strumenti appropriati. Implementare la scienza aperta significa dotarsi di infrastrutture, servizi ma anche di politiche, di raccomandazioni, di studi e di analisi di percorsi formativi, di best practice. Il gruppo di lavoro IOSSG intende,in questo ambito, portare il suo contributo alla diffusione e adozione della cultura della scienza aperta in Italia.

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

1. Il contesto

Nel promuovere e sostenere la scienza aperta, cioè un approccio alla ricerca scientifica basato sulla collaborazione, sull'apertura, sulla trasparenza, sulla condivisione degli strumenti medesimi e dei risultati e sulla disseminazione in accesso aperto di questi ultimi, la Commissione Europea, ha definito le politiche di accesso ai risultati della ricerca (pubblicazioni e dati della ricerca) prodotti in progetti di ricerca da essa finanziati. Ha stabilito, alla fine del 2013, che tutti i beneficiari dei progetti di ricerca afferenti al programma di finanziamento Horizon 2020 che pubblicano i risultati in riviste peer reviewed sono tenuti a disseminarli in accesso aperto, secondo due strategie: la via verde (deposito del post-print dell'articolo in un repository aperto) o la via oro (pubblicazione su riviste peer reviewed ad accesso aperto). (1) Dal 2017 anche i dati della ricerca che validano i risultati presentati nelle pubblicazioni (underlying data) vanno dotati di un piano di gestione e depositati, in un data repository, resi accessibili e fruibili, secondo i principi FAIR (Findable, Accessibile, Interoperable, Reusable), principi di rintracciabilità, accessibilità, interoperabilità e riusabilità/riproducibilità dei dati della ricerca. (2)

Nell'aprile del 2016, la Commissione ha presentato una bozza di progetto su EOSC – European Open Science Cloud Initiative, un'iniziativa o per meglio dire un processo per l'implementazione in Europa della cultura dei dati e della loro condivisione e dei relativi servizi basata su un'architettura di data infrastructure commons, e su una leggera e agile governance al fine di rispondere ai bisogni dei ricercatori e all'avanzamento della scienza e della conoscenza nel contesto di una ricerca sempre più data driven. La EOSC Declaration resa pubblica nell'ottobre scorso riassume efficacemente le finalità di questo processo. (3) In alcuni paesi europei, già da un po'di anni, sono stati messi in piede servizi nazionali volti a sostenere la cultura dei dati e i suoi sviluppi, a fornire un servizio di data stewardship, cioé supporto e strumenti per l'interoperabilità, la gestione dei dati della ricerca, per la elaborazione del piano di gestione. Tali servizi sono rivolti alle istituzioni di ricerca, ai ricercatori, al personale tecnico che vi operano e coprono vari aspetti riguardanti l'intero ciclo di vita dei dati, la raccolta, l'analisi, la gestione, la cura, il trattamento e la conservazione dei dati medesimi, le licenze d'uso, così come la creazione di materiali e corsi di formazione.

In Italia la situazione si presenta molto frammentaria, sebbene esistano infrastrutture di servizi operanti a livello nazionale e comunità di ricerca che si sono organizzate nella loro disciplina, non c'è alcun coordinamento ed attualmente manca una politica a livello nazionale relativamente ai dati della ricerca. Di fatto, non viene erogato nessun servizio a livello nazionale e raramente a livello istituzionale, volto a fornire conoscenze, modelli organizzativi, strumenti necessari per affrontare queste problematiche, ad eccezione di alcune comunità scientifiche che hanno una lunga esperienza di raccolta, conservazione e gestione dei dati.

2. Nascita del gruppo IOSSG

Affrontare le sfide di una scienza sempre più data intensive è fortemente sentito nelle università italiane che si confrontano con richieste di supporto interno quando la Commissione Europea così come altri finanziatori internazionali della ricerca richiedono ai

beneficiari dei progetti di ricerca di mettere a disposizione i dati della ricerca risultanti da ricerche finanziate e permetterne l'accesso e il riuso. Diventa, allora, necessario e indispensabile che le istituzioni di ricerca dei singoli paesi, particolarmente se non sono supportati da una politica nazionale, provvedano a fornire l'assistenza e le infrastrutture necessarie ai propri ricercatori nel contesto di una scienza globale.

Per rispondere a queste esigenze e per dare delle risposte all'impatto che l'iniziativa EOSC avrà localmente, è nato nella primavera del 2016 un gruppo di lavoro in ambito universitario, un gruppo trasversale di esperti, responsabili dell'ufficio ricerca, informatici, bibliotecari e responsabili degli archivi aperti della ricerca, appartenenti al Politecnico di Milano, alle Università degli Studi di Bologna, di Milano, di Padova (prima in qualità di uditore e successivamente di membro), di Torino, di Trento, di Ca 'Foscari Venezia. Il gruppo ha deciso di collaborare informalmente, di mettere a fattore comune le conoscenze e le competenze che stavano acquisendo per dare risposte concrete e supporto ai ricercatori nella gestione dei dati e fornire raccomandazioni e linee guida alle rispettive amministrazioni.

A promuovere questa collaborazione è stato il Cineca, in qualità di NOAD - National Open Access Desk per l'Italia, del progetto europeo OpenAIRE (4) finalizzato all'implementazione delle politiche a sostegno della scienza aperta della Commissione, e che in questo ruolo, svolge anche attività di disseminazione, training, promozione e per questa ragione, il gruppo ha chiesto al NOAD di coordinarne i lavori per il primo biennio.

3. Le attività

L'obiettivo principale del gruppo è sostenere le parti interessate (amministrazione universitaria, ricercatori, uffici di ricerca, servizi IT e bibliotecari) nell'affrontare le recenti sfide che la data driven science pone alle istituzioni accademiche e di ricerca, con particolare attenzione, ai processi della ricerca, al ciclo di vita e gestione dei dati, ai servizi e alle infrastrutture di supporto.

L'elaborazione di una politica istituzionale sui dati e, in particolare, di un piano di gestione dei dati e la sua implementazione richiede affrontare in modo dettagliato:

- la descrizione dei data set,
- gli standard e i metadati che verranno utilizzati,
- gli aspetti prettamente gestionali (data management, documentazione e cura dei dati),
- la sicurezza e eventuale confidenzialità dei dati,
- la condivisione, l'accesso, il riuso,
- gli aspetti legali e di proprietà intellettuale,
- la definizione di responsabilità nelle varie fasi del ciclo dei dati.

È indispensabile affrontare in primis, il contesto politico e di governance che i dati, la fornitura di servizi e di infrastrutture a livello locale pongono, e rispondere fornendo strumenti e ponendo le basi per l'acquisizione di diverse e articolate competenze per la gestione dei dati.

Le attività del gruppo consistono nell'elaborare linee guida, toolkit, raccogliere buone pratiche, condurre analisi, studi, indagini e produrre materiali per la formazione e condi-

videre i risultati e i materiali del lavoro di gruppo con tutte le comunità scientifiche interessate e il pubblico in generale.

Il gruppo ha iniziato le sue attività a metà 2016 ed ha già prodotto un modello di politica istituzionale di gestione dei dati di ricerca e una checklist dettagliata per supportare la stesura di un piano di gestione dei dati di ricerca, liberamente accessibili sul sito del gruppo. (5)
I lavori si basano su documenti, raccomandazioni elaborate da altri atenei all'estero, in particolare, sullo scambio con l'università di Vienna e soprattutto sui materiali e il toolkit prodotto dal progetto europeo LEARN di cui la stessa università di Vienna è partner. (6)
Entrambi i documenti sono in italiano e contestualizzati al quadro giuridico e di ricerca italiano.

All'inizio di settembre 2017 il gruppo ha deciso di adottare l'acronimo IOSSG, Italian Open Science Support Group, di continuare le attività e di aprire la partecipazione ad altri esperti in qualità di osservatori quali Anna Maria Tammaro del DILL- International Master in Digital Library Learning e Simone Sacchi di LIBER, ad una piccola azienda italiana attualmente impegnata in soluzioni di gestione dei dati di ricerca e di Paolo Budroni in qualità di esperto dell' E-infrastructure Austria con cui il gruppo aveva già lavorato per la stesura di un modello di policy. IOSSG è, inoltre, aperto alla collaborazione con altri gruppi simili operanti in Italia e all'estero nel mondo della ricerca e dell'università. Ha già stabilito un contatto con il corrispondente gruppo austriaco AOSSG, Austrian Open Science Support Group.

A settembre il gruppo ha definito il programma di lavoro dettagliato per il 2018 che ha come focus l'impatto a livello locale delle politiche di EOSC ed è incentrato su queste tematiche:

- elaborazione di linee guida per l'attuazione di una politica istituzionale nella gestione dei dati di ricerca ponendo attenzione alla parte giuridico-legale, in particolare alla protezione dei dati sensibili e al riuso dei dati,
- creazione di un unico punto di accesso, che aggreghi competenze e risorse diverse al fine di costruire un efficiente servizio di sostegno trasversale alla ricerca,
- definizione dei requisiti dei data repository per la conformità agli attuali standard, ai principi di certificazione e a principi FAIR, studio e proposte di soluzioni di infrastrutture a livello locale per la gestione e la conservazione dei dati della ricerca in un'ottica aperta e federata, basata su tecnologia open source,
- campagna di formazione e di promozione della scienza aperta tramite la creazione di materiali di didattici, organizzazione di workshop per la diffusione della conoscenza di tutti gli aspetti relativi alla gestione dei della ricerca e ai principi FAIR.

I risultati dei lavori, una volta conclusi, vengono messi a disposizione online, liberamente accessibili e riutilizzabili sul sito del gruppo. (7)

4. Conclusioni

IOSSG auspica che le attività portate avanti quest'anno e anche negli anni a venire possano dare un contributo effettivo alla creazione e alla diffusione della cultura della scienza aperta in Italia e all' adozione di un modello EOSC anche localmente, tramite la fornitura

di strumenti, modelli di politiche, best practice, studi e soprattutto favoriscano un maggior coordinamento tra i diversi attori: infrastrutture di servizi, infrastrutture di ricerche, comunità di ricerca, istituzioni universitarie e centri/ istituti di ricerca e pongano le basi per l'adozione di una visione e una politica nazionale sulla scienza aperta.

Riferimenti bibliografici

(1) EC - 2020 Programme – Guidelines to rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, vers.3.2 (March 2017) http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

(2) Wirkinson MD, Dumontier M, et al . (2017) The Fair Guiding Principles for scientific data management and stewardship, Nature. https://www.nature.com/articles/sdata201618#bx2

(3) EOSC - https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

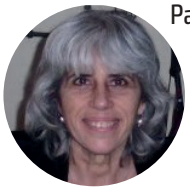
(4) OpenAIRE – http://www.openaire.eu

(5) IOSSG – Modello di Policy per la gestione dei dati della ricerca e Data Management Checklist (2017) https://sites.google.com/view/iossg/materiali-prodotti?authuser=0

(6) LERU Project – http://learn-rdm.eu/

(7) IOSSG – http://sites.google.com/view/iossg

Autori



Paola Gargiulo p.gargiulo@cineca.it

Responsabile delle attività di formazione, disseminazione sui temi relativi alla scienza aperta con particolare riferimento alle politiche dell'EC in qualità di National Open Access Desk del progetto europeo OpenAIRE ,di cui Cineca è partner.

DOI: 10.26314/GARR-Conf17-proceedings-18

Nuovi servizi di timing over fibre su reti di trasporto ottico

Davide Calonico

Istituto Nazionale di Ricerca Metrologica

Abstract. La distribuzione di segnale di tempo campione è importante sia in ambito scientifico che industriale. Nuove tecniche di disseminazione di segnali di T/Fsu fibra ottica oggi offrono le migliori performance, la maggiore resilienza e sono ormai accessibili in termini di costo. Qui si presenta lo stato dell'arte in Europa e in Italia e si analizzano alcuni passi per la diffusione più capillare di nuovi servizi.

Keywords. Timing, Orologi atomici, Tempo su Fibra

Introduzione

Nella nostra società c'è una domanda crescente di reti di sincronizzazione che garantiscano un riferimento accurato e stabile di tempo e frequenza (T/F): dagli operatori di telecomunicazioni, alle smart grid elettriche, al settore finanziario per rispondere con nuove regolamentazioni EU, agli utenti scientifici.

Oggi la distribuzione T/F nella maggioranza delle applicazioni si basa sul broadcasting satellitare, per esempio il tempo distribuito da un Global Navigation Satellite System (GNSS), di cui il Global Positioning System (GPS) è forse il maggiormente noto.
Per le prestazioni, la migliore accuratezza possibile per il T/F distribuito da un GNSS arriva a 2-5 ns, tuttavia, questo risultato si ottiene con ricevitori molto specifici e competenze spesso relegate a pochi istituti di metrologia primaria e di geodesia. L'accademia e l'industria può contare su apparecchiature GNSS meno accurate, ma che garantiscono un'incertezza di 100 ns a costi contenuti.

D'altra parte, il GNSS soffre nell'integrità del segnale, poiché la debole potenza ricevuta dai satelliti rende un pericolo concreto lo spoofing e l'hacking o i disturbi legati alle condizioni del tempo spaziale. Infine, oggi il servizio GNSS più diffuso, il GPS, è distribuito da un ente militare, pertanto non offre alcun livello garantito di servizio, problematica che sarà superata con il pieno sviluppo del GNSS europeo, il progetto Galileo.

In generale, le piattaforme di distribuzione del tempo troverebbero rilevante beneficio nella capillarizzazione delle tecniche di diffusione del tempo campione via fibra ottica. Le tecnologie oggi a disposizione rendono la disseminazione via fibra estremamente interessante non solo per le realtà scientifiche più esigenze, ma anche per l'accademia in generale e per le industrie. Il tutto con un portafoglio di tecniche in grado di soddisfare diverse richieste rispetto alla coppia costi/performance.

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

1. Le tecniche di distribuzione del tempo in fibra ottica

Esistono diverse tecniche per trasmettere il segnale di T/F in fibra ottica. Qui riassumeremo in breve le tecniche più interessanti per l'uso in reti di trasporto ottico. La tecnica più performante, ma difficile nell'integrazione, è quella cosiddetta coerente, che trasporta il segnale di un laser a frequenza ultra-stabile. E' cruciale che la fase della luce laser sia imperturbata e la sua stabilità non sia degradata dalle fluttuazioni del mezzo vetroso, dovute a vibrazioni meccaniche o fluttuazioni di temperatura. Per ovviare a questi disturbi naturalmente presente, una parte della luce viene retro-riflessa dal punto di arrivo verso il sito di emissione. Nel sito di partenza, la luce retro-riflessa è confrontata con quella di origine, pervenendo a una misura del rumore aggiunto dalla fibra nel tragitto di andata e ritorno.

Con tecniche optoelettroniche è possibile cancellare il rumore una volta misurato. L'assunto di base è la reciprocità di cammino in andata e ritorno. La tecnica pertanto raggiunge la massima efficacia con l'uso di una fibra singola in bidirezionalità. Proprio quest'ultimo elemento rende questo metodo poco integrabile con le reti trasmissive, che sono di tipo unidirezionale, nel senso che si hanno una fibra in Tx e una in Rx. Per far convivere la tecnica coerente bidirezionale con le reti ordinarie è necessario quindi che a ogni nodo con apparati della rete ci sia un by-pass che permetta al segnale bidirezionale di non essere compromesso di dispositivi, generalmente unidirezionali. L'amplificazione per compensare le perdite è affidata spesso ad amplificatori di tipo EDFA ma bidirezionali, anche se ottimi risultati sono stati dimostrati con tecniche Raman o Brillouin, e soprattutto con l'amplificazione Raman ci può essere una forte integrazione, anche se questi amplificatori non sono quelli maggiormente usati.

I livelli di accuratezza che si raggiungono sono inferiori alle parti per 10^{19}, raggiunti in long-haul (>1000 km) con tempi di misura di 1000-5000 secondi.

La seconda tecnica è molto più integrabile con le reti di trasporto ottiche e si caratterizza come un'evoluzione del Precision Time Protocol (PTP) già noto in ambiente telecomunicazionista e ritrovabile in diverse realizzazioni commerciali dei vendor. L'evoluzione in esame applica a PTP alcuni accorgimenti software e hardware che ne migliorano nettamente la prestazione portandola a livelli di metrologia primaria, con accuratezze di tempo a livello del sub-nanosecondo e stabilità sotto 10^{-13} già con 1 secondo di misura, che migliora le distribuzioni satellitari di 4 ordini di grandezza. Un'evoluzione disponibile è nota come protocollo White Rabbit, (per esempio, www.ohwr.org/projects/white-rabbit), inventato al CERN di Ginevra e diffuso attualmente in diverse esperienze scientifiche e tra diversi istituti metrologici. Il cuore dell'evoluzione PTP risiede nella capacità migliorata di quantificare l'asimmetria dei canali Tx e Rx, visto che una soluzione di protocollo utilizza la rete esattamente come per il traffico dati. La migliore valutazione dell'asimmetria è data da modelli di analisi e anche da un hardware di comparizione di tempo tra segnali di Tx ed Rx molto più accurata. La potenzialità di migliorare la tecnica White Rabbit è molto ben presente soprattutto a livello hardware.

2. Prospettive e punti di attenzione

Se guardiamo alla possibilità di una distribuzione capillare di Tempo in Fibra Ottica, sono

diversi i punti di attenzione che richiedono un impegno di ricerca e realizzativo. Innanzitutto, occorre ben distinguere le distribuzioni ad alto costo da quelle invece commercialmente accessibili al segmento maggiore di utenti.

Alla prima categoria pertengono le trasmissioni coerenti in fibra con bi-direzionalità su singola fibra e relativa compensazione del rumore di fase. L'integrazione di reti ottiche per la connettività dati con simili infrastrutture per il T/F richiede una forte interazione con i carrier e con l'infrastruttura, con importanti inserimenti di apparecchiatura dedicata, come elementi di rigenerazione della portante, amplificatori ottici particolari etc. Sebbene dunque questa modalità sembri confinata a un ristretto numero di utenti molto specifici, in prospettiva l'integrazione di dispositivi ottici idonei e l'evoluzione delle moderne architetture di rete non sembra escludere una scalabilità della tecnica coerente.

Se guardiamo invece alle soluzioni di protocollo, quelle cioè che migliorano le tecniche tipo PTP, come il White Rabbit, aumentiamo notevolmente le capacità in termini di accuratezza e stabilità del segnale di tempo, coniugandola con un'alta scalabilità e un'integrazione praticamente già in corso sulle piattaforme ordinarie di telecomunicazioni.
Qui le prospettive di sviluppo riguardano la capacità di caratterizzare le asimmetrie di ritardi nei percorsi Tx ed Rx, possibilmente con misure in tempo reale, che permettano di garantire i livelli di prestazione costanti nel tempo.

La coniugazione di accuratezza a livelli del picosecondo, con velocità di misura, resilienza agli attacchi ed elevati SLA sono sicuramente alla portata di queste tecniche che potranno fornire una piattaforma formidabile da affiancare al consueto servizio di traffico dati.

3. Conclusioni

In Italia la distribuzione di T/F tramite fibra ottica ha ottenuto risultati di ricerca molto competitivi, e un'infrastruttura di ricerca dedicata di 2000 km realizzata da INRIM è operativa. Il futuro di questa rete sarà quello di collegarsi ad analoghe europee, che stanno sviluppandosi in Francia, Germania, UK, Polonia, Repubblica Ceca. L'interesse per la metrologia e la scienza è ben dimostrato, ma la maturità tecnologica porta oggi anche l'accademia e l'industria a poter usufruire dei vantaggi offerti. Restano ancora alcune sfide aperte, in particolari un'integrazione sodisfacente con le reti ottiche di trasporto dati che renderebbero capillare la distribuzione in fibra ottica dei segnali di tempo.

Riferimenti bibliografici

Clivati, C et al. (2016) "Measuring absolute frequencies beyond the GPS limit via long-haul optical frequency dissemination" Opt. Exp. 24, 865-1875

Clivati, C. et al. (2015) "A Coherent Fiber Link for Very Long Baseline Interferometry" IEEE Trans. on UFFC, 62 1907-1912

D. Calonico, M. Inguscio, F. Levi, (2015) "Light and the distribution of time, 2015 European Physics Letters 110 4000

D. Calonico et al., (2014) "High accuracy coherent optical frequency transfer over a doubled 642 km fiber link" Applied Physics B, 117, pp 979-986

F. Torres-Gonzalez, E. Marin-Lopez, J. Diaz, (2016) “Beyond PTP technologies: Scalability Analysis of the White-Rabbit Technology for Cascade-Chain Networks”. IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication, Stockholm, Sweden, 7-9 September, pp. 89-94

E.F. Dierikx, et al., (2016) “White Rabbit Precision Time Protocol on Long Distance Fibre Links,” IEEE Trans. UFFC, 63(7), 945-952.

Autori

Davide Calonico d.calonico@inrim.it

Fisico, PhD al Politecnico di Torino. Si occupa all'INRIM di campioni atomici di frequenza e distribuzione di T/F su fibra ottica. Ha realizzato il primo campione ottico italiani ad atomi di Itterbio ultrafreddo e un'infrastruttura di ricerca in fibra ottica per Tempo e Frequenza di 2000 km che collega le principali città italiane da Torino a Matera e si proietta verso l'Europa con un collegamento con la Francia.

DOI: 10.26314/GARR-Conf17-proceedings-19

OpenCitations: enabling the FAIR use of open citation data

Silvio Peroni[1], David Shotton[2]

[1]DASPLab, DISI, University of Bologna, Bologna, Italy, [2]Oxford eResearch Centre, University of Oxford, Oxford UK

Abstract. Citations are the bridges that enable people to pass from one scholarly work (e.g. a conference paper) to others (e.g. journal articles and book chapters). At present, the unrestricted travel over the entire network of bridges by using existing services requires one to pay an expensive fee, which is affordable only by rich professionals – such universities or other research institutes. The general populace is excluded.
In this paper, we introduce the OpenCitations Corpus, an open repository of scholarly citation data available in RDF and published according to the FAIR principles, which is an attempt to provide open bridges between scholarly works.
This version: https://w3id.org/people/essepuntato/papers/oc-garr2017/2017-11-10.html
Last version: https://w3id.org/people/essepuntato/papers/oc-garr2017.html

Keywords. OpenCitations, OpenCitations Corpus, OCC, FAIR Data Principles

Citation data: the story so far

Sharing scholarly data to foster their reuse is one of the main goals of the current practices in Data Science and, more generally, in scholarly communication. While some parties are trying to boycott the unscrupulous reuse of experimental data – for instance see the "research parasites" issue (Longo, Drazen 2016) that generated a strong response from the scientific community –, others have recently proposed common practices and guidelines to support a better sharing and reuse of scholarly data in order to support secondary data analysis.

In particular, a FORCE11 (https://force11.org) working group has recently proposed the FAIR Data Principles (Wilkinson et al. 2016). The acronym FAIR stands for Findable, Accessible, Interoperable, and Reusable, which are the main leading principles that, if applied, would facilitate the discovery, access, integration, and analysis of scholarly knowledge by humans and machines (https://www.force11.org/group/fairgroup/fairprinciples).
Citation data are among the scholarly data to which the application of FAIR principles would benefit the whole scholarly community, particularly since analysis of citation events is one of the main ways of finding key publications on a particular topic. Citation data are also important for addressing institutional goals, such as the assessment of the quality of research by means of metrics and indicators calculated from citation databases. However, the cruel reality that the most authoritative citation indexes, i.e. Scopus (https://www.scopus.com) and Web of Science (http://webofscience.com), do not follow the FAIR principles, primarily because they can only be accessed by paying significant

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

subscription fees, which may amount to tens of thousands of euros annually per institution. In the current age, in which Open Access is considered a necessary practice in research, "citation data now needs to be recognized as a part of the Commons [...] and placed in an open repository" (Peroni et al. 2015).

The Initiative for Open Citations (I4OC, https://i4oc.org) has been recently launched to promote the aforementioned idea. I4OC is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data, initially by encouraging scholarly publishers to make open the article reference lists they already deposit to Crossref (https://crossref.org). The intent is to have citation data that are structured, separable, and open. Among the I4OC founders, one specific organization, OpenCitations (http://opencitations.net), has the further objective of employing Semantic Web technologies to create an open repository of the scholarly citation data published in RDF according to the FAIR principles.

The rest of the paper is organised as follows. In Section 1 we briefly summarise the story of OpenCitations. In Section 2 we describe the OpenCitations Corpus, and we focus particularly on how it complies with the FAIR data principles. Finally, in Section 3, we conclude the paper sketching out some future works.

1. OpenCitations

OpenCitations (Shotton 2013; Peroni et al. 2015) has formally started in 2010 as a one-year project funded by JISC (with a subsequent extension), with David Shotton as director, who at that time was working in the Department of Zoology at the University of Oxford. The project was global in scope, and was designed to change the face of scientific publishing and scholarly communication, since it aimed to publish open bibliographic citation information in RDF (Cyganiak et al. 2014) and to make citation links as easy to traverse as Web links. The main deliverable of the project, among several outcomes, was the release of an open repository of scholarly citation data described using the SPAR (Semantic Publishing and Referencing) Ontologies (Peroni 2014), and named the OpenCitations Corpus (OCC), which was initially populated with the citations from journal articles within the Open Access Subset of PubMed Central.

At the end of 2015 Silvio Peroni joined OpenCitations as co-director and technical manager, with the aim of setting up a new instantiation of the Corpus based on a new metadata schema and employing several new technologies to automate the ingestion of fresh citation metadata from authoritative sources. The current instantiation of the OCC is hosted by the Department of Computer Science and Engineering (DISI) at the University of Bologna, and since the beginning of July 2016 has been ingesting, processing and publishing reference lists of scholarly papers available in Europe PubMed Central. Additional metadata for these citations are obtained from Crossref and (for authors) ORCID.

2. The OpenCitations Corpus: a FAIR-compliant resource

The OpenCitations Corpus (OCC) is a database of open citation data, made available in RDF (Cyganiak et al. 2014). At the end of 2015, a formal collaboration between the Uni-

versity of Oxford and the University of Bologna was initiated to build from that initial Oxford prototype, setting up a new instantiation of the OCC based on a revised metadata schema (Peroni, Shotton 2016), and employing several new technologies to automate the daily ingestion of fresh citation metadata from authoritative sources (Peroni et al. 2017). The OCC is now one of the largest truly open collection of RDF-based citation data available on the Web, and includes more than 11.5 million citation links to around 6 million cited resources (as of 10 November 2017).

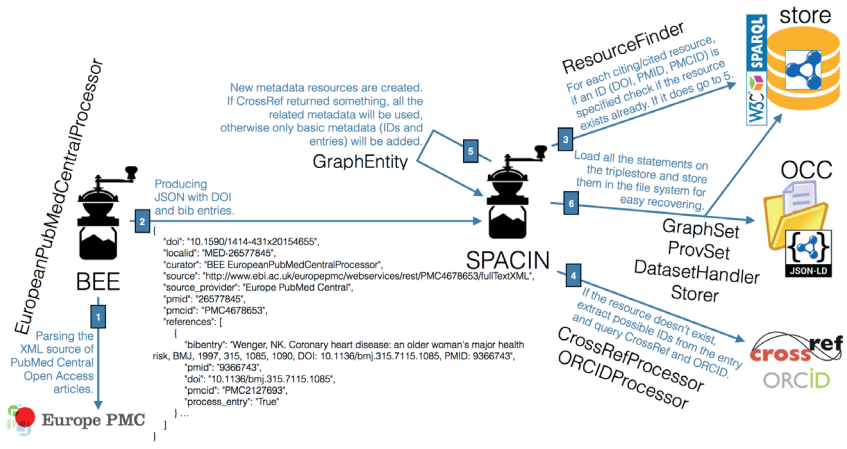


Fig. 1
The ingestion workflow implemented by the OpenCitations Corpus

The ingestion of new data into the OCC, briefly summarised in Figure 1, is curated by two Python scripts, the Bibliographic Entries Extractor, a.k.a. BEE, and the SPAR Citation Indexer, a.k.a. SPACIN. Both of these are available on the OpenCitations GitHub repository and are released as open source code according to the ISC Licence.

BEE is responsible for the creation of JSON files containing reference lists from articles in the OA subset of PubMed Central (retrieved by using the Europe PubMed Central API). SPACIN processes each JSON file created by BEE, retrieves additional metadata information about all the citing/cited articles described in it by querying the Crossref API and the ORCID API, and finally stores all the data in the OCC triplestore, which is a Blazegraph instance that makes available a SPARQL endpoint (Harries, Seaborne 2013) with the full text search enabled for all the entities included in the entire OCC.

In the following subsections we describe the tools and technologies used by the OCC to comply with the FAIR principles.

2.1 To be Findable

The OpenCitations Corpus (http://opencitations.net) uses w3id.org, a service run by the W3C Permanent Identifier Community Group (http://www.w3.org/community/perma-id/) to provide a secure, permanent URL re-direction service for Web applications, as the root URL for assigning persistent identifiers to all its entities. For example, https://w3id.org/oc/corpus/br/7295288 (which resolves to http://opencitations.net/corpus/br/7295288.html) identifies the paper referenced in (Peroni et al. 2015) within the OCC.

All the citation data are described according a specific metadata model (Peroni, Shotton 2016) that is based on SPAR (Semantic Publishing and Referencing) Ontologies (http://www.sparontologies.net) and other standard vocabularies. All the data within the OCC are queryable by means of the OpenCitations SPARQL endpoint (https://w3id.org/oc/sparql), and dumps of the entire database are uploaded monthly to Figshare (https://figshare.com). In addition, all the original sources from which the information within the OCC has been obtained are linked by means of provenance information according to PROV-O (Lebo et al. 2013).

2.2 To be Accessible

All the data in the OCC can additionally be retrieved by using their unique identifiers – which are Uniform Resource Locators (URLs) – via the Hypertext Transfer Protocol (HTTP) (Fielding, Reschke 2014), as in the example given in the previous paragraph. The resource metadata are made available either in human-readable HTML or in a variety of machine-readable forms (RDF/XML, Turtle or JSON-LD) via content negotiation. All the citation data within the OCC can be accessed independently from the current Web existence (or absence) of the original publications from where they have been obtained.

2.3 To be Interoperable

The data within the OCC are described by means of the Resource Description Framework (RDF) (Cyganiak et al. 2014), which is the main data model for representing information in machine-readable form on the Web, and this enables qualified links from/to any entity included in the Corpus. The data are modelled in RDF according to a set of ontologies, grouped together and formalized in the OpenCitations Ontology (https://w3id.org/oc/ontology), as summarised in the diagram in Figure 2. These ontologies themselves follow the FAIR principles.

2.4 To be Re-usable

The entities described within the OCC includes: citing and cited bibliographic resources (conference papers, book chapters, journal articles, etc.) and their containers (academic proceedings, books, journals, etc.), the formats in which they have been embodied (digital vs. print, first and ending pages, etc.), the names and roles of relevant bibliographic agents related to these resources (author, editor, publisher, etc.), the literal textual content of each reference in the reference list of each citing bibliographic resource, and all the identifiers (DOI, ORCID, etc.) employed to identify these bibliographic resources and the agents involved. Provenance information is associated with all the entities in the OCC via PROV-O (Lebo et al. 2013) (e.g. https://w3id.org/oc/corpus/br/7295288/prov/se/1), and changes in the data can be tracked by means of SPARQL UPDATE queries (Peroni et al. 2016). All the data included in the OCC are available under a Creative Commons public domain dedication (CC0, https://creativecommons.org/publicdomain/zero/1.0/legalcode).

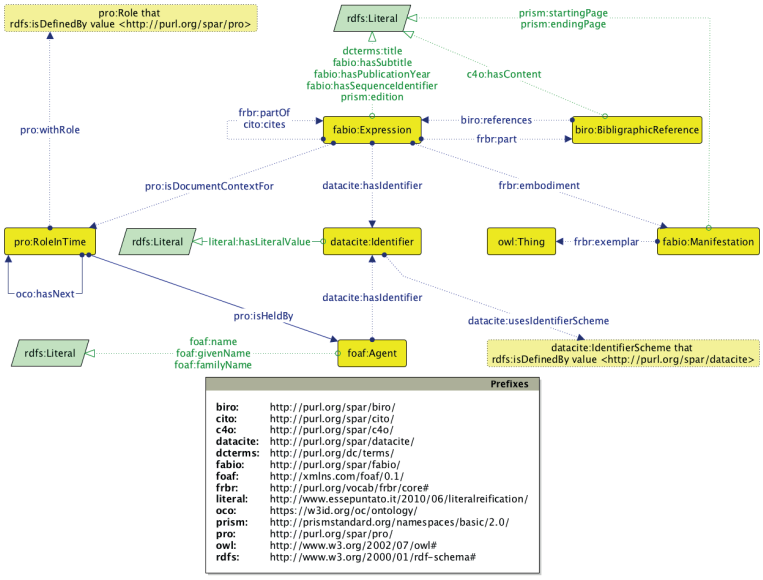


Fig 2
A Graffoo diagram (Falco et al. 2014) describing the OpenCitations Ontology

3. Conclusions

In this contribution we have introduced the tools and technologies that ensure that the OpenCitations Corpus – an RDF database of open scholarly citation data – is fully compliant with the FAIR Data Principles. As an immediate future development of the project, made possible by means of recent funding from the Alfred P. Sloan Foundation (https://sloan.org), we will extend the current infrastructure and the rate of data ingest. Our goal is to increment the daily ingestion rate of citation data from ~500,000 citations per month to ~500,000 citations per day.

Riferimenti bibliografici

Richard Cyganiak, David Wood, Markus Lanthaler (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. W3C. https://www.w3.org/TR/rdf11-concepts/

Riccardo Falco, Aldo Gangemi, Silvio Peroni, Fabio Vitali. (2014). Modelling OWL ontologies with Graffoo. In Proceedings of ESWC 2014 Satellite Events: 320–325. https://doi.org/10.1007/978-3-319-11955-7_42

Roy Fielding, Julian Reschke (2014). Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing. Request for Comments: 7230. IETF. https://tools.ietf.org/html/rfc7230

Steve Harris, Andy Seaborne. (2013). SPARQL 1.1 Query Language. W3C Recommendation 21 March 2013. W3C. https://www.w3.org/TR/sparql11-query/

Timothy Lebo, Satya Sahoo, Deborah McGuinness (2013). PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013. W3C. https://www.w3.org/TR/prov-o/

Dan L. Longo, Jeffrey M. Drazen. (2016) Data sharing. New England Journal of Medicine, 374: 276–277. DOI: https://doi.org/10.1056/NEJMe1516564

Silvio Peroni (2014). The Semantic Publishing and Referencing Ontologies. In Semantic Web Technologies and Legal Scholarly Publishing: 121-193. https://doi.org/10.1007/978-3-319-04777-5_5

Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation, 71 (2): 253–277. DOI: https://doi.org/10.1108/JD-12-2013-0166

Silvio Peroni, David Shotton (2016). Metadata for the OpenCitations Corpus. Figshare. DOI: https://doi.org/10.6084/m9.figshare.3443876

Silvio Peroni, David Shotton, Fabio Vitali (2016). A document-inspired way for tracking changes of RDF data. In Proceedings of Drift-a-LOD 2016: 26–33 http://ceur-ws.org/Vol-1799/Drift-a-LOD2016_paper_4.pdf

Silvio Peroni, David Shotton, Fabio Vitali (2017). One year of the OpenCitations Corpus: Releasing RDF-based scholarly citation data into the Public Domain. In Proceedings of ISWC 2017: 184–192 https://doi.org/10.1007/978-3-319-68204-4_19

David Shotton (2013). Open citations. Nature, 502 (7471): 295–297. DOI: https://doi.org/10.1038/502295a

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Authors

Silvio Peroni silvio.peroni@unibo.it

Silvio Peroni is an Assistant Professor at the Department of Computer Science and Engineering, University of Bologna, Italy. He is Co-Director of OpenCitations, a founding member of the Initiative for Open Citations, and one of the main developers of the SPAR (Semantic Publishing and Referencing) Ontologies. His work on Semantic Publishing topics has been recently published by Springer in a book entitled Semantic Web Technologies and Legal Scholarly Publishing.

David Shotton david.shotton@oerc.ox.ac.uk

David Shotton is Co-Director of OpenCitations, a founding member of the Initiative for Open Citations, and one of the main developers of the SPAR (Semantic Publishing and Referencing) Ontologies. Originally a cell biologist, he has for the last decade pioneered the field of Semantic Publishing. He is a founding member of Force11, a scholarly community advancing the future of research communication and e-scholarship through the effective use of information technology.

DOI: 10.26314/GARR-Conf17-proceedings-20

Open Science, dati FAIR e l'Osservatorio Virtuale

Marco Molinaro, Fabio Pasian

Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Trieste

Abstract. Il ciclo di vita dei dati astrofisici non deve esaurirsi nel ciclo di progettazione, produzione, analisi e pubblicazione dei risultati. Vi sono varie ragioni a supporto: le pubblicazioni basate su dati d'archivio sono di volume comparabile a quelle del periodo proprietario; la ricerca è sempre più basata su dati multi-strumento, multi-banda, osservativo/numerici, multi-messenger; un'analisi esauriente delle attuali moli dati è fattibile solo da grosse collaborazioni o riutilizzando i dati; il finanziamento della ricerca astrofisica è spesso pubblico e tali devono essere i suoi risultati. Per permettere che le risorse dati astrofisiche siano efficacemente utilizzate occorre uno sforzo di omogeneizzazione dei metodi di ricerca e accesso alle risorse dati stesse. Esistono linee guida e tecnologie che si occupano di questo aspetto e sono sostenute dalla Comunità Europea. Qui spieghiamo come l'Osservatorio Virtuale sia un'implementazione di quelli che sono oggi noti come FAIR principles.

Keywords. Osservatorio Virtuale, Astrofisica, FAIR data, Open Science, Interoperabilità

Introduzione

La Comunità Europea sta spingendo affinché la ricerca scientifica costruisca una cultura del dato e della sua interoperabilità secondo il concetto dell'Open Science e i principi FAIR (Findable, Accessible, Interoperable, Re-usable, Wilkinson et al. 2016). A questo scopo ha iniziato il progetto EOSC (European Open Science Cloud) la cui fase iniziale è finanziata attraverso il progetto Horizon 2020 EOSCpilot. La Dichiarazione EOSC definisce in dettaglio gli intenti di tale iniziativa e fa chiari riferimenti ai principi FAIR per quanto riguarda le implementazioni.

Nel dominio dell'astrofisica quanto dettato sia dai principi FAIR che dalle prime analisi del progetto EOSC sono da tempo oggetto di discussione e implementazione all'interno della comunità dell'Osservatorio Virtuale (VO, Virtual Observatory) che si raccoglie attorno all'iniziativa globale dell'IVOA (International Virtual Observatory Alliance) di cui il progetto VObs.it è il rappresentante per l'Italia, mentre a livello comunitario le coordinazione avviene sotto il nome di EURO-VO e attualmente confluisce nel pacchetto DADI del progetto H2020 ASTERICS.

L'iniziativa VO, con gli standard tecnologici sviluppati dall'IVOA, rappresenta lo stato dell'arte per le buone abitudini nella gestione dei dati in astronomia e risponde ai requisiti di accessibilità e interoperabilità dei dati necessari per un efficace sfruttamento dei dati acquisiti delle osservazioni e prodotti delle simulazioni numeriche in campo astrofisico.
Nel seguito di questo articolo verrà fornita una descrizione dell'architettura dell'IVOA, §1, della sua connessione diretta ai quattro fondamenti dei principi FAIR, §2, e infine si scenderà nei dettagli più specifico su come gli standard VO coprano lo scenario complessivo

della data FAIRness §3. Nelle conclusioni verranno discusse le differenze fra FAIR e VO.

1. L'architettura dell'IVOA

L'architettura dell'IVOA (Arviset et al. 2010) rappresenta, con tre livelli sovrapposti, lo schema di funzionamento del VO. I livelli, numerati 0, 1 e 2, vanno dal generale al particolare.

Il livello 0 (Figura 1) rappresenta una descrizione di alto livello delle componenti del VO. Vi si riconoscono due strati orizzontali principali, il Resource Layer (risorse disponibili) e lo User Layer (utenti): il primo a condividere le risorse (dati, servizi, ...), sharing, il secondo a utilizzarle, using.

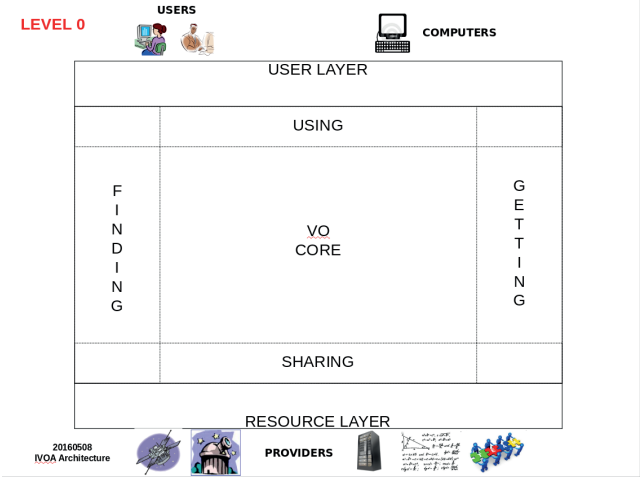


Fig. 1
Architettura dell'IVOA, livello 0, descrizione generale dei concetti e componenti di base

La connessione fra questi due strati è garantita dalle due sezioni verticali (getting e finding), che rappresentano, rispettivamente, i canali di accesso alle e di scoperta delle risorse disponibili da parte degli utenti. Il tutto è permesso da quello che è il cuore dell'architettura (VO core) che non è altro che la base dell'interoperabilità, parola chiave per la comunità VO che, proprio sotto il nome di Interoperability Meeting, si riunisce due volte all'anno per discutere e migliorare le specifiche tecniche.
Già al livello 0 troviamo un primo riferimento alle parole chiave dei principi FAIR, la F di findable, di cui parleremo meglio in seguito al §2, mentre abbiamo poco fa accennato alla I di interoperable.

Il livello 1 dell'architettura IVOA (Figura 2) aggiunge al livello precendente alcune definizioni nel gergo VO e descrive meglio le sue componenti; così il Resource Layer si definisce in funzione delle reali risorse che i providers mettono a disposizione: spazio dati, collezioni di dati e di metadati, risorse di calcolo e servizi di vario genere, mentre lo User Layer mostra che l'utilizzo delle risorse può avvenire tramite applicazioni di vario tipo, grafiche, a riga di comando e integrate nelle pagine web (questo avviene fondamentalmente perché il protocollo base di comunicazione del VO è l'HTTP).

I concetti di finding e getting vengono riscritti in funzione delle aree delle specifiche tecniche che se ne occupano: il Registry (finding) che è la parte del VO che si occupa di mantenere un elenco strutturato e annotato delle risorse disponibili attraverso le tecnologie VO, e i Data Access Protocols (getting) che definiscono le specifiche di accesso alle risorse (con chiaro accenno alla A di accessible dei principi FAIR).

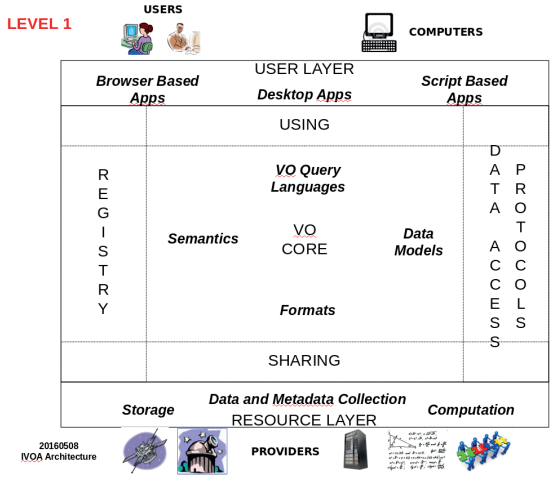


Fig. 2
Architettura dell'IVOA, livello 1, descrizione più dettagliata delle componenti principali

Anche il nucleo centrale di interoperabilità viene dettagliato meglio, proponendo la modellistica delle risorse, la semantica (nelle annotazioni delle risorse e nei dati stessi), i formati dei dati e i linguaggi di ricerca come nuclei fondamentali per garantire l'interoperabilità delle risorse.
Il livello 2 (Figura 3) non fa altro che posizionare nello spazio di competenza i vari standard definiti dall'IVOA per garantire l'interoperabilità delle risorse, così fornendo una panoramica, piuttosto dettagliata e densa, di formati, protocolli, modelli, sistemi di re-

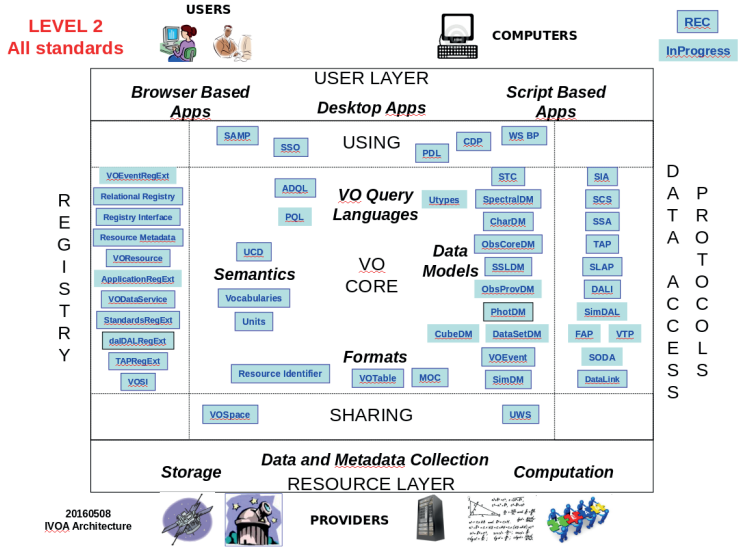


Fig. 3
Architettura dell'IVOA, livello 2: ovvero dove si posizionano tutti gli standard definiti dall'Alliance

gistrazione delle risorse, vocabolari, linguaggi e quanto altro serva in funzione dei casi d'uso portati all'attenzione dell'IVOA. Si vedrà meglio, nel §3, a cosa corrispondano questi standard all'interno dei principi FAIR.

2. I principi FAIR e l'architettura IVOA

Descrivendo l'architettura dell'IVOA abbiamo già incontrato alcuni punti di connessione con i principi FAIR a cui la Comunità Europea fa riferimento nella sua volontà di costruzione di una European Open Science Cloud. L'unica lettera, se vogliamo, mancante, è la R di re-usable.

Tuttavia la comunità astrofisica si è riunita nel VO proprio per permettere un migliore e più efficiente utilizzo dei dati e per scambiarsi informazioni sulle migliori tecniche per descriverli e renderli accessibili e interoperabili. Quindi la riutilizzabilità dei principi FAIR è direttamente connessa al modo in cui il Resource Layer viene costruito a partire dalle specifiche del VO e investe un po' tutta l'architettura del VO, da chi distribuisce i dati a chi li riutilizza, passando per i metodi di ricerca e accesso agli stessi.

Un modo di descriverlo è rappresentato dalla Figura 4, in cui, unendo i primi due livelli dell'architettura VO, e tenendo conto di quanto descritto nel § 1, si mostra come il lavoro svolto dall'IVOA nel dominio dell'astrofisica nei suoi 15 anni di vita altro non sia che quello che oggigiorno è stato definito dall'Open Science e dai concetti dei principi FAIR.
Non tutto quello che è identificato dai principi FAIR è tuttavia parte del VO, nè il VO si ferma alla sola data FAIRness, come si cercherà di mostrare in dettaglio e discutere nella prossima sezione e nelle conclusioni.

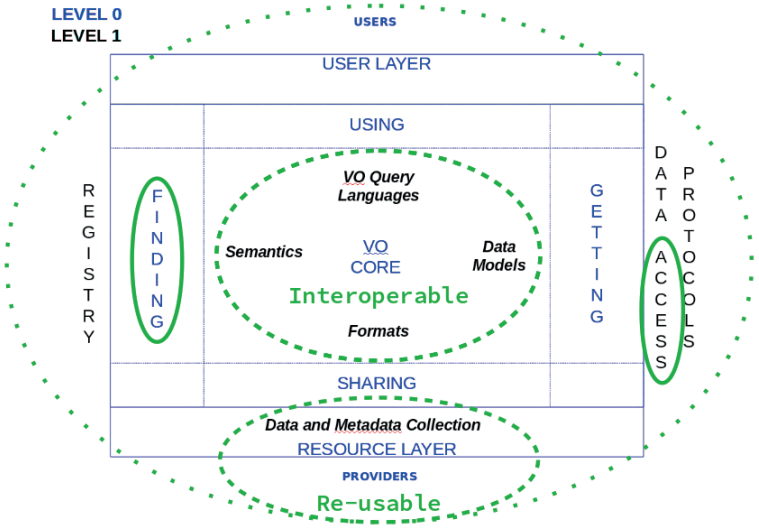


Fig. 4
Rimaneggiamento dei livelli 0 e 1 dell'architettura IVOA con sovrapposti i concetti chiave dei principi FAIR, schematicamente collocati sopra le componenti che più li rappresentano

3. Gli standard dell'IVOA a copertura dei principi FAIR

Andiamo ora in dettaglio, cercando di vedere quanto il VO e i principi FAIR siano simili. Per farlo partiamo riportando i principi FAIR così come riportati in Wilkinson et al. 2016 nella box 2. Li riportiamo in lingua inglese per comodità, anche perché riteniamo che

una traduzione, visti i termini tecnici, possa più nuocere che risultare utile. Nel seguito identificheremo i principi con l'etichetta lettera-numero (es: A1.2) con cui gli elementi dell'elenco puntato iniziano.

To be findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

To be accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

To be interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

To be re-usable:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R2. (meta)data meet domain-relevant community standards.

Quello che si nota subito è il ricorrere continuo del termine metadata, ovvero tutte le annotazioni ai dati che permettono a quest'ultimo di essere compreso da chi il dato deve utilizzarlo anziché apparirgli come un insieme di numeri o descrizioni testuali.

I metadati sono parte del cuore non solo dei principi FAIR, ma anche del VO, dove sono parte fondamentale dei Data Model, per collegare i concetti astratti alle loro rappresentazioni fisiche (compito del Data Model Working Group dell'IVOA) e sono utilizzati dai protocolli e dalle descrizioni delle risorse per permettere un sistema di ricerca, accesso e utilizzo del dato che possa essere automatizzato quanto più possibile e, allo stesso tempo, fornisca un modo di interpretare correttamente il contenuto d'informazione ricercato o recuperato dal data provider.

Ma andando punto per punto nell'elenco possiamo stabilire meglio le connessioni con le specifiche tecniche dell'IVOA. Le specifiche, qui di seguito riportate virgolettate nel testo, sono tutte disponibili presso il Document Repository dell'IVOA all'indirizzo http://www.ivoa.net/documents.

3.1 Findable

F1 richiede l'uso di unique and persistent identifiers; a livello VO esiste una specifica, "I-

VOA Identifiers" che si applica a qualunque oggetto o risorsa (termine generico) si voglia sia identificabile e raggiungibile (collezione di dati, singolo dataset, organizzazione, ...). L'unicità globale di questi identifier è garantita dal modo in cui sono costruiti. Quello che manca, parzialmente è la persistenza indefinita, poiché è legata ai singoli provider essendo l'IVOA uno sforzo collaborativo. Parzialmente perché la specifice del VO per la descrizione di base delle sue risorse sta andando nella direzione di includere i DOI fra i possibili identifier.

F2 parla di metadata richness. Quasi ogni specifica del VO descrive quali metadati andrebbero associati alle collezioni, dati, servizi cui sono associati. Restando strettamente all'interno del Registry, ovvero del repository dove confluiscono e vengono mantenute tutte le descrizioni delle risorse del VO, esiste la specifica "VOResource" che descrive i metadati e la serializzazione di base che ogni risorsa documentata del Registry deve avere. Tale specifica è poi ampliata da estensioni specializzate in funzione del tipo di risorsa che debba essere descritta: "Standards Registry Extension", "Simple Data Access Layer Registry Extension", "VOResource Schema Extension for Describing Collections and Services", "Table Access Protocol Registry Extension". Questo vale per insiemi di dati e servizi, e comprende informazioni sulla creazione, cura, preservazione, localizzazione della risorsa, riferimenti bibliografici, estensione delle collezioni di dati in termini di coordinate spaziali, temporali, dello spettro elettromagnetico, disponibilità del servizio e policy di accesso. Tutti questi dettagli sono disseminati attraverso il registry, mentre altri metadati sono annotati alla risposta dei servizi per descrivere i dati stessi, oltre che i servizi. Per le annotazioni, molto lavoro rientra nelle specifiche del Semantics WG, dove gli UCD "An IVOA standard for Unified Content Descriptors" con relativo "Controlled Vocabulary" e lo standard sulle "Units in the VO" si uniscono ai vocabolari e al loro sistema di manutenzione per permettere l'annotazione di qualsiasi campo dati si voglia descrivere, senza contare le annotazioni specifiche dei Data Model.

F3 è coperto a livello VO dal Resource Registry parzialmente descritto nel paragrafo precedente. In più si può aggiungere che fra le sue interfacce ce n'è una che è basata sullo standard OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting), rendendo le risorse astrofisiche interoperabili (fino a un certo livello) anche oltre il prorpio dominio di ricerca. I metadati di base su cui l'OAI-PMH e lo standard "VOResource" si basano sono quelli noti come Dublin Core , utilizzati universalmente per la metadescrizione di risorse. Da notare che, oltre all'OAI-PMH, i Registry IVOA possono esporre un'interfaccia basata sun un sistema relazionale ("Registry Relational Schema") per una più efficace ricerca condizionale delle risorse desiderate.

F4 richiede che l'identifier sia parte dei metadati, cosa richiesta per ogni risorsa VO. Gli "IVOA Identifier" (IVOID) sono praticamente delle URI specializzate che possono essere risolte nelle rispettive risorse.

In generale, le capacità di ricerca delle risorse in campo VO copre in modo completo le richieste legate ai principi FAIR, forse perfino andando oltre i dettami in quanto si occupa di descrivere non solo i dati, ma i servizi stessi di accesso agli stessi e qualsiasi altra risorsa collegata alle collezioni di dati astrofisici.

3.2 Accessible

A1 ripercorre, per il VO (Registry in particolare), quanto già visto per F3 e F4. Esiste una specifica, "IVOA Registry Interfaces" che descrive i protocolli di accesso alle risorse identificate del rispettivo IVOID, ed esiste un protocollo specializzato (oltre all'OAI-PMH), "Registry Relational Schema", per un accesso ottimizzato con capacità di ricerca filtrata a tutte le risorse del Registry. Oltre a questo, a partire da ciascun IVOID di risorsa dati, è possibile risalire agli eventuali servizi (standard o meno) di accesso alle risorse dati descritte. Questo è fatto attraverso i protocolli definiti dal Data Access Layer WG che descrivono l'accesso a cataloghi di sorgenti astrofisiche, immagini, dati spettrali e ogni altro tipo di risorsa disponibile, incluse serie temporali, sistemi di notifica di eventi transienti, e dati provenienti da simulazioni numeriche. Il punto A1.1 è automaticamente coperto in quanto tutte le specifiche VO sono open e pubbliche, mentre il punto A1.2 è solo parzialmente coperto perché sistemi di autenticazione e autorizzazione sono ammessi e descritti sulle singole risorse (si veda il "Single Sign-On Profile"), ma non esiste un reale sistema interoperabile fra tutti gli attori che richiedano data policy comuni. Esiste anche un protocollo, il "Credential Delegation Protocol", che si occupa di descrivere come le credenziali possano essere passate da un servizio o risorsa ad un altro, permettendo così un'interoperabilità che arriva fino al livello delle risorse di calcolo.

A2 è vero per il VO sebbene non sia auspicabile. La specifica "IVOA Support Interfaces" definisce, in particolare, il sistema grazie al quale una risorsa descrive la sua availability, ovvero il suo stato di attività al momento della richiesta. I metadati del Registry sono presenti e accessibili anche se le risorse specifiche sono temporaneamente non disponibili. Se una risorsa non è disponibile per lungo tempo è prassi etichettarla inactive mantenendo i metadati, ma senza distribuirli.

L'accessibilità di dati e metadati è uno dei punti chiave per l'IVOA, il cui lavoro parte sempre da requisiti, sotto forma di casi d'uso, che descrivono come una determinata risorsa possa essere trovata all'interno del mondo sfaccettato del VO e, una volta individuata, come possa essere acceduta e utilizzata, il tutto in maniera trasparente per l'utente, ovvero senza doverne conoscere ubicazione fisica reale e sistema di gestione specifico.

3.3 Interoperable

I1 descrive la necessità di un linguaggio comune per descrivere le informazioni contenute nei dati esposti. L'IVOA sta terminando lo sviluppo di un metalinguaggio, il "VO-DML - A consistent Modeling Language for IVOA Data Models", proprio per uniformare non solo le annotazioni delle proprie risorse, ma il modo in cui tali annotazioni nascono, ovvero i modelli di dominio più specifico. Esiste inoltre uno di questi modelli, l'"Observational Data Model Core Components", che definisce proprio gli elementi base comuni a qualsivoglia osservazione astrofisica si voglia esporre in rete (definendo inoltre le linee guida per l'uso di un protocollo per l'accesso a tale insieme di annotazioni).

I2 prevede che i vocabolari utilizzati siano a loro volta FAIR. I vocabolari del VO sono degli standard, in quanti tali open, descritti come risorse a loro volta all'interno del Registry, accessibili in vari formati convenzionali (non interni al VO, ma generici quali SKOS o

RDFa) quindi garantendone sia l'interoperabilià interna ed esterna al dominio di appartenenza che il riutilizzo.

La connessione interna, fra metadati e metadati e fra metadati e dati, richiesta da I3, è garantita proprio dall'ambiente complessivo descritto dagli standard dell'IVOA. Le annotazioni che una specifica impone nella descrizione di un servizio rimanda a namespace specifici che permettono di risalire all'insieme dei metadati utilizzati. Tutte le specifiche legate ai gruppi di Semantics e Data Model lavorano proprio in quest'ottica.
L'interoperabilità delle risorse, intese come generiche risorse dati o servizi standard per gli stessi, è al cuore delle iniziative dell'IVOA. Questo fa sì che quanto definito dai principi FAIR sia di fatto già attuato in campo VO, probabilmente anche oltre i dettami stessi.

3.4 Re-Usable

R1 richiede, nuovamente, ricchezza di metadati e ridondanza degli stessi. Ogni servizio che segua le specifiche del VO e sia registrato espone i propri metadati sia attraverso il Registry ("VOResource" e sue estensioni) che direttamente attraverso il sistema di capabilities previsto dalle specifiche "Data Access Layer Interface" e "IVOA Support Interfaces". Le risorse che rappresentano esplicitamente delle collezioni di dati hanno i loro metadati esposti in maniera globale attraverso il Registry e trasmettono i loro dati, con formato predefinito in forma di "VOTable Format Definition", accompagnati da un insieme di annotazioni così come richiesto dai vari protocolli e standard. Tutto questo usando i modelli, i descrittori e i vocabolari descritti, ad esempio, nella sezione §3.3 dedicata all'interoperabilità.
R1.1 specifica che debba esistere un chiara espressione delle licenze applicabili ai dati. Una prima considerazione da fare qui è che le risorse del VO partono generalmente non dall'essere semplicemente open ma dall'essere solitamente pubbliche, fatto salvo un possibile iniziale periodo proprietario che ne garantisca un'indagine prioritaria da parte di chi ha richiesto l'osservazione o il run di simulazione. Fatto salvo questo, a livello di Registry c'è la possibilità di definire delle licenze per le risorse, possibilmente facendo riferimento (tramite URI globali esterni al mondo VO) alle licenze solitamente utilizzate, quali possono essere le Creative Commons o le GPL.

R1.2 entra nel merito della tracciabilità del dato, dalla sua origine, passando per le possibili manipolazioni allo stesso, gli attori coinvolti e tutto quanto definisca la storia del dato come risorsa. Al di là delle informazioni di base reperibili attraverso l'uso del Dublin Core a livello di registrazione delle risorse VO, è ormai pronto per la standardizzazione il "Provenance Data Model", che si occuperà proprio degli aspetti sopra citati all'interno del VO, prendendo il via dall'analogo standard del World Wide Web Consortium (W3C) per garantirne l'interoperabilità verso risorse cross-domain fuori dal campo specifico dell'astrofisica.

R1.3, infine, richiede a chi espone risorse dati, nel nostro caso in astrofisica, nient'altro che quanto l'IVOA ha cercato di fare nei suoi 15 anni di vita e sta proseguendo a fare anche oggi.

Il ri-utilizzo dei dati e delle risorse, ovvero la possibilità di individuare le risorse più adatte al proprio lavoro di ricerca o indagine, parte della necessaria collaborazione fra più strumenti, sistemi di indagine e osservatori necessari alla ricerca astrofisica. L'infra-

struttura informativa definita dall'IVOA ne prende semplicemente atto e definisce le linee guida per poterla attuare.
Un punto base per capire quanto la ri-utilizzabiltà dei dati sia importante nel VO lo definiscono i suoi standard di formato ("VOTable Format Description", "HEALPix Multi-Order Coverage Map", "Hierarchical Progressive Survey" e "Simple Application Messaging Protocol"), che uniscono in un sistema unico dati, metadati e applicazioni e permettono uno scambio fra le parti che garantisce la comprensione sia da parte di attori diversi che in diversi momenti, per esempio dopo un salvataggio su un sistema locale per un successivo riutilizzo.

4. Conclusioni

Nel capitolo §3 abbiamo descritto come le specifiche e l'architettura dell'IVOA non siano altro che un'implementazione che ha precorso i dettami dell'Open Science e dei principi FAIR per il dominio di ricerca dell'astrofisica.

Al di là di quanto l'IVOA ha fatto e sta facendo bisogna ricordare che il VO stesso prende spunto da una realtà, quella del formato FITS (Flexible Image Trasnport System, Pence et al. 2010) che già alla fine degli anni '70 preparava un formato di scambio dati completo di annotazioni e metadati, incluso un minimo modello interno per gli stessi. Questo formato ha giocato e gioca un ruolo attivo nell'accomunare la comunità astrofisica e l'IVOA ne ha sempre tenuto conto in fase di sviluppo dei propri standard, per non divergere da quanto la comunità già conosceva e utilizzava. Questo è vero al punto che i principi F2, I1, R1, R1.2 e R1.3 sono già rispettati proprio da questo formato, che definisce quindi già un piccolo sistema di scambio dati FAIR in sé stesso.

Restano ovviamente delle differenze fra quanto esiste nel mondo VO dell'astrofisica e i principi FAIR. Da un lato si può dire che tutti i dati esposti tramite il VO sono FAIR, dall'altro però bisogna specificare che non tutti i dati astrofisici sono raggiungibili attraverso il VO e che la conformità ai principi FAIR non è sufficiente per l'astrofisica.

Per far sì che i dati astrofisici diventino tutti FAIR è necessario uno sforzo implementativo da parte dei centri dati e provider dell'astrofisica (istituzioni, organizzazioni, enti di ricerca, ...), ad esempio inserendo uno strato di traduzione per l'accesso ai propri dati che rispetti gli standard VO. Questo è possibile se gli attori in gioco e l'IVOA lavorano di concerto nello sviluppo e miglioramento degli standard.

Ma è anche necessario andare oltre la FAIRness dei soli dati. É importante che servizi di calcolo specifici possano essere agganciati alle risorse dati stesse. In astrofisica questo è necessario per fornire immagini ridotte, ritagliate sulle necessità dell'utenza, per generare cataloghi di sorgenti e oggetti, per permettere che le moderne tecniche di machine learning possano interagire direttamente coi dati. Questo è possibile se l'IVOA contribuisce a definire degli standard per la descrizione, reperibilità e utilizzo di queste risorse e se i provider dei dati si pongono nell'ottica di fornire capacità di calcolo sopra i dati stessi.

Riferimenti bibliografici

Christophe Arviset, Severin Gaudet and the IVOA TCG, "IVOA Architetcure – version

1.0", IVOA Note, 23 Novembre 2010

Pence, W. D.; Chiappetti, L.; Page, C. G.; Shaw, R. A.; Stobie, E., "Definition of the Flexible Image Transport System (FITS), version 3.0", Astronomy and Astrophysics, Volume 524, id.A42, Dicembre 2010

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons (2016), "The FAIR guiding principles for scientific data management and stewardship", Nature – Scientific Data, 3, DOI: 10.1038/SDATA.2016.18

https://ec.europa.eu/research/openscience/index.cfm

https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

http://www.openarchives.org/OAI/openarchivesprotocol.html

http://dublincore.org/documents/dcmi-terms/

Autori

Marco Molinaro marco.molinaro@inaf.it

Laureato in Fisica, con indirizzo in Astrofisica, si occupa ormai da oltre un decennio di data management e data science presso l'INAF-OATs e in connessione al centro Italiano Archivi Astronomici (IA2). É membro dell'IVOA e del suo Technical Coordination Group, dove attualmente ricopre il ruolo di vice-chair del Data Access Layer Working Group.

Fabio Pasian fabio.pasian@inaf.it

Astronomo Ordinario presso INAF-OATs, è stato a capo dell'archivio HST di ESO, PI degli archivi del TNG, manager del DPC di Planck e coinvolto attualmente nell'SGS di EUCLId. Da sempre coordinatore delle attività VO italiane e referente in tal senso a livello europeo (EURO-VO) e internazionale (come membro dell'Executive Committee dell'IVOA).

DOI: 10.26314/GARR-Conf17-proceedings-21

Accesso ai dati astronomici e radioastronomici: Autenticazione e Autorizzazione in INAF

Franco Tinarelli [1], Sonia Zorba[2], Cristina Knapic[2]

[1] INAF Istituto di Radioastronomia, [2] INAF Osservatorio Astronomico di Trieste

Abstract. L'Istituto Nazionale di Astrofisica gestisce i dati prodotti dalle osservazioni di una serie di telescopi e radiotelescopi, i dati vengono archiviati nei DB gestiti dal servizio IA2. Per l'accesso ai dati è stata sviluppata una suite di applicazioni composta da un modulo di autenticazione chiamato RAP (Remote Authentication Portal) che permette l'autenticazione con eduGAIN, Google, Facebook, LinkedIn, X.509 e con account registrati localmente. La suite è completata da connettori per l'interazione con Grouper, un tool Java EE sviluppato da Internet2 per la gestione dei gruppi e delle identità.

Keywords. Autenticazione, Autorizzazione, Grouper, Account-Linking, eduGAIN

Introduzione

L'Istituto Nazionale di Astrofisica gestisce i dati prodotti dalle osservazioni di una serie di telescopi (Asiago, TNG e LBT) e radiotelescopi (Medicina, Noto e SRT). Essi vengono archiviati nei DB gestiti dal servizio IA2. Per l'accesso ai dati è stata sviluppata una suite di applicazioni, in collaborazione tra IRA (Istituto di Radioastronomia) e IA2 (Archivi astronomici Italiani). La suite è composta da un modulo di autenticazione chiamato RAP (Remote Authentication Portal) che permette l'autenticazione con eduGAIN, Google, Facebook, LinkedIn, X.509 e con account registrati localmente. La suite permette inoltre l'account-linking ed ha un connettore per l'interazione con Grouper, un tool Java EE sviluppato da Internet2 per la gestione dei gruppi e delle identità.

1. RAP

RAP (Remote Authentication Portal) è un'applicazione web scritta in PHP ed è completamente indipendente dalle applicazioni che lo usano come autenticatore. L'applicazione chiamante viene registrata in un file che contiene l'indirizzo di call-back per ritornare i dati dell'utente che si è autenticato.

Le principali funzionalità del programma sono:

- autenticazione con diversi metodi;
- account-linking;
- registrazione in MySQL o LDAP;
- editing dei profili registrati.

Ciascuna delle funzionalità può essere attivata o disattivata a piacimento e se disattivata viene nascosta nell'interfaccia utente. Il meccanismo di autenticazione viene reso più sicu-

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

ro tramite l'associazione della richiesta di autenticazione ad un token, inviato all'applicazione chiamante che lo userà come chiave per richiedere le informazioni di autenticazione, con conseguente eliminazione di dati transienti e token immediatamente dopo l'invio.

Fig. 1
RAP: l'interfaccia utente

La registrazione degli utenti può essere effettuata direttamente da RAP su proprie tabelle associate all'applicazione chiamante o remotamente su DB della stessa applicazione. Analogamente all'indirizzo di call-back anche le informazioni specifiche dei DB, vengono associate all'applicazione chiamante in un file di configurazione dei client.
RAP può utilizzare indifferentemente un DB relazionale o LDAP per la registrazione degli utenti sia in locale che in remoto per la funzionalità di account-linking.
L'utilizzo di LDAP permette, attraverso una procedura di gestione, di accreditare gli utenti al login via SSH su workstation che lo utilizzino come sistema di autenticazione. Successivamente la stessa funzionalità può esser estesa all'utilizzo di Kerberos per quelle applicazioni che ne richiedessero l'utilizzo.

RAP è Open Software e può essere adattato e inserito in una propria applicazione, come realizzato dal team di IA2. Le attuali versioni di RAP sono ospitate su server Apache (httpd). Specifici moduli di Apache sono stati configurati per effettuare la validazione dei certificati X.509 e per realizzare l'autenticazione SAML utilizzando uno Shibboleth Service Provider.

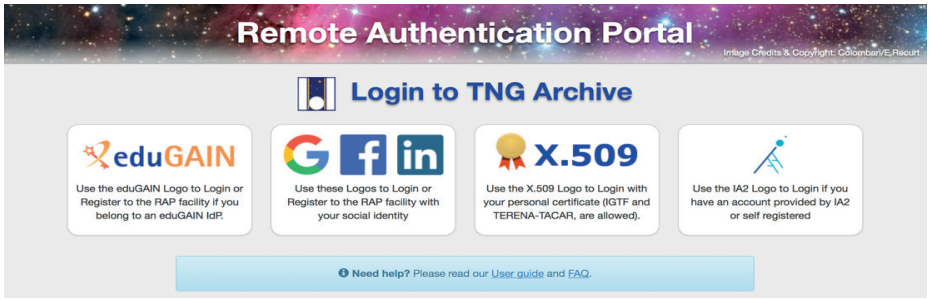


Fig. 2
IA2: Accesso al Data Base del telescopio TNG

2. Account-Linking

L'account-linking è stato realizzato con due possibili implementazioni alternative. La prima prevede, all'atto della registrazione, l'invio via e-mail di un codice univoco che identifica l'utente. Se l'utente desidera unire due suoi account può quindi inserire questi codici all'interno dell'interfaccia di RAP. La seconda implementazione prevede che l'utente effettui un login su una pagina di gestione del suo account e ricerchi altri utenti registrati ai quali inviare una "richiesta di join". L'utente target della richiesta riceverà un messaggio e-mail con un link di conferma, contenente un token univoco che identifica la richiesta di join. L'account-linking avviene solo se l'utente apre il link e clicca su un pulsante di conferma.

3. Grouper

Grouper è stato scelto da IA2 per organizzare le autorizzazioni d'accesso alle risorse fornite tramite i propri servizi in quanto strumento maturo e già utilizzato con successo da altre organizzazioni che operano nell'ambito della ricerca. Esso inoltre ha il vantaggio di fornire un'interfaccia web che consente di delegare agli utenti alcune delle operazioni di amministrazione dei gruppi. Grouper non è nativamente in grado di gestire l'account linking, e la relativa autorizzazione per questo è stato necessario personalizzarne alcune componenti, in modo da renderlo compatibile con il modello dati utilizzato da RAP per rappresentare gli utenti.

Grouper memorizza le informazioni relative a gruppi e permessi all'interno di un suo database, detto registry. L'installazione di Grouper utilizzata da IA2 si appoggia attualmente su un database MySQL, tuttavia, poiché Grouper è basato sulla tecnologia ORM Hibernate, si potrebbero utilizzare indistintamente molte altre tipologie di RDBMS.
Grouper è stato installato su server Tomcat.

4. Il Connettore

La suite è completata dal connettore tra RAP e Grouper, sviluppato dal team IA2, che permette l'autenticazione su Grouper tramite l'utilizzo di un sistema multi protocollo che non era nativamente supportato da Grouper.

Grouper con la modifica apportata alla sua nativa basic authentication, importa da RAP le diverse identità possedute dall'utente che si è autenticato come un'unica, per gestire i gruppi di cui è amministratore. RAP espone queste informazioni attraverso un servizio REST protetto da password.

Due moduli di Grouper sono stati personalizzati per poter dialogare con questo web service: il Source Adapter e l'Authentication Filter. La creazione di questi componenti custom avviene estendendo delle classi Java, modificando i file XML della configurazione di Grouper e infine ricompilando Grouper.

Nel gergo di Grouper un Source Adapter è un componente che può essere interrogato per ricavare informazioni riguardo un insieme di utenti. Ogni installazione di Grouper può avere uno o più Source Adapter. I Source Adapter messi a disposizione da Grouper permettono di interrogare LDAP o database relazionali. Il Source Adapter scritto da IA2 interroga invece il servizio REST di RAP, che restituisce le informazioni in formato JSON.

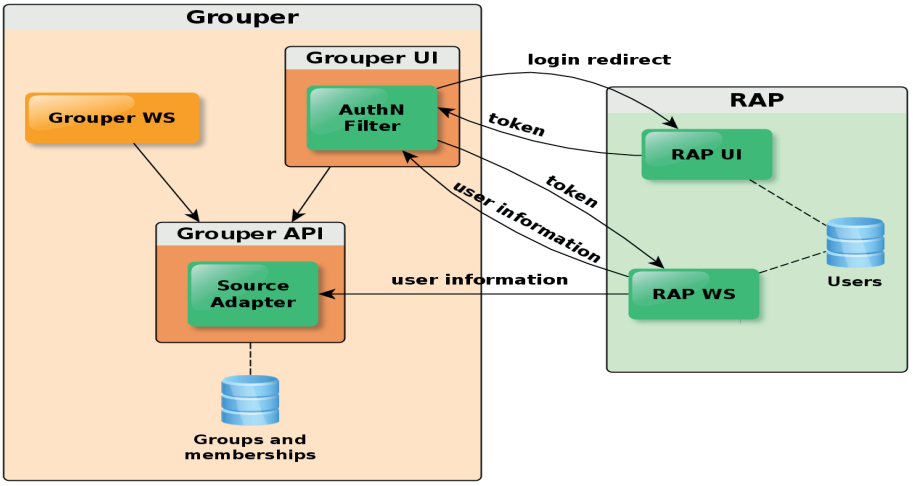


Fig. 3
RAP + Grouper: schema funzionale

In questo modo un insieme di identità sulle quali è stata effettuata una procedura di account-linking viene interpretato da Grouper come un'entità atomica. Nella Grouper UI, a fianco del nome di ogni utente vengono elencate le sue diverse identità.

L'Authentication Filter è un servlet filter che va configurato nel deployment descriptor della Grouper UI. Verifica la presenza di un utente associato al cookie di sessione, in caso contrario effettua un redirect su RAP e si autentica allo stesso modo degli altri client.

5. Conclusioni

L'implementazione di un meccanismo di autenticazione multi protocollo (SAML2.0, OAuth2, X.509), la registrazione automatica su RDBMS, LDAP e Kerberos, l'account-linking delle identità e la gestione di gruppi di utenti permette oggi l'accesso ai dati prodotti dagli strumenti osservativi di INAF e permetterà in futuro l'accesso alla impressionante mole di dati prodotta dal radiotelescopio SKA.

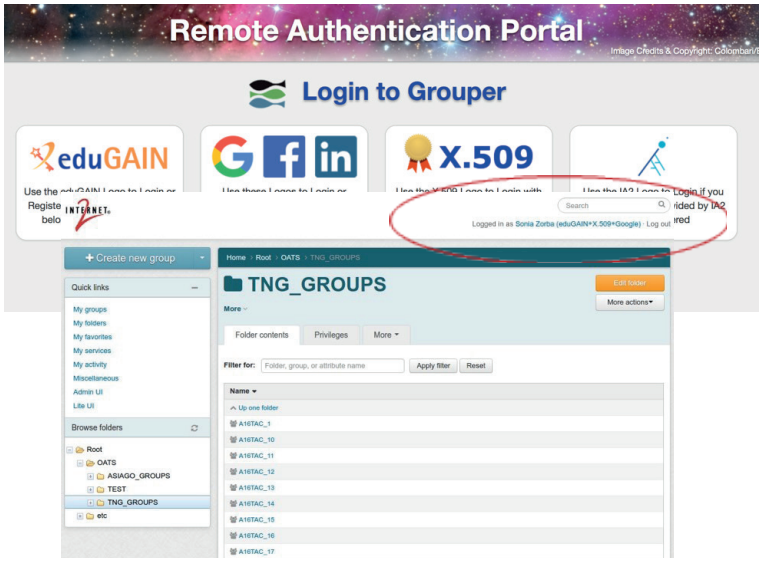


Fig. 4
Login in Grouper utilizzando RAP

Riferimenti bibliografici

F. Pasian, S. Bertocco, A. Bignamini, A. Costa, G. Jerse, C. Knapic, M. Molinaro, E. Sciacca, G. Taffoni, F. Tinarelli, F. Vitello, S. Zorba (2017), "Authentication & Authorization Technology Benchmarking Report", Asterics Project

Autori

Franco Tinarelli f.tinarelli@ira.inaf.it

System e Network Manager dell'Istituto di Radioastronomia dal 1988. Ha sviluppato software per la schedula delle osservazioni VLBI. Partecipa al work package SKA TM per il quale ha sviluppato il primo prototipo di A&A e il software RAP.

Sonia Zorba zorba@oats.inaf.it

Dal 2015 lavora come full-stack web developer presso il Centro Italiano Archivi Astronomici (IA2). Sviluppa interfacce per l'accesso ai dati e tool di supporto, sia ad uso interno che nell'ambito del Virtual Observatory.

Cristina Knapic knapic@oats.inaf.it

Tecnologo presso INAF-OATs, si occupa degli archivi astronomici italiani (IA2) dal 2008. Ha sviluppato sistemi di archiviazione distribuita, interfacce e servizi web compatibili con il Virtual Observatory. Attualmente si occupa di vari work packages sia di SKA che dei SKA Regional Centers partecipando al progetto AENEAS.

DOI: 10.26314/GARR-Conf17-proceedings-22

Life (of big storage) in the fast lane

Ivan Andrian, Roberto Passuello, Iztok Gregori, Massimo Del Bianco

Elettra Sincrotrone Trieste

Abstract. Parkinson's law: work expands so as to fill the time available for its completion.Corollary applied to computers: Data expands to fill the space available for storage. As time goes by, your storage will become too small. And too slow. And the pace is accelerating. What can we do about it? This paper presents Elettra Sincrotrone Trieste's experience in managing scientific data generated by its two lightsources, using state-of-the-art echnology and tools, taking them to the limit just to discover some shortfalls and weaknesses

1. Facing the problem

Lightsources like Elettra, a third-generation synchrotron radiation facility, or FERMI (acronym for Free Electron laser Radiation for Multidisciplinary Investigations), the new seeded free electron laser (FEL) accelerator operating next to Elettra, are using a lot of detectors which produce etherogeneous data in form of streams of floating point numbers or raw images. Analysis like Computed Tomography (CT) scans are being used more and more frequently in many beamlines which are striving to increase the quality of their experiments on samples. This increase in quality usually can be performed by having better radiation light characteristics, by increasing the number of images taken per second and, last but not least, by using bigger detectors. From the historical "kilopixel" sensors, we are now in the era of various megapixel (MP) detectors, usually ranging from 1MP to 13MP, operating at higher frequencies compared to the past times (e.g., upgrading from 5 to 10Hz or even 120Hz, now entering the Khz range). In time-resolved studies (4DCT), several tens of datasets can be collected in sequence, yielding TB of data to be stored and managed (10 TB/day with Elettra and up to 100 TB/day with the planned Elettra 2.0).

Advanced algorithms and processes are being developed to handle this huge amount of acquired data before it is even stored and used for further analysis [DiamondTC]: however, even reducing the amount of data, the result can be in the order or TB/day per active beamline which means that 1PB of data per year is nothing less than reality right now.

However, it must be stressed that not all these data are here to stay. Every investigation needs to be analysed and, eventually, part of the data will be deleted because useless or redundant. Lossless data compression algorithms can highly reduce the size of images.

The amount of storage needed for these kind of jobs is not the only problem: as anticipated, the sampling frequency is increasing which, also considering the bigger size of data, brings to the high throughput requirements of the storage system that will handle the data itself.

Huge, fast and... cheap, of course! These are the easy requirements for the storage sy-

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

stems at any lightsource facilities these days.

2. Evolution of storage at Elettra

Prehistoric era: minicomputers

Elettra was built in the early nineties, starting operating with its first beam in 1993. At that time, the main storage facility for scientific data (as well as the technology for data analysis) was based on a number of DEC AlphaServer 2100 and VAXstation machines running OpenVMS and, lately, Tru64 Unix (Fig. 1). When the use of personal computers became common, these were widely used as storage systems at the beamlines against the DEC servers; as a side effect, this led to a very etherogeneous situation with no centralised standard of access for the data. After a number of years this anarchic situation came to an end: a more organised approach was mandatory, and the needs for a high performance, reliable, centralised storage were born.

Fig. 1
DEC AlphaServer 2100

Middle age: SANs

When the hype of data storage anarchy was gone leaving only the dark face of the trend, the new Storage Area Network products looked like the cure for any disease. EMC2 entered the Elettra datacenter with a CX4-240 machine and everybody was happy with its assistance, internal redundancy and scalability. After a bit of use, however, some important limits of the system became clear to the operators: even if rock solid, so that a good lifetime of about ten years could be foreseen, the performances were not as needed. Expandibility was one key factor of the system, however it came at a cost: every original spare part, even if similar or even identical (i.e., rebranded) to many other products on the unbranded market, costed twice. When it came the time of a SAN expansion, right after the end of the included support period, the estimated cost of support renewal and new disks was absurdly high. A quick market research suggested that it was cheaper to buy a new complete solution based on Commodity Off The Shelf (COTS) hardware.

A new hope: DFS on COTS

At the beginning of the new millennium the server area of the PC market was quickly

evolving in technology and power, while the prices were going down. Linux was becoming more stable and powerful, gaining many features ready for the enterprises, in particular for the research centres where the IT departments have always been interested in state-of-the-art technology. The evolution of the storage at Elettra passed through dedicated fileservers based on the x86_64 architecture with a lot of disks onboard, managed by dedicated RAID controllers and usually exporting the volumes via NFS to the data acquisition workstations at the beamlines. Pretty soon the increasing number of such servers was causing problems in terms of mainteinability, not counting the issues when the volumes needed to be expanded, operation that could often lead to downtimes due to physical installation of new hardware. At that time, however, open source Distributed Filesystems (DFS) where production ready.

A Big Bench to accomodate all the data

Experiments with distributed filesystems at Elettra were performed in the past. The initial choice of IBM's General Parallel FileSystem on Linux (GPFS, now rebranded as IBM Spectrum Scale [GPFS]) was promising, with only minor stability problems when performing maintenance operations on problematic volumes. GPFS is always been a commercial product, but at the time it was offered at no cost for research facilities. Suddenly, IBM changed this license and the costs were not sustainable for our laboratory; it was then decided to move to an Open Source DFS. After some tests the choice went to Gluster [gluster], a promising technology initially developed by Gluster Inc. and then bought by Red Hat. Gluster exports one or more underlying filesystems to a cluster; both spanned and replicated (or mixed) configurations can be made, providing expandibility and redundancy. The simplified specifications of the Elettra "Bigbench" storage cluster commissioned in 2011 were the following:

- 7 Supermicro 4U servers;
- 24 x 3TB SAS disks per server (72TB raw per node);
- 2-copies mirrored volumes between the servers to cover possible failures or maintenance downtimes;
- total of 504 TB raw, 252 TB net;
- RAID 6 with dedicated controller on every server for in-node disk failure protection;
- LVM + XFS (lately, ZFS) as underlying gluster brick structure;
- dedicated 10Gbps network between the cluster nodes, 1Gbps network to the client workstations;
- glusterFS accessed by some XEN virtual machines used as frontend peers for the data acquisition machines to which the volumes were exported via NFS.

Before moving the servers into production some tests were performed on the machines in order to get some performance benchmarks. Considering the RAID controller, LVM and XFS, the iozone tool gave results around 1GB/s both in reading and writing, not considering the controller cache. However, performances were affected by the pile of layers, in particular Gluster and, obviously, by the network itself with its theoretical limit of 10Gb/s. Analysing the whole architectural setup as depicted in fig. 2, it was decided to get rid of

two layers: LVM and the RAID controller. This was made possible at first by moving to a different filesystem for the bricks: from XFS to ZFS. The performances of the latter, however, were negatively affected by the RAID hardware connecting the disks.

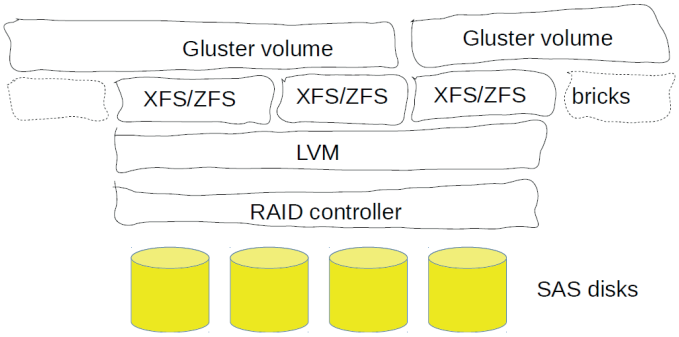


Fig. 2
Bigbench architecture v1

Tests done with disks directly attached to a simple HBA demonstrated that the ZFS performances were at least at the same level of the RAID+XFS. Added to this, ZFS can act as an LVM with its concept of pools and volumes, so this layer became useless. The resulting configuration, used for additional nodes added to the cluster, is presented in fig. 3.

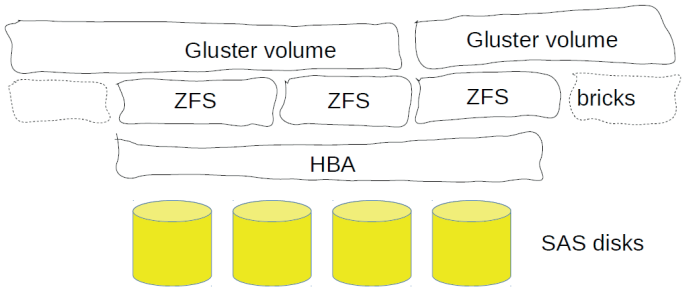


Fig. 3
Bigbench architecture v2

This solution resulted to be successful and was able to accommodate the needs, in terms of space and speed, of most of the beamlines at the time. However, some strange behaviours were noticed by some beamlines in particular job configurations. The performances dropped down dramatically from time to time: the bottleneck was eventually recognised in Gluster itself. In fact, the Gluster architecture suffers of some limitations, in particular due to its userspace nature for GlusterFS and very poor performance when managing a great number of small files (due to its metadata in the filesystem-only feature). Modular expansion can be performed permitting horizontal scaling, with the constraints, for example, to keep the duality of servers if mirror volumes are to be implemented. These drawbacks, together with the introduction of new detectors with higher storage speed needs, brought the IT department to the decision of a new change in technology.

Relaxing constraints on a Sofa

Object-based storage, in contrast to filesystem-based storage, was in development for years and many projects were looking promising. Around the end of 2015 the CEPH distri-

buted storage system turned into being stable and some internal tests validated it for its adoption at Elettra. The new storage cluster, Sofa, was born with these specifications:

- 4 3U Supermicro servers (plus other 4 added during a first minor upgrade);
- 20 x 6TB 7200RPM SAS disks per server (120TB raw per node);
- journaling dedicated on 4 SSD per node (one SSD serves 5 HDD);
- CEPH configured in replica 2 between the OSDs (storage units), with copy constrained to be on different nodes;
- total of 480 TB raw (960TB after the upgrade);
- additional "caching" tier for lower CEPH volumes on 3 servers with:
- 20 x 2.5" 600GB 15000RPM SAS disks per server
- journaling on 4 x 800GB HGST Ultrastar SN100 NVMe
- 40Gbps dedicated network for the cluster nodes, 1Gbps,10Gbps and 40Gbps to the clients;
- volumes exported via RBD to KVM virtual machines (Proxmox), used as frontend peers for the data acquisition machines to which the volumes are exported via NFS or SMB.

This solution proved to be a big success in terms of flexibility in allocating storage to clients, horizontal scaleup capability, high iops and I/O throughput. Even if we haven't been able yet to extensively take the system to its I/O limits with data read and written from the clients, during a CEPH self-healing repair due to a number of offline OSDs we were able to see I/O rates above 2.7GBps in combined reading and writing (Fig. 4). However, during an operation session of FERMI's most demanding endstation, an incoming throughput of more than 800MBps has been seen.

Still, we were able to face some issues not foreseen at the time of design. In particular, the choice of having the journaling of 5 HDD OSDs on one SSD was not a good choice for two reasons. We were confident that SSD were very reliable and their speed was optimal for the journaling role, as suggested by the CEPH documentation itself [CEPHSSD]. However, we didn't consider both the total writes limit of SSDs (e.g., a Kingston SE50S37 of 100GB is guaranteed with 310TB of Total Writes – TBW – at 3 Drive Writes per Day – DWPD) and that when an SSD fails it affects 5 OSDs. After only three months of operation we realised that the expected lifetime of our SSDs was one year at maximum – not a great deal in the long term for a critical system like this. In only 4 days we also experienced a sudden failure of half of the SSD installed on three servers due to a manufacturing defect that was present in one production lot. As a result, in that short period we had OSDs in two server being put offline at the same time because of their unavailable journals. The number and location of the remaining OSDs were too small to guarantee the operation, and the system went into a state of mulfunctioning. Fortunately, the design of CEPH is rock solid so that, just after replacing the SSDs and a bit of sysadmin's tricks to help the data relocation routines, the OSDs came back online and the storage was put into service again with no data loss.

A bigger and better Sofa

We are now expanding the CEPH storage cluster, adding 8 new servers with 24 x 10TB

SAS HDD (240TB per server), getting rid of all the SSDs and upgrading to the latest stable version of CEPH (Luminous), which brings Bluestore into stable state. Bluestore, the new data writing method, will guarantee higher performances even without SSD journaling. The total raw space will increase to ~ 3PB, wich means 1PB net with 3 replicas of the data (in different nodes). The volumes will be exported both in RBD and in CEPHFS.

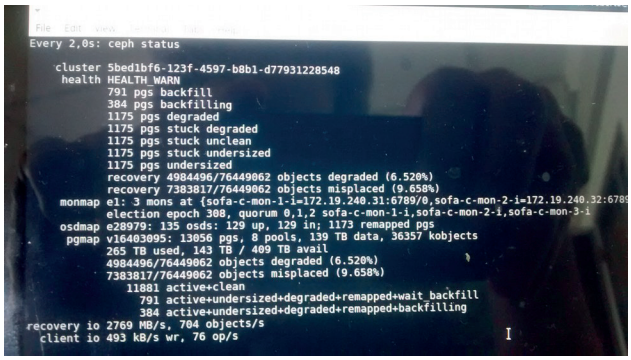


Fig. 4
State of the CEPH cluster during a recovery operation

Acknowledgement

We would like to thank A. Curri which was the original designer of Bigbench and Sofa, and The Eagles for inspiring the title of this paper.

References

[DiamondTC] Robert C. et al., A high-throughput system for high-quality tomographic reconstruction of large datasets at Diamond Light Source, 2015

[GPFS] https://www.ibm.com/us-en/marketplace/scale-outfile-and-object-storage
[gluster] http://www.gluster.com

[CEPHSSD] http://docs.ceph.com/docs/master/rados/configurati on/ osd-config-ref/?highlight=ssd

DOI: 10.26314/GARR-Conf17-proceedings-23

Gestione distribuita dei dati sperimentali da prove su tavola vibrante per la protezione sismica di murature storiche

Irene Bellagamba[1], Francesco Iannone[2], Marialuisa Mongelli[2], Silvio Migliori[2], Giovanni Bracco[2]

[1] Consortium GARR, [2]Centro Ricerche ENEA.

Abstract. Allo scopo di rafforzare la collaborazione tra esperti della comunità scientifica che operano nel settore della protezione sismica del costruito storico, l'ENEA ha recentemente sviluppato, nell'ambito del progetto C.O.B.R.A., un'architettura denominata ENEA Staging Storage Sharing (E3S), per l'archiviazione, la condivisione e l'analisi dei dati sperimentali prodotti da differenti labora-tori di ricerca dell'ENEA distribuiti geograficamente. In tale architettura, sono stati sviluppati tool di visualizzazione grafica dei dati, servizi web e infrastrut-ture cloud storage, per la condivisione e l'analisi in tempo reale dei dati sperimentali acquisiti durante prove su tavola vibrante condotte presso il laboratorio SITEC (Sustainable Innovation TEChnologies) dell'ENEA. Il presente lavoro intende mostrare l'architettura applicata nell'ambito di una campagna sperimentale su tavola vibrante per l'analisi del comportamento sismico di una muratura tipicamente usata nei borghi storici del centro Italia.

Keywords. real time data streaming; scientific data sharing; distributed data management; seismic pro-tection; shaking table tests.

Introduzione

L'architettura E3S è stata sviluppata ed integrata, nell'ambito del progetto COBRA (Sviluppo e diffusione di metodi, tecnologie e strumenti avanzati per la COnservazione dei Beni culturali, basati sull'applicazione di Radiazioni e di tecnologie Abilitanti) e supporta l'intero processo di gestione dei dati, sia acquisiti sperimentalmente che post-elaborati e prodotti da analisi strutturali ad elementi finiti (FE).

I dati sperimentali sono acquisiti da un sistema optoelettronico, 3DVision, che utilizza una costellazione di telecamere NIR (Near Infra Red) per la misura nel tempo delle posizioni di numerosi marcatori retroriflettenti disposti sulle strutture in prova.

Per ogni step di prova i dati vengono scritti all'interno di un file in formato standard C3D temporaneamente memorizzato in un'area del disco locale del sistema di acquisizione e successivamente sincronizzati con le aree di storage dipartimentali dell'infrastruttura ENEA. Quest'ultime sono basate sul filesystem AFS distribuito geograficamente su tutti i centri di ricerca ENEA, utilizzato per la condivisione dei dati su WAN e sul filesystem parallelo GPFS utilizzato dai sistemi di calcolo ad alte prestazioni (HPC).

Un servizio web, denominato "DySCo Logbook", è stato sviluppato per gestire sia l'inserimento dei metadati in database relazionale, sia il processo di streaming dei dati sperimentali acquisiti durante le prove. Gli utenti autorizzati possono così accedere di-

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

rettamente ai dati per la calibrazione da remoto di modelli FE delle strutture in prova, sfruttando le potenzialità dei codici di calcolo disponibili sull'infrastruttura HPC CRESCO (Computational Research Centre for Complex Systems), al fine di migliorare le simulazioni successive e l'affidabilità dei modelli per futuri test sperimentali.
L'efficacia dell'architettura viene mostrata nell'ambito di una campagna di test sismici eseguiti all'interno di un progetto di cooperazione e trasferimento scientifico-tecnologico Italia-USA, finalizzati alla valutazione del comportamento dinamico di strutture murarie e alla verifica dell'efficacia di sistemi di rinforzo innovativi.

1. L'architettura E3S

L'architettura E3S consente l'archiviazione fisica dei dati scientifici sull'infrastruttura ENEA garantendone l'integrità e la sicurezza nonché la condivisione su area geografica. Essa è basata su tre componenti funzionali principali (Figura 1): il "Gateway Node" (GWN), il "Middleware Node" (MWN) e il "GPU Application Server" (GAS).

E3S consente principalmente di archiviare temporaneamente i dati acquisiti in aree di staging locali sincronizzate con aree di storage, fornendo un servizio denominato sync-storage. Le aree di staging vengono gestite da un componente funzionale denominato Gateway. Grazie all'utilizzo di tecnologie basate su cloud storage i dati vengono archiviati in sicurezza anche quando il link di rete risulta non essere disponibile.

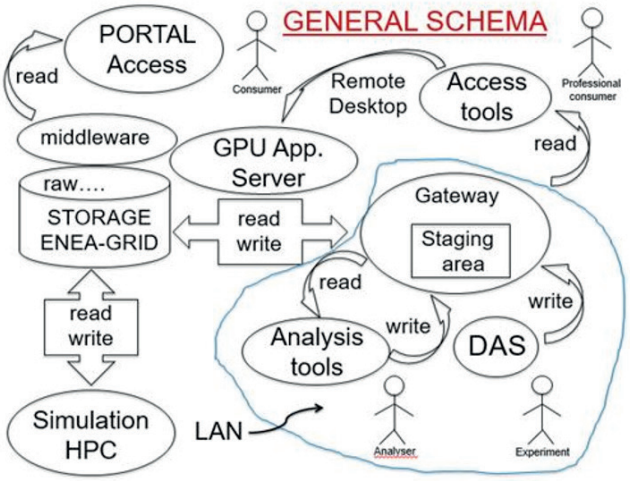


Fig. 1
Schema E3S

L'accesso ai dati archiviati nelle aree di staging è reso possibile grazie all'utilizzo di strumenti di analisi e visualizzazione che sfruttano il modello client/server per il servizio di accesso remoto ai dati. Oltre alla funzionalità di staging, il GWN gestisce la sincronizzazione delle suddette aree con le aree di storage del filesytem della cella AFS: enea.it, utilizzato per la condivisone dei dati su rete geografica, o in alternativa con le aree di storage del filesystem parallelo GPFS per l'accesso dei sistemi HPC.
Sia AFS che GPFS permettono a tutti i sistemi computazionali di ENEA di condividere i

filesystem fornendo servizi di accesso alle aree di storage. Tali servizi vengono gestiti dal secondo componente funzionale dell'architettura, denominato Middleware che consente inoltre di effettuare il data-sharing. Il componente funzionale denominato GPU-Application Server fornisce un ambiente grafico integrato per l'esecuzione di applicazioni su piattaforme hardware basate su GPU accessibili remotamente. L'acceso al GAS da parte di un'utenza specializzata permette l'elaborazione dei dati condivisi nelle aree di storage con specifici criteri di autorizzazione attivabili da sistemi sicuri di autenticazione.

2. Condivisione delle prove sperimentali su tavola vibrante

La campagna di prove sperimentali su tavola vibrante è stata eseguita su due paramenti murari, uno in blocchi di tufo e l'altro in pietra, rappresentativi dei materiali degli edifici storici tipici del centro Italia.

Le murature sono state sottoposte a 5 input sismici ad intensità crescente, imposti mediante una tavola vibrante a 6 gradi di libertà, e i dati sperimentali sono stati acquisiti dal sistema 3DVision, costituito da una costellazione di 4 videocamere e 10 telecamere NIR ad alta risoluzione, con frequenza di campionamento di 200 Hz. Esso è in grado di acquisire le traiettorie di numerosi marcatori disposti sulla superficie dei provini in corrispondenza delle aree critiche, individuate mediante l'esecuzione di un'analisi FE preliminare (Figura 2).

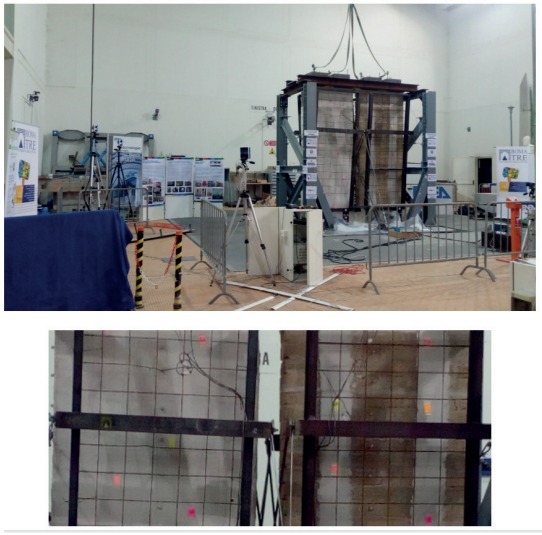


Fig. 2
Setup di prova e sistema 3DVision

Le prove sperimentali sono state condivise in tempo reale con i partner del progetto (UniRoma3 e Università di Miami), ed esperti che operano nel settore della protezione sismica. Il processo di condivisione parte dall'acquisizione del dato sperimentale e dalla sua immediata memorizzazione in un file C3D. Mediante l'interfaccia browser "Insert/modify" dell'applicazione web "DySCo Logbook", basata su piattaforma LAMP, l'addetto all'esecuzione della prova lancia un applicativo sviluppato in Java che esegue la lettura dei

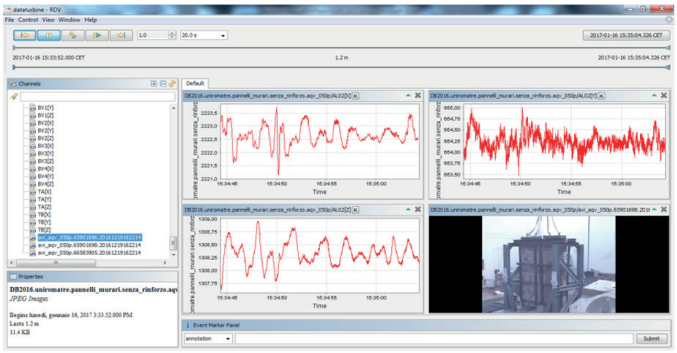


Fig. 3
RDV DataTurbine

C3D e inserisce sia i segnali che i video della prova nel motore per lo streaming real time denominato DataTurbine (DT). L'utente remoto, mediante la relativa interfaccia grafica RDV (Real Time Data Viewer) di DT (Figura 3), può visualizzare in near real-time le traiettorie dei marcatori disposti sulle strutture e i video della prova.

L'interfaccia "Insert/modify" è accessibile solo all'interno della LAN del laboratorio al fine di garantire la sicurezza e l'integrità del dato sperimentale, mentre l'interfaccia "View" del Logbook è accessibile anche su WAN e consente di visualizzare e trasferire remotamente le traiettorie dei marcatori anche successivamente all'esecuzione dei test sperimentali. Tramite la stessa interfaccia è possibile visualizzare tutte le informazioni relative al singolo step di acquisizione, inoltre grazie alle WebGL API in javascript:Three.js disponibili in rete, è possibile visualizzare la posizione all'interno di una geometria 3D dei marcatori, affiancati dalla relativa label, per permettere all'utente remoto di selezionare in completa autonomia il marcatore di suo interesse e visualizzarne le traiettorie sia nel domino del tempo che in quello delle frequenze (Figura 4 e Figura 5). I dati acquisiti vengono archiviati nelle aree di storage mediante un processo batch di sincronizzazione con le aree di staging e condivisi attraverso il MWN tramite un server "OwnCloud", che ha come backend storage le aree dipartimentali in condivisione.

3. Conclusioni

Il sistema E3S, replicabile e adattabile ad ogni altro tipo di strumentazione e laboratorio, è completamente integrato nelle procedure preesistenti di acquisizione, visualizzazione e

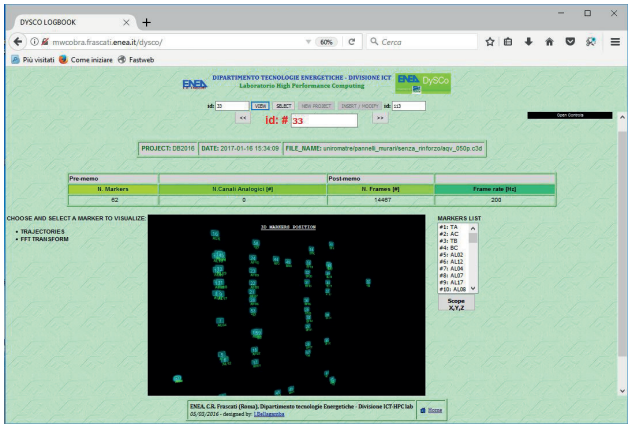


Fig. 4
Logbook Dysco

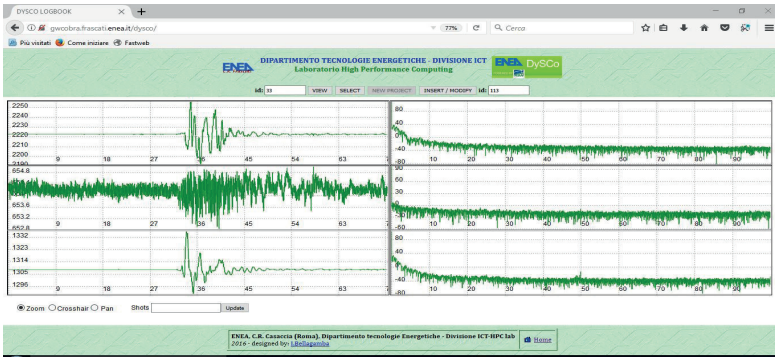


Fig. 5
Traiettorie nel dominio del tempo (sx) e delle frequenze (dx)

archiviazione dati, sfruttando le potenzialità dei filesystems dell'ENEA. Questo nuovo approccio di "sperimentazione condivisa" consente di rafforzare la cooperazione scientifica nel campo della protezione sismica, riducendo i tempi necessari all'esecuzione delle prove e all'elaborazione dei risultati da parte di utenti remoti, in completa autonomia e indipendentemente dalla propria posizione geografica. Inoltre facilita il processo di calibrazione a seguito di ciascuno step di prova dei modelli FE, in quanto i risultati sperimentali disponibili in near real-time, grazie alle funzionalità di streaming distribuito via DataTurbine, possono essere utilizzati per analizzare più velocemente il livello di danno subìto dalle strutture e calibrare automaticamente i modelli FE tenendo conto della loro reale risposta simica.

Riferimenti bibliografici

Abate D., Ambrosino F. et al. (2010-2011), "ENEA-GRID Infrastructure". Proceedings of High performance computing on CRESCO Infrastructure: research activities and results.

Iannone F., Bellagamba I., Bracco G., Calosso B., Giovanetti G., Migliori S., Mongelli M., Perozziello A., Pierattini S., Quintiliani A., Ambrosino F., Di Mattia D., Funel A., Guarnieri G., Ponti G., Simoni F., Steffé M. (2017), "A Staging Storage Sharing System for Data Handling in a Multisite Scientific Organization". Proceedings of the 3th CS3 Workshop on Cloud Services for File Synchronisation and Sharing SurfSARA, Amsterdam.

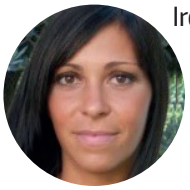
Mongelli M., De Canio G., Roselli I., Baldini M., Colucci A., Di Biagio F., Picca A., Tatì A., Cancelliere N., Coniglio L. and Ghersi A. (2011), "Experimental tests of reinforced concrete buildings and ENEA DySCo Virtual Laboratory", Proceedings of the 5th International Conference on Structural Health Monitoring of Intelligent Infrastructure (SHMII-5), (11-15/12), Cancún, México.

Mongelli M, Roselli I, De Canio G, Ambrosino F. (2016), "Quasi real-time FEM calibration by 3D displacement measurements of large shaking table tests using HPC resources", Advances in Engineering Software.

DOI: 10.1016/j.advengsoft.2016.07.005 30.

G. Ponti et al. (2014), "The role of medium size facilities in the HPC ecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure", Proceedings of the 2014 International Conference on High Performance Computing and Simulation, HPCS 2014, art. no. 6903807, pp. 1030-1033.

Autori



Irene Bellagamba irene.bellagamba@enea.it

Ing. Civile. Nel 2016 ha ottenuto una borsa di studio GARR "Orio Carlini", durante la quale si sta occupando dello sviluppo di una piattaforma web per lo streaming, lo storage e la gestione distribuita dei dati scientifici prodotti durante prove sperimentali su tavola vibrante per la protezione sismica, condotte presso il laboratorio SITEC dell'ENEA Casaccia.

Marialuisa Mongelli marialuisa.mongelli@enea.it

Ing. Chimico, PhD in metallurgia. Dal 2004, ricercatore ENEA, ha maturato la sua esperienza nel campo della protezione, conservazione e tutela del patrimonio culturale. Nel laboratorio HPC si occupa della definizione di modelli numerici, dalla ricostruzione 3D fotogrammetrica all'analisi agli elementi finiti per lo studio del comportamento dinamico di macroelementi strutturali o singole opere d'arte.

Francesco Iannone francesco.iannone@enea.it

Fisico. Dal 1993, ricercatore ENEA nel settore delle energie rinnovabili, in organico presso il laboratorio HPC, si occupa dello sviluppo di sistemi di acquisizione e gestione dati di impianti per la fusione nucleare.

Silvio Migliori silvio.migliori@enea.it

Ing. Nucleare. Direttore della divisione ICT dell'ENEA. Ha maturato una forte esperienza nello sviluppo di nuove tecnologie nel campo del supercalcolo scientifico e nella messa a punto di sistemi avanzati per la realizzazione di infrastrutture ICT (calcolo scientifico, ambienti virtuali).

Giovanni Bracco giovanni.bracco@enea.it

Fisico. Dal 1981 ricercatore ENEA, attualmente responsabile del laboratorio HPC. Si occupa dello sviluppo e della gestione del Cluster CRESCO e dell'infrastruttura di calcolo ENEAGRID.

DOI: 10.26314/GARR-Conf17-proceedings-24

Condivisione dei dati sui beni culturali: DIGILAB, l'esperienza di ARIADNE e di E-RIHS

Franco Niccolucci[1], Carlo Meghini[2], Achille Felicetti[1], Luca Pezzati[3]

[1] PIN, Università degli Studi di Firenze, [2] CNR-ISTI, [3] CNR-INO

Abstract. Sulla base dell'esperienza del progetto ARIADNE, che ha realizzato un catalogo dei dataset archeologici in Europa, l'infrastruttura europea di ricerca E-RIHS (European Research Infrastructure on Heritage Science) sta realizzando un sistema integrato di gestione dei dati relativi alle scienze del patrimonio culturale, denominato DIGILAB, che coprirà i settori della conservazione, del restauro e in generale della ricerca sui beni culturali. Questa comunicazionene illustrerà gli aspetti principali, che realizzano i principi FAIR e pubblicano i dati relativi a un settore finora largamente inaccessibile on-line.

Keywords. DIGILAB, Infrastrutture dati, Conservazione, Restauro, Beni Culturali

Introduzione

Il progetto ARIADNE (www.ariadne-infrastructure.eu), coordinato da PINUniversità di Firenze, ha realizzato un catalogo dei dataset archeologici in Europa che raccoglie i metadati di circa due milioni di dataset (Meghini et al. 2017). Questi comprendono report, immagini, database relativi alle ricerche archeologiche in tutti i paesi europei, indicizzati secondo uno schema dati comune. Il catalogo di ARIADNE rende così reperibili, accessibili e interoperabili dati forniti da oltre trecento istituzioni distribuite sul territorio europeo, e pubblicati da oltre venti centri di archeologia digitale. La ricerca nel catalogo è organizzata per luogo, periodo temporale e tipologia di contenuto dei dataset, oltre che a testo libero sui metadati di catalogo.

Le difficoltà linguistiche – i dati testuali che sono la grande maggioranza dei contenuti utilizzano una quindicina di lingue diverse – sono state affrontate con la creazione di vocabolari multilingue specializzati. Il sistema ha richiesto anche la creazione di un sistema di allineamento dei diversi periodi storici, collegato alla localizzazione geografica: com'è noto, ad esempio, l'età del ferro in Italia termina nel III sec. a.C. mentre in Gran Bretagna si conclude con la conquista romana (I sec. d.C.) e in Irlanda conle invasioni vichinghe (V sec. d.C.).

Il consorzio ARIADNE continua la sua attività anche dopo la fine del finanziamento europeo, sia raccogliendo nuovi elementi per il suo catalogo sia aggiornando il sistema di gestione. Infatti, il catalogo è stato inizialmente realizzato utilizzando uno schema dati apposito (ARIADNE Catalog Data Model, ACDM). Successivamente, il progetto PARTHENOS (www.parthenos-project.eu), un cluster di progetti e iniziative digitali in campo umanistico, ha riformulato il modello come estensione di un'ontologia standard, il CIDOC CRM, producendo così il PARTHENOS Entities Data Model (PEM) (Aloia et al. 2017). Il PEM è idoneo a gestire metadati relativi a dataset e servizi per un ampio spettro di discipline nel campo delle scienze umane. Il gruppo di lavoro originario di ARIADNE, che ha partecipato attivamente allo sviluppo di PEM, ha quindi prodotto un mapping da ACDM a PEM e ha realizzato la conversione

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

dei dati, che sono ora catalogati secondo lo schema più aggiornato. Anche se PEM è centrato sui dati relativi alle scienze umane (linguistica, digital humanities, archeologia, architettura, storia, conservazione, e così via), è estensibile con profili specializzati per i singoli settori che ne migliorano l'efficacia, mantenendo allo stesso tempo la compatibilità interdisciplinare.

L'esperienza di E-RIHS (www.e-rihs.eu), un'infrastruttura di ricerca a guida CNR già inserita nella roadmap ESFRI (European Strategy Forum on Research Infrastructures) e attualmente nella fase di preparazione di un ERIC (European Research Infrastructure Consortium), nasce invece da una serie di progetti per la ricerca sulle scienze della conservazione, inizialmente concepiti senza la componente digitale. Attraverso una sequenza decennale di progetti europei (EU-ARTECH, CHARISMA e infine IPERION CH, tuttora in corso), tutti coordinati dal CNR, quest'iniziativa si è sviluppata basandosi sull'accesso a laboratori mobili (MOLAB), fissi (FIXLAB) e archivi (ARCHLAB) operanti nel campo della ricerca sulla conservazione del patrimonio; dopo l'incontro con ARIADNE e l'approvazione di E-RIHS comprende ora anche un'infrastruttura digitale (DIGILAB).

1. DIGILAB

DIGILAB è basato, come il catalogo di ARIADNE, sulla "federazione" di archivi locali, dove ricercatori e operatori del settore depositano i risultati digitali della propria attività. Il servizio principale è il catalogo, simile a quello di ARIADNE e basato su un profilo di PEM adattato alle specifiche esigenze della comunità scientifica di riferimento. La funzione di acquisizione (ingestion) dei metadati utilizza la stessa tecnologia di quelle di ARIADNE, come pure la funzionalità di ricerca nel catalogo che si basa su Elasticsearch.

DIGILAB è ispirato ai principi FAIR: rende i dati reperibili attraverso il suo catalogo; ne provvede l'accesso attraverso un sistema d'identità federata; ne supporta l'interoperabilità sia a livello di metadati organizzati secondo PEM, sia a livello di dati, ove possibile, proponendo uno schema basato su un profilo specializzato di CRMsci (Doerr et al. 2015), l'estensione dello standard CIDOC CRM per la gestione dei dati scientifici, e compatibile anche con CRMba (Ronzino et al. 2016), l'estensione CRM per l'architettura, e CRMarcheo (Doerr et al. 2016), l'analoga estensione per l'archeologia. Per quanto riguarda il riuso, DIGILAB propone un approccio basato sul Virtual Research Environment (VRE) fornito da D4Science (Assante et al. 2016), offrendo in ambiente cloud una serie servizi utili ai ricercatori del settore. Questi servizi coprono una gamma di attività tipiche della ricerca, quali ad esempio la creazione di grafici e diagrammi su dati numerici di analisi chimiche o fisiche, l'elaborazione di immagini multispettrali e l'arricchimento dei metadati dei testi (ad esempio relazioni tecniche), utilizzando sistemi di Natural Language Processing (NLP) e Named Entity Recognition (NER) con l'uso di vocabolari specializzati. Alcuni di questi servizi sono attualmente già in fase di test. DIGILAB fornirà agli utenti anche servizi di corredo, quali un formulario per la compilazione on-line del Data Management Plan (DMP), ormai obbligatorio per i progetti finanziati dalla Commissione Europea, e un sistema chiavi in mano per la realizzazione di archivi locali, personalizzabile secondo le esigenze dell'utente ma basato sullo schema dati standard del progetto.

L'organizzazione decentrata degli archivi consente di non affrontare nell'immediato il problema dello storage dei dati, che è affidata ai nodi locali di E-RIHS o alle rispettive comunità di

ricercatori, mentre DIGILAB necessita di risorse modeste in quanto deve contenere soltanto i relativi metadati. Tuttavia, si pone il problema di organizzare in modo efficiente il nodo italiano, assicurando le necessarie risorse a livello nazionale per il deposito dei dati prodotti dalla ricerca nel settore.

DIGILAB è una realizzazione di PIN-Università di Firenze, già coordinatore di ARIADNE, e CNR, attraverso i suoi istituti ISTI per la parte informatica e INO, coordinatore di E-RIHS, che organizza la partecipazione di vari istituti di scienze dei beni culturali (ISTM, ICVBC e altri) e di istituti specializzati del MIBACT come l'OPD.

2. Piano di lavoro e successivi sviluppi

DIGILAB è attualmente in fase sperimentale; se ne stanno sviluppando le componenti e verificando le funzionalità su una serie di archivi di test. I moduli NLP e NER fanno parte di un dimostratore inserito in EOSCpilot (www.eoscpilot.eu), un progetto pilota sulla realizzazione di EOSC (European Open Science Cloud).

I moduli necessari a DIGILAB richiedono tutta una gamma di attività differenti: alcune di tipo tecnologico, come l'implementazione del VRE; altre di tipo semantico, come la realizzazione dei vari schemi di dati che derivano e insieme impattano sui protocolli sperimentali adottati da E-RIHS; altre ancora di carattere organizzativo, come l'identità federata. I moduli tecnologici saranno rilasciati via via che verranno prodotti, e ci si aspetta di avere le prime funzionalità disponibili per la fine del 2017. Per gli inizi del 2018 sarà invece rilasciata la prima versione della parte semantica.

Una prima versione completa dell'intero sistema sarà disponibile entro il 2018, procedendo in parallelo con la creazione e l'adattamento degli archivi locali, di cui allo stato attuale solo alcuni sono in funzione. In ogni modo, è previsto che il sistema sia pienamente funzionante quando anche E-RIHS si costituirà come ERIC. Per ottenere quest'obiettivo, è indispensabile che le risorse necessarie a livello nazionale siano rese pienamente disponibili, in modo da offrire ai partner europei, e probabilmente a un'audience mondiale, il sistema avanzato di gestione dei dati della ricerca sopra descritto.

Riferimenti bibliografici

Aloia N. et al. (2017), D5.2 Design of the Joint Resource Registry. Available at

http://www.parthenosproject. eu/Download/Deliverables/ 5.2_Report_on_design_Joint_Resource_Registry.pdf

Assante M., Candela L., Manghi P., Pagano P., Castelli D. (2015), Providing research infrastructures with data publishing. ERCIM News, Issue 100, January 2015.

Doerr M., Kritsotaki A., Rousakis Y. (2015), Definition of the CRMsci. An Extension of CIDOC-CRM to support scientific observation. Available at:

http://www.ics.forth.gr/isl/CRMext/CRMsci/docs/CRMsci1.2.3.pdf

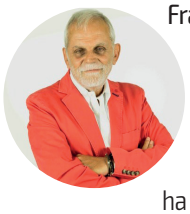
Doerr M., Felicetti A., Hermon S. et al. (2016), Definition of the CRMarchaeo. An Extension of CIDOC CRM to support the archaeological excavation process. Available at:

http://www.ics.forth.gr/isl/CRMext/CRMarchaeo/docs/CRMarchaeo_v1.4.pdf

Meghini C. et al. (2017), ARIADNE: A research infrastructure for archaeology. ACM Journal on Computing and Cultural Heritage 10(3):1-27, August 2017.

Ronzino P., Niccolucci F., Felicetti A., Doerr M. (2016), CRMba a CRM extension for the documentation of standing buildings. Int. J. on Digital Libraries 17(1): pp. 71-78.

Autori



Franco Niccolucci franco.niccolucci@gmail.com

Franco Niccolucci è il direttore del laboratorio VAST-LAB presso il PIN di Prato, e il coordinatore del progetto europeo PARTHENOS, cluster delle infrastrutture di ricerca nel settore Digital Humanities e Cultural Heritage. Ha coordinato l'infrastruttura ARIADNE nel campo dell'archeologia digitale, che ha realizzato un registry di oltre 2.000.000 di dataset archeologici. In passato ha coordinato vari progetti europei nel campo dei beni culturali. Matematico come formazione, è stato professore all'Università di Firenze fino al 2008 e successivamente, fino al 2013, direttore dello Science and Technology in Archaeology Research Center presso il Cyprus Institute a Nicosia, Cipro.

Carlo Meghini carlo.meghini@isti.cnr.it

Carlo Meghini è ricercatore capo presso il CNR-ISTI e coordinatore del gruppo Digital Libraries del laboratorio NeMIS di ISTI. La sua area di ricerca comprende la progettazione concettuale, le digital library e le infrastrutture di ricerca per le scienze umane. È coinvolto in progetti europei dal 1988, nelle aree delle digital library e della digital preservation. È stato il coordinatore della CSA PRELIDA per la preservation di Linked Data. Dal 2007 partecipa alla creazione di Europeana, la digital library europea, curando gli aspetti scientifici del progetto. A partire dal progetto ARIADNE, è coinvolto nella ricerca e nello sviluppo di infrastrutture di ricerca. Ha pubblicato oltre 90 articoli scientifici in riviste, libri e conferenze internazionali.

Achille Felicetti achille.felicetti@pin.unifi.it

Achille Felicetti è un archeologo attivo fin dal 2004 nello sviluppo e l'applicazione di nuove tecnologie per la codifica, la condivisione e l'integrazione di dati e nella definizione di ontologie e strumenti semantici per la modellazione di informazioni nell'ambito dei Beni Culturali. Ha partecipato a diverse iniziative europee e coordinato gruppi di ricerca e sviluppo in progetti quali ARIADNE, per l'interoperabilità di dati archeologici, PARTHENOS ed EOSCpilot, per la definizione di servizi e tecnologie di Natural Language Processing. È parte del team di sviluppo di CRMarchaeo, l'estensione CIDOC CRM per la codifica di informazioni di scavo.

Luca Pezzati luca.pezzati@cnr.it

Luca Pezzati è un fisico e uno specialista di ottica. Dal 1995 è con INO-CNR (l'Istituto Nazionale di Ottica del Consiglio Nazionale delle Ricerche d'Italia) dove è attualmente ricercatore senior. Ha avviato il Gruppo Beni Culturali (INA) nel 1998 e lo ha coordinato per 14 anni. Ha gestito molti progetti di ricerca per CNR nel campo delle tecnologie ottiche applicate al patrimonio culturale. È coordinatore di E-RIHS, l'Infrastruttura europea di ricerca per la scienza del patrimonio e del progetto d'integrazione IPERION CH. Ha coordinato il nodo nazionale di DARIAH ERIC, DARIAH-IT, dal 2013 al novembre 2016.

DOI: 10.26314/GARR-Conf17-proceedings-25

SensorWeb Hub as an interoperable research data infrastructure for low-cost sensor data sharing

Tiziana De Filippis, Leandro Rocchi, Elena Rapisardi

CNR-Ibimet, National Research Council, Institute of Biometeorology

Abstract. Data accessibility, discovery, re-use, preservation and, particularly, data sharing are the key to promote the open innovation approach to research studies and enhance interdisciplinary analysis. The CNR-Institute of Biometeorology (CNR-Ibimet) developed an open source and interoperable platform, called SensorWebHub (SWH) to manage both mobile and fixed meteorological and environmental sensors data that integrate the existing monitoring networks in urban and agricultural research initiatives. The infrastructure has been developed to ensure access, management and preservation of data within and across research teams; it is a technical support service for coordinated management of data and encourages reuse and collaboration. SWH is a bottom-up collaborative initiative to share real-time raw research data and pave the way for an open innovation approach in the scientific research, so contributing to the processes of the production and dissemination of research data.

Keywords. Low-cost sensors, research data, open data, open source.

Introduction

In the last years, RDIs' (Research Data Infrastructures) landscape offers networks and international programs aiming to build reliable and interoperable research data e-infrastructures to promote the open innovation approach to research studies and enhance interdisciplinary analysis. RDIs are designed to follow different data policies and reference models. Many of European Environmental RDIs, with more than half of them prioritized in the roadmap of ESFRI (the European Strategy Forum On Research Infrastructures), are run by Institutions having as own duty data storage and management. Others are a legal entity or in the process of doing so or are linked to research programs funded by H2020 EU. The current state of RDIs development varying from operational to preparatory phase project (Zaho et al. 2015).

However, for national or sub-national research projects, there is a scarcity of open data infrastructures to store, share and manage a huge amount of local research data. So, small research institutions or projects are disadvantaged by the lack of common spatial data infrastructure and specific expertise that is crucial factors to reduce open data access fragmentation within and across research units. As a matter of fact, data collected by local research activities are often the missing piece of the puzzle as they are not easily findable, or accessible. In this context the CNR-Institute of Biometeorology developed an open source and interoperable platform, called SensorWebHub (www.sensorwebhub.org), to manage both mobile and fixed meteorological and environmental sensors data, that inte-

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

grate the existing monitoring networks in urban and agricultural research initiatives. The data, collected through innovative low-cost and open source sensor devices, are processed and published using OGC (Open Geospatial Consortium) services and geospatial data standards. This infrastructure is currently focused on the following sensor data categories: Agrometeo, Meteo, Urban Climate, Renewable Energy, Indoor.

1. Objective

The aim of this work is to deploy an interoperable and open data infrastructure in order to help the scientific community to share relevant and timely data and services.

This initiative arises from the fact that researchers mainly store their data, as well as intermediate products processed for environmental and agro-meteorological investigations, in personal archives. However, if this data were shared they could be used for further applications in other research fields. Data, climate products, informatics procedures, code and pre-processed data on a geographical area of interest could thus be reused, reducing the time and human resources necessary for further investigations. The availability of an interoperable research data infrastructure to store and manage data could also facilitate and encourage the adoption of a data sharing approach.

On this premise, SWH has been developed with the aim:

- to support the participating approach to monitoring urban environment;
- to share research data acquired by low-cost sensors from no-conventional networks and fixed and mobile sensors;
- to test new analysis procedures and integrated approaches using multi sources data;
- to develop new web geoprocessing tools;
- to encourage the development of user-friendly interfaces for different stakeholders.

The challenge is also to attract more internal researchers to share their data and quality checked climate products easily through an available data infrastructure, for further interdisciplinary investigations.

2. SensorWeb Hub infrastructure

SensorWeb Hub infrastructure and functions were designed to create a participatory environmental monitoring system where the data collected with innovative low-cost and open source sensor devices are processed and published using OGC (Open Geospatial Consortium) services and geospatial data standards.

SWH web application manages both mobile and fixed open source and low-cost sensor platforms, to integrate the existing monitoring networks (De Filippis et al., 2015).

This interoperable infrastructure is currently focused on the following sensor data categories: Agrometeo, Meteo, Urban Climate, Renewable Energy, Indoor.

The infrastructure is composed by: 1) the AirQino Sensor-Boxes (ASB) as main components of the sensor network; 2) a central GeoDataBase (GeoDB), the PostgreSQL/PostGIS, for data storage and management; 3) a GIS engine and a WebGIS application for viewing, querying and performing data analysis; 4) the specific web services for data flux management. The components (Fig.1) are organized in typical client-server architecture

and interact from the sensing process to the representation of the results to the end-users adopting the OGC® SWE (Sensor Web Enablement) common standards.

Specific web services were developed using JavaEE technology. They work as web service callback and reads all sensing data, performs a data quality check on Arduino devices and stores them in the GeoDB. Using PostGIS functions, the geographic information is transformed from NMEA RMC standard into point elements for PostgreSQL. UML (Unified Modeling Language) as formal language adopted in the ISO TC/211 (https://committee.iso.org/tc211) context for geomatic data description has been used for formal data model definition. The web interface and functions have been developed using J2EE technology with Java Server Faces and PrimeFaces library for GUI (Graphic User Interface) customization. Using any desktop or mobile browser, all collected data can be visualized in near-real time in table or chart format, or tracks, and spot values on a Google mashup.

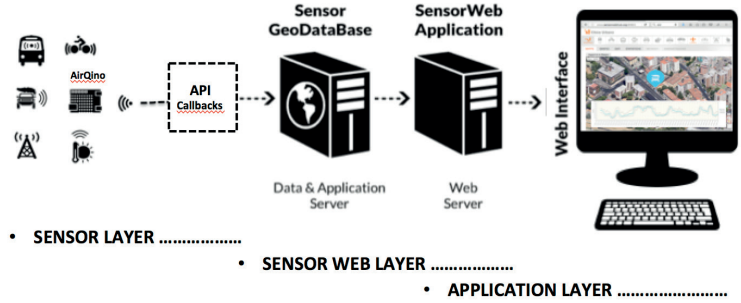


Fig. 1
SensorWeb Hub
infrastructure components

3. Results

SWH's data and web services are now available at http://www.sensorwebhub.org/. The whole framework source code of underlying SWH application is deployed on GitHub platform at https://github.com/n3tmaster/SensorWebHub. At present, the platform manages data from field campaigns on urban climate monitoring, agro-meteorological survey and renewables energy related to the funded project at local scale. It is a work in progress initiative open to manage and sharing further raw data coming from different research projects. The infrastructure supports multiple research communities and individuals. The data, accessible through the web interface and also by standard web services (API and RestFul services), in the first instance are sharing within CNR-Ibimet research units (e.g. Geomatic and ICT, agricultural sustainability, air quality monitoring.) but could be downloaded also from other users (practitioners, students and external researchers) taking into account the recommendations in their proper use. SWH has multiple scopes and responds to different researchers' needs. From ICT & Geomatic point of view, it is conceived for implementing and testing interoperable OGC standards and RDIs reference models. The educational goal is to disseminate the best practices on data management and sharing and support the researchers in adopting RDIs guidelines in the life cycle of their research activities. It also facilitates the exchange within the research's team and offers easy solutions for a remote control of non-conventional environmental monitoring networks. Furthermore, SWH offers a real data set to develop customized web applications for a

general public (citizens, farmers, decision makers students, developers). In addition, the adoption of interoperable web services facilitates data sharing and their reuse in order to enable real interdisciplinary innovation.

4. Conclusions

SWH infrastructure is designed to manage further Ibimet-CNR environmental data and services derived from advances in research in applied meteorology and climatology. The use of open source tools and standardized interoperable web services ensure sustainability in the development and deployment of web applications with geo-referenced data and customized territorial analysis that could be connected to other interoperable RDIs.
Lastly, the availability of an interoperable and open source infrastructure enhances both the timeliness and quality of information provided and offers a technical bridge that enables open sharing of data following the guidelines and principles of the research data infrastructure actions, under the umbrella of RDA (Research Data Alliance https://www.rd-alliance.org).

References

De Filippis, T., Rocchi, L., Rapisardi, E., 2015. SensorWeb Hub infrastructure for open access to scientific research data. Geophysical Research Abstracts, Vol. 17, EGU2015-7847-2, EGU General Assembly.

ESFRI 2016 Public roadmap 2018 guide dated 9th December 2016. https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri. Last accessed 18/09/2017

Zhao, Z., Martin, P., Grosso, P., Los, W., de Laat, C., Jeffrey, K., Hardisty, A., Vermeulen, A., Castelli, D., Legre, Y., Kutsch, W., 2015. Reference Model Guided System Design and Implementation for Interoperable Environmental RIs, DOI:10.1109/eScience. 2015.41.

Authors

Tiziana De Filippis t.de.filippis@ibimet.cnr.it
PhD, permanent researcher in agro-meteorology at CNR-IBIMET of Florence. She is currently Project Manager of SensorWeb Hub infrastructure based on the paradigms of open source, interoperability, and open data.

Leandro Rocchi l.rocchi@ibimet.cnr.it
Computer Science specialist at National Research Council - Institute of Biometeorology. Expert developer of agrometeorological SDI, web services, web applications, and models implementation. Expert designer of GeoDatabase and mobile solutions.

Elena Rapisardi e.rapisardi@gmail.com
Degree in Political Science, PhD in Earth Sciences on Natural Hazards Risk communication. She deals with information architecture, web contents writing and scientific communication.

DOI: 10.26314/GARR-Conf17-proceedings-26

WeatherLink una piattaforma per l'integrazione e la visualizzazione dei dati meteo

Riccardo La Grassa[1], Marco Alfano[1,2], Biagio Lenzitti[1], Davide Taibi[3]

[1]Dipartimento di Matematica e Informatica, Università degli Studi di Palermo, [2]Anghelos Centro Studi Sulla Comunicazione, [3]Consiglio Nazionale delle Ricerche, Istituto per le Tecnologie Didattiche

Abstract. WeatherLink è una piattaforma per la raccolta, l'elaborazione e la pubblicazione di dati relativi a misure atmosferiche. WeatherLink si compone di una parte server che memorizza l'enorme quantità di dati provenienti dalle diverse board dislocate nel territorio che fungono da client. Ogni board tramite i suoi sensori raccoglie dati su temperatura, pressione atmosferica, qualità dell'aria e li invia al server. Il server si occupa della memorizzazione dei dati, e offre anche delle API che possono essere interrogate per estrarre i dati ed effettuare elaborazioni successive. I dati vengono esposti come dati aperti nel formato JSON e JSON-LD. Le API sono state utilizzate per la realizzazione di una Web Application che consente il monitoraggio delle board e la visualizzazione dei dati aggregati o per singola board attraverso grafici relativi ai diversi parametri metereologici catturati dai sensori.

Keywords. Open Data, IoT, Data integration and visualization, monitoraggio ambientale.

Introduzione

WeatherLink è una piattaforma per la raccolta, l'elaborazione e la pubblicazione di dati di natura metereologica. L'idea di WeatherLink è quella di creare una fitta rete di stazioni meteo amatoriali e non, al fine di integrare e rendere disponibili per successive elaborazioni i dati, in formato aperto, provenienti da diverse sorgenti. Come mostrato in figura 1, la piattaforma si compone di due elementi principali: i) il server che riceve dalle stazioni i dati me-

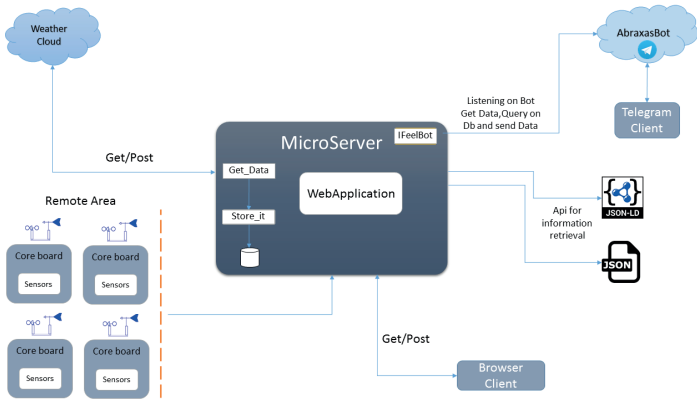


Fig. 1
Architettura di WeatherLink

Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

teo acquisiti dai sensori ed espone, tramite servizi implementati con metodi REST, i dati in formato aperto; ii) le stazioni meteo, realizzate tramite Board specifiche, sono dislocate nel territorio e raccolgono i dati mediante sensori di temperatura, umidità, pressione.

Il server è stato implementato attraverso diversi moduli Python per l'acquisizione dei dati e la loro memorizzazione. Per quanto riguarda le board, il codice è stato scritto interamente in C. Il codice è stato reso pubblico in modo da consentirne il riuso . Gli utenti che vogliono contribuire alla rete di sensori WeatherLink possono registrare sul server la propria stazione meteo e installare sulle board il codice necessario. L'utente, inoltre, può accedere sia ai dati della propria board in locale, per via dell'aggiuntivo sviluppo del frontend lato board, che ai dati storicizzati sul server. Uno dei punti di forza di WeatherLink è offrire un pieno servizio a chiunque voglia utilizzare le API, esponendo dati in formato aperto (JSON e JSON-LD) e senza restrizioni temporali. Questo approccio rende WeatherLink differente da altre piattaforme disponibili in rete che limitano il riuso dei dati o ne restringono l'accesso solo a un ristretto arco temporale. Sulla base dei dati resi disponibili dal server WeatherLink è stata costruita una Web Application appositamente progettata per la visualizzazione dei dati meteo, descritta in dettaglio nella seguente sezione.

1. Data Visualization

La Web Application è stata realizzata con l'utilizzo del framework Django, ed è basata sul paradigma MVC (Model View Control) che ha permesso la creazione di modelli di dati efficienti utilizzati per lo storage su database. Una volta creati i modelli dei dati, specifici comandi estrapolano il codice SQL necessario per eseguire le query su un database di tipo SQLite. Inoltre si è preferito avere due database indipendenti, uno gestito da Django per la gestione delle autenticazioni degli utenti, ed uno per la memorizzazione dei dati provenienti dalle stazioni meteo alimentato anche da moduli python in esecuzione periodica per il recupero dei dati da fonti di dati esterni come OpenWeatherMap (OWM) . Nello specifico i moduli scritti in python, Get_data e Store_it, hanno il compito di recuperare i dati da OWM e di memorizzarli in db_weather.sqlite. Una volta registrati su OpenWeatherMap e ottenuto un token, si sono utilizzate le API messe a disposizione della piattaforma OWM (con licenza CC-BY-SA), al fine di creare un proprio storico da utilizzare per successive elaborazioni e da integrare con i dati raccolti dalle stazioni meteo. La finalità di WeatherLink, è quella, di creare un proprio network di weather station, integrare i dati con altre fonti di dati aperti relativi alle misurazioni meteo e fornire, dopo successiva aggregazione ed elaborazione, nuove visualizzazione dei dati.

In fase di progettazione del database, si sono considerati diversi aspetti legati alla pertinenza dei dati nel contesto meteorologico, come la gestione dei diversi tipi di dati, o le problematiche su possibili errori dovuti all'invio di dati parziali. Un utente ha la facoltà di creare più stazioni meteo legate al proprio account, registrando ogni stazione meteo sul server WeatherLink e ottenendo uno specifico token. Uno dei principali servizi offerti da WeatherLink è la possibilità di visualizzare in tempo reale i dati raccolti dalla propria stazione meteo direttamente plottata su OpenStreetMap (OSM).
La piattaforma presenta due tipologie di plotting: Mappa e Grafici. Per la visualizzazione

Mappa è stata utilizzata la libreria folium (leaflet derivata per python) per plottare degli oggetti di tipo circle direttamente su mappa OpenStreetMap (OSM). Inizialmente, si effettua un recupero tramite query, nel database db_weather.sqlite dei dati più recenti, ma l'utente può estendere l'arco temporale da visualizzare. La figura 2 mostra un esempio dei grafici che possono essere visualizzati dalla Web Application di WeatherLink.

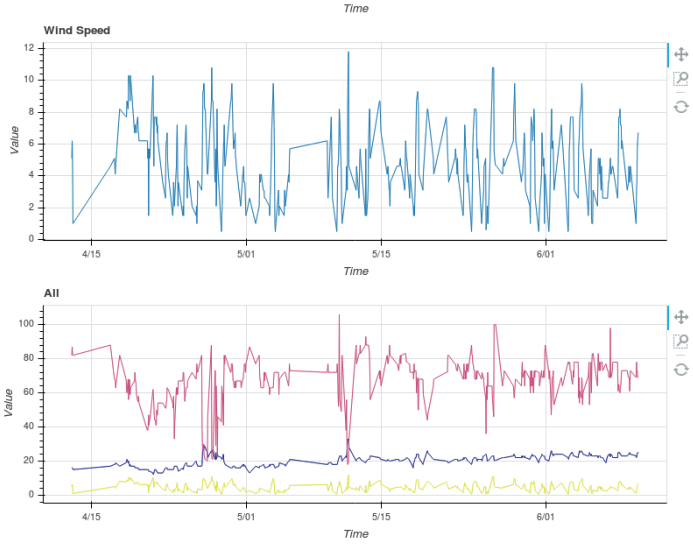
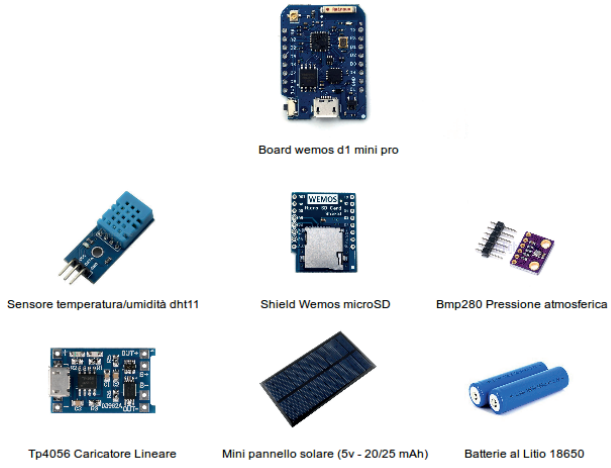


Fig. 2
Esempio di visualizzazione dei dati

2. Realizzazione e dati tecnici

L'architettura di WeatherLink è composta da un server centrale e numerose boards che compongono la rete di stazioni meteo. Come mostrato in Figura 3, al momento la rete è composta da board Wemos d1 mini pro con Esp8266 (Datasheet per ESP8266), sensore DHT11 (Temperatura/Umidità) e Bmp 280 (Pressione). Inoltre le board sono dotate di Caricatore Lineare Tp4056, Mini pannello Solare (5v - 20/25 mAh) e Batteria al litio. I2C/

Fig. 3
Stazione meteo e sensori utilizzati



OneWire/SPi sono i tre possibili protocolli per la comunicazione con sensori esterni. La board ha solo un canale analogico e 11 digitali, Flash 16 MB, 80KB di ram e 80/160 MHz di clocks.
Il flusso di esecuzione implementato su ciascuna board è il seguente:

- Setup del web Server e connessione alla rete.
- Inizializzazione dei vari sensori e della porta d'ascolto
- Acquisizione dati ogni 5 secondi e invio al server. (Funzionalità one shoot)

Inoltre in locale gli utenti possono accedere ai dati raccolti tramite normale protocollo HTTP (attraverso la funzionalità di web server locale). In base alla richiesta fatta dagli utenti (tramite HTTP GET), la board risponde con l'informazione desiderata per ciascun sensore.

Recentemente, una nuova versione è in fase di testing per ridurre drasticamente i consumi energetici da parte della board, introducendo un pannello fotovoltaico, un caricatore lineare con modalità trickle, ed una batteria al litio. Il software è stato modificato, eliminando la funzionalità di web server locale, e decrementando il consumo di corrente ponendo la board in modalità deep-sleep ogni qualvolta si effettui un invio dei dati (one shoot). Il risveglio dalla modalità, viene effettuato per mezzo di un interrupt causato da un timer rtc interno. La board è stata pensata per operare sia in rete che in zone sprovviste di copertura wireless. In quest'ultimo caso la board implementa un file system e conserva i dati dei sensori in una microSD.

3. Conclusioni e sviluppi futuri

La piattaforma WeatherLink è stata progettata con l'obiettivo di integrare dati provenienti da stazioni meteo realizzate attraverso apparecchiature IoT, con altre sorgenti di dati pubblicati in formato aperto, in modo da creare una fonte completa di dati meteo su cui effettuare nuove elaborazioni e visualizzazioni.

Al momento si è in procinto di costruire una rete di stazioni di rilevamento sparse per la città di Palermo che conterranno, oltre al sensore per il rilevamento della temperatura e umidità, un rilevatore del livello di qualità dell'aria in modo che si riesca a fornire una sorta di indice di "qualità ambientale" delle varie zone della città. Inoltre sono in via sviluppo dei connettori per i dati, disponibili in formato aperto, messi a disposizione da ARPA Sicilia . In questo modo i dati raccolti dalle stazioni meteo verranno integrati non solo con i dati di OWM ma anche altre fonti di dati aperti spianando la strada ad ulteriori elaborazioni (Bannayan 2008, Holmstrom 2016).

Riferimenti bibliografici

Bannayan M., Hoogenboom G. (2008). Weather analogue: A tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach, In Environmental Modelling & Software, 23, (6), 703-713.

Datasheet per ESP8266, disponibile in rete: http://espressif.com/sites/default/files/documentation/0a-esp8266ex_datasheet_en.pdf

Holmstrom H., Liu D., and Vo C. (2016). Machine Learning Applied to Weather Foreca-

sting. Stanford University.

http://www.arpa.sicilia.it/temi-ambientali/monitoraggi-ambientali

https://github.com/pulsar2468/IoT-project

https://www.djangoproject.com

https://openweathermap.org

https://www.youtube.com/watch?v=pKFKAl9XQC4

Autori

Riccardo La Grassa riccardo.lagrassa@community.unipa.it

Laureato in scienze informatiche nel 2015, da alcuni anni si occupa di progetti in cui vengono applicate le tecnologie IoT, embedded programming e machine learning in diversi settori. Recentemente ha collaborato con esperti del CNR-ISSIA sull'interfacciamento hardware delle board programmabili. Attualmente è laureando magistrale in informatica, la tesi di laurea riguarda la realizzazione di un sistema intelligente per la distribuzione di energia elettrica mediante tecniche di machine learning.

Marco Alfano marco.alfano@anghelos.org

Ingegnere Elettronico e Dottore di Ricerca in Ingegneria Elettronica, Informatica e delle Telecomunicazioni con un'esperienza professionale trentennale nel campo dell'ICT maturata (sia in Italia che all'estero) in diversi ambiti lavorativi quali quello della pubblica amministrazione, aziendale, universitario e di ricerca e con diverse mansioni quali manager di struttura, coordinatore di progetto, consulente, sistemista, progettista, programmatore, docente e tutor. Svolge ricerca nel campo Open Data e servizi socio-sanitari.

Biagio Lenzitti biagio.lenzitti@unipa.it

Ricercatore presso il Dipartimento di Matematica e Informatica dell'Università degli studi di Palermo. Attualmente docente del corso di programmazione per il corso di laurea in Informatica e già docente dei corsi di Reti di calcolatori e Sistemi Operativi. Responsabile locale di vari progetti Europei (ReBus, Fetch, Trice) le sue attività di ricerca sono andate dal Calcolo parallelo ai linguaggi Pittorici fino al Pattern Recognition. Negli ultimi anni si è interessato di E-Learning e di Smart Health.

Davide Taibi davide.taibi@itd.cnr.it

Ricercatore a tempo indeterminato dell'Istituto per le Tecnologie Didattiche del Consiglio Nazionale delle Ricerche. Le sue attività di ricerca sono focalizzate principalmente sull'applicazione delle tecnologie innovative nel campo dell'e-learning, con particolare riguardo all'apprendimento con dispositivi mobili, il Web semantico e i Linked Data, gli standard per la progettazione dei processi educativi. Professore a contratto presso l'Università degli Studi di Palermo per il corso di Tecniche per la gestione degli Open Data.

DOI: 10.26314/GARR-Conf17-proceedings-27

Motivating carsharing services open-data mandatory APIs

Andrea Trentini, Federico Losacco

Dipartimento di Informatica – Università degli Studi di Milano

Abstract. Modern countries car traffic is nowadays fought by regulations: tolls, dedicated/narrow lanes, low speed limits, reduced parking availability, etc. Some help can come from carsharing, i.e., pools of shared vehicles to be rented for short periods of time. The authors scraped carsharing websites for a couple of years, uniformed data and then queried and graphed the dataset, a summary of results is presented. Carsharing vendors should publish - openly, mandatorily, through standardized APIs - the state of their pool because this data can be used to analyse the overall traffic behaviour in town and study impacts, effectiveness and costs.

Keywords. Application Programming Interface, urban congestion, open data, public accountancy.

Introduction

Air pollution became a problem for human beings with the industrial revolution (Seinfeld and Pandis, 2012; Lave and Seskin, 2013; Steinle et al., 2013). Afterwards many governments begun legislating (US-EPA, 2013) to reduce industrial emissions. Then, air pollution slowly began to decrease as new generations of technologies replaced older ones (Figure 1).

Fig. 1
Historical trend of SO2
(source: ARPA Lombardia)

While pollution is being reduced, many governments and administrations are now addressing the "congestion factor" with regulations (Vanderbilt, 2009) to lower private traffic. A few years ago "carsharing" was introduced with the rationale that: 1) parking space is saved since a single parked vehicle serves a lot of users, not a single one (of a private

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

car); 2) because of the higher (>private car, <taxicab) costs per use, many carsharing users, would use it wisely (maybe mixing it with standard public transport), reducing the overall "car mileage load" on the city; 3) it may further lower pollution (Firnkorn and Müller, 2011). A review on carsharing studies can be read in (Jorge and Correia, 2013).

GPSs and smartphones track down available cars and lead users to precise parking locations. Detailed location data are, of course, stored and secured into vendors' systems but some data are online and can be used to analyse the overall traffic behaviour in town. Current online data is referred to the real-time situation only and it is often behind "web-stacles" (Trentini, 2014). Given the usefulness of this data, the authors propose the initiation of a standardization and "open-data-ization" process. I.e., governments should force vendors to publish data (both real-time and historical) through public, documented and well-defined APIs (Application Programming Interfaces).

1. The quest for data (web-scraping)

In Milan (ITALY) there are five carsharing vendors: 1) Share'NGo (http://www.sharengo.it); 2) Twist-Car (http://twistcar.it, discontinued 17-Nov-2015); 3) Enjoy (http://enjoy.eni.com); 4) Car2Go (http://www.car2go.com); 5) GuidaMi (http://www.guidami.net). Some vendors do publish data about car positions and availability which the authors using wget in shell scripts or small python programs periodically run on a server.

2. Data analysis

The following analyses are based on data between September 2015 and March 2016, they show an evident daily and weekly patterns, i.e.:

1) every day is "cyclical", e.g., night differs from day and office hours are peaks;

2) weekly effects: a) Mon-Fri days are similar to one another; b) Saturday and Sunday are similar; c) Saturday and Sunday differs from Mon-Fri workdays;

3) there are, of course, random fluctuations.

Data were divided into weeks and for every week Q-Q plots (Quantile-Quantile, to check if samples are statistically similar) were created. A Q-Q plot for every day pair (Mon-Tue, Mon-Wed, Mon-Thu, etc.) in every week was generated, originating (7 on 2) = 21 Q-Q plots per week (examples in Figures 3 and 5). This procedure statistically confirmed the evident (and expected) pattern difference between Mon-Fri and Sat-Sun. Abbreviations: H = Holiday; NH = Non Holiday.

2.1 Available cars

The total number of free cars at a given time t is the sum of parked cars. When this value is low it means that many cars are in use. Figure 2 shows the usage for two typical days (from 00:00 to 23:59), one weekday and one weekend day. The "late Saturday" (Sunday early morning) usage is evident in the right graph. Other notable general remarks are: 1) night-time (between 2AM and 7AM) is a peak of free cars; 2) the usage peak (least number of free cars) is between 6PM and 9PM; 3) morning (between 8AM and noon) usage is lesser than afternoon (noon to 7PM) usage; 4) night-time peak sports a shorter timespan

(between 6AM and 8AM) w.r.t. the Monday-to-Friday night-time peaks; 5) there is a usage peak between midnight and 2AM, often more substantial than the afternoon peak. In Figure 3 the Q-Q plots of two day pairs taken as example: 2015-10-05 against 2015-09-29 (Tue against Mon), near the y = x line → statistically similar; 2015-11-22 against 2015-11-17 (Sun against Tue), far from the y = x line → statistically different. I.e., (as expected!) users' behaviour on Sunday is different from a weekday.

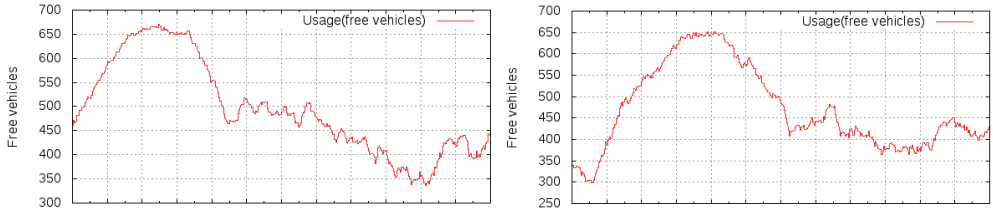


Fig. 2
Free cars average: typical weekday (Wed) vs. weekend (Sun)

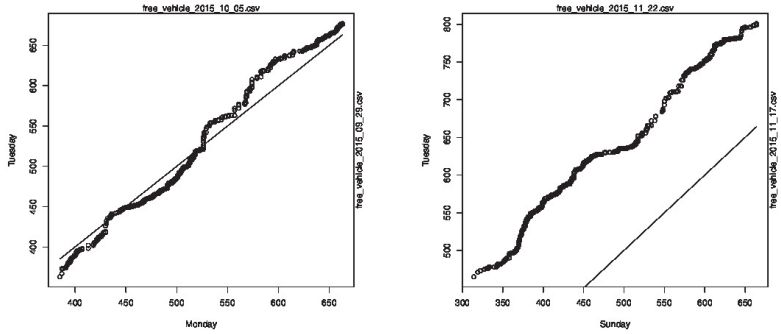


Fig. 3
Usage Q-Q plots: typical NH-NH vs. H-NH

2.2 Overall (average) distance from city centre

Notable remarks in Figure 4 ("Mon to Fri" left, "Weekend" right, they represent typical days from the dataset, X-axis is from 00:00 to 23:59) are: 1) the "lung" effect: average distance decreases during the day (8AM to 6PM) and increases during the rest of the day, i.e., people move into the city during daytime and move out of the city otherwise; 2) there is a concentration peak between 11AM and 2PM, i.e., during lunchtime users are closer to the centre; 3) between 8AM and 11AM there is a sudden out-movement of people closely followed by an in-movement, i.e., many users move, but not exactly at the same time; In Figure 5 the Q-Q plots of two day pairs taken as example: 2015-10-22 against 2015-10-21 (Wed against Thu). Again, users' behaviour on Sunday is different from a normal weekday.

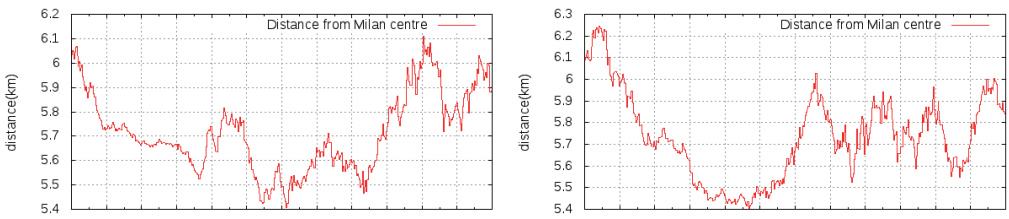


Fig. 4
Average distances: typical weekday (Wed) vs. weekend (Sun)

It is also interesting to show the average distance of vehicles "per vendor" (Figure 6): 1) Enjoy is always more distant than the others (Car2go and Sharengo); 2) Sharengo (electrical only) was always the closest; 3) there has been a "getting farer" trend in Sharengo cars between 2015 and 2016.

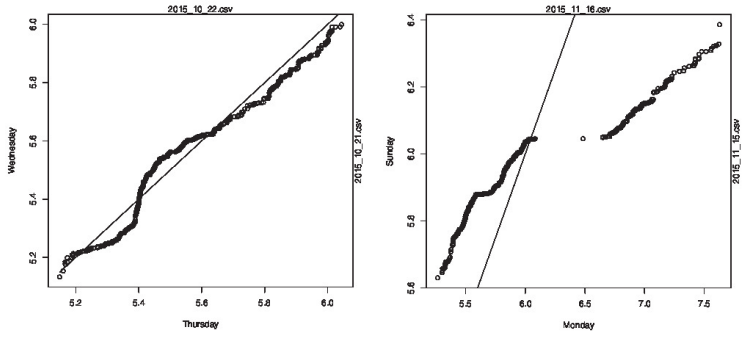


Fig. 5
Distances Q-Q plots: typical NH-NH vs. H-NH

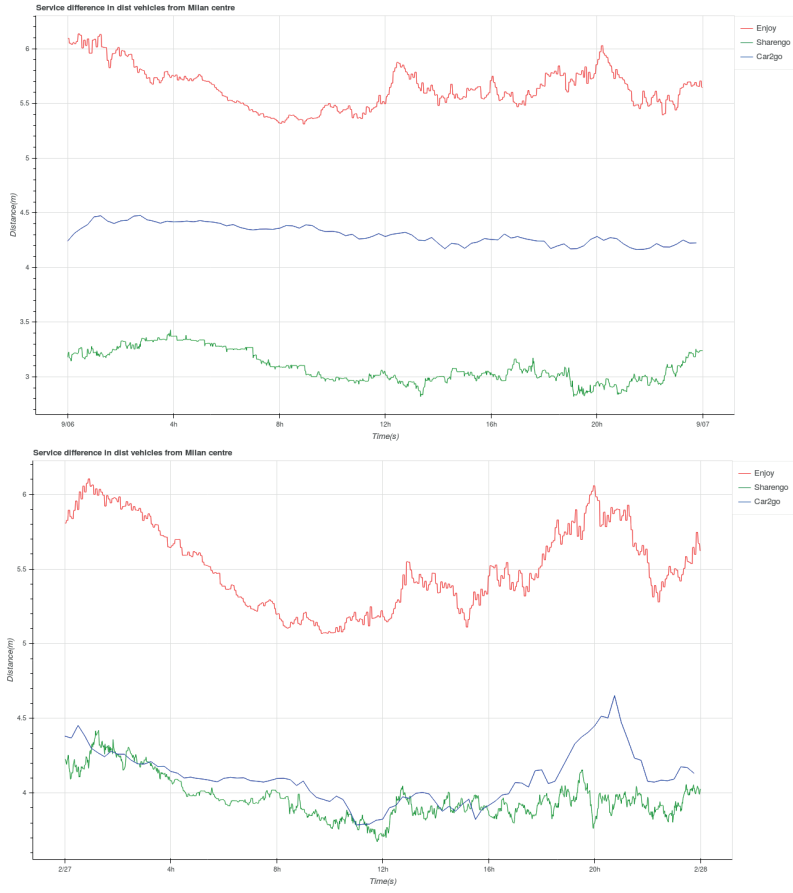


Fig. 6
Difference in average distances "per vendor" (top: 2015/09/06, bottom: 2016/02/27)

2.3 Movements

A "movement" is simply a vector (it can be plotted) between a car "catch" and a "release". The database of movements contains the records: current fuel, delta fuel, delta km, delta

time, end date, end pos lat, end pos lon, id vehicle, kmh, start date, start pos lat, start pos lon, type vehicle. Figure 7 shows a very small time-frame (10 minutes) of car movements, this kind of graph may be used to find movement patterns and/or visually identify outliers.

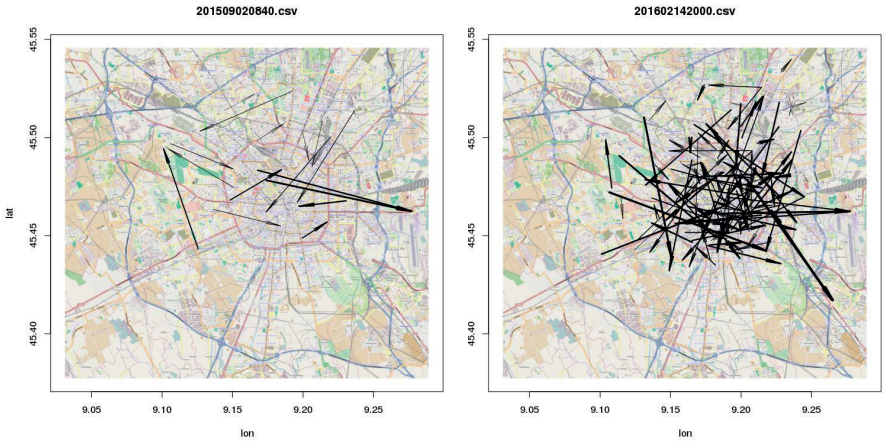


Fig. 7
At 2015/09/02 08:40 (left), two cars moving from the city centre towards the airport (rightmost arrows) in the morning;
at 2016/02/14 20:00 (right) many cars moving in the evening

3. Mandatory APIs proposal

The rationale of this proposal is based on these:
1) a resource (a large pool of cars) is converted from being private to "public" (private firms actually) control;
2) the private form of the resource is artificially limited (by regulations and policies) to push users towards the shared pool;
3) the cost burden is loaded on the public (because firms need to get revenues);
4) costs for users can be higher or lower than pre-sharing, based on usage profile;
5) workflow is heavily changed (from private to shared use);
6) environment impact can be lower if the whole system is well adapted to the actual needs of users;
7) at present, carsharing management is in the hands of private firms which do not publish enough data to let third parties implement independent analyses.

The authors' point of view is that some form of mandatory public access to status (both real-time and historical) data should be put into place by law, to give citizens the right to public accountability also in the field of carsharing. The ideal reification of this "third party" access would be the implementation of standardized APIs to be mandatorily supplied by carsharing vendors. This constraint should be introduced in every public tendering procedure.

Far from being a detailed and complete proposal (a standardization process - such as (Hovey and Bradner, 1996) - should be activated) a suggestion for such an API could be defined using a FSA (Finite State Automaton) representation of a single car in the carsharing system, using state transitions and dynamic and static properties: Dynamic: position (lat - long); fuel gauge; parked/taken anonymized user ID (if taken). Static: car ID; type (electric, diesel, gasoline, LPG, ...); odometer; usage fare; owner. A "start of discussion" function list proposal is listed below. The functions could be implemented via HTTP+REST

(Fielding and Taylor, 2000).

All the follwing functions should offer both the real-time and the historical version, i.e., they should also accept a "time-frame parameter such as in the last two examples:

- getStaticVehicles() => list of all "parked" vehicles;
- getMovingVehicles() => lista of all "non parked" vehicles;
- getInfo(id vehicle) => status of a single vehicle;
- getInfoService(service) => combined info on service
- getPath(id vehicle,interval) => returns the list of all paths followed by a vehicle in the time interval;
- getUserPath(id user,interval) => returns the list of all paths followed by a single user during the interval.

References

Roy T Fielding and Richard N Taylor. Architectural styles and the design of nework-based software architectures. University of California, Irvine Doctoral dissertation, 2000.

Jörg Firnkorn and Martin Müller. What will be the environmental effects of new free-floating carsharing systems? The case of car2go in Ulm. Ecological Economics, 70(8):1519–1528, 2011.

Richard Hovey and Scott Bradner. The organizations involved in the ietf standards process. 1996.

Diana Jorge and Gonçalo Correia. Carsharing systems demand estimation and defined operations: a literature review. European Journal of Transport and Infrastructure Research, 13(3):201–220, 2013.

Lester B Lave and Eugene P Seskin. Air pollution and human health, volume 6. Routledge, 2013.

John H Seinfeld and Spyros N Pandis. Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, 2012.

Susanne Steinle, Stefan Reis, and Clive Eric Sabel. Quantifying human exposure to air pollution—moving from static monitoring to spatio-temporally resolved personal exposure assessment. Science of the Total Environment, 443:184–193, 2013.

Andrea Trentini. Lombardy epa obtorto collo data and anti-pollution policies fallacies. Journal of e-Learning and Knowledge Society, 10(2), 2014.

US-EPA. History of the clean air act. http://www.epa.gov/air/caa/amendments.html, 2013.

Tom Vanderbilt. Traffic: Why We Drive the Way We Do (and What It Says About Us). New York Times, 2009.

Authors

Andrea Trentini andrea.trentini@unimi.it

Computer Science PhD, Researcher at Dipartimento di Informatica (Computer Science Department) of Università degli Studi di Milano. He teaches "Programming 101", "Digital Citizenship and Technocivism" and "Embedded Systems". Free/Libre/Open Source Software advocate and knowledge sharing activist. He proudly defines himself a "hacker" (with the

original term meaning). Founder of ArduinoAfternoon and HackLabCormano.

Federico Losacco federico.losacco@studenti.unimi.it

Computer scientist and student at Dipartimento di Informatica (Computer Science De-partment) of Università degli Studi di Milano. Very fond of tech-nology and Free/Libre/ Open Source Software addict. He worked as a develo-per on embedded systems, mobile systems and web backend/frontend.

DOI: 10.26314/GARR-Conf17-proceedings-28

A.Da.M. 1.0 (Archaeological Data Management): un'applicazione al servizio dell'archeologia per la gestione dei dati di scavo e ricognizione

Antonio Corvino[1], Nicodemo Abate[1], Fabio Giansante[2]

[1]Università degli Studi Suor Orsola Benincasa, [2]Spindox

Abstract. L'idea di A.Da.M. (Archaeological Data Management) nasce dall'esigenza, avvertita e maturata in anni di lavoro sul campo, di avere uno strumento agevole e nello stesso tempo veloce per la gestione dei dati, eliminando così le operazioni ridondanti, cercando di arrivare alla creazione di uno standard vicino quanto più possibile alle necessità di ogni singolo contesto. Ovviamente, il mobile rappresenta il mezzo preferenziale per la creazione di uno strumento simile: maneggevole, con buona autonomia e facilmente trasportabile. Lo strumento è studiato per il singolo archeologo (studente, ricercatore, libero professionista) e per gli enti (ditte, cooperative, università). Tramite diversi livelli di autority l'applicazione permette la gestione del lavoro in singolo o in team, su uno o più progetti. All'interno di un unico applicativo è così possibile accedere a tutti gli strumenti che solitamente un archeologo adopera sul cantiere per registrare dati ed informazioni.

Keywords. Data Management, Conservazione dei dati, Riuso dei dati, cloud storage, Archeologia.

Introduzione

"L'archeologia è distruzione": queste parole riecheggiano nelle aule di tutte le università del mondo, durante il corso di "Metodologia della ricerca archeologica", esame cardine su cui è fondata l'intera formazione degli aspiranti archeologi (Carandini 2000). La finalità della ricerca archeologica coincide con la lettura dell'azione antropica all'interno di un determinato contesto - o nei rapporti tra contesti - nel tempo e nello spazio (Manacorda 2008).

"L'archeologo non scava oggetti, ma esseri umani": così Sir Mortimer Wheeler, riassume il lavoro dell'archeologo che per la comprensione dell'evoluzione storica - ovvero per il suo fine ultimo - ha bisogno di distruggere, compromettendo inevitabilmente la possibilità di poter ripristinare uno status ante quem a cui fare appello nello sfortunato caso in cui la meta non venga raggiunta con successo (Barker 1981).

I danni legati alla rimozione degli accumuli sono, in qualche modo, limitati dalla pratica di documentare, rigorosamente, ogni evidenza riscontrata tramite schede, rilievi (recentemente anche con l'ausilio del 3D) e fotografie. Il fine è ridurre la perdita di informazioni insita nell'attività di scavo e poter, a posteriori, ragionare sul susseguirsi di eventi in un determinato contesto.

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

1. La documentazione archeologica e l'idea di A.Da.M.

Allo stato attuale, in Italia, la documentazione archeologica segue le norme dettate dall'ICCD (Istituto Centrale per il Catalogo e la Documentazione) del MiBACT (Ministero dei Beni e delle Attività Culturali e del Turismo), che fornisce precisi dettami per la catalogazione dei contesti archeologici, dei reperti (mobili ed immobili) e delle evidenze identificate durante le ricognizioni.

Tuttavia, accade spesso che le direttive fornite dall'ICCD vengano riformulate, in un format user friendly tale da velocizzare le operazioni di raccolta e registrazione dei dati. All'apparente uniformità, quindi, si contrappone una reale e, piuttosto diffusa, frammentazione di modalità per la registrazione e la catalogazione delle informazioni.

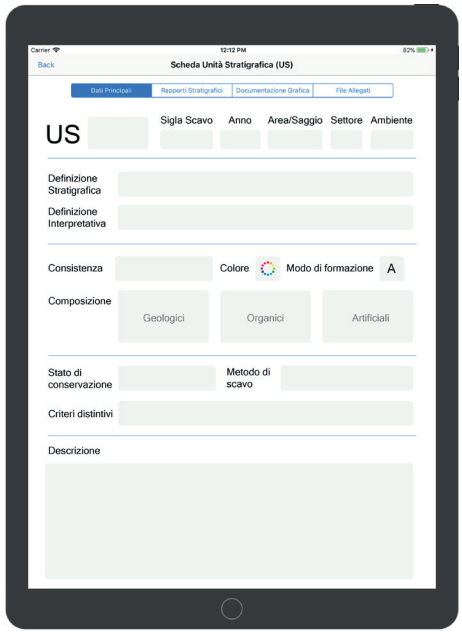
L'idea di A.Da.M. (Archaeological Data Management) nasce dall'esigenza, avvertita e maturata in anni di lavoro sul campo, di avere un supporto agevole e veloce per la gestione dei dati, eliminando, così, le operazioni ridondanti e cercando di arrivare alla creazione di uno standard vicino quanto più possibile alle necessità di ogni contesto. Ovviamente, il mobile (smartphone e tablet) rappresenta il mezzo preferenziale per la creazione di uno strumento simile: maneggevole, con buona autonomia e facilmente trasportabile.

A.Da.M. nasce per la gestione delle informazioni pertinenti ai propri lavori: cantieri, ricognizioni e laboratori/magazzini. L'applicazione è studiata per il singolo archeologo (studente, ricercatore, libero professionista) e per gli enti (ditte, cooperative, università). Infatti, tramite diversi livelli di autority (Figura 1) è possibile lavorare in singolo o in team, su uno o più progetti.

Fig. 1
Esempio di Login e autority

All'interno di un unico applicativo si rendono accessibili tutti gli strumenti che solitamente un archeologo adopera sul cantiere per registrare dati ed informazioni: diverse schede pre-impostate (sia di contesto che di reperto); GPS per la localizzazione; bussola; fotocamera; strumento di disegno; Munsell Color System per i colori; diario di scavo; note vocali;

etc. (Figura 2).
Alle normali operazioni si aggiungono funzioni accessorie di indubbia utilità come la possibilità di generare report di dati (elenchi e grafici) e collegarsi direttamente - sfruttando le funzioni del database e dei punti GPS - a software di database e GIS.



Figu. 2
Esempio di Scheda di
Unità Stratigrafica (US) in A.Da.M.

2. Architettura del progetto – Servoy Framework

L'architettura dell'applicativo fa uso del framework Servoy (wiki.servoy.com). La suite dei prodotti Servoy è basata su Java e può girare su tutte le piattaforme più popolari tra cui Windows, Mac OS, Linux e Solaris.

Servoy può connettersi ad ogni base dati presente su tutte quelle piattaforme che supportano una connessione JDBC, offrendo la possibilità di costruire un'infrastruttura SERVER-CLIENT adatta al presente contesto applicativo.
Consideriamo CLIENT tutti i dispositivi (tablet, smartphone, etc.) che forniscono la GUI interattiva dedicata all'utente per la registrazione dei dati.

Quando invece si parla di SERVER si fa riferimento ad un Application Server che fornisce l'accesso all'applicazione, gestisce la connessione al database e la concorrenza in ambiente multi-utente.

Servoy Server è un vero prodotto three-tier, che garantisce alte prestazioni, maggiore sicurezza ed una più semplice gestione del prodotto stesso.

La raccolta di dati ed informazioni all'interno di A.Da.M. è pensata per funzionare anche in contesti in cui l'aggiornamento cloud diretto è impossibile. Infatti, spesso, l'archeologo si trova ad operare in condizioni in cui vi è la totale assenza di copertura telefonica e rete dati. Per questo motivo, A.Da.M. utilizza lo storage interno del dispositivo per

immagazzinare e registrare le informazioni raccolte, per poi, tramite Servoy, riversarle, eliminando le ridondanze, all'interno di un database più ampio (Figura 3).

Fig. 3
Interoperabilità tra A.Da.M.
ed altre piattaforme

3. Conclusioni

L'apporto del supporto tecnologico rappresenta un passo decisivo ed innovativo in un settore estremamente restio all'ammodernamento: la baseline è il raggiungimento di uno strumento che permetta di automatizzare operazioni consuetudinarie, portando la gestione ed il trattamento delle informazioni (immissione, conservazione, accessibilità, replicabilità, condivisione) a livelli non raggiungibili con le attuali modalità, senza distaccarsi mai eccessivamente dalle azioni canoniche che l'operatore compie ogni giorno.

A.Da.M. fornisce tutti gli strumenti che un archeologo – a qualsiasi livello – utilizza quotidianamente, consentendo di conservare e recuperare facilmente i dati raccolti tramite salvataggi locali (su dispositivo) ed upload su database centralizzati. La personalizzazione dei format è minima ma essenziale, permettendo all'operatore di gestire al meglio il flusso di lavoro, in entrata (registrazione delle informazioni) ed in uscita, grazie all'automatizzazione di report, grafici, elenchi, tabulati, matrix (Harris 1983). Tempo e denaro, in archeologia, sono elementi di un'equazione che non può non essere presa in considerazione.

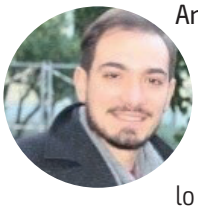
La gestione dell'autority consente, inoltre, di condividere progetti per lavori in team, fornendo privilegi a determinati utenti e garantendo – lì dove desiderata – un'accessibilità limitata ai dati ed alle operazioni da compiere.
La data di rilascio per l'applicazione A.Da.M., nella sua versione 1.0, è prevista la primavera del 2018.

Il Project Plan non esclude in alcun modo eventuali collaborazioni od integrazioni con altri sistemi simili, qualora esistenti, tali da avviare un processo di standardizzazione il cui bisogno è estremamente sentito.

Riferimenti bibliografici

Barker P. (1981), Tecniche dello scavo archeologico, Longanesi & Co., Milano 1981.
Carandini A. (2000), Storie dalla terra. Manuale dello scavo archeologico, Einaudi, Torino.
Harris E.C. (1983), Principi di stratigrafia archeologica, Carocci, Roma.
Manacorda D. (2008), Lezioni di archeologia, Laterza, Roma/Bari.

Autori



Antonio Corvino corvino.antonio@ymail.com

Archeologo e dottorando, XXXI ciclo col titolo di "Humanities and technologies", presso l'Università degli Studi Suor Orsola Benincasa di Napoli. Dal 2014 collabora con la Cattedra di Letteratura Latina Medievale dell'Università degli Studi Suor Orsola Benincasa di Napoli. I suoi studi partono dall'archeologia classica, per poi focalizzarsi, attraverso lo studio della filologia medievale, sulle Digital Humanities. É membro attivo del PRIN 2015 "ALIM'.

Nicodemo Abate abate.nicodemo@gmail.com

Archeologo. Collabora con la Cattedra di Archeologia Cristiana e Medievale dell'Università degli Studi Suor Orsola Benincasa di Napoli. I suoi studi si focalizzano sull'archeometallurgia dei contesti medievali, sul rilievo, modellazione 3D, creazione di App per i BB.CC. e sullo sviluppo di piattaforme GIS e WebGIS per il trattamento dei dati.

Fabio Giansante fabio.giansante89@gmail.com

Laurea in Informatica presso l'Università degli Studi di Napoli Parthenope. Si occupa prevalentemente di creazione App per sistemi Android/iOS. Attualmente sviluppatore software presso Spindox.

DOI: 10.26314/GARR-Conf17-proceedings-29

Mettere in campo servizi per Smart City a Messina con #SmartME

Dario Bruneo[1], Salvatore Distefano[1,2], Francesco Longo[1], Giovanni Merlino[1], Antonio Puliafito[1]

[1] Università di Messina, [2] Kazan Federal University

Abstract. Una Smart City è un'area urbana in cui le tecnologie dell'informazione e della comunicazione (ICT) sono impiegate per migliorare la qualità della vita dei cittadini in settori quali, per fare un esempio, mobilità, sorveglianza urbana e gestione dell'energia. In questo articolo presentiamo il progetto #SmartME, che mira a creare un'infrastruttura ed un ecosistema di servizi "intelligenti", sfruttando dispositivi, sensori ed attuatori distribuiti nella città di Messina.

Keywords. Keywords. Smart City, infrastruttura, Cloud, #SmartME, servizi.

Introduzione

Una Smart City è qualcosa di più di un semplice insieme di oggetti connessi a Internet, in quanto "unisce tecnologia, governo e società per abilitare le seguenti caratteristiche: un'economia intelligente, una mobilità intelligente, un ambiente intelligente, persone intelligenti, vita intelligente, governance intelligente" 1. Proiettando questa definizione nel contesto tecnologico, una Smart City può essere quindi considerata un ecosistema di infrastrutture e servizi con l'obiettivo di implementare le suddette caratteristiche. L'obiettivo è quindi quello di stabilire un ecosistema omogeneo in cui più applicazioni possano adattarsi a un ambito metropolitano, sottolineando così un'infrastruttura ICT aperta e condivisa fatta di sensori, attuazione, reti, elaborazione e risorse di archiviazione.

Seguendo un approccio di apprendimento attraverso esempi concreti, in questo articolo riportiamo un caso di studio che stiamo implementando a Messina attraverso il progetto #SmartME2. L'obiettivo principale del progetto #SmartME è quello di stabilire un'infrastruttura ed un ecosistema di servizi basati su dispositivi, sensori e attuatori già presenti sul posto. In questa maniera, abbiamo spostato il nostro focus sul lato software, adottando uno specifico framework in grado di risolvere problemi di interoperabilità, rete, sicurezza ed altro ancora. Uno dei principali vantaggi del framework proposto è la programmabilità: chiunque, se autorizzato, può iniettare ed eseguire proprie applicazioni e servizi, sfruttando la stessa infrastruttura Smart City, che è quindi condivisa da diverse applicazioni e servizi, anche nello stesso momento.

1. Panoramica

Il progetto #SmartME (Bruneo et al, 2016) è un'iniziativa finanziata attraverso una campagna di crowdfunding, che mira a trasformare Messina in una Smart City (Merlino et

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

al, 2015). L'obiettivo principale è quello di dispiegare le risorse IoT su tutto il territorio del comune di Messina, creando così un laboratorio virtuale a cui più attori possono contribuire con le proprie risorse, e sulle quali possono sviluppare applicazioni e servizi per ricerca, affari ed attività amministrative. Di conseguenza, uno dei principali contributi e novità del progetto #SmartME è la creazione di una nuova forma di Smart City, condivisa e partecipata da una moltitudine di contributori, in cui chiunque, dai cittadini alle amministrazioni pubbliche, dai negozi e dalle imprese agli edifici privati, possono condividere le loro strutture hardware per assemblare l'infrastruttura fornita. Ciò è reso possibile grazie al framework di gestione Stack4Things3 che consente di registrare e gestire queste risorse, fornendo anche meccanismi di personalizzazione e modalità di fruizione per il loro effettivo sfruttamento, basate su un modello di provisioning di tipo Cloud. In questo modo, diversi servizi sono stati sviluppati, coinvolgendo diversi stakeholder nel progetto #SmartME, come meglio descritto nelle seguenti sezioni.

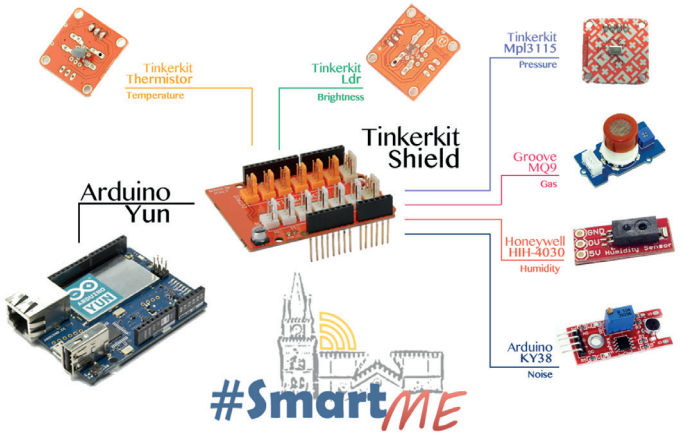


Fig. 1
Un esempio tipico di nodo #SmartME

L'infrastruttura #SmartME è composta da dispositivi che forniscono funzionalità di sensori e/o attuatori. La figura 1 riporta la composizione di un tipico nodo #SmartME. Arduino YUN è un calcolatore a scheda singola alimentato da un microcontrollore Atmel ATmega32u4 ed un system-on-a-chip Atheros AR9331. La scheda #SmartME YUN è stata equipaggiata con una shield Tinkerkit che ospita una serie di sensori a basso costo.

I nodi #SmartME sono programmati per inviare periodicamente i campioni a un dataset nel sistema di archiviazione CKAN (Comprehensive Knowledge Archive Network)4, un sistema OpenSource basato sul Web per l'archiviazione e la distribuzione di dataset. Il recapito dei dati viene eseguito tramite l'interfaccia API REST. I dati vengono utilizzati dal portale pubblico, che offre una vetrina user-friendly attraverso la quale i cittadini possono esplorare i nodi #SmartME e dare una sbirciata ai dati raccolti. Inoltre, i datastore CKAN offrono la possibilità sia ai cittadini che ai servizi di terze parti di eseguire interrogazioni complesse sui dati, così come di recuperare serie storiche.

Il cuore del progetto #SmartME è alimentato dal framework Stack4Things (Merlino

et al, 2014), che permette di gestire nodi IoT seguendo un modello di provisioning su richiesta ed orientato ai servizi, spostando il paradigma IoT verso il Cloud, da un lato, per fornire funzionalità di controllo e gestione ai nodi IoT e, dall'altra, per estendere il paradigma Cloud aggiungendo capacità pervasive di interazione con il mondo fisico.

2. Applicazioni e Servizi

Diverse applicazioni e servizi sono stati sviluppati sull'infrastruttura #SmartME. Alcuni di essi sono brevemente descritti nel seguito.

#SmartME Parking: un servizio realizzato in collaborazione con ParkSmart Srl6, una startup che mira a risolvere i problemi di parcheggio con una soluzione innovativa che sfrutta il paradigma Edge computing, con sensori distribuiti sulla città ed unità computazionali che elaborano i dati ed inviano al Cloud solo le informazioni relative all'occupazione del parcheggio e nient'altro.

#SmartME Lighting: la piattaforma #SmartME integra una soluzione a basso costo per la raccolta dati e il controllo remoto di sistemi di illuminazione di aree pubbliche, private e industriali, sviluppati da Meridionale Impianti Spa7. Ogni lampada è dotata di un dispositivo elettronico che attiva / disattiva la lampada e monitora i parametri di consumo principali come tensione, corrente e potenza assorbita.

#SmartME Pothole: un esempio di applicazione mobile, accoppiata ad un servizio Web come back-end per la mappatura ed altre attività di visualizzazione, seguendo il paradigma del mobile crowdsourcing. Lo scopo del sistema consiste nel rilevare buche e altri elementi di pericolo sulla superficie delle strade pubbliche. L'app calcola le variazioni nei valori campionati per la norma del vettore di accelerazione: intuitivamente, quando si imbatte in una buca sulla strada o, più in generale, viaggiando su una superficie stradale deformata, questi cambiamenti possono rivelarsi significativi. La presenza di una condizione potenzialmente critica viene quindi contrassegnata insieme alle corrispondenti coordinate geospaziali.

#SmartME Taxi: rappresenta l'integrazione di un'applicazione di gestione della flotta dei taxi sviluppata da Arkimede Srl9 nel portale #SmartME. Si tratta di un'applicazione mobile basata su Android in esecuzione su un tablet a disposizione del tassista e che invia periodicamente la posizione e la velocità del taxi al datastore di CKAN. Il portale #SmartME visualizza i dati su una mappa in tempo reale e consente agli utenti di ottenere informazioni sullo stato dei taxi (per esempio, se disponibili od impegnati).

3. Conclusioni

In questo documento, abbiamo presentato il progetto #SmartME insieme a tutti i principali ambiti nei quali sono stati realizzati. È stata fornita anche una descrizione delle tecnologie adottate ed il framework per l'orchestrazione e gestione dei servizi intelligenti.

Riconoscimenti

Gli autori ringraziano ParkSmart S.r.l., Meridionale Impianti S.p.a., SmartMe.io S.r.l., and Arkimede S.r.l. per la fruttuosa collaborazione nell'ambito del progetto #SmartME.

Riferimenti bibliografici

D. Bruneo, S. Distefano, F. Longo, and G. Merlino (2016), An IoT testbed for the Software Defined City vision: the #SmartME project, IEEE Int. Conf. on Smart Computing - SMARTCOMP, PP 1–6.

G. Merlino, D. Bruneo, F. Longo, A. Puliafito, and S. Distefano (2015), Software Defined Cities: a novel paradigm for Smart Cities through IoT clouds, IEEE 12th Intl. Conf. on Ubiquitous Intelligence and Computing - UIC-ATC-ScalCom, PP 909–916.

G. Merlino, D. Bruneo, S. Distefano, F. Longo, and A. Puliafito (2014), Stack4things: Integrating IoT with OpenStack in a Smart City context, IEEE Smart Computing Workshops - SMARTCOMP, PP 21–28.

http://smartcities.ieee.org/about.html

http://smartme.unime.it/

http://stack4things.unime.it

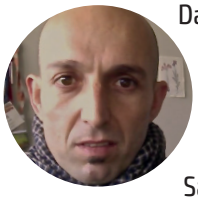
http://ckan.org

https://www.ethereum.org

http://parksmart.it

http://www.merimp.com

Autori



Dario Bruneo dbruneo@unime.it

Professore associato di sistemi embedded presso l'Università di Messina. I suoi interessi di ricerca comprendono Internet of Things (IoT), smart cities, Cloud computing e valutazione delle prestazioni.

Salvatore Distefano sdistefano@unime.it

Professore associato presso l'Università di Messina e fellow professor presso l'Università Federale di Kazan. I suoi interessi di ricerca includono Cloud computing, IoT, crowdsourcing, big data e qualità del servizio.

Francesco Longo flongo@unime.it

Ricercatore presso l'Università di Messina. I suoi interessi di ricerca includono valutazione delle prestazioni dei sistemi distribuiti, virtualizzazione di sistemi IoT ed il machine learning applicato ad ambienti intelligenti.

Giovanni Merlino gmerlino@unime.it

Assegnista di ricerca all'Università di Messina. I suoi interessi di ricerca includono Cloud ed Edge computing, Internet of Things, Network Virtualization, Mobile Crowdsensing.

Antonio Puliafito apuliafito@unime.it

Professore ordinario di ingegneria informatica e direttore del Centro Informatico di Ateneo presso l'Università di Messina. I suoi interessi di ricerca includono i sistemi paralleli e distribuiti, le reti wireless ed il Cloud computing.

DOI: 10.26314/GARR-Conf17-proceedings-30

SemplicePA: SEMantic instruments for PubLIc administrators and CitizEns

Martina Miliani[1], Anna Gabbolini[1], Lucia C. Passaro[2], Francesco Sandrelli[1], Alessandro Lenci[2] , Roberto Battistelli[1]

[1]Eti3 s.r.l., [2]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica Università di Pisa

Abstract. La trasformazione digitale italiana sta procedendo a rilento rispetto a quella Europea, con un digital divide che penalizza soprattutto i comuni più piccoli e gli open data che faticano ad essere pienamente valorizzati, con il risultato che il Foia è ancora ben lungi dall'essere applicato come dovrebbe. Eppure non mancano le iniziative civiche, alcune stimolate dalle stesse amministrazioni. Così come non mancano le tecnologie di eccellenza, sviluppate all'interno di start-up che collaborano con università statali e centri di ricerca. SemplicePA nasce in questo contesto con lo scopo di fornire uno strumento utile alla cittadinanza e alle amministrazioni a partire dall'analisi semantica computazionale di un archivio spesso sconosciuto, l'Albo Pretorio.

Keywords. Keywords. Conservazione e condivisione dei dati, Natural Language Processing, Machine Learning, Big Data.

Introduzione

Otto paesi su dieci, in Europa, hanno attivato una regolamentazione sugli open data. L'Italia, che si trova sotto la media europea, è tra i "follower" delle buone pratiche, con un "Mezzogiorno notevolmente indietro" [1]. Tra le cause, il grande divario tra le piccole e le grandi amministrazioni. In base all'osservatorio dell'Istat, le province autonome e l'85,5% dei comuni sopra i 60.000 abitanti possiedono un ufficio dedicato all'Information & Communication Technology (ICT), ovvero poco più dell'1% del totale dei comuni [2]. Anche i risultati del primo monitoraggio sull'applicazione del Freedom of Information Act (Foia) sono tutt'altro che positivi: il 73% degli utenti non ha ricevuto risposta e un diniego su tre era invece illegittimo [3]. Anche per questo, accanto all'Agenzia per l'Italia Digitale (AgID), sta lavorando il Team per la Trasformazione Digitale che ha una diversa concezione delle informazioni possedute dalla PA: "I dati sono nostri e li gestiamo insieme" [4]. Eppure in Italia si registrano già alcune iniziative sull'uso degli open data. Talvolta sono gli stessi enti pubblici a indire contest per premiare il migliore utilizzo dei dati aperti: a vincere l'hackathon sul tema della disabilità indetto dal Comune di Lecce, è stato il censimento delle barriere architettoniche in città [5]. Inoltre, sono tantissime anche le iniziative civiche, come il progetto di crowdfunding Ricostruzione Trasparente [6], il cui obiettivo è quello di tenere traccia di tutti gli atti pubblici che consentano di esercitare un controllo diffuso sugli attori della ricostruzione in seguito al terremoto del 2016 avvenuto nel Centro Italia. Si tratta di dati rilasciati dalle pubbliche amministrazioni, che sono stati poi rielaborati

e resi disponibili ai cittadini in modo del tutto nuovo. Purtroppo, sono ancora tante le risorse non adeguatamente valorizzate, come ad esempio, l'Albo Pretorio, l'archivio degli atti di ciascun comune.

La prima legge che sancisce la trasformazione digitale dell'Albo Pretorio risale al gennaio 2009 e giunge a pieno regime nel 2013 [7]. Nonostante la trasformazione sia avvenuta, cercare un atto all'interno dell'Albo sembra possibile soltanto conoscendo in anticipo il documento stesso: gli atti sono ricercabili solo in un arco di tempo ristretto, in genere 15 giorni, e nei siti di molti comuni per recuperare un provvedimento è necessario conoscerne la data, l'organo che lo ha emanato, l'oggetto o il suo numero identificativo. A mancare sono soprattutto le relazioni tra i singoli documenti, non solo tra quelli di uno stesso comune, ma anche, e soprattutto, tra comuni differenti.

1. Un motore di ricerca semantico

SemplicePA [8] è nato nel 2015 con l'obiettivo di valorizzare i contenuti degli atti registrati nell'Albo Pretorio di tutti i comuni italiani, di rendere navigabili le informazioni e soprattutto le relazioni che tra esse intercorrono. Il fine è quello di creare un Albo Pretorio Nazionale che consenta la navigazione dei contenuti attraverso un motore di ricerca semantico, in grado di estrarre informazioni significative, quali nomi di aziende, organizzazioni, persone e luoghi e mostrare le loro relazioni attraverso strumenti di visual analytics.

In Italia, Cogito [9] si basa sull'analisi semantica dei testi sfruttando un'ampia banca dati che vede raggruppate più ontologie differenti, costruite anche in diverse lingue. È nato invece all'Università di Pavia lo strumento Facility Live, che mostra il suo valore nei domini più ristretti: l'ontologia dietro a motori di ricerca come questo è molto più "specializzata" e per questo anche precisa e puntuale nel recupero delle informazioni richieste dall'utente [10]. Légilocal è un motore di ricerca semantico che in Francia si occupa della gestione degli atti, dedicando anche un framework apposito per la loro stesura, in modo che siano facilmente leggibili ed elaborabili dal motore di ricerca (Amardeilh, 2013). Sugli enti locali in Italia si è specializzato il motore di ricerca Sophia Semantic Search, che riconosce le entità elencate all'interno dei documenti e li classifica per similarità [11]. SemplicePA è un ambiente dotato di vari componenti in grado di arricchire i documenti amministrativi con diversi tipi di informazioni semantiche che ne consentono, oltre che l'indicizzazione, anche la codifica, la ricerca e la navigabilità, in una prospettiva del tutto nuova e in linea con il paradigma di trasformazione digitale sostenuto da AgID. I componenti principali del sistema sono descritti di seguito.

2. Estrazione delle entità

All'interno di ogni documento sono individuate diverse entità: persone, luoghi, aziende, organizzazioni, importi, date e indirizzi email ma anche elementi più specifici dei provvedimenti amministrativi come riferimenti legislativi e ad altri atti, partite iva, codici identificativi di gara e codici fiscali.

Al fine di estrarre questi nuclei informativi, il sistema sfrutta un modulo di analisi linguistica del testo (Dell'Orletta et al. 2014) e l'estrazione della terminologia (Passaro e

Lenci, 2016). L'estrazione vera e propria delle entità avviene integrando due approcci diversi, uno basato su regole e un altro su modelli di "machine learning". L'approccio a regole è basato su algoritmi che contengono precise espressioni regolari sull'estrazione. Ad esempio una porzione di testo sarà estratta e classificata come partita iva se costituita da un codice di undici cifre che rispetta precisi parametri. L'altro approccio, descritto in dettaglio in Passaro et al., 2017, è stato sviluppato da Eti3 in collaborazione con il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa. In questo caso, la probabilità che i termini estratti siano delle entità è dedotta dalla distribuzione delle parole all'interno di un corpus di training annotato manualmente con le entità rilevanti per il dominio della pubblica amministrazione. Infine, un modulo di "normalizzazione" si occupa di riportare le varie entità a una forma univoca standard per astrarre rispetto alle forme grafiche in cui una stessa entità viene citata all'interno dei vari documenti.

3. Ontologia

L'ontologia su cui si basa SemplicePA è costruita con un metodo bottom-up e top-down, da un lato attraverso l'individuazione automatica dei termini di dominio (Passaro e Lenci, 2016) e la loro espansione sfruttando metodi di semantica distribuzionale (Baroni e Lenci, 2010), e dall'altro attraverso la loro classificazione da parte di esperti di quel dominio. Inoltre, i documenti sono stati organizzati per aree tematiche sfruttando algoritmi di Topic Modeling basati su Latent Dirichlet Allocation (Blei, 2003; Blei 2012), un modello generativo bayesiano in grado di individuare gli "argomenti latenti" dei documenti, che sono rappresentati da una distribuzione di probabilità delle parole e dei documenti. I testi, quindi, vengono classificati sia in base alla presenza dei termini dell'ontologia, sia in base agli argomenti estratti automaticamente sfruttando LDA, che permette di cogliere le diverse aree tematiche degli atti amministrativi.

4. Network Analysis

Le relazioni tra le entità presenti nei documenti sono calcolate dalla piattaforma sulla base della compresenza all'interno degli atti. Una sezione è appositamente dedicata alla visualizzazione di reti in cui i nodi sono le entità estratte, le relazioni sono gli archi che le collegano e il peso degli archi è dato dal numero di documenti in cui le entità collegate sono presenti contestualmente. Si può partire da una persona, un'organizzazione o un'azienda

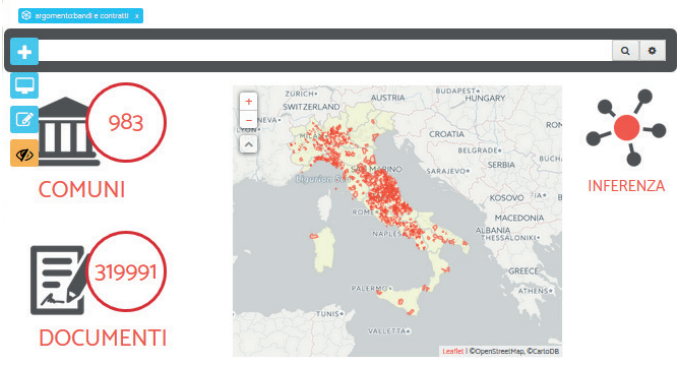


Fig. 1
La homepage di SemplicePA: un esempio di ricerca sui documenti processati associati all'argomento "bandi e contratti".

e decidere di visualizzare in una rete diversi tipi entità ad essa "collegate" (ancora aziende, persone, organizzazioni o atti stessi). In alternativa, è possibile visualizzare le relazioni tra gli elementi a partire da un gruppo di documenti selezionati.

5. Altri Strumenti

Tra gli altri strumenti offerti dalla piattaforma, una mappa, caricata automaticamente da OpenStreetMap [12] all'invio di ciascuna query, segnala i comuni e i luoghi citati all'interno dei documenti (Figura 1); le entità che appaiono all'interno dei documenti restituiti all'utente sono ordinate per frequenza, in modo da fornire una panoramica generale dei contenuti. Inoltre, per ciascun documento, sono rappresentati graficamente in una rete i riferimenti agli altri atti, allo scopo di ricostruire l'iter di pubblicazione dei documenti che si rifanno ad un unico procedimento amministrativo. Per una maggiore navigabilità, in fondo alla pagina di consultazione dell'atto sono presentati i documenti simili e una sezione di visual analytics mostra i trend delle pubblicazioni degli atti nel tempo, in base ai vari argomenti. Uno strumento aggiuntivo, infine, mette in contatto gli utenti connessi alla piattaforma attraverso una chat.

6. Conclusioni

SemplicePA nasce per valorizzare gli atti amministrativi di tutta Italia, grazie all'applicazione delle tecnologie del linguaggio più innovative che rendono le informazioni contenute nei documenti strutturate e navigabili. SemplicePA consente di consultare con facilità gli atti pubblicati anche ai comuni cittadini, mentre agli addetti ai lavori fornisce strumenti di analisi e di visualizzazione di dati che consentono una maggiore comprensione della realtà amministrativa. Gli strumenti di SemplicePA migliorano l'efficienza e l'organizzazione dell'ente in maniera del tutto automatica, e forniscono concretezza a obiettivi fondamentali quali trasparenza, Foia e anticorruzione. Un modo nuovo di interpretare la digitalizzazione della PA, che genera nuova conoscenza e la mette a disposizione di cittadini, amministratori e funzionari, perché siano più partecipi e più efficienti. L'innovazione tecnologica di SemplicePA contribuisce a più importanti innovazioni di carattere culturale, politico e sociale che dal territorio possono determinare un virtuoso miglioramento della PA.

Riferimenti bibliografici

Amardeilh F., Bourcier D., Cherfi H., Dubai, C.H., Garnier A., Guillemin-Lanne S., Mimouni N., Nazarenko A., Paul È., Salotti S., Seizou M., (2013), The Légilocal project: the local law simply shared, JURIX, PP 11-14.

Baroni M., & Lenci A. (2010). Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4), PP 673-721.

Blei, D. M., Ng A. Y., Jordan I. M., (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, (3), PP 993-1022.

Blei, D. M. (2012). Probabilistic Topic Models, Communications of the ACM, (55:4), PP 77-84.

Dell'Orletta F., Venturi G., Cimino A., & Montemagni S. (2014). T2k2: a system for automatically extracting and organizing knowledge from texts. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014).

Passaro L. C., Lenci A. (2016), Extracting Terms with EXTra, Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives, Geneva, Editions Tradulex, PP 188-196.

Passaro L. C., Gabbolini A., Lenci A. (2017) INFORMed PA: A NER for the Italian Public Administration Domain. In Proceedings of the 4th Italian Conference on Computational Linguistics (CLiC-it 2017). Forthcoming.

1. Luca Tremolada, "L'Europa dei dati. Otto paesi su dieci hanno regole sugli open data, Il Sole 24 Ore, 5 Aprile 2017 (http://goo.gl/ZqVAfG) [Accesso: 09/11/2017].
2. Istat, "Le tecnologie dell'informazione e della comunicazione nella pubblica amministrazione locale", 2015 (http://goo.gl/7q4Loq) [Accesso: 09/11/2017].
3. Diritto di Sapere, "Ignoranza di Stato" (http://goo.gl/K4Cgx1) [Accesso: 09/11/2017].
4. Raffaele Lillo, "Data & Analytics Framework", Medium, 13 Febbraio 2017 (http://goo.gl/CtQXNy) [Accesso: 09/11/2017].
5. Piersoft, "Lecce, Luoghi accessibili per disabilità varie e di interesse", Umap, 30 Aprile 2016 (http://goo.gl/NJq6g7) [Accesso: 09/11/2017].
6. Ricostruzione Trasparente (http://ricostruzionetrasparente.it) [Accesso: 09/11/2017].
7. Qualità PA, "Albo Pretorio Online" (http://goo.gl/LxNLV7) [Accesso: 09/11/2017].
8. SemplicePA (http://www.semplicepa.it/) [Accesso: 09/11/2017].
9. Cogito, Expert System (http://goo.gl/CWJE2o) [Accesso: 09/11/2017].
10. Luca Piana, "Facility LIve, start-up italiana che sfida Google", L'Espresso, 14 Dicembre 2015 (http://goo.gl/sjxblF) [Accesso: 09/11/2017].
11. Celi, Language Technology, (http://goo.gl/2XxCqU) [Accesso: 09/11/2017].
12. OpenStreetMap (https://goo.gl/V96Vsw) [Accesso: 09/11/2017].

Autori

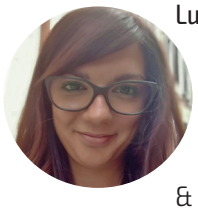
Martina Miliani martina.miliani@semplicepa.it

Giornalista pubblicista, per quattro anni ha vissuto a Palermo dove ha lavorato come cronista. Laureanda del corso Magistrale di Informatica Umanistica dell'Università di Pisa, collabora con ETI3 occupandosi prevalentemente di user experience nell'ambito del progetto SemplicePA.

Anna Gabbolini anna.gabbolini@eti3.it

Avvocato, con esperienza nel dominio della Pubblica Amministrazione e nell'ambito della gestione e rendicontazione di progetti R&S. In ETI3 si occupa di coordinamento UX, di analisi dei procedimenti automatizzabili, di elaborazione di tassonomie di base per implementazione delle ontologie, della definizione dei requisiti d'uso per applicativi di analisi

semantica e linguistica e della loro verifica.



Lucia C. Passaro lucia.passaro@for.unipi.it

È assegnista di ricerca presso il presso il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa, e membro del CoLing Lab. I suoi interessi di ricerca vanno dall'Affective computing, all'estrazione di informazione da corpora. Altri ambiti di interesse sono il text mining, la semantica distribuzionale, l'information retrieval, la Business & Competitive Intelligence.

Francesco Sandrelli francesco.sandrelli@eti3.it

Si occupa della ricerca e dello sviluppo delle tecnologie open source per varie aziende IT. Si occupa anche della definizione dell'architettura dei sistemi e della definizione delle tecnologie di riferimento. Partendo da una formazione scientifica e da un dottorato in Fisica, ha trasformato la passione per l'informatica in un'esperienza di oltre 10 anni come sviluppatore e Project manager.

Alessandro Lenci alessandro.lenci@unipi.it

Professore associato di linguistica presso il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa, dove dirige il Laboratorio di Linguistica Computazionale (CoLing Lab). Ha come principali aree di ricerca la semantica distribuzionale, lo sviluppo di strumenti e risorse per il trattamento automatico della lingua ed estrazione delle informazioni da testi.

Roberto Battistelli roberto.battistelli@eti3.it

Dalla sua esperienza come amministratore comunale è nata l'idea degli strumenti volti ad implementare la knowledge awareness nella Pubblica Amministrazione e, quindi, nell'ambito dei soggetti privati. In ETI3 si occupa di analisi dei requisiti e delle criticità, di progettazione e di sviluppo, della verifica dei risultati e dei test, del coordinamento e della collaborazione con i soggetti istituzionali.

DOI: 10.26314/GARR-Conf17-proceedings-31

I-Media-Cities, una piattaforma multidisciplinare per l'analisi e l'annotazione di materiale video

Simona Caraceni, Michele Carpenè, Mattia D'Antonio, Giuseppe Fiameni, Antonella Guidazzoli, Silvano Imboden, Maria Chiara Liguori, Margherita Montanari, Giuseppe Trotta, Gabriella Scipione

Cineca

Abstract. Il progetto europeo I-Media-Cities ha come obiettivo quello di dare nuove possibilità di accesso al Patrimonio audiovisivo delle cineteche e archivi europei. Grazie ad innovativi strumenti di ricerca, decodifica automatica di immagini in movimento ed annotazioni automatiche di metadati, diventa possibile esplorare le città da prospettive finora inedite.

Keywords. Accesso libero ai dati, Condivisione dei dati e cloud storage, Research Support Issues, Conservazione dei dati.

Introduzione

I-Media-Cities è un progetto finanziato dalla Comunità Europea che coinvolge 9 archivi cinematografici europei, 5 istituti di ricerca, 2 fornitori tecnologici e uno specialista di modelli di business digitali (https://imediacities.eu/). L'obiettivo principale è la messa in condivisione, l'accesso e la valorizzazione dei contenuti audiovisivi presenti negli archivi per favorire la ricerca nel settore delle Digital Humanities. Il progetto ruota attorno alla storia delle città coinvolte e prevede la realizzazione di una piattaforma innovativa attraverso la quale è possibile raccogliere, analizzare, integrare e condividere l'enorme quantità di opere audiovisive provenienti dagli archivi - dalla fine del XIX secolo in poi - che descrivono le città in tutti gli aspetti, tra cui la trasformazione fisica e le dinamiche sociali. La piattaforma prevede la realizzazione di diversi e-environment che verranno utilizzati da ricercatori ed innovatori per la ricerca e per altri scopi creativi

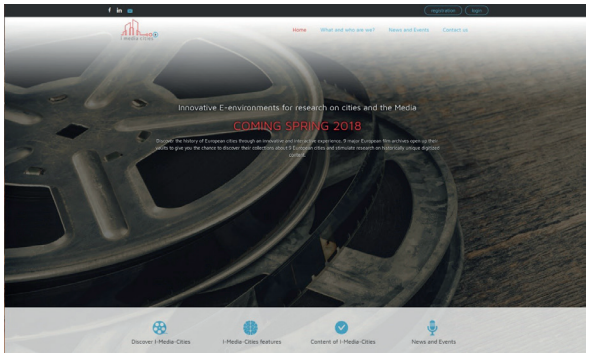


Fig. 1
Home page del sito I-Media-Cities

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

(figura 1). Ciò consentirà nuovi approcci alla ricerca nell'ambito di Digital Humanities, Scienze Sociali ed industrie creative.

Il progetto triennale - attualmente alla fine del suo secondo anno - ha raggiunto una importante milestone nello sviluppo della piattaforma che rappresenta la spina dorsale di tutto il progetto. L'aspetto innovativo consiste nella possibilità di eseguire annotazioni automatiche sui contenuti multimediali che sono già stati arricchiti, a livello di documento, con metadati provenienti dagli archivi e che verranno ulteriormente annotati manualmente.

La piattaforma riceve i contenuti in forma di video e immagini, a cui vengono associati i metadati forniti in file XML dagli archivi audiovisivi. Una volta che i contenuti e i metadati vengono caricati all'interno della piattaforma, diversi strumenti di analisi video automatici, orchestrati attraverso una pipeline, li analizzano ed estraggono informazioni e metadati, in formato XML o in JSON, a livello di fotogramma o a livello di scena.

1. Annotazione e tagging automatico

La piattaforma IMC deve essere in grado di estrarre le annotazioni automatiche tramite strumenti di analisi video forniti dal partner tecnico Fraunhofer per Digital Media Technology IDMT; (Https://www.idmt.fraunhofer.de/en.html). I recenti progressi nel campo della computer vision e del machine learning consentono di estrarre informazioni e metadati interessanti da immagini e video più velocemente di quanto potrebbe fare qualsiasi essere umano, cosa particolarmente importante con la crescente quantità di dati digitali o digitati. Fraunhofer, responsabile dello sviluppo e della configurazione degli strumenti di automatic tagging, propone l'utilizzo di diversi strumenti per eseguire in automatico la segmentazione video temporale, il rilevamento del movimento della telecamera, la misurazione della qualità video ed il rilevamento e riconoscimento degli oggetti. Tutti questi strumenti sono stati integrati nella piattaforma e orchestrati attraverso una

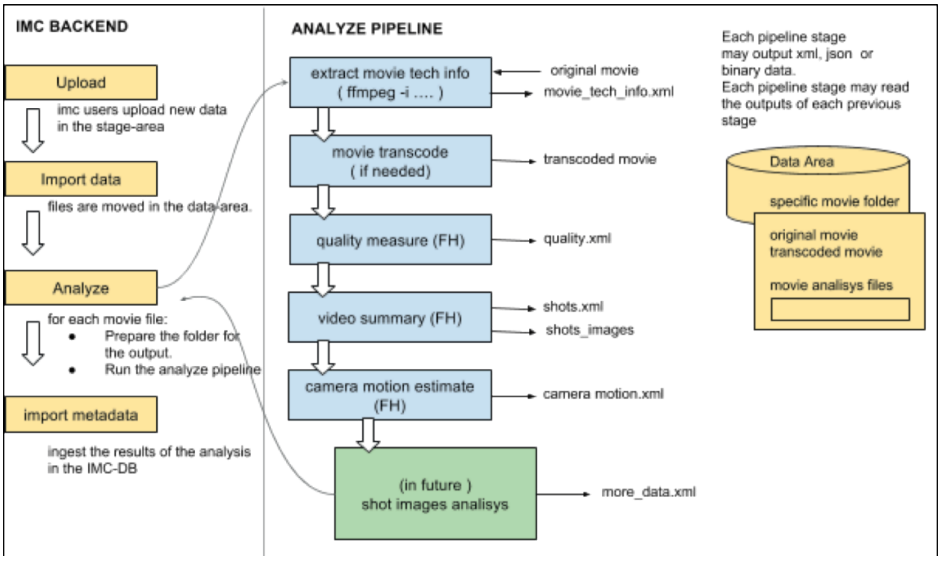


Fig 2
IMC pipeline

catena di procedure da parte di CINECA.

2. Annotazione e tagging manuale

Per agevolare il tagging manuale di un così vasto repertorio è stato predisposto un vocabolario controllato costituito da liste di termini a tre livelli di classificazione che coprono i topics definiti come essenziali per lo studio multidisciplinare del territorio e della storia urbana attraverso il materiale multimediale, per citarne alcuni: Architettura, Sviluppo urbano, Storia dei veicoli di trasporto, Sistema del traffico, Turismo, Eventi e Celebrazioni.
Poiché il progetto I-Media-Cities adotta un approccio Linked Data per il metadata model, si sta ricercando l'eventuale versione Linked Data per ogni lista di termini specifica del vocabolario per arricchire con informazione resa disponibile da differenti domini di conoscenza. Ad esempio informazioni sullo stile architettonico possono essere annotate sia con termini dal Vocabolario del progetto sia recuperandole da un thesaurus o un'ontologia di dominio che le mette a disposizione come Linked Data. Inoltre nel sistema di annotazione è implementato il collegamento con Geonames (http://www.geonames.org/) per annotare le coordinate geografiche di ogni edificio importante, monumento o landmark individuato.

Il sistema di annotazione descritto fino ad ora prevede l'inserimento del tag a livello di ogni singolo segmento del video analizzato. Per i metadati sia tecnici che descrittivi che corredano ogni contenuto multimediale nel suo complesso sono stati selezionati sia lo standard di meta datazione sia i vocabolari specifici usati dallo European Film Gateway project (http://www.europeanfilmgateway.eu/).

L'obiettivo del modello di annotazione di I-Media-Cities è, quindi, fornire un collegamento verso i domini della conoscenza, gettando le basi per un sistema di annotazioni semantico interconnesso ad altri set Open Data disponibili via Web (come il summenzionato Geonames o Wikidata).

Naturalmente l'uso estensivo di vocabolari controllati e di servizi Linked Open Data sarà comunque affiancato dalla possibilità di aggiungere annotazioni a testo libero.

3. Interfaccia utente

Il framework back-end di I-Media-Cities fornisce funzionalità di "ricerca e browsing" che elaborano le richieste provenienti dall'interfaccia utente attraverso il front end del portale web e, conseguentemente, recuperano le informazioni relative agli elementi contenuti, che sono state caricate nel repository Neo4J (https://neo4j.com/), sotto forma di oggetti grafici.

I metadati raccolti possono essere presentati agli utenti attraverso diverse modalità: geolocalizzati sulle mappe oppure mostrando eventi lungo una linea temporale o presentando informazioni strutturate in un albero interattivo. Una caratteristica interessante è il "reverse storyboard" che, recuperando dal film completo l'immagine iniziale di ogni singolo shot, fornisce un riepilogo leggibile o stampabile di un filmato, incluse le istantanee di riprese, arricchite con le annotazione di movimento di camera e la trascrizione automatica del discorso parlato, se presente.

Lo storyboard video (figura 3) include il numero della scena a cui si riferisce, l'intervallo di frame, un'immagine key frame della scena, informazioni temporali sulla scena, come start, timecode e durata e, come detto, i parametri di movimento della fotocamera relativi alla scena.

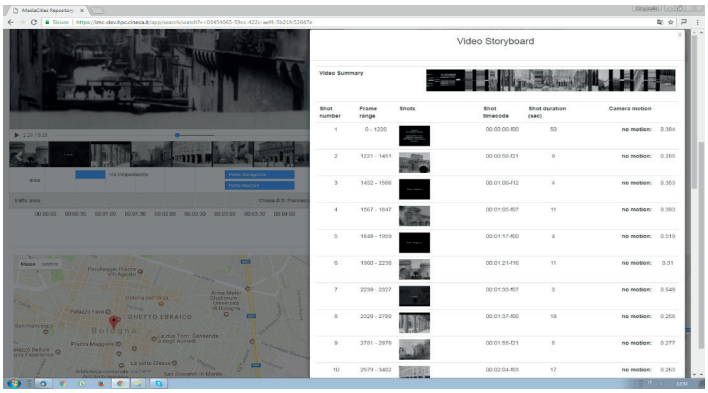


Fig 3
Funzionalità di reverse storyboard che mostra le scene, i fotogrammi, ed i luoghi riferiti al materiale audiovisivo

La visualizzazione e le funzionalità messe a disposizione nel portale web di IMC saranno differenziate a seconda delle tipologie di utenza. Al momento si prevedono servizi pensati per i ricercatori e servizi dedicati al pubblico in generale. I ricercatori avranno l'opportunità di pubblicare i risultati in forma di raccolte dinamiche di contenuti A/V e immagini fisse. Lo strumento di raccolta dinamica, denominato Virtual Collection Creator, è stato progettato per creare nuovi modi di navigare e presentare i contenuti per le diverse aree di ricerca definite nel progetto. Il Virtual Collection Creator simulerà, infatti, spazi espositivi tridimensionali in cui raccogliere e presentare immagine statiche e in movimento e documenti risultato delle ricerche sulla piattaforma.

4. Conclusioni

Il progetto I-Media-Cities, sviluppato con una metodologia di tipo Agile (https://it.wikipedia.org/wiki/Metodologia_agile), proseguirà nel suo percorso iterativo con un continuo confronto fra sviluppatori della piattaforma e curatori dei contenuti, tra fasi di sviluppo e feedback, verso la realizzazione di una piattaforma in grado di usare algoritmi e crowdsourcing rispettivamente per il tagging automatico e manuale, utilizzando anche le risorse di calcolo più opportune come, per esempio, quelle basate su GPU. L'adattamento dinamico consentirà infatti anche a questo processo di procedere eventualmente in maniera ricorsiva, in grado di accogliere via via sempre nuovi algoritmi e annotazioni, per una piattaforma capace di porsi come un nuovo ecosistema digitale a disposizione di ricercatori e cittadini alla scoperta di una comune identità europea attraverso lo studio degli ambienti urbani.

Riferimenti bibliografici

Baraldi Lorenzo, Grana Costantino, Cucchiara Rita, "Recognizing and presenting the storytelling video structure with deep multimodal networks" IEEE Transactions on mul-

timedia, pp. 1 -14 , 2016

Baraldi Lorenzo, Grana Costantino, Cucchiara Rita, “Scene-driven retrieval in edited videos using aesthetic and semantic deep features” Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, New York, USA, pp. 23 -29 , 6-9 Giugno 2016, 2016
Doi: 10.1145/2911996.2912012

Capoccioni Carlo, Porcari Giovanni, il Muvir: il primo museo virtuale delle banche operanti in italia, in negri clementi gianfranco, economia dell’arte: Proteggere, gestire e valorizzare le opere d’arte, EAGEA, Milano, 2017

Autori

Simona Caraceni s.caraceni@cineca.it
PhD alla Plymouth University in Aesthetics and Technology, si occupa di progetti per Musei e Beni Culturali per il Visit-Lab del Dipartimento di SuperCalcolo, Applicazioni ed Innovazione del Cineca

Michele Carpené m.carpen@cineca.it
Laureato presso la Facoltà di Scienze dell'Informazione all'Università degli Studi di Bologna, specializzatosi in autonomation. Software Engineer con esperienza in C++, Java e PHP. Presso CINECA su progetto internazionali presso il Middleware team e nello sviluppo di applicativi interni.

Mattia D'Antonio m.dantonio@cineca.it
Master in Informatica presso l'Università di Sapienza (Roma, IT) e un dottorato di ricerca in Biochimica, Biologia Molecolare e Bioinformatica presso l'Università di Bari, dal 2007 lavora presso CINECA nel gruppo middleware HPC per la gestione dei dati. La sua attività principale è la costruzione e la gestione di servizi di calcolo ad alte prestazioni, database e flussi di lavoro automatizzati di analisi collaborando in numerosi progetti di ricerca nazionali ed internazionali.

Giuseppe Fiameni g.fiameni@cineca.it
Laureato in Scienze Informatiche presso l'Università di Bologna ed è un dottorando all'Università di Modena e Reggio Emilia. Guida il gruppo Servizi Dipartimento di SuperCalcolo, Applicazioni ed Innovazione del Cineca. Attualmente è coinvolto in diversi progetti europei, tra cui il progetto Human Brain, EUDAT2020 e PRACE, in attività di ricerca e sviluppo nel campo della gestione e dell'elaborazione dei dati.

Antonella Guidazzoli a.guidazzoli@cineca.it
MD in Ingegneria Elettronica; MD in Storia, “cum laude”. Attività: visualizzazione scientifica, archeologia virtuale, digital cultural heritage. Responsabile del Visit Lab (Visual Information Technology Laboratory) del Dipartimento di SuperCalcolo, Applicazioni ed Innovazione del Cineca.

Silvano Imboden s.imboden@cineca.it
MD in Computer Science. Senior architect e sviluppatore al Cineca; supervisor tecnico della

computer grafica.

Maria Chiara Liguori m.liguori@cineca.it
Laureata in Scienze Politiche ed in Storia Contemporanea, con un dottorato di ricerca in Storia. Coordina progetti presso il Visit-Lab del Dipartimento di SuperCalcolo, Applicazioni ed Innovazione del Cineca.

Margherita Montanari m.montanari@cineca.it
Laureata in Sociologia. Lavora presso Cineca dal 1990, ha partecipato allo sviluppo di sistemi di supporto alle decisioni e applicazioni per la gestione e la rappresentazione di domini di conoscenza per gli aspetti di estrazione analisi e classificazione di concetti, predisposizione di thesauri e tassonomie. Attualmente partecipa alle attività di analisi dei requisiti , gestione metadati e test di funzionalità delle applicazioni sviluppate.

Giuseppe Trotta g.trotta@cineca.it
Laurea in Informatica prima di entrare nel progetto mEDRA dal 2008 occupandosi del repository metadati esistente e ai diversi servizi web costruiti su di esso. È entrato a far parte della CINECA nel 2016 nel "team metadati' del dipartimento HPC. È principalmente coinvolto nella progettazione e nello sviluppo di repository di metadati e applicazioni semantiche.

Gabriella Scipione g.scipione@cineca.it
Coordina da vari anni all'interno del dipartimento HPC in CINECA il team di sviluppo che si occupa principalmente di metadata management, rights management, open access e persistent identifiers, . E' responsabile da diversi anni di numerosi progetti europei, fra i quali I-Media-Cities, FORWARD (http://project-forward.eu/), RDI (http://www.rdi-project.org/)Arrow, Arrow Plus (www.arrow-net.eu), VOA3R (www.voa3r.eu), mEDRA (https://www.medra.org/).

DOI: 10.26314/GARR-Conf17-proceedings-32

Un innovativo graphic matching system per il recupero di informazioni di contenuto in database digitali di manoscritti antichi

Nicola Barbuti[1], Stefano Ferilli[2], Tommaso Caldarola[3]

[1]Università degli Studi di Bari Aldo Moro, Dipartimento di Studi Umanistici,
[2]Università degli Studi di Bari Aldo Moro, Dipartimento di Informatica,
[3]D.A.BI.MUS. S.r.l.

Abstract. Il paper descrive il sistema di graphic matching ICRPad M-Evo, sviluppato con l'obiettivo di consentire agli studiosi di humanities di effettuare ricerche su grandi database di manoscritti storici applicando ai data humanities l'approccio metodologico definito dal "quarto paradigma" del data science (data intensive scientific discovery – Gordon Bell, 2012). Secondo tale approccio, gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di pattern estratti direttamente da database di grandi dimensioni.

Keywords. Graphic Matching, Data Humanities, Digital Recognition

Introduzione

Nel presente intervento si descrive l'innovativo sistema di graphic matching ICRPad, che utilizza un algoritmo sviluppato con l'obiettivo di consentire agli studiosi di humanities di effettuare ricerche su grandi database di manoscritti storici applicando ai data humanities l'approccio metodologico definito dal "quarto paradigma" del data science (data intensive scientific discovery – Gordon Bell, 2012). Secondo tale approccio, gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di pattern estratti direttamente da database di grandi dimensioni.

A oggi, infatti, i database digitali a disposizione degli studiosi del CH utilizzano processi di interrogazione che replicano il medesimo approccio metodologico di tipo tradizionale, il cui presupposto indispensabile è l'elaborazione preliminare di ipotesi precise sulle quali si vanno poi a formulare le query. Un approccio che, con lo sviluppo di database sempre più ampi e complessi, risulta ormai inadeguato a soddisfare pienamente i bisogni di chi li interroga.

1. ICRPad M-Evo

L'algoritmo utilizzato nel modulo M-Evo di ICRPad è stato sviluppato avendo quale obiettivo la costruzione di uno strumento tecnologico che consentisse agli studiosi di paleografia di avvalersi nelle proprie ricerche dei database digitali esistenti, interrogandoli sia secondo metodi di approcci tradizionali (primo e secondo paradigma), sia utilizzando

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

l'approccio definito dal quarto paradigma, del tutto nuovo nel dominio di riferimento, di modo da poter inferire nuove o inattese ipotesi di ricerca dall'analisi dei dati risultati dall'interrogazione dei database.

L'algoritmo si basa sul concetto di shape contour recognition, che consente di evitare laboriose attività manuali o complessi training preliminari per la segmentazione del layout e il riconoscimento delle regioni grafiche. L'utente seleziona direttamente sul layout di un'immagine da lui preliminarmente scelta una regione grafica, che l'algoritmo codifica come lo shape model da utilizzare quale chiave di ricerca per recuperare regioni omografe o graficamente simili in una o più immagini di destinazione.

Per eseguire il matching con le immagini di destinazione, l'algoritmo utilizza non i valori in scala di grigio dell'immagine, ma i pixel della forma che costituisce il modello scelto dall'utente e il parametro del numero di livelli della piramide che ne strutturano la rappresentazione iconica.

In tal modo, il processo di interrogazione del modulo M-Evo consente la massima efficacia nella ricerca e, contestualmente, le più ampie potenzialità di effettuarla sia secondo metodi tradizionali che secondo il quarto paradigma, in quanto:

- permette di collegarsi real time come client a n database esistenti on line le cui immagini sono fruibili liberamente, grazie alla funzione di selezione e scelta di "repository" prevista nel sistema;
- consente di visualizzare ed esplorare le immagini contenute nei diversi database per valutare eventuali elementi di interesse, anche secondo scelta casuale, da selezionare per creare shape models da utilizzare quali chiavi di ricerca;
- consente di variare, modulare e personalizzare in qualsiasi momento i parametri di setting per la ricerca, la quantità e la qualità delle risposte, in relazione alle attese di maggiore o minore quantità di dati da rilevare (soglie di deformazione, etc.);
- consente di creare gli shape models in tempo reale secondo le esigenze dell'utente: visualizzate una o più immagini, egli può selezionare le regioni di interesse direttamente sulle immagini e modellarle secondo le sue necessità (fermarsi a un singolo grafo, comprendere più grafi, un'intera parola, etc.); un tool di rilevazione delle rumorosità dell'immagine gli consente di verificare i livelli di "sporcizia" che potranno in qualche modo compromettere l'affidabilità della ricerca (Figura 1);
- consente di personalizzare le ricerche salvando le regioni selezionate e utilizzate come modelli per la ricerca in apposita repository di sistema.

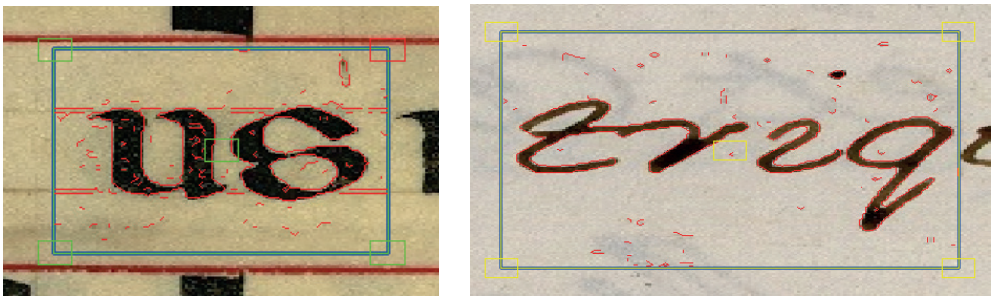


Fig 1
Creazione dei modelli

2. Risultati della sperimentazione

Sono stati eseguiti numerosi test per verificare le funzionalità del sistema e la sua validità. In particolare, sono state effettuate simulazioni prendendo in considerazione gli ambiti di ricerca paleografica. Si è simulato un approccio metodologico di ricerca tramite l'utilizzo di database digitali basato sul quarto paradigma, secondo il quale il paleografo sceglie di analizzare oggetti digitali contenuti in alcuni tra i più importanti (Biblioteca Apostolica Vaticana, British Library, Trinity College, BNF, John Rylands Library of Manchester, etc.), senza formulare preliminarmente una precisa ipotesi da cui partire, ma volendo valutare le possibili ipotesi deducibili dalle risposte alle query che andrà a fare.

Tra i vari test validi, si descrive in questa sede quello eseguito su due manoscritti greci (A e B) contenuti nel codice Sinaitico conservato presso la British Library, considerati opera di due diversi amanuensi, in quanto ha prodotto risultati a nostro parere di particolare interesse.

Il test è stato effettuato allo scopo di verificare se, lanciando query su un campione significativo di immagini scelte casualmente da entrambi i codici, i risultati consentissero di formulare ipotesi di ricerca diverse da quelle comunemente formulate dai paleografi.
Sono stati selezionati sulle immagini alcuni grafi secondo criterio casuale, utilizzati come modelli per le query poi lanciate sulle immagini. Sono state quindi analizzate le istanze restituite dalle diverse interrogazioni, delle quali si descrive di seguito, per necessità di sintesi, il solo risultato relativo al grafo "psi":

- positivi (grafi omografi al psi): 75%, di cui 50% nel ms A e 25% nel ms B
- falsi positivi: 25%, di cui 5% nel ms A e 20% nel ms B;

da attenta analisi dei falsi positivi sono risultati i seguenti elementi di interesse:

- grafi quasi omografi al "psi": 20%, di cui 5% in ms A e 15% in ms B, tutti riproducenti la lettera "phi", con tratto delle curvature perfettamente sovrapponibile alle corrispondenti del "psi";
- grafi approssimativamente omografi: 5%, tutti in ms B, tutti riproducenti la lettera "y", con alcuni tratti delle curvature sovrapponibili alle corrispondenti del "psi".

Le istanze positive possono essere di per sé sufficienti per elaborare un'ipotesi di ricerca finalizzata a dimostrare che, diversamente da quanto a oggi comunemente riconosciuto, i due manoscritti possano essere opera del medesimo amanuense. Ha invece costituito risultato del tutto inatteso la restituzione di un'ampia percentuale di grafi "falsi positivi" aventi tratti del tutto omografi rispetto al grafo scelto come modello. Un dato, questo, che renderebbe quasi inevitabile sia intraprendere ricerche più complesse e approfondite, anche "analogiche", finalizzate a verificare l'ipotesi di cui sopra, sia formulare altre ipotesi, quali:

- che i due manoscritti siano stati prodotti da mani diverse nello stesso scriptorium, nel quale però si utilizzava un canone estremamente rigido;
- che siano stati prodotti dalla stessa mano in tempi diversi e in scriptoria differenti, nei quali si utilizzava il medesimo canone ma con alcune leggere varianti;
- che il medesimo canone di particolare rigore sia stato utilizzato in un determinato scriptorium con leggerissime modifiche nel corso del tempo (secoli?), ovviamente da amanuensi diversi.

3. Conclusioni

In questo documento abbiamo descritto le caratteristiche di ICRPad M-Evo, un sistema brevettato di graphic matching per il riconoscimento digitale dei manoscritti che propone un nuovo approccio alla ricerca e al recupero delle informazioni di contenuto nelle biblioteche digitali. Questo approccio si basa sull'applicazione ai data humanities del quarto paradigma dei data science per lo sviluppo della conoscenza nel campo scientifico, che è alla base dell'informatica scientifica. Il processo di formazione si basa sull'algoritmo di corrispondenza descritto, che utilizza il riconoscimento della forma senza alcun processo di segmentazione. Si seleziona una regione appropriata che automaticamente crea il modello grafico da utilizzare per la ricerca all'interno di data base di immagini.

Riferimenti bibliografici

Adamek, T., O' Connor, E. N., & Smeaton, A. F. (2007). Word matching using single closed contours for indexing handwritten historical documents. In International Journal of Document Analysis and Recognition (IJDAR), Volume 9, Issue 2-4, (pp. 153-165).

Barbuti, N., & Caldarola, T. (2012). An innovative character recognition for ancient book and archival materials: A segmentation and self-learning based approach. In M. Agosti, F. Esposito, S. Ferilli, N. Ferro (Ed.), Communications in Computer and Information Science. Vol. 354: Digital Libraries and Archives, IRCDL 2012, Heidelberg: Springer, (pp. 261-270).

Bar-Yosef, I., Mokeichev, A., Kedem, K., & Dinstein, I. (2008). Adaptive shape prior for recognition and variational segmentation of degraded historical characters. Pattern Recognition, vol. 42(12), 3348-3354.

Bulacu M., & Schomaker L. (2007). Automatic Handwriting Identification on Medieval Documents. In ICIAP 2007: 14th International Conference on Image Analysis and Processing (pp. 279-284).

Cheriet, M. [et al.] (2009). Handwriting recognition research: Twenty years of achievement... and beyond, Pattern Recognition, vol. 42, 3131–3135.

Dalton, J., Davis, T., & van Schaik, S. (2007). Beyond Anonymity: Paleographic Analyses of the Dunhuang Manuscripts. Journal of the International Association of Tibetan Studies, No. 3, 1–23.

Fischer, A., Wüthrich, M., Liwicki, M., Frinken, L., Bunke, H., Viehhauser, G., & Stolz, M. (2009). Automatic Transcription of Handwritten Medieval Documents. In Proceedings of 15th International Conference on Virtual Systems and Multimedia (pp. 137-142).

Fischer, A., & Bunke, H. (2011). Character prototype selection for handwriting recognition in historical documents. In Proceedings of 19th European Signal Processing Conference, EUSIPCO (pp. 1435–1439).

Gordo, A., Llorenz, D., Marzal, A., Prat, F., & Vilar, J. M. (2008). State: A Multimodal Assisted Text-Transcription System for Ancient Documents. In DAS '08. Proceedings of 8th IAPR International Workshop On Document Analysis Systems (pp. 135-142).

Herzog R., Neumann B., & Solth A. (2011). Computer-based Stroke Extraction in Histori-

cal Manuscripts, Manuscript Cultures. Newsletter No. 3, (pp. 14-24).

Indermühle, E., Eichemberger-Liwicki, M., Bunke, H. (2008). Recognition of Handwritten Historical Documents: HMM-Adaptation vs. Writer Specific Training. In Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, Montreal, Quebec, Canada (pp. 186-191).

Krtolica, R. V., & Malitsky, S. (2012). Multifont Optical Character Recognition Using a Box Connectivity Approach (EP0649113A2). Retrieved May, 20, 2012 from http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en_EP

Le Bourgeois, F., & Emptoz, H. (2007). DEBORA: Digital AccEss to BOoks of the RenaissAnce. IJDAR, vol. 9(2-4), 193-221.

Le Bourgeois, F., & Emptoz, H. (2009). Towards an Omnilingual Word Retrieval System for Ancient Manuscripts. Pattern Recognition, vol. 42(9), 2089-2105.

Leydier, Y., Le Bourgeois, F., & Emptoz, H. (2005). Textual Indexation of Ancient Documents. In Proceedings of the 2005 ACM Symposium on Document Engineering (pp. 111-117).

Nel, E.-M., Preez, J. A., & Herbst, B. M. (2009). A Pseudo-skeletonization Algorithm for Static Handwritten Scripts. International Journal on Document Analysis and Recognition (IJDAR) 12, 47–62.

Rath, M. T., Manmatha, R.A., & Lavrenko, V. (2004). Search Engine for Historical Manuscript Images. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. (369-376).

Srihari, S., Huang, C., & Srinavasan, H. (2005). A Search Engine for Handwritten Documents. In Document Recognition and Retrieval XII, vol. 154, no. 3. (pp. 66-75).

Stokes, P. A. (2009). Computer-aided Palaeography, Present and Future, in M. Rehbein [et al.] (Eds.), Codicology and Palaeography in the Digital Age, Schriften des Instituts fur Dokumentologie und Editorik, Band 2, Norderstedt: Book on Demand GmbH.

Toselli, A. H., Romero, V., Pastor, M., & Vidal, E. (2010). Multimodal Interactive Transcription of Text Images. Pattern Recognition, vol. 43(5), 1814-1825.

Autori

Nicola Barbuti nicola.barbuti@uniba.it

Ricercatore Universitario Confermato in Archivistica, Bibliografia e Biblioteconomia presso il Dipartimento di Studi Umanistici (DiSUM) dell'Università degli Studi di Bari Aldo Moro. Svolge attività di ricerca e docenza in scienze biblioteconomiche e dell'informazione, digital cultural heritage, digital humanities. È Responsabile scientifico UNIBA nella Scuola a Rete Nazionale DiCultHer. È Coordinatore del Polo Apulian DiCultHer. È co-inventore del software ICRPad.

Stefano Ferilli stefano.ferilli@uniba.it

Professore Associato per il settore INF/01 presso il Dipartimento di Informatica dell'Università degli Studi di Bari Aldo Moro. Attualmente è Direttore del Centro Interdipartimentale di Logica ed Applicazioni. La sua attività scientifica si focalizza su temi inerenti l'acquisizione automatica di conoscenza espressa in formalismi simbolici, in particolare sui fondamenti logici ed algebrici dell'apprendimento automatico di concetti e sul confronto di descrizioni, elaborando modelli e metodi per la loro applicazione, fornendone realizzazioni ed applicazioni a domini del mondo reale. Collabora all'implementazione del software ICRPad.

Tommaso Caldarola t.caldarola@dabimus.com

Senior Software Architect esperto nella definizione e implementazione di procedure per il controllo di qualità per il system testing, per la scrittura di documentazione tecnica, per le modalità di bug trace e object management per una corretta gestione delle componenti sw finalizzata a facilitarne il riuso, gestione dei processi di configuration, patching & versioning management. È co-inventore del software ICRPad.

ISBN 978-88-905077-7-9