Consortium GARR

Conferenza GARR_18
Selected papers

DATA REVOLUTION

Cagliari, 3-5 ottobre 2018

Consortium GARR

Conferenza GARR 2018
Selected papers

DATA
REVOLUTION
Cagliari, 3-5 ottobre 2018

ISBN 978-88-905077-8-6
DOI 10.26314/GARR-Conf18-proceedings

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

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

Prima stampa: Aprile 2019
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 2018 sono disponibili all'indirizzo:
www.garr.it/conf18

Indice

Comitato di programma

Claudio Allocchio - GARR
Giuseppe Attardi - GARR
Annalisa Bonfiglio - CRS4
Paolo Budroni - Universität Wien
Simonetta Buttò - ICCU
Mauro Campanella - GARR
Stefano Cappa - IRCCS Fatebenefratelli
Donatella Castelli - CNR ISTI
Massimo Cocco - INGV
Antonio Di Lorenzo - ENEA
Gianni Fenu - Università degli Studi di Cagliari
Filippo Lanubile - Università degli Studi di Bari
Marcello Maggiora - CNR IEIIT e Politecnico di Torino
Alberto Masoni - INFN
Gelsomina Pappalardo - CNR
Andrea Quintiliani - ENEA
Federico Ruggieri - GARR
Riccardo Smareglia - INAF
Federica Tanlongo - GARR
Federico Zambelli - CNR e Elixir IIB

Tutte le presentazioni e maggiori informazioni sono disponibili sul sito dell'evento: www.garr.it/conf18

DOI: 10.26314/GARR-Conf18-proceedings-01

Il Geoportale Nazionale per ricerca e la tutela del patrimonio archeologico

Valeria Acconcia[1], Annalisa Falcone[1], Paola Ronzino[2]

[1] Istituto Centrale per l'Archeologia - MIBAC, [2] PIN, Polo Prato

Abstract. Il presente lavoro descrive un'attività congiunta promossa dall'Istituto Centrale per l'Archeologia (ICA), in collaborazione con un gruppo di ricercatori del VAST-LAB del PIN, Polo Universitario Città di Prato, e supportata dal Ministero per i Beni e le Attività Culturali (MiBAC), per lo sviluppo di un Geoportale nazionale per l'archeologia. L'obiettivo che questo lavoro si propone di raggiungere è quello di implementare un portale che divenga un hub per la ricerca archeologica, con lo scopo di raccogliere, rendere accessibile e diffondere la conoscenza del patrimonio archeologico italiano. In questa prospettiva, una delle principali attività per il design della piattaforma è stata la raccolta degli user requirements attraverso l'invio di un questionario ad un gruppo di esperti di dominio e ai funzionari delle soprintendenze. Il risultato ottenuto da questo sondaggio ci ha permesso di definire la metodologia da seguire per la raccolta dei dati esistenti e per la produzione dei nuovi dati da parte dei fornitori di contenuti, in modo da rendere la pubblicazione dei dati attraverso il geoportale una procedura standard e permettere l'interoperabilità dei dati spaziali.

Keywords. Web-GIS, research Infrastructures, dataset integration

Introduzione

Il progetto del Geoportale Nazionale per l'Archeologia (GNA), promosso dall'accordo stipulato tra l'Istituto Centrale per l'Archeologia (ICA), l'ICCU (MiBAC) e il PIN è finalizzato alla realizzazione di una piattaforma digitale on line che si configurerà come punto di accesso e di interscambio di dataset archeologici, per la ricerca e la conoscenza dei dati relativi al patrimonio archeologico sul territorio italiano.

L'esigenza di predisporre un sistema di consultazione in rete dei dati territoriali a livello nazionale, si colloca all'interno di una discussione a livello europeo che ha messo in luce una chiara esigenza di coordinamento nella gestione dei dati spaziali, riferiti soprattutto ai dati provenienti dalla ricerca archeologica (McKeague et al. 2012). In Italia, tale necessità è emersa in modo chiaro nel corso dell'ultimo decennio a seguito di una riflessione preliminare che ha avuto l'obiettivo di orientare e definire con chiarezza le strategie operative e le tecnologie informatiche da utilizzare. Il GNA rappresenta, infatti, l'esito di un percorso avviato dalle due commissioni paritetiche del 2007 e 2009. La prima (Carandini 2008), "Commissione Paritetica per la realizzazione del Sistema Informativo Archeologico delle città italiane e dei loro territori", ha avuto come risultato l'individuazione di strumenti condivisibili e la definizione di alcune precise linee di intervento per assicurare l'opportuna conoscenza necessaria alla tutela e alla valorizzazione del patrimonio archeologico. La seconda commissione (Azzena et al 2012), "Commissione Paritetica per lo sviluppo e

la redazione di un progetto per la realizzazione del sistema informativo territoriale del patrimonio archeologico italiano", ha proposto la realizzazione del Sistema Informativo Territoriale Archeologico Nazionale (SITAN) con l'obiettivo di definire una metodologia per l'acquisizione e l'elaborazione di dati territoriali sulla base di procedure stabilite in modo omogeneo, coinvolgendo università, uffici del Ministero e enti locali.

Al fine di realizzare il progetto del GNA, sono stati attivati diversi canali di collaborazione volti alla ingegnerizzazione di un sistema inclusivo non solo dei dati relativi agli interventi di tutela del MIBAC, ma anche dei risultati delle indagini sul campo condotte da università e altri enti di ricerca, in un'ottica di data-sharing e collaborazione tra Ministero, mondo accademico e altri soggetti che a diverso titolo operano sul campo, contribuendo alla conoscenza e alla protezione del patrimonio archeologico nazionale

1. La gestione dei dati spaziali in Italia

Un recente censimento delle risorse disponibili online, condotto dall'ICA, ha segnalato la presenza in Italia di un elevato numero di database online, WebGis e geoportali nati come risultato di diverse attività che hanno coinvolto il settore pubblico, privato e accademico. La diversità degli obiettivi per cui tali strumenti sono stati realizzati, ad es. archeologia preventiva, studio della distribuzione degli insediamenti, mappa dei vincoli, ecc., ha inevitabilmente dato vita a sistemi strutturati sulla base di modelli di dati differenti, precludendo la possibilità di effettuare delle ricerche integrate sui dati disponibili. Alcuni sistemi sono stati creati per far fronte alle criticità generate da eventi straordinari, come ad esempio il Geoportale del patrimonio culturale della regione Emilia-Romagna (Di Cocco 2014), nato all'indomani del terremoto del 2012, il cui obiettivo principale è stato quello di registrare e documentare gli edifici interessati dal terremoto, e successivamente aggiornato con la documentazione di tutti i beni archeologici, edifici con interesse storico, musei e archivi. Altri sistemi sono il risultato di progetti orientati verso una logica di apertura, condivisione e interoperabilità di dati, pur essendo stati sviluppati da soggetti con fini istituzionali diversi come ad esempio il Sistema Informativo Territoriale Archeologico di Roma (SITAR), nato con lo specifico obiettivo di supportare la tutela (Serlorenzi et al 2012) e il Progetto MAPPA: Metodologie Applicate alla Predittività del Potenziale Archeologico (Anichini et al 2012), sviluppato dal Laboratorio di Cultura Digitale dell'Università di Pisa a partire da quesiti integrati tra ricerca e tutela.

A livello generale, il sistema VIR (Vincoli in Rete) sviluppato dall'Istituto Centrale per il Restauro (ISCR), rappresenta al momento il progetto più esteso e complesso per quanto riguarda i dati territoriali in capo al Ministero: esso consente l'accesso per la consultazione e gestione degli atti di protezione del patrimonio culturale, e la localizzazione delle aree soggette a restrizioni ambientali da parte degli uffici del Ministero, utilizzando una piattaforma georeferenziata destinata ad accogliere anche i dati presenti nel sistema generale del catalogo SigecWEB, a sua volta sviluppato dall'ICCD. Altri sistemi sono stati creati con lo scopo di condividere dati di ricerca, come il WebGis Marmora Phrygiae, di recente reso disponibile online da CNR-IBAM, che è stato sviluppato utilizzando strumenti open-source con lo scopo di archiviare e condividere dati acquisiti durante il progetto di

ricerca (Di Giacomo et al 2018). Va infine sottolineato il grande numero di portali WebGis sviluppati a livello regionale come risultato di attività derivanti da disposizioni stabilite dal Codice di beni culturali, che prevede la collaborazione con le regioni nella definizione dei piani regionali, oltre a quelli su scala territoriale più ridotta, sviluppati da singoli comuni o consorzi di comuni.

È soprattutto per far fronte alla situazione frammentaria ed eterogenea dei vari sistemi disponibili in rete, la quale impedisce non solo l'interoperabilità dei dataset, ma anche la loro scoperta, che l'ICA, con il supporto tecnico di un gruppo di ricercatori del VAST-LAB sta lavorando al design e allo sviluppo del Geoportale Nazionale per l'Archeologia.

2. Linee di sviluppo del Geoportale Nazionale per l'Archeologia

Una delle prime attività svolte ai fini della progettazione del GNA è stato il sopracitato censimento delle risorse disponibili online. Parallelamente a questa analisi, il gruppo di lavoro (PIN-ICA) ha condotto una survey per la raccolta degli user requirements, sottoponendo un gruppo di esperti di dominio ad un questionario, basato sul metodo semplificato Cockburn (Cockburn 2000) per la raccolta degli user needs. Il questionario mirava a conoscere i motivi della consultazione di un ipotetico Geoportale, l'obiettivo della consultazione, l'ambito in cui si inquadra l'attività dell'utente, la tipologia dell'utente e suggerimenti per lo sviluppo della piattaforma. I risultati della survey, hanno riportato che l'archeologia preventiva risulta essere uno degli ambiti prioritari nella ricerca dei dati, con il 60% delle risposte, mentre l'analisi del paesaggio archeologico ha ricevuto il 25% delle risposte. Le attività connesse all'archeologia pubblica e alla ricerca accademica hanno interessato il 15% degli intervistati. Dall'analisi delle risorse disponibili e sulla base delle richieste e dei suggerimenti espressi dagli intervistati, il GNA sarà implementato, in una sua prima fase, come catalogo per l'aggregazione e il recupero di tutte le risorse e servizi disponibili online, seguendo il paradigma del portale ARIADNE, il quale sarà successivamente integrato con funzionalità che permetteranno di gestire i dati spaziali, come previsto nell'ambito del progetto ARIADNEplus (Niccolucci 2018). Il catalogo permetterà di scoprire i dataset disponibili online attraverso una ricerca a testo libero oppure attraverso una selezione su mappa o su barra temporale. I dataset che soddisfano la richiesta saranno poi consultati direttamente sul sito proprietario, sulla base delle relative condizioni di accesso. In questa prima fase, il portale raccoglierà e renderà fruibili, con diversi gradi di accessibilità attraverso il catalogo delle risorse, i dati relativi all'archeologia preventiva e quelli provenienti da interventi archeologici eseguiti dagli uffici del Ministero, dei dataset pubblicati dalle università e dalle autorità locali e dagli enti periferici, eterogenei per formazione e consultabilità.

3. L'archeologia preventiva e le concessioni di scavo

Nella seconda fase del progetto, il GNA permetterà non solo il reperimento dei webGIS e dei geoportali esistenti attraverso il catalogo delle risorse e dei servizi, ma anche l'archiviazione e la gestione di nuovi dataset. Il portale infatti potrà accogliere come sezione aggiuntiva, il collegamento a uno sportello online permanente, attualmente in fase di

elaborazione da parte dell'ICA, attraverso cui potranno essere presentate le richieste per il rilascio delle concessioni di scavo (come previsto dagli articoli 88 e 89 del d.lgs. 42/2004 e dalle circolari in materia), con l'obiettivo di snellire la procedura, garantendo trasparenza, parità di trattamento e tempi certi nella conclusione del procedimento. Oltre al rilascio delle concessioni, lo sportello sarà deputato alla ricezione della documentazione di fine scavo, che dovrà comprendere, tra le altre cose, la geolocalizzazione dell'area indagata che, in automatico, confluirà all'interno del Geoportale restituendo un'immagine in tempo reale delle indagini archeologiche in regime di concessione sul territorio nazionale.

Nel GNA l'integrazione delle informazioni geografiche all'interno dello stesso portale, sarà estesa a livello del singolo record. Questa funzione sarà resa disponibile adottando gli standard proposti dalla direttiva INSPIRE per condividere set di dati contenenti dati spaziali e permetterne l'accesso da parte di altre organizzazioni pubbliche. In particolare, si richiederà che i database spaziali che saranno integrati nel GNA dovranno essere pubblicati come Web Map Services (WMS), utilizzando lo standard ISO 19128 WMS 1.3.0. In alcuni casi sarà previsto il servizio di download attraverso i Web Feature Services (WFS). Questi servizi permetteranno la fruizione dei dati spaziali detenuti da diverse istituzioni a cui si potrà accedere in remoto (sul proprio ambiente di lavoro) oppure attraverso il sistema fornito dal GNA, visualizzando i vari dati geografici ed i relativi attributi sulla medesima mappa.

4. Linee guida e standard minimi di interoperabilità

In linea con gli obiettivi statutari dell'ICA, inoltre, si prevede che il GNA diventi uno strumento per definire e diffondere linee guida e standard cui adeguarsi per elaborare banche dati in formato open, operando una mediazione tra le esigenze di uniformità, necessarie alla creazione di un punto di accesso univoco alle informazioni territoriali in campo archeologico, e di indipendenza della ricerca. In questi termini, sarà necessario proporre un livello minimo di interoperabilità, che non confligga con gli obiettivi di ciascun soggetto partecipante rappresentando, anzi, la chiave per incoraggiare gli enti di ricerca, gli enti pubblici territoriali e le stesse strutture operative del MiBAC ad adeguarsi agli standard e a partecipare al GNA.

Dal GNA, inoltre, potranno essere attivate funzioni di download di documenti, sotto forma di pacchetti GIS pre-impostati, templates e micro-manuali operativi, in un'ottica di divulgazione di buone pratiche utili per l'omologazione nell'acquisizione e archiviazione del dato.

5. Conclusioni

Il presente contributo descrive le attività in corso presso l'Istituto Centrale di Archeologia, in collaborazione con un gruppo di ricercatori del VAST-LAB (PIN), per la creazione di un geoportale pensato come punto d'accesso e piattaforma di interscambio per molteplici fonti documentarie, in particolare i dataset archeologici con una componente geografica. Al momento, la situazione italiana è rappresentata da una forte mancanza di omogeneità, comprendendo da un lato soluzioni altamente avanzate (sviluppate sia a livello ministeriale

che universitario) che permettono la fruizione dei dati aperti, con servizi conformi alla normativa INSPIRE, e dall'altro lato, la presenza di numerosi WebGis e geoportali nati come risultato di singoli progetti e quindi non conformi a standard comuni. Questa situazione frammentaria e la predisposizione di molti istituti a contribuire con i propri dati ha fornito lo spunto all'ICA per sviluppare un'infrastruttura spaziale nazionale per l'integrazione di dati archeologici. L'adozione di standard ampiamente accettati, come quello del progetto ARIADNE e della legislazione INSPIRE, rappresenta un primo passo verso l'integrazione di dati spaziali non solo a livello nazionale ma anche a livello europeo, esplorando il potenziale per la pubblicazione, condivisione e riutilizzo di set di dati geospaziali archeologici.

Riferimenti bibliografici

McKeague P., Corns A, and Shaw R. (2012), Developing a spatial data infrastructure for archaeological and built heritage, International Journal of Spatial Data Infrastructures Research 7, 2012, pp. 16-37. http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/239

Carandini A, (2008), Archeologia Classica. Vedere il tempo antico con gli occhi del 2000, p.200, Torino.

Azzena G., Campana S., Carafa P., Gottarelli P., (2012), Il Sistema Informativo Territoriale Archeologico Nazionale – SITAN, SITAR. Sistema Informativo Territoriale Archeologico di Roma. Potenziale archeologico, pianificazione territoriale e rappresentazione pubblica dei dati. Atti del II Convegno (Roma Palazzo Massimo, 9 novembre 2011), Roma, pp. 41-45.

Di Cocco I., (2014), Il WebGIS dei beni culturali: dall'emergenza alla quotidianità", in R. Gaudioso (a cura di), Terreferme. Emilia 2012. Il patrimonio culturale oltre il sisma, Milano, pp. 43-46.

Serlorenzi M., Lamonaca F., Picciola S., Cordone C., (2012), Il Sistema Informativo Territoriale Archeologico di Roma: SITAR, Archeologia e Calcolatori, 23, pp. 31-50

Anichini F., Fabiani F., Gattiglia G., Gualandi M.L., (2012), ProgettoMAPPA. Metodologie Applicate alla Predittività del Potenziale Archeologico, I, Roma.

Di Giacomo G., Ismaelli T., Scardozzi G., (2018), Marmora Phrygiae. Il geodatabase delle cave di marmo e dei cantieri costruttivi di Hierapolis di Frigia: archeologia, archeometria e conservazione, Edizioni CNR-IBAM, ISBN 9788889375211

Cockburn A. (2000), Writing Effective Use Cases. Boston, USA: Addison-Wesley Professional.

Niccolucci F., (2018), Integrating the digital dimension into archaeological research: the ARIADNE project. Post-Classical Archaeologies, 8, pp. 63-74.

Autori

Valeria Acconcia - valeria.acconcia@beniculturali.it

Valeria Acconcia studied at Rome University "La Sapienza", specialized in Etruscology, worked at the University of Chieti-Pescara as post-doc and as professor. She is now official of the Ministry of Cultural Heritage (MiBAC), at the Central Institute for Archaeology. She deals with digitalization of National Public Heritage and with the procedures of protection; with standard guidelines of archaeological methodologies and with online publishing.

Annalisa Falcone - annalisa.falcone@beniculturali.it

Annalisa Falcone studied archaeology at Rome University "La Sapienza". She holds a specialization and a PhD in Roman Provinces Archaeology; since 2009 she is a member of the Italian Archaeological Mission at Elaiussa Sebaste in Turkey. She worked as official at the Museum Pole of Umbria and Museum Pole of Molise, where she has been director of the Archaeological Museum of Venafro; she is now official of the Ministry of Cultural Heritage (MIBAC), at the Central Institute for Archaeology. She deals with digitalization of National Public Heritage and with the procedures of protection; with standard guidelines of archaeological methodologies and with online publishing.

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, and in the ARIADNEplus project as Project Manager and in the activities related to archaeological and geographical datasets integration.

DOI: 10.26314/GARR-Conf18-proceedings-02

A (rock and) rolling ecosystem for next gen DaaS

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

Elettra Sincrotrone Trieste

Abstract. The requirements of large computing infrastructures are dictated by the applications that should be supported. Albeit this is a very general definition, that is what happened when Elettra designed and implemented its current ICT infrastructure serving the needs of beamline experiments. In practical terms, the network, computing, and storage requirements were defined by the specific needs of each beamline-laboratory of the institute and this lead to a complex ecosystem of different software, hardware, workflows and policies. This paper describes the past, present and future of this mixture of technology that for simplicity we refer to as "infrastructure". The next-generation is not only going to be defined by the institute's need but also from those proposed by a larger informal consortium of similar entities, formed by participating in multiple very large scale projects with several partners. The core and common part of these projects in relation to the infrastructure can be outlined as following: a common cloud-based ecosystem that support Data Analysis as a Service covering data acquisition & management, policies, HPC, storage, networking, remote access and all the related specific technologies. This common core sets a road-map for the ICT infrastructure of Elettra for the forthcoming years.

Keywords. Data Analysis, Distributed Storage, Big Data, CEPH, Kubernetes

1. Growing up a baby storage

The basic needs of any modern scientific experiment always include some flexible way of storing all the data produced by detectors, control devices and related instruments. This brings to consider at first the storage systems that will manage the collected data; the evolution of the particular solution now in service at Elettra is presented here by borrowing the story of a fictious human being.

1.1 Infancy

When the SOFA Storage System (Andrian et al. 2017) was born at Elettra, its parents were, of course, very proud of the newborn. At the same time, they were aware of the huge responsibility they had towards their baby, whose life, since the very first day, was expected long, brilliant and plenty of success. In fact, children are expected to grow up, fortify, become more clever and get as much notions and lessons as years go by. SOFA was no exception: weighting only 480TB (raw) at its first cry, it quickly doubled to 960TB trying hard to break the next barrier of the Petabyte.

1.2 Adolescence

Strength and resilience are two characteristics that every being should gain with time and effort. That's why the initial "factor 2" data replication policy evolved to a safer "factor 3".

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

This guarantees data integrity against hardware failures of high impact (like two complete cluster nodes put offline). Maturity also brought a new, more stable, release of CEPH, called Luminous, that superseded the initial simple and buggy Hammer. The upgrade of the FileStore approach of using the OSDs (relying on a basic filesystem like XFS on a disk partitioned as usual) to the more modern and efficient BlueStore (direct access to block device by CEPH) is something that requires virtually replacing every single disk in the cluster. This would have led to a lengthy and performance inefficient process. That is why it was decided that every new OSD in the cluster would be set up with BlueStore, while all the previous FileStore-defined OSD's would be converted with low priority during times when the storage cluster is not used intensively by the users.

1.3 Maturity

The Elettra 2.0 project (WWW 1) has been approved in 2017 and its financing secured for five years starting from 2018. The upgrade plan includes a new storage ring design that can be adapted to the existing Elettra storage ring tunnel, building and infrastructure. The new magnetic lattice will increase the brilliance and coherence of the X-ray beam by a factor of 20. The upgrade to Elettra 2.0 will enable new science and the development of new technologies. In particular, the science case supporting the Elettra storage ring upgrade represents a major step forward in synchrotron research, which complements and integrates the new possibilities offered by FERMI, the Free Electron Laser already in operation. As expected, this will have a direct impact on IT and will be at the same time a great challenge. The estimated data grow ratio for the next three years is currently around 4PB/year (Pugliese 2013). That's why the capacity expansion of SOFA has been designed as a rolling process: plan to add new storage nodes every year in order not to run out of storage for experimental data.

1.4 Memory

As time goes by, data will pile up in memory and its future use and need could not be easily predictable. SOFA is a disk-based, multi replicated object storage for fast read and write access. Storing huge amounts of "sleepy" data would be uneconomic to say the least. That's why a new extension of the system for "nearly online" data is under procurement. According to the new defined data retention policy, experimental data will be stored in the fast online system (CEPH) for three years. After that period it will be moved to a slower technology, still allowing users to retrieve it locally (filesystem paradigm) or even from remote sites (through the VUO interface). Tape cartridges have proved to be cheaper that disks per same data size, with the only side effects of longer data access times and the need of using different protocols and tools for the data interface. However, also these systems have evolved over time and now abstraction layers are available as "gateways" to data. This means that it is possible to create a virtual filesystem or even an object storage interface that doesn't care if the final destination is based on disks, flash memories, tapes or future technologies still to be released. The exact final solution for SOFA has not been selected yet. At the moment of this writing there are various candidates both hardware (tape

libraries) and software (abstraction layer). The tape libraries currently under analysis, all based on the LTO-8 standard and with an initial raw target of 5PB meant to be extended yearly, are:

- DELL EMC2 ML3
- Quantum Scalar i6
- Spectra Logic T950(v)
- Spectra Logic Spectra Stack
- IBM (model tbd)

The software abstraction layer is of no less importance. This need itself is nothing new: in the past years the GRID community, for example, has developed a complete solution based mainly on IBM GPFS, IBM TSM (now Spectrum Pro) and a custom-made software, GEMSS (Pier Paolo Ricci et al. 2012), whose main components are presented in figure 1. Even if this solution works perfectly for the purposes and contexts it was designed for, it is too demanding for a new isolated solution like ours. IBM licenses would probably not fit in the budget themselves and not all the features of this solution are necessary for our requirements.

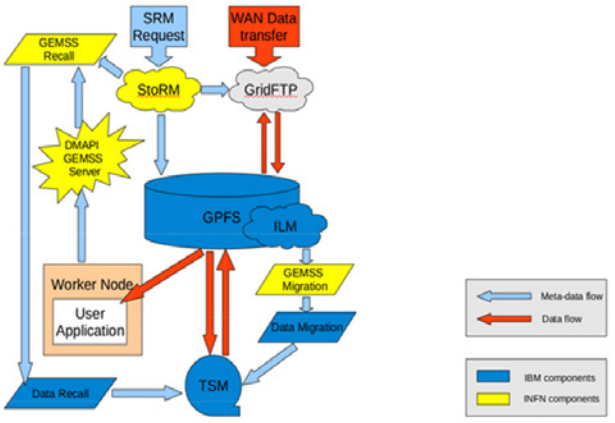


Fig. 1
Scheme of the GEMSS components

After a market analysis, which took into consideration both commercial and Open Source software, four candidates were selected:

- LizardFS – Open Source, (WWW 2) – is a distributed Filesystem that can operate both on disks and on tapes. Optional commercial subscriptions are available to secure the needed support against bugs and incidents, but are mandatory to get the development required to cope with the latest tape technologies (LTO-8, LTFS, etc.). Access to the data is possible by using a userspace mount on Linux systems (POSIX-compatible filesystem);
- OpenArchive – Open Source, (WWW 3) – is a mature and at a state-of-the-art archive solution that supports disks and tapes. Access to the data is made through protocols like NFS, CIFS and FTP;

- QStar Archive Storage Manager – Commercial, (WWW 4) – virtualises the access to storage devices of a wide spectrum (Disk Array, Object Storage, Tape Libraries, Optical Disk Libraries, WORM and Cloud), offering NFS and S3 interfaces;
- Spectra BlackPearl appliance – Commercial, (WWW 5) – which offers access only through S3 APIs and dedicated clients.

Again, the final choice hasn't been taken yet, but the most likely candidates so far are OpenArchive and Qstar ASM. The completion of this project extension to SOFA is planned to be done by first half of 2019.

2. Data Analysis As A Service (DaaS)

Data analysis is becoming a key factor for the success of modern experiments. This also relates to the huge growth in size of the dataset produced nowadays. Processing of Big Data needs capable HPC clusters and smart algorithms that can reduce data without losing information and processing time. Access to the data is also critical, since HPC requires high throughputs that can be guaranteed only inside a local area network (at Elettra, for instance, the beamline workstations connect to the cluster no faster than 10Gb/s, the in-cluster communication is now at 40Gb/s and will be upgraded to 100Gb/s). At the same time, experiments need to take place in various different facilities around the world: the data generated cannot be easily and effectively moved among laboratories for a later analyses. Remote access to a local set of data, tools, and processing is pursued by several projects Elettra is part of.

2.1 CALIPSOplus

This is the main project the activities of which are already ongoing. Its aim is to remove barriers for access to accelerator-based light sources in Europe and in the Middle East. In numbers it counts a 10M EUR budget, 19 partners, 3 associated partners and several observer members. It expects to provide more than 82,500 hours of trans-national access to 14 synchrotrons and 8 free electron lasers plus trans-national access tailor-made to SMEs. Its running time is 2017-2021 and its JRA2 work package (Demonstrator of a Photon Science Analysis Service) is what dictates the DaaS component for Elettra (WWW 6).

2.2 PaNOSC

This is a very large project of 20M EUR budget involving CERIC-ERIC, where Elettra is a member of. CERIC-ERIC serves as a single entry point to some of the leading national research infrastructures in eight European countries, enabling the delivery of innovative solutions to societal challenges in the fields of energy, health, food, cultural heritage and more. The PaNOSC Photon and Neutron community provides curated Open Data from a large variety of experiments and generates Petabytes of data every year. With specific focus on the European Open Science Cloud (WWW 7) the PaNOSC consortium aims at realising the EOSC concepts for finding and analysing scientific data produced from its facilities.

2.3 LEAPS

The League of European Accelerator-based Photon Sources (WWW 8) is a strategic consortium initiated by the Directors of the Synchrotron Radiation and Free Electron Laser (FEL) in Europe (figure 2). Its primary goal is to actively and constructively ensure and promote the quality and impact of the fundamental, applied and industrial research carried out at their respective facility to the greater benefit of European science and society. As expected computing is one of the main elements of the proposed strategy. The infrastructure described in this paper should be compatible for the first phase of LEAPS related projects.

Fig. 2
The LEAPS
Consortium

3. Storage and data analysis: a story of a marriage

The requirements dictated by the collaboration projects Elettra is part of, suggest that it's not possible any more to think of moving data between laboratories. It is then necessary to bring the Data Analysis tools closer to the data itself. Standardisation of data formats and programs is an ongoing process that every facility is pursuing, but that is not sufficient alone. To speed up this process it is clear that complete ready-to-use packages need to be developed, and an effective and easy way of deployment should be created. Technologies like virtualisation (KVM, ...), containerisation (LXC, Docker, ...), orchestration (Swarm, Kubernetes, ...) can now be thought as the building blocks of a new Data Analysis layer on top of existing HPC/Storage clusters. The blueprint presented in figure 3, for example, details the architecture under development in the CALIPSOplus JRA2 collaboration. When completed, a scientist will be able to connect from remote to the Data Analysis facilities of one of the collaboration members via a dedicated Access Portal, set up a computation environment by pulling ready-made containers from one of the repositories (Github-like

services), link the computational nodes (containers) to the Data Catalogue and Storage and finally orchestrate this on-demand computational cluster by way of Kubernetes. An important aspect of this scenario is that it will be performed by way of a self service DaaS infrastructure, with no need of intervention of any data or computing operator/scientist.

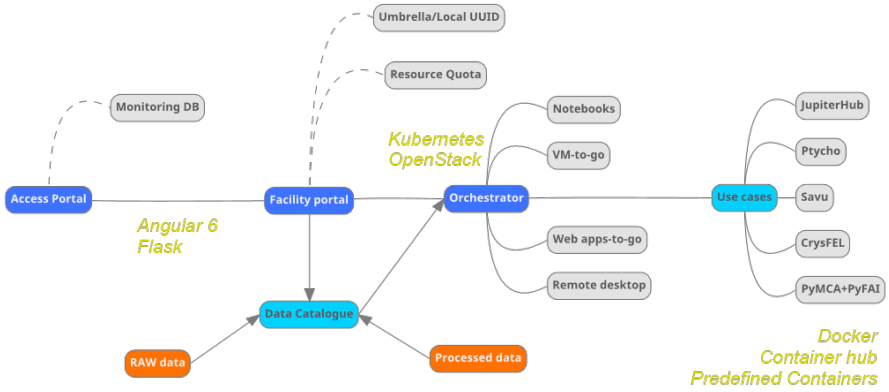


Fig. 3
CALIPSOplus JRA2
DaaS Blueprint

4. Conclusions

The ITC infrastructure may be seen as a complex ecosystem of software, hardware, workflows and policies. These are traditionally designed according to the requirements of the environment where they are going to be used for. In the case of synchrotron facilities like Elettra, they mostly depend on the needs of its beamlines and labs. There is though a special case that we report in this paper. Due to the participation in large international projects and consortiums, which are also considerable sources of funds, the future ICT must take into account their demands. Our ICT infrastructure will be influenced by the direct and indirect participation in CALIPSOplus, PaNOSC and LEAPS, among other projects of smaller size. At the same time there is an ongoing major upgrade project for the Elettra 2.0 new storage ring which will give us new IT-related challenges to face.

References

Ivan Andrian et al. 2017, Life (of Big Storage) in the Fast Lane, Proceedings of 2017 GARR Conference, pp. 117-122, DOI:10.26314/GARR-Conf17-proceedings-22

Roberto Pugliese 2013, "Relazione Tecnica Accompagnatoria Progetto DIESEL 2.0", Internal Elettra document.

Pier Paolo Ricci et al. 2012, J. Phys.: Conf. Ser. 396 042051.

WWW 1, http://www.elettra.eu/lightsources/elettra/elettra-2-0.html

WWW 2, https://lizardfs.com

WWW 3, https://www.openarchive.net

WWW 4, http://www.qstar.com/index.php/archive-manager/

WWW 5, https://spectralogic.com/products/blackpearl/

WWW 6, http://www.calipsoplus.eu

WWW 7, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

WWW 8, https://www.leaps-initiative.eu

Authors

Ivan Andrian - ivan.andrian@elettra.eu

Ivan Andrian (degree in Information and Computer Science at the University of Udine) started his career as researcher in System Administration and Network Engineering at Elettra Sincrotrone Trieste. Over the years, where he also worked as developer of mixed web and mobile systems, he specialised in Linux distributed systems. He is currently Chair of JACoW.org, an international collaboration project for the Open Access to conference proceedings in the field of particle accelerators.

George Kourousias - george.kourousias@elettra.eu

George Kourousias (PhD) is a researcher in Elettra Sincrotrone Trieste, Scientific Computing team. His research is in X-ray microscopy, X-ray Florescence, CDI/Ptychography, and Data Science including Neural Networks. With more than 40 publications in international peer reviewed journals and a Patent in image compression, he leads AI3, an Elettra project funding, and he is co-supervising PhD students on Imaging and AI. George is in the founding team of Singularity University Trieste Chapter.

Iztok Gregori - iztok.gregori@elettra.eu

Iztok Gregori is a Linux System Administrator in the Elettra ICT Systems and Services team. He has years of experience in High Performance Computing, Virtualization Technologies and Storage infrastructures. He supervises the Elettra Scientific Storage and its computing resources.

Massimo Del Bianco - massimo.delbianco@elettra.eu

Massimo Del Bianco is a Network Engineer taking care of the network at Elettra Sincrotrone Trieste, where he participates in the ICT Systems and Services team as Network and Systems Administrator, also managing the peripheral firewall; he collaborates in managing the virtualisation and storage clusters with particular attention to high bandwith networks optimisation.

Roberto Passuello - roberto.passuello@elettra.eu

Roberto's been a network – and UNIX – System Administrator employed in research institutions since 1996. He's always been a Linux enthusiast and is currently Head of ICT Systems and Services at Elettra Sincrotrone Trieste. He spends his spare time practicing yoga, staring at the Universe and looking after his three kids.

DOI: 10.26314/GARR-Conf18-proceedings-03

Un framework semantico per la ricerca e l'esecuzione di codice sorgente

Mattia Atzeni, Maurizio Atzori

Università degli Studi di Cagliari

Abstract. Questo articolo introduce un framework semantico per la ricerca automatica e la conseguente esecuzione di codice sorgente open source. A tal fine, abbiamo recentemente introdotto CodeOntology, un approccio ideato per applicare lo stack tecnologico fornito dal Semantic Web, nel dominio dell'ingegneria del software. In questo modo, l'informazione estratta dal codice sorgente può essere facilmente collegata con risorse esterne, in maniera da permettere interessanti analisi e ricerche semantiche che sarebbero altrimenti impossibili. Inoltre, proponiamo un approccio algoritmico che si basa su CodeOntology, al fine di selezionare un insieme di metodi rilevanti, che vengono successivamente combinati per tradurre una specifica, espressa in linguaggio naturale, in codice sorgente basato sul paradigma orientato agli oggetti

Keywords. Comprensione del Linguaggio Naturale, Semantic Web, Question Answering

Introduzione

Nonostante il bisogno di crescente automazione da parte di una vasta gamma di utenti differenti (Atzeni & Atzori, 2018c), la libera disponibilità online di un'enorme quantità di codice sorgente esibisce ancora un considerevole, ma al tempo stesso inespresso, potenziale per lo sviluppo di tecniche di intelligenza artificiale, finalizzate a sfruttare questa notevole quantità di informazione. Pertanto, il lavoro descritto in questo articolo ambisce a risolvere il problema di fornire un framework per la ricerca e l'esecuzione di componenti software, tramite l'introduzione di approcci semantici che permettano di tradurre complesse interrogazioni e comandi espressi in linguaggio naturale in un appropriato codice sorgente.

Questo problema di ricerca può essere scomposto in due sotto-problemi principali: fornire una rappresentazione semantica dei sistemi software e dei linguaggi di programmazione orientati agli oggetti ed utilizzare la base di conoscenza risultante, per rispondere in maniera automatica ad interrogazioni espresse in linguaggio naturale. Il primo sotto-problema può essere risolto tramite l'uso di CodeOntology (Atzeni & Atzori, 2017a), una risorsa che consiste di tre contributi principali:

- un'ontologia OWL 2 che rappresenta il dominio dei linguaggi di programmazione orientati agli oggetti;
- un parser che permette di serializzare codice sorgente in triple RDF;
- un dataset contenente circa 2.5 milioni di triple RDF, generate tramite l'applicazione del parser sul codice sorgente del progetto OpenJDK 8.

Ciò permette di eseguire interrogazioni espresse nel linguaggio SPARQL per scopi diffe-

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

renti, quali l'analisi statica del codice sorgente, il rilevamento di design pattern ed anti-pattern, la ricerca semantica di codice (Atzeni & Atzori, 2017b).

Seguendo questa linea di ricerca, il presente articolo definisce un approccio semantico (Atzeni & Atzori, 2018a) che sfrutta la conoscenza estratta da CodeOntology per eseguire comandi in linguaggio naturale e fornire una risposta a domande complesse che richiedono la generazione di codice ad hoc.

1. Descrizione del sistema

Questa sezione descrive un approccio semantico per la traduzione in codice di comandi espressi in linguaggio naturale (Atzeni & Atzori, 2018b). Il sistema si basa su CodeOntology per la ricerca di metodi definiti in OpenJDK, tali che la loro firma sia compatibile con i valori specificati dall'utente. I risultati restituiti da questa query sono successivamente ordinati per selezionare il candidato che meglio corrisponde alla specifica in linguaggio naturale. L'ordinamento dei candidati selezionati si basa sia su misure semantiche che su indicatori sintattici, come spiegato meglio nella Sezione 2. Data una descrizione in linguaggio naturale, il sistema esegue inizialmente il parsing delle dipendenze. (Manning et al. 2014), come mostrato nella Figura 1.

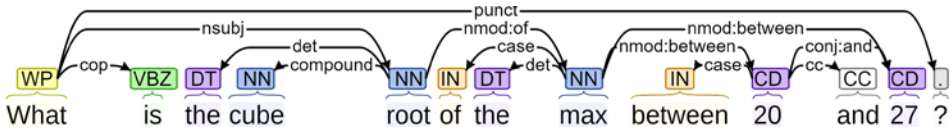


Fig. 1
Risultato del parsing delle dipendenze per una domanda ricevuta in input

Il grafo risultante viene trasformato in un albero, tale che l'insieme dei nodi N possa essere partizionato in due sottoinsiemi $\mathcal{L}$ e $\mathcal{M}$, dove:

- $\mathcal{L}$ è il sottoinsieme corrispondente agli argomenti letterali;
- $\mathcal{M}$ contiene i nodi che corrispondono ad espressioni che denotano l'invocazione di un metodo;
- ogni nodo $i \in \mathcal{L}$ è una foglia;
- $N = \mathcal{L} \cup \mathcal{M}$ e $\mathcal{L} \cap \mathcal{M} = \emptyset$.

Ogni nodo $i \in \mathcal{M}$ è etichettato con un ranking di metodi $\mathcal{R}_i$, cioè una sequenza

$$(m_1, s_1),\ldots,(m_n,s_n)$$

tale che:

- m_i : Method per ogni $i = 1\ldots n$
- $s_i \in [0,1]$ per ogni $i = 1\ldots n$
- $i < j \Rightarrow s_i \geq s_j$

I metodi candidati sono selezionati tramite l'esecuzione di interrogazioni su CodeOntology, basate sul tipo dei parametri e sul tipo di ritorno atteso. Successivamente, ad ogni metodo viene associato uno score reale, usando le features descritte nella Sezione 2.

A questo punto, si rende necessario risolvere un problema di programmazione lineare intera, che prevede di massimizzare il seguente obiettivo, dove $x_{ij}=1$ se il j-esimo metodo

del ranking $\mathcal{R}_i$ è selezionato, ed $x_{ij} = 0$ altrimenti:

$$\sum_{i \in \mathcal{M}} \sum_{(m_{ij}, s_{ij}) \in \mathcal{R}_i} x_{ij} \cdot s_{ij}$$

con i seguenti vincoli:

- $x_{ij} \in \{0,1\}$;
- $\sum_j x_{ij} = 1$, per ogni $i \in \mathcal{M}, 1 \leq j \leq |\mathcal{R}_i|$;
- la combinazione dei metodi selezionati può essere compilata.

Per migliorare ulteriormente l'approccio algoritmico descritto, si applica un metodo euristico che agisce sulla struttura dell'albero e adotta una ricerca greedy per massimizzare la seguente funzione:

$$z(\mathcal{T}_k) = \frac{1}{|\mathcal{M}_k|} \cdot \sum_{i \in \mathcal{M}_k} \sum_{(m_{ij}, s_{ij}) \in \mathcal{R}_i^k} x_{ij} \cdot s_{ij} - \lambda \cdot NTED(\mathcal{T}_k, \mathcal{T}_o),$$

dove $\mathcal{T}_k$ è l'albero all'iterazione k, $\mathcal{M}_k$ è l'insieme dei nodi non letterali in $\mathcal{T}_k$ e $\mathcal{R}^k{}_i$ è il ranking associato al nodo i in $\mathcal{T}_k$. Inoltre, λ è una costante ed $NTED(\mathcal{T}_k, \mathcal{T}_0)$ è la tree edit distance normalizzata tra $\mathcal{T}_k$ e $\mathcal{T}_0$.

2. Esperimenti

L'ordinamento dei metodi Java è stato valutato tramite un dataset contenente domande estratte da StackOverflow. Sono state sperimentate diverse combinazioni delle seguenti misure sintattiche e semantiche:

- similarità di Levenshtein normalizzata, calcolata tra il nome del metodo e la specifica in linguaggio naturale;
- n-gram overlap con il nome della classe dichiarante;
- n-gram overlap con il commento relativo al metodo considerato;
- similarità del coseno tra il vettore medio (Mikolov et al. 2013) associato al commento del metodo ed il vettore medio associato alla descrizione in linguaggio naturale;
- frazione dei link a DBpedia condivisi tra il commento del metodo e il comando in linguaggio naturale, annotati tramite l'uso di TagMe (Ferragina & Scaiella, 2012).

La combinazione di queste features permette di raggiungere una Mean Average Precision pari a 0.92 sul dataset menzionato in precedenza. L'approccio è stato ulteriormente valutato su un benchmark contenente 120 domande relative all'esecuzione di espressioni matematiche e manipolazione di stringhe. É possibile applicare una soglia $t \in [0,1]$ sul valore obiettivo, in modo da rilevare le domande che non possono essere elaborate correttamente dal sistema. La Tabella 1 confronta il nostro approccio con i risultati ottenuti da WolframAlpha.

3. Conclusioni

Questo articolo ha introdotto un approccio che fa uso di CodeOntology, al fine di suppor-

Tab. 1
Risultati sperimentali

	CodeOntology	WolframAlpha
Numero di domande	120	120
Domande elaborate	116	108
Risposte corrette	109	98
Precisione (globale)	0.91	0.82
Precisione (domande elaborate)	0.94	0.91

tare la ricerca e l'esecuzione di metodi, tramite semplici interrogazioni espresse in linguaggio naturale. Gli esperimenti sono stati condotti su OpenJDK, ma l'approccio è generale e può essere applicato per qualunque linguaggio di programmazione.

Riferimenti bibliografici

M. Atzeni, M. Atzori (2018a), What is the Cube Root of 27? Question Answering over CodeOntology, The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, California, USA.

M. Atzeni, M. Atzori (2018b), Translating Natural Language to Code: an Unsupervised Ontology-Based Approach, 2018 IEEE International Conference on Artificial Intelligence and Knowledge Engineering, Laguna Hills, California, USA.

M. Atzeni, M. Atzori (2018c), Towards Semantic Approaches for General-Purpose End-User Development, 2nd IEEE International Conference on Robotic Computing, IRC 2018, Laguna Hills, California, USA.

M. Atzeni, M. Atzori (2017a), CodeOntology: RDF-ization of Source Code, The Semantic Web – ISWC 2017: 16th International Semantic Web Conference, ISWC 2017, Vienna, Austria, pp. 20–28.

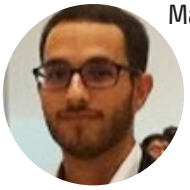
M. Atzeni, M. Atzori (2017b), CodeOntology: Querying Source Code in a Semantic Framework, 16th International Semantic Web Conference (Posters & Demo), Vienna, Austria.

P. Ferragina, U. Scaiella (2012), Fast and accurate annotation of short texts with Wikipedia pages., IEEE Software, 29(1), pp. 70–75.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, D. McClosky (2014), The Stanford CoreNLP natural language processing toolkit, Association for Computational Linguistics (ACL), System Demonstrations, pp. 55–60.

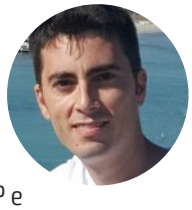
T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean (2013), Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, Curran Associates, pp. 3111–3119

Autori



Mattia Atzeni - m.atzeni38@studenti.unica.it

Mattia Atzeni è studente di dottorato presso l'Università degli Studi di Cagliari. Ha conseguito la Laurea Magistrale in Informatica presso la stessa università, ad Aprile 2018. I suoi principali interessi di ricerca includono la comprensione del linguaggio naturale, l'intelligenza artificiale e il machine learning.



Maurizio Atzori - atzori@unica.it

Maurizio Atzori è Professore Associato presso l'Università degli Studi di Cagliari. Ha conseguito il dottorato presso l'Università di Pisa nel 2006 ed è stato membro del KDD Lab presso l'Istituto di Scienza e Tecnologia dell'Informazione del CNR. È stato visiting researcher presso la Purdue University (Indiana, USA), la Sabanci University (Istanbul, Turchia) e presso la UCLA (Los Angeles, USA). I suoi principali interessi di ricerca includono AI, NLP e grafi di conoscenza.

DOI: 10.26314/GARR-Conf18-proceedings-04

La prima infrastruttura di data management per le nanoscienze

Rossella Aversa, Stefano Cozzini

CNR-IOM Istituto Officina dei Materiali

Abstract. In questo articolo presentiamo la prima infrastruttura di data management per le nanoscienze. L'infrastruttura, denominata IDRP (Information and Data management Repository Platform), è stata sviluppata dal CNR-IOM (Istituto di Officina dei Materiali) di Trieste all'interno di NFFA-EUROPE (Nanoscience Foundries & Fine Analysis). Questo progetto europeo Horizon 2020, coordinato dal CNR-IOM, coinvolge 20 partner ed ha lo scopo di fornire una infrastruttura di ricerca distribuita per la comunità di nanoscienze. La IDRP è nata con l'obiettivo di gestire ed archiviare in maniera FAIR la grande varietà di dati generati dalla strumentazione NFFA-EUROPE offrendo un accesso standardizzato.

Keywords. Data Repository, Metadati, Nanoscienze, SEM, Reti neurali

Introduzione

I dati scientifici, generati da esperimenti o da simulazioni numeriche, vengono prodotti ad un ritmo sempre maggiore. I concetti di riproducibilità, di interoperabilità, così come di open data, stanno assumendo un ruolo centrale nella ricerca. In effetti molti campi di ricerca dipendono quasi interamente dalla disponibilità e dalla condivisione di dati globali forniti tramite data repository aperti. I principi FAIR (Wilkinson et al. 2016) offrono linee guida per rendere i dati reperibili (Findable), accessibili (Accessible), interoperabili (Interoperable) e riutilizzabili (Reusable). Riconoscendo l'importanza della condivisione e del riutilizzo di dati scientificamente validi, il progetto NFFA-EUROPE (www.nffa.eu) ha implementato la prima piattaforma di data management, denominata IDRP, per la comunità di nanoscienze. L'obiettivo principale della IDRP è quello di automatizzare la raccolta di dati scientifici provenienti dai diversi strumenti presenti nelle strutture europee che partecipano al progetto, identificando i metadati corretti per consentire loro di essere ricercabili in conformità ai principi FAIR. Come primo esempio, ci siamo concentrati sui dati del microscopio a scansione elettronica (SEM). Questo è uno strumento estremamente versatile che viene utilizzato quotidianamente nelle nanoscienze e nelle nanotecnologie per esplorare la struttura dei materiali con risoluzione spaziale fino a 1 nanometro. Gli scienziati che lavorano con tecniche di microscopia elettronica sono particolarmente interessati a strumenti in grado di identificare e riconoscere automaticamente alcune caratteristiche specifiche all'interno delle immagini SEM. A tale scopo abbiamo cercato il modo di applicare algoritmi di classificazione di immagini nel campo delle nanoscienze.

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

1. L'architettura

L'architettura del prototipo è mostrata in Fig.1. La IDRP è stata pensata come archivio di metadati relativi ai dati raccolti negli esperimenti fatti dagli utenti NFFA e salvati su storage/repository locali. Questo approccio permette di non interferire con la modalità di acquisizione dati presso le facility. Servizi di gestione dati come quello descritto in questo articolo permettono la registrazione automatica di tali metadati all'interno della IDRP.

L'infrastruttura è stata sviluppata sulla cloud OpenStack del CNR-IOM. I servizi sono configurati come macchine virtuali indipendenti l'una dall'altra. Le istanze più importanti sono il portale di accesso (portal.nffa.eu) e la IDRP per la gestione dei metadati (idrp.nffa.eu). Ci sono poi una istanza dedicata allo storage locale (datashare.iom.cnr.it) ed una per il servizio di classificazione delle immagini (sem-classifier.nffa.eu). Una volta effettuato il login nel portale di accesso, appare il link alla IDRP, dove si registrano e si gestiscono i metadati relativi ai dati che si trovano negli storage locali. Questi ultimi sono molto diversificati nei vari istituti; al CNR-IOM abbiamo adottato un sistema di data management basato su NextCloud. Abbiamo poi iniziato ad aggiungere dei servizi di analisi dati online, in aggiunta a quelli locali. Il primo ad essere stato sviluppato è un tool di classificazione di immagini SEM, che annota le immagini e aggiunge questa annotazione come metadato, oltre a quelli strumentali già presenti.

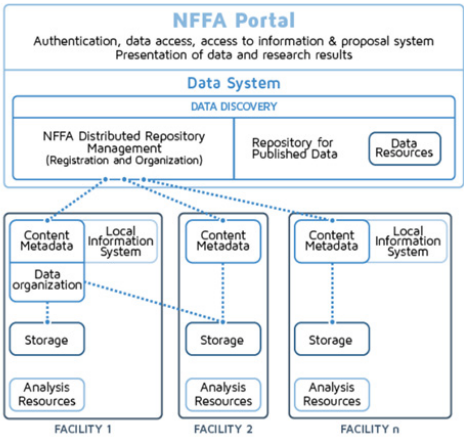


Fig. 1
Architettura della infrastruttura. Dall'alto verso il basso: attraverso il portale NFFA si accede alla IDRP, dove vengono registrati i metadati relativi ai dati prodotti alle varie facility, che vengono salvati negli storage locali.

2. Classificazione di immagini SEM

La classificazione, il riconoscimento di immagini, ed altri algoritmi basati su deep learning, ampiamente applicati in molte aree, non sono ancora sfruttati appieno in nanoscienze. Queste tecniche possono essere uno strumento potente, in particolare quando all'analisi scientifica è associata la gestione dei dati. Una rete neurale per classificare le immagini SEM offre molti vantaggi ai ricercatori in nanoscienze: le annotazioni automatiche, che evitano la necessità di classificare manualmente ciascuna immagine registrata; un database che consenta agli scienziati di trovare una categoria specifica di immagini SEM; la possibilità di adattare la rete neurale per svolgere compiti specifici relativi all'elaborazione dell'immagine, ad esempio quantificare la frazione di nanotubi coerentemente allineati

nelle immagini SEM (Modarres et al. 2017).

2.1 Creazione del dataset

Un set di 18577 immagini SEM è stato manualmente annotato e convalidato da un gruppo di scienziati, i quali hanno convenuto su una suddivisione in 10 categorie, per formare il primo dataset di immagini SEM classificate, disponibile pubblicamente (Aversa et al. 2018). Le categorie sono state stabilite tenendo conto delle caratteristiche visive delle immagini piuttosto che di classi astratte relative a nozioni specifiche. Un'immagine rappresentativa per ciascuna delle categorie scelte è mostrata in Fig.2.

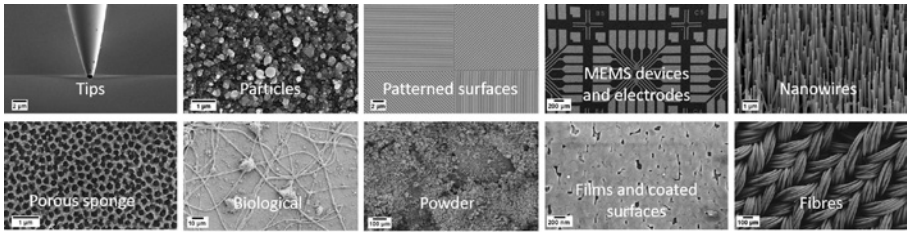


Fig. 2
Categorie del dataset SEM

2.2 Training della rete neurale

Il dataset annotato è stato usato per eseguire il training delle più note reti neurali convoluzionali (Aversa et al. 2019); l'architettura adottata dopo aver confrontato i risultati di accuratezza è Inception-v3 (Szegedy et al. 2015).

2.3 Classificatore online

Il risultato finale è un sito pubblico (sem-classifier.nffa.eu), la cui pagina iniziale è mostrata in Fig.3, dove l'utente carica una o più immagini SEM, la rete neurale fa inferenza e mostra il risultato più probabile. Cliccando su ciascuna immagine si possono vedere i dettagli dell'inferenza; si può anche decidere di selezionare una diversa categoria o di inserirne una nuova oltre alle 10 presenti al momento, che non sono rappresentative di tutto ciò che viene visto con il SEM, ma sono basate sulle immagini a nostra disposizione.

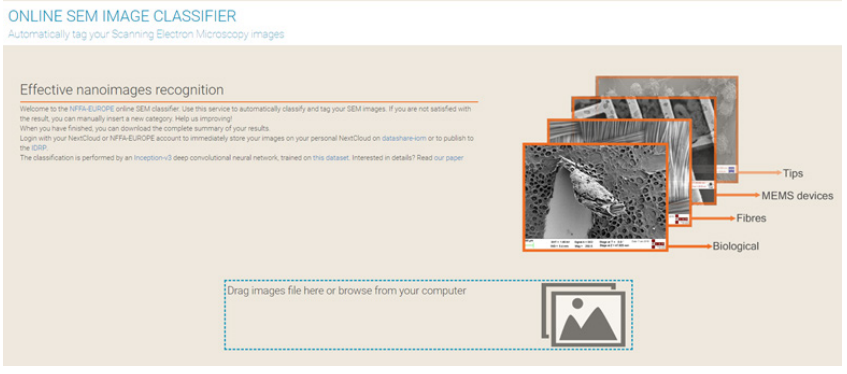


Fig. 3
Pagina iniziale del sito di classificazione di immagini SEM

4. Conclusioni

Abbiamo sviluppato un'infrastruttura per gestire i dati di nanoscienze, ed un sito di analisi online per le immagini SEM, dal quale l'utente può attivare il caricamento nello storage locale. Tutti i metadati vengono automaticamente estratti dalle immagini e possono essere registrati sulla IDRP, dove diventano ricercabili e quindi riutilizzabili dalla comunità scientifica. In futuro verranno aggiunte nuove reti neurali e verrà esteso il dataset, sia in termini di numero di immagini che di categorie.

Riferimenti bibliografici

Aversa R., Modarres M.H., Cozzini S., Ciancio R., Chiusole A. (2018), The first annotated set of scanning electron microscopy images for nanoscience, Scientific Data 5 (180172)

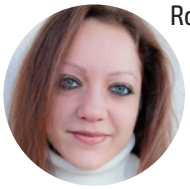
Aversa R., De Nobili C., Cozzini S., (2019) Deep learning for Nanoscience SEM image recognition, in prep.

Modarres M.H., Aversa R., Cozzini S., Ciancio R., Leto A., Brandino G.P. (2017), Neural Network for Nanoscience Scanning Electron Microscope Image Recognition, Scientific Reports 7 (13282)

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2015), Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567v3

Wilkinson M.D., et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (160018)

Autori



Rossella Aversa - aversa@iom.cnr.it

Rossella Aversa si è laureata nel 2011 in Astrofisica e Cosmologia all'università di Bologna e nel 2015 ha ottenuto il dottorato in Astrofisica alla SISSA di Trieste. Ha poi frequentato il Master in HPC a Trieste, diplomandosi nel 2016. Da allora è assegnista di ricerca al CNR-IOM di Trieste e lavora per il progetto NFFA-EUROPE.

Stefano Cozzini - cozzini@iom.cnr.it

Stefano Cozzini è tecnologo presso il CNR-IOM con oltre 20 anni di esperienza nelle gestione di infrastrutture IT per HPC e data management. Ha gestito diversi progetti di ricerca a livello internazionale e sta coordinando le attività di data management dei progetti NFFA-EUROPE ed EUSMI. È inoltre coinvolto nel master in HPC e nella laurea magistrale in "Data Science and Scientific Computing", entrambi promossi da diversi enti di Ricerca a Trieste, tra cui CNR-IOM.

DOI: 10.26314/GARR-Conf18-proceedings-05

Alla ricerca dell'arca perduta. Ovvero: dov'è il digital cultural heritage?

Nicola Barbuti[1], Stefano Ferilli[2]

[1] Università degli Studi di Bari Aldo Moro, Dipartimento di Studi Umanistici
[2] Università degli Studi di Bari Aldo Moro, Dipartimento di Informatica

Abstract. L'identificazione del digital cultural heritage (DCH) asserita nell'art. 2 delle Conclusioni del Consiglio dell'UE del 21 maggio 2014 sul Patrimonio culturale come risorsa strategica per un'Europa sostenibile (2014 / C 183/08) rende indispensabile ripensare i processi di digitalizzazione e di co-creazione digitale, al fine di individuare approcci e metodi che consentano di individuare cosa e quanto possa essere riconoscibile come patrimonio culturale tra le risorse digitali create fino a oggi e in produzione.
Questo paper intende proporre un approccio metodologico innovativo ai processi di digitalizzazione e di co-creazione delle entità digitali che, garantendone fin dalla progettazione la conservazione a lungo termine, conferisca loro il ruolo e la funzione di memoria dell'Evo Digitale contemporaneo e, in tal modo, li caratterizzi quali potenziale DCH.

Keywords. Digital Cultural Heritage, Digital Libraries, R[4], Metadati descrittivi.

Introduzione

L'art. 2 delle Conclusioni del Consiglio dell'UE del 21 maggio 2014 sul "Patrimonio culturale come risorsa strategica per un'Europa sostenibile (2014 / C 183/08)" recita:
"Cultural heritage consists of the resources inherited from the past in all forms and aspects - tangible, intangible and digital (born digital and digitized) …".
Tale identificazione rende quanto mai urgente e inderogabile ripensare i processi di digitalizzazione e di co-creazione digitale, al fine di individuare approcci e metodi che consentano di identificare cosa e quanto possa essere definito Digital Cultural Heritage (DCH) tra le risorse digitali create fino a oggi e in produzione.
In questo paper si propone un approccio metodologico innovativo ai processi di digitalizzazione e di co-creazione delle entità digitali che, garantendone fin dalla progettazione la conservazione a lungo termine, conferisca loro il ruolo e la funzione di memoria dell'Evo Digitale contemporaneo e, in tal modo, le caratterizzi quali potenziale DCH.

1. Alcune questioni per nulla secondarie

La nuova definizione di DCH data nelle Considerazioni solleva numerose domande cui rispondere, tra le quali:

- quali entità digitali possiamo definire risorse culturali, e secondo quali parametri?
- quante entità digitali possiamo identificare come DCH tra quelle oggi esistenti?
- quante entità culturali digitali sopravvivono tra quelle prodotte in passato, e andando a

ritroso qual è l'arco cronologico entro cui le si può considerare tali?
- come le riconosciamo e quali aspetti consideriamo? I processi? I risultati prodotti? Entrambi?
- le entità culturali digitali devono aver cessato le funzioni per cui sono nate, o possiamo considerare tali anche risorse oggi ancora fruibili secondo le funzioni per cui sono attualmente utilizzate?
- quali figure professionali saranno abilitate a valutare il DCH, e di quali conoscenze/competenze/abilità dovranno essere provviste?

Oltre le sopra elencate, molte altre criticità necessitano di risposte urgenti e scientificamente attendibili, in tempi peraltro stretti.

Diverse entità digitali oggi esistenti potrebbero essere classificate come DCH, ma si tratterebbe di una valutazione arbitraria in quanto non rispondente ad alcuna sistematizzazione.

2. Verso la definizione del DCH

Analizzando le entità digitali oggi esistenti inquadrate in ambito culturale, ne abbiamo dedotto una prima classificazione nelle seguenti tre macro-categorie, che nell'insieme potrebbero essere rappresentative del DCH:

- Born Digital Heritage: entità native digitali che registrano nei contenuti processi, metodi e tecniche digitali rappresentativi della co-creazione digitale delle comunità contemporanee, da salvaguardare, riusare e preservare nel tempo quale potenziale memoria storica e fonte di conoscenza per le generazioni future;
- Digital FOR Cultural Heritage: processi, metodi e tecniche della digitalizzazione finalizzata alla co-creazione di entità culturali digitali riproducenti nei contenuti e nei metadati descrittivi entità culturali analogiche tangibili e intangibili (digital libraries, musei virtuali, database demo-etno-antropologici, etc.);
- Digital AS Cultural Heritage: entità digitali derivate dalla digitalizzazione e dalla dematerializzazione che registrano nei contenuti approcci culturali, processi, metodi e tecniche rappresentativi della loro evoluzione, da salvaguardare, riusare e preservare nel tempo valorizzandole quale potenziale memoria storica e fonte di conoscenza per le generazioni future.

Da questa classificazione consegue che, tra le tante, una questione di primo livello da affrontare con la massima urgenza per la definizione del DCH sia conferire la funzione di memoria alle entità digitali e, in particolare, ai metadati che ne sono componente essenziale.

Riguardo a questi ultimi, riteniamo che i FAIR Principles: Findable, Accessible, Interoperable, Readable definiscano requisiti non del tutto sufficienti a garantire questa funzione, sebbene, a oggi, i metadati siano l'unica componente a poterla avere, soprattutto negli elementi descrittivi.

Nella nostra analisi abbiamo preso in considerazione gli schemi di metadati descrittivi di risorse digitali di diversa tipologia (Europeana, World Digital Library, Library of Congress, ecc.), soffermandoci in particolare sui diversi processi di digitalizzazione e co-creazione

digitale e sulla funzione assegnata ai metadati descrittivi. Abbiamo facilmente rilevato come tutte le risorse esaminate siano accomunate dalla medesima scarsa attenzione per gli elementi descrittivi dei metadati.
Partendo da tale constatazione, abbiamo elaborato un ampliamento della R dei FAIR Principles in R4, aggiungendo i seguenti requisiti:
- Re-usable: la riusabilità garantisce la sopravvivenza delle entità digitali in quanto i diversi usi nel tempo ne determinano la funzione culturale di memoria;
- Relevant: la persistenza di un'entità digitale nel tempo si caratterizza per le possibili trasformazioni della funzione che il riuso comporta, ed è perciò requisito indispensabile affinché essa, nata con un certo scopo non necessariamente culturale, essendo riutilizzata più volte con funzioni anche differenti evolva in portatrice di memoria e, quindi, in risorsa culturale
- Reliable: l'affidabilità è strettamente connessa alla capacità dell'entità digitale di rappresentare la propria evoluzione, in quanto in essa sono registrati i processi validati e certificati che ne hanno caratterizzato di volta in volta forme e funzioni nel corso del tempo;
- Resilient: la resilienza è requisito indispensabile alle entità digitali perché possano essere recuperabili e riusabili nel tempo, e continuino sia a farsi conoscere nella loro funzione originaria, sia a fornire un servizio pur nell'evoluzione delle loro funzioni da strumentali a culturali.

A nostro parere, questi requisiti, convenientemente applicati alla strutturazione dei metadati descrittivi fin dalla fase di progettazione, garantirebbero alle entità digitali le caratteristiche necessarie a trasformarne la funzione da meramente strumentale a memoria, consentendo così di distinguere le risorse digitali culturali dai consumer data.
Riteniamo, perciò, che la necessità di conferire o meno ai dati digitali i requisiti di R4 debba regolare sia l'approccio metodologico e tecnologico ai processi di co-creazione e di digitalizzazione, sia soprattutto la struttura dei metadati e la descrizione dei contenuti di ciascuna risorsa born digital o digitalizzata.

3. Conclusioni

Il riconoscimento europeo del DCH insieme al patrimonio culturale tangibile e intangibile apre una nuova fase dell'Era Digitale, in quanto sancisce definitivamente il ruolo di patrimonio culturale delle entità digitali. Si rende perciò indispensabile procedere a una sistematizzazione che contribuisca a identificare quali entità digitali possano essere considerate DCH.
Questo paper ha inteso delineare una prima proposta di definizione del DCH classificandolo in tre macro-categorie: il Born Digital Heritage, il Digital FOR Cultural Heritage e il Digital AS Cultural Heritage.
La necessità di conferire alle entità digitali la funzione di memoria indispensabile a evolverle in DCH ha poi indotto a proporre un ampliamento della R dei FAIR Principles in R4, alla luce della considerazione che le entità digitali, e in particolare i metadati soprattutto negli elementi descrittivi, debbano essere Re-usable, Relevant, Reliable and Resilient.

Dalle riflessioni sopra esposte riteniamo si possa inferire la seguente proposta di identificazione del DCH:
Il Digital Cultural Heritage è l'ecosistema di processi, entità, fenomeni Born Digital e Digitized che, essendo stati certificati e validati come conformi ai requisiti R4 tempo, fin dalla loro genesi siano potenziali testimonianze, manifestazioni ed espressioni dei processi evolutivi che identificano e connotano ciascuna comunità, contesto socio culturale, ecosistema semplice o complesso dell'Evo Digitale, assumendo la funzione di memoria e fonte di conoscenza per le generazioni future.

Riferimenti bibliografici

https://www.go-fair.org/fair-principles/

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG0614%2808%29

http://www.interpares.org/

Agenzia per l'Italia Digitale (AgID), Presidenza del Consiglio dei Ministri, Linee guide sulla conservazione dei documenti informatici,, Versione 1.0 – dicembre 2015, pp. 45 ss.

(http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def.pdf).

L. Bailey, Digital Orphans: The Massive Cultural Black Hole On Our Horizon, Techdirt, Oct 13th 2015 (https://www.techdirt.com/articles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml).

S. Cosimi, Vint Cerf: ci aspetta un deserto digitale, Wired.it, 16 febbraio 2015 (http://www.wired.it/attualita/2015/02/16/vint-cerf-futuro-medievale-bit-putrefatti/).

L. Duranti, E. Shaffer (ed. by), The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage, UNESCO Conference Proceedings, 26-28 September 2012, Vancouver (http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf)

V. Gambetta, La conservazione della memoria digitale, [Rubano], Siav, 2009.

M. Guercio, Gli archivi come depositi di memorie digitali, "Digitalia", Anno III, n. 2, ICCU Roma, 2008, pp. 37-53.

M. Guercio, Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali, Roma-Bari, Laterza, ed. 2013.

B. Lavoie, R. Gartner, Preservation Metadata (2nd edition), DPC Technology Watch Report, 03 May 2013, DPC Technology Watch Series (http://www.dpconline.org/docman/technology-watch-reports/894-dpctw13-03/file).

Library of Congress, PREMIS – Preservation Metadata: Implementation Strategies, v. 3.0 (http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf)

G. Marzano, Conservare il digitale. Metodi, norme, tecnologie, Milano, Editrice Bibliografica, 2011.

OCLC. PREMIS (PREservation Metadata: Implementation Strategies) Working Group, 2005 (http://www.oclc.org/research/projects/pmwg/).

Autori

Nicola Barbuti - nicola.barbuti@uniba.it

Nicola Barbuti è 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

Stefano Ferilli è 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.

DOI: 10.26314/GARR-Conf18-proceedings-06

Analisi dei dati di H.E.R.M.E.S. Pathfinder: una costellazione di nanosatelliti per l'astrofisica delle alte energie

Carlo Cabras[1], Alessandro Riggio[1], Luciano Burderi[1], Andrea Sanna[1], Tiziana Di Salvo[2]

[1] Università degli studi di Cagliari, [2] Università degli studi di Palermo

Abstract. HERMES Pathfinder è un progetto che riguarda lo sviluppo ed il lancio di un nuovo tipo di osservatorio a raggi X composto da uno sciame di nano-satelliti, nato dalla collaborazione tra diverse Università e Centri di Ricerca italiani. Il suo scopo è di rivelare, osservare e localizzare con precisione la posizione nel cielo di eventi astrofisici brevi e molto luminosi, tra i quali i Gamma Ray Burst. La natura distribuita di HERMES pone delle sfide tecniche relative al raccoglimento e l'analisi real-time dei dati.
In questo contributo, presentiamo la missione HERMES e discutiamo le sfide principali, con particolare enfasi su algoritmi innovativi per la misura dei ritardi temporali tra i nano-satelliti, specialmente in presenza di segnali molto deboli.

Keywords. Nanosatellites, Gamma-ray bursts, Gravitational waves, Constellation of satellites, Scintillators

1. Introduzione

1.1 L'astronomia multimessaggero

Grazie alla rilevazione diretta delle onde gravitazionali, siamo entrati nell'era dell'astronomia multimessaggero, ovvero la possibilità di osservare lo stesso evento astronomico attraverso la sua emissione di onde elettromagnetiche e gravitazionali (GW).
Il 17 agosto 2017, un tale evento (GW170817, Abbott et al. 2017a,b) è stato osservato sia come onda gravitazionale che come lampo gamma e, grazie alla precisa determinazione della sua posizione, è stato possibile osservarne l'evoluzione con i principali telescopi radio, ottici e X.

1.2 I lampi gamma

I lampi gamma, o Gamma-Ray Burst (GRB), sono esplosioni brevi ed altamente energetiche di origine extragalattica. Sono gli eventi più luminosi che si conoscono e possono durare da 10 millisecondi fino a un migliaio di secondi (Abdo et al. 2009a,b, Bloom et al. 2009, Flores et al. 2013, Zhu et al. 2013).
L'evento fisico all'origine dei GRB si ipotizza sia l'esplosione di una supernova (per i GRB lunghi) e la coalescenza di sistemi binari formati da due stelle di neutroni o una stella di neutroni e un buco nero (GRB brevi).
Questi ultimi sono forti emettitori di onde gravitazionali e quindi rilevabili dagli osservatori

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

LIGO e Virgo. I GRB, pur essendo estremamente brillanti, hanno una durata molto breve che attualmente non permette di determinarne la posizione con una precisione adeguata così da poter puntare i telescopi per osservarne il cosiddetto afterglow.

2. HERMES

HERMES (High Energy Rapid Modular Experiment Scintillator) è un osservatorio a tutto cielo di raggi X (2keV - 1 MeV) in grado di determinare con precisione ($\lesssim 1°$) la posizione nel cielo di eventi fortemente energetici come i GRB.
È costituito da una costellazione di nano-satelliti in orbita bassa (LEO, 500 km di quota), ognuno dei quali è in grado di rilevare autonomamente i raggi X e gamma.
La costellazione (Fig. 1) ha come suo punto di forza la ridondanza: al crescere del numero dei componenti, cresce la robustezza dell'intero sistema.

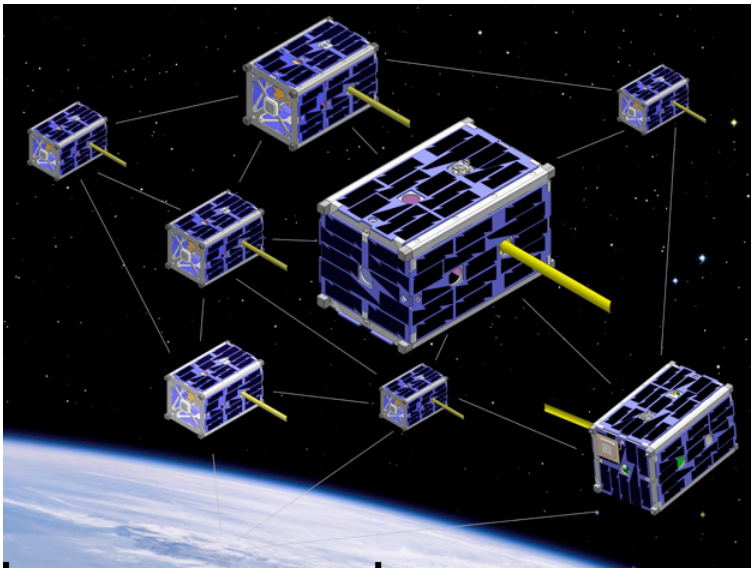


Fig. 1. Rappresentazione artistica di una costellazione di cubesat. (NASA)

Il nano-satellite è progettato per essere semplice, economico e veloce da costruire; è di tipo 3U (10x10x30 cm) ed è in grado di rilevare i raggi X/gamma grazie a degli innovativi rivelatori a stato solido SDD (Silicon Drift Detector) accoppiati a cristalli scintillatori che costituiscono il payload del nanosatellite (Fig. 2). Ognuno è munito di GPS al fine di determinare la posizione con sufficiente accuratezza (< 100 m).

3. Localizzazione dei GRB

La posizione dei GRB nel cielo sarà stimata per triangolazione attraverso la misura dei brevi ritardi (< 0,05 s) tra le rilevazioni dei diversi satelliti dovuti alla diversa distanza di ogni nano-satellite dalla sorgente. Per misurarli il rilevatore deve avere un'ottima risoluzione temporale (< 10 microsecondi) e la loro posizione deve essere nota con una grande accuratezza (< 100 metri): si potrà così determinare la posizione dei GRB che sarà tanto più precisa quanto è più alto il numero di satelliti.

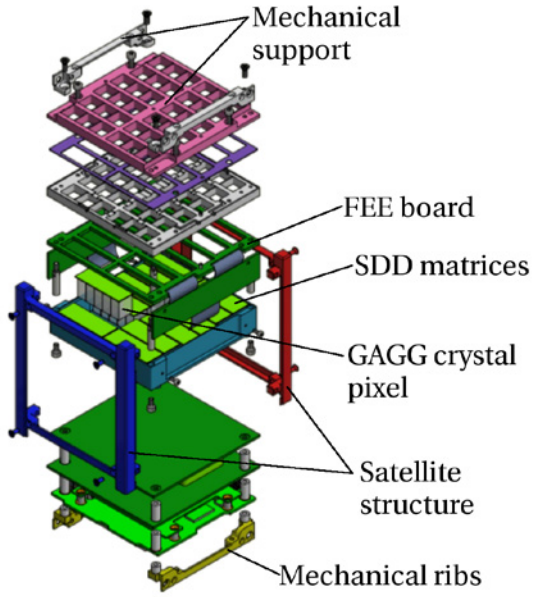


Fig. 2
Esploso del payload del nano-satellite. (Fuschino et al. 2018)

Per determinare il ritardo tra due satelliti che osservano lo stesso GRB, il metodo migliore è la Cross-Correlation Function (CCF) (Papoulis 1962), il cui costo computazionale non dipende dal numero dei fotoni ma solo dalla risoluzione temporale delle curve di luce. Mentre per i GRB luminosi non ci sono problemi, nel caso di segnali deboli la CCF diventa troppo rumorosa (Fig. 3, dati in nero). Per superare questo problema abbiamo sviluppato il metodo kernel che pesa la distanza tra i singoli fotoni usando una funzione kernel (Fig. 3, dati in rosso). Il metodo è promettente ma pesante, stiamo cercando di ottimizzare l'algoritmo parallelizzandolo ed implementandolo su GPU.

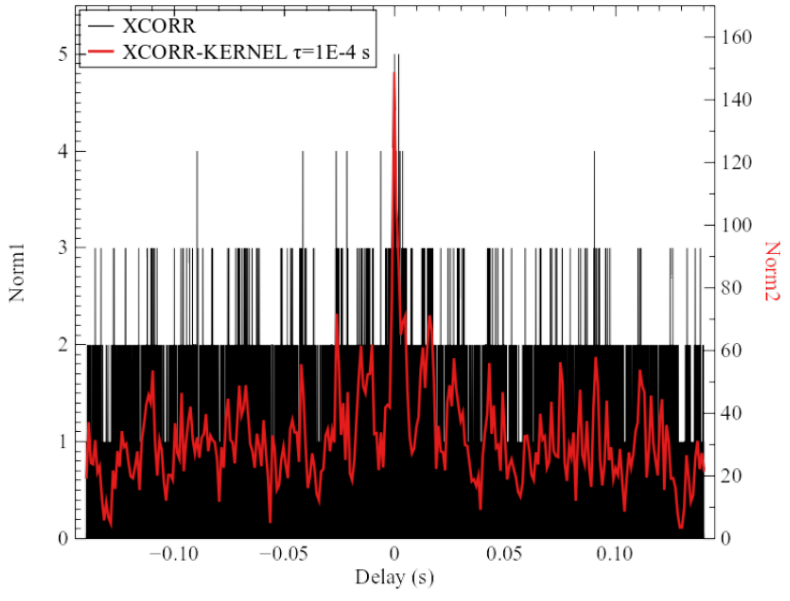


Fig. 3.
Comparazione tra diverse tecniche per determinare i ritardi temporali.
In nero: funzione di cross-correlazione.
In rosso: funzione kernel.

4. I dati

Un aspetto critico della missione è costituito dalla trasmissione e analisi dei dati. Per essere efficace, la notifica di un evento GRB deve arrivare alla comunità scientifica in un tempo dell'ordine di un'ora dal GRB.

Con la configurazione iniziale di 6 nano-satelliti, le uniche comunicazioni possibili sono con la superficie terrestre, dato che essi non hanno abbastanza potenza per comunicare tra di loro. Avendo poche stazioni riceventi è difficile che tutti i satelliti potranno inviare i dati in tempi brevi: si potrebbe creare una fitta rete di piccole antenne riceventi a terra e interconnesse o porre i nanosatelliti su orbite polari sincrone al sole (SSO) e sfruttare antenne presenti sulle isole Svalbard.

Nel caso invece del raggiungimento della configurazione a pieno regime (decine di satelliti), la distanza media diventa si ridurrà rendendo possibile l'intercomunicazione tra i satelliti. Saranno implementati algoritmi che renderanno i satelliti sensibili agli eventi transitori X/gamma, comparando il flusso di dati che si sta ottenendo con i dati di background rilevati in precedenza. L'evento rilevato viene salvato in un buffer e poi inviato a terra tramite pacchetti di telemetria.

5. Conclusioni

HERMES darà alla comunità scientifica uno strumento per studiare i GRB con una precisione senza precedenti, aprendo le porte alla comprensione di fenomeni cruciali nel settore dell'astrofisica e della fisica fondamentale. La fase corrente prevede la realizzazione di 3 nano-satelliti che saranno lanciati nel 2020 e che dimostreranno la fattibilità dell'approccio adottato e le sue performance.

Riferimenti bibliografici

[1] Abbott et. al. (2017) Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A, The Astrophysical Journal Letters, 848:L13 (27pp), 2017 October 20, https://doi.org/10.3847/2041-8213/aa920c

[2] Abbott et. al. (2017) Multi-messenger Observations of a Binary Neutron Star Merger, The Astrophysical Journal Letters, 848:L12 (59pp), 2017 October 20, https://doi.org/10.3847/

2041-8213/aa91c9

[3] Abdo et. al. (2009) Fermi observations of high-energy gamma-ray emission from GRB 080916C, Science. 2009 Mar 27;323(5922):1688-93. doi: 10.1126/science.1169101

[4] Abdo et. al. (2009) A limit on the variation of the speed of light arising from quantum gravity effects, Nature volume 462, pages 331-334 (19 November 2009)

[5] Bloom et. al. (2009) Observations of the naked-eye GRB 080319B: implications of nature's brightest explosion, The Astrophysical Journal, 691:723-737, 2009 January 20, doi:10.1088/0004-637X/691/1/723

[6] Fuschino et. al. (2018) HERMES: An ultra-wide band X and gamma-ray transient monitor on board a nano-satellite constellation. Nuclear Instrumentation and Methods

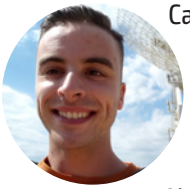
in Physics Research, A. https://arxiv.org/abs/1812.02432

[7] Flores et al. (2013), GCN Circ. 14491, https://gcn.gsfc.nasa.gov/other/130427A.gcn3

[8] Papoulis, A. (1962) The Fourier Integral and Its Applications. New York: McGraw-Hill, pp. 244-245 and 252-253, 1962.

[9] Zhu et. al. (2013) GCN Circ. 14471, http://gcn.gsfc.nasa.gov/gcn3/14471.gcn3

Autori



Carlo Cabras - carlocabras21@gmail.com
Studente del secondo anno del Corso di Laurea Magistrale in Informatica ed appassionato di astronomia, in passato si è occupato dello sviluppo di software per l'analisi di dati astrofisici.

Alessandro Riggio - alessandro.riggio@dsf.unica.it
Ricercatore TD/B presso il Dipartimento di Fisica dell'Università di Cagliari, si occupa dello studio di sorgenti transienti X galattiche. In particolare si occupa dello studio delle variabilità spettrali e temporali di pulsatori X al millisecondo.

Luciano Burderi - burderi@dsf.unica.it
Professore associato presso il Dipartimento di Fisica dell'Università di Cagliari. Il suo campo di ricerca è l'astrofisica delle alte energie. Si concentra sullo studio di sistemi binari galattici che ospitano un oggetto compatto (stella di neutroni o buco nero) che accresce materia da una stella compagna. L'attività di ricerca si basa su un approccio teorico e di osservazione, utilizzando i dati provenienti dai satelliti a raggi X.

Andrea Sanna - andrea.sanna@dsf.unica.it
Ricercatore TD/A presso il Dipartimento di Fisica dell'Università di Cagliari. I suoi attuali interessi di ricerca sono proprietà spettrali e temporali di binarie a raggi X di piccola massa, con particolare attenzione alle proprietà delle stelle di neutroni a rotazione rapida che accrescono la materia dagli inizi di compagni di massa ridotta.

Tiziana Di Salvo - tiziana.disalvo@unipa.it
Professore associato presso il Dipartimento di Fisica e Chimica dell'Università di Palermo. La sua attività di ricerca concerne l'astrofisica delle alte energie, e in particolare sullo studio di sistemi binari X galattici.

DOI: 10.26314/GARR-Conf18-proceedings-07

Un modello per la compilazione di Data Management Plan per l'archeologia

Sara Di Giorgio[1], Paola Ronzino[2]

[1]ICCU, Istituto Centrale per il Catalogo Unico delle biblioteche italiane per le informazioni bibliografiche
[2]PIN Scrl Servizi didattici e scientifici per l'Università di Firenze

Abstract. Il contributo presenta il template per la creazione di Data Management Plan (DMP), sviluppato dal progetto PARTHENOS con l'obiettivo di offrire uno strumento che soddisfi i requisiti degli istituti finanziatori dei progetti di ricerca, fornendo al contempo uno strumento che permetta ai ricercatori di documentare la propria ricerca. Il modello, che trae origine dall'iniziativa europea dell'Open Science e dai principi FAIR, è stato adattato alle esigenze dei ricercatori nel settore archeologico, ed è stato testato da un gruppo di esperti del dominio. Il contributo presenta gli ultimi aggiornamenti apportati al template e include una panoramica sugli strumenti implementati per facilitarne la compilazione e produrre un modello che sia machine-actionable.

Keywords. Data Management Plan, FAIR, Open Science, archeologia

Introduzione

I Data Management Plan (DMP) sono degli strumenti che permettono di raccogliere informazioni su come i dati saranno creati, definendone i piani per la loro condivisione e conservazione nel tempo, indicando eventuali restrizioni da applicare ai dati di ricerca nelle varie fasi del progetto.

La politica lanciata dalla Commissione Europea in materia di Open Innovation e Open Science ha dato luogo a una serie di iniziative conl'obiettivo di rendere i dati di ricerca scientifica accessibili e di garantire il riutilizzo, la ridistribuzione e la replicabilità della ricerca (1).

In questo contesto, nel 2015 sono stati sviluppati i FAIR principles (2, 3), i quali fungono da linea guida per i ricercatori che intendano migliorare la riusabilità dei loro dati fornendo una serie di indicazioni affinchè i dati siano reperibili, findable, accessibili, accessible, interoperabili, interoperable e riutilizzabili, reusable.

Per perseguire tale obiettivo, il programma Horizon 2020 ha pubblicato le linee guida sulla gestione dei dati FAIR e ha fornito ai beneficiari di finanziamenti H2020 il modello FAIR del DPM di H2020 (4), da presentare entro i primi sei mesi dall'inizio di un progetto. Partendo dalle linee guida di H2020 sulla gestione dei dati FAIR, PARTHENOS ha elaborato un modello per la definizione di un DMP specificatamente pensato per la comunità archeologica (5).

I DMP si presentano generalmente sotto forma di documenti compilati dai ricercatori

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

e nella maggior parte dei casi la procedura di creazione è eseguita manualmente oppure mediante l'utilizzo di strumenti online (6). I file ottenuti utilizzando le soluzioni digitali si presentano spesso in formati testuali, come file PDF o DOC, orientati all'human readability, cioè ottimizzati per essere letti in modo naturale dalle persone. Solo in alcuni casi gli strumenti disponibili permettono di ottenere in uscita un formato machine readable, come JSON o XML, cioè i cui dati sono codificati per essere processati in modo automatico dai calcolatori.

Il presente contributo include una panoramica sugli ultimi aggiornamenti del template e degli strumenti sviluppati per rendere il modello facilmente utilizzabile e machine-actionable.

1. PARTHENOS DMP per la Open Research in archeologia

PARTHENOS (7) è un progetto H2020 che mira a rafforzare la coesione della ricerca nel settore degli studi linguistici, dei beni culturali, della storia, dell'archeologia attraverso un gruppo tematico di infrastrutture di ricerca europee, favorendo l'interoperabilità dei dati e formulando standard e politiche comuni in materia di gestione dei dati.

In PARTHENOS, la comunità archeologica è rappresentata da ARIADNE (8), una rete europea che è stata finanziata tra il 2013 e il 2017, e che ha sviluppato un'infrastruttura di ricerca per l'integrazione dei dati archeologici distribuiti negli archivi digitali di tutta Europa, consentendone la scoperta e l'accesso per mezzo del suo portale (9).

Un'indagine condotta tra gli esperti del consorzio e successivamente estesa ad altri esperti del dominio, ha fornito al gruppo di lavoro di PARTHENOS un quadro completo di standard e buone pratiche per la creazione, l'archiviazione e la condivisione dei dati utilizzati dalla comunità archeologica, i quali sono stati utilizzati come punto di partenza per preparare l'elenco dei requisiti e linee guida offerte ai ricercatori per la compilazione del DMP.

Il modello del DMP di PARTHENOS mantiene la stessa struttura di quello di H2020, in modo da utilizzare un modello oramai familiare ai ricercatori. Rispondendo alle domande contenute nel DMP, i ricercatori forniscono informazioni su:

- i dati prodotti dalla ricerca;
- gli standard usati per strutturare i dati;
- il sistema di identificazione unica e persistente dei dati;
- la metodologia di archiviazione dei dati;
- la metodologia di condivisione dei dati con informazioni sulle condizioni di accesso e di riuso dei dati.

Rispetto al DMP di H2020, nel modello DMP di PARTHENOS sono forniti degli elenchi in cui è possibile selezionare standard e flussi operativi adottati in archeologia, offrendo spunti per la compilazione.

La necessità di supporto alla compilazione di un DMP è stata fortemente sottolineata da un gruppo di esperti, i quali hanno risposto a un'indagine condotta dal team di Data Management del progetto OpenAIRE e dal gruppo di esperti sui dati FAIR (10). L'obiettivo di questo sondaggio è stato quello di raccogliere feedback da un gruppo di

esperti che hanno valutato il template DMP di H2020 al fine di identificare eventuali lacune e raccogliere suggerimenti per il miglioramento.

2. Sviluppi tecnici del DMP di PARTHENOS

Una delle attività in fase di finalizzazione prevede la creazione di collegamenti tra le linee guida del PARTHENOS DMP e gli altri strumenti sviluppati da PARTHENOS e già disponibili online sul sito del progetto. In particolare, saranno possibili i collegamenti con i contenuti del Wizard e lo Standardization Survival Kit (SSK). Si tratta di due strumenti online realizzati per offrire un supporto pratico ai ricercatori che non hanno familiarità con il processo di gestione dei dati e che così possono conoscere in modo interdisciplinare, come i dati dovrebbero essere raccolti, elaborati, archiviati e condivisi con altri ricercatori. Il Wizard offre una serie di documenti di policy sulla gestione dei dati organizzati per aree tematiche di ricerca e lo SSK, attraverso l'uso di scenari di ricerca, guida i ricercatori nella selezione e nell'uso degli standard più appropriati per le loro particolari discipline e flussi di lavoro.

Il PARTHENOS DMP offrirà quindi una vasta gamma di elementi da utilizzare per la compilazione del DMP, adatti anche a utenti meno esperti, grazie ai ricchi contenuti offerti dai vari strumenti.

Per agevolare la compilazione del DMP di PARTHENOS, è stato sviluppato un applicativo ad hoc, la cui progettazione ha tenuto in considerazione sia le necessità pratiche dei ricercatori, sia l'attuale evoluzione tecnologica che i documenti digitali stanno subendo.

L'obiettivo finale è quello di ottenere dei machine-actionable DMP (11), cioè dei DMP le cui informazioni possano essere elaborate e rese comprensibili automaticamente dai calcolatori, e che siano allo stesso tempo dei documenti interoperabili, aggiornabili e condivisibili all'interno della comunità degli stakeholder. L'interfaccia messa a disposizione online è stata studiata per agevolare la compilazione del DMP mediante l'utilizzo di soluzioni intuitive e user friendly.

Le domande a cui il ricercatore è invitato a rispondere sono suddivise in pagine successive, dotate di una barra di avanzamento comune che si presenta come il punto di riferimento principale per l'utilizzatore. La visualizzazione complessiva delle varie parti che compongono il modello guida passo a passo l'utente, indicando approssimativamente il tempo necessario per finirli.

Ogni pagina raggruppa domande tematiche simili, distinte in obbligatorie e facoltative, arricchite da pop-up informativi di aiuto alla compilazione. Nel caso in cui alcuni dei punti obbligatori per la sottomissione del DMP non siano stati compilati, il loro numero sarà visualizzato in rosso nella barra di avanzamento. Al termine della procedura di compilazione sarà possibile scaricare le informazioni in esso contenute in formato PDF, oppure in JSON .

Il file JSON è fondamentale all'interno dell'applicativo poiché offre all'utente la possibilità di salvare una versione del proprio lavoro. La compilazione del questionario, infatti, può essere interrotta in qualsiasi momento scaricando il file JSON contenente i dati

correnti. Tale file potrà essere ricaricato all'interno dell'interfaccia online per proseguire e terminare il lavoro.

Se invece la compilazione del questionario è definitiva, i dati contenuti nel file costituiranno una versione del DMP utile per eventuali successive revisioni o aggiornamenti.

La progettazione degli sviluppi futuri dell'applicativo è finalizzata a rendere i dati contenuti all'interno dei DMP condivisibili e interoperabili tra quelle comunità di ricerca che adotteranno delle soluzioni comuni per agevolare la cooperazione tra i propri sistemi. Rendere dei documenti interoperabili significa fare in modo che le informazioni in essi contenuti possano essere scambiate tra sistemi diversi in maniera completa e affidabile. Per questo è necessario considerare sia l'aspetto sintattico sia quello semantico dei dati.

I calcolatori possono trattare e gestire in maniera sintattica la maggior parte delle informazioni, se queste sono codificate in formati standard come XML o JSON, ma non sono in grado di interpretarle e "comprenderle" se queste non sono modellate utilizzando dei vocabolari controllati e degli standard condivisi.

I DMP generati nella prima versione dell'applicativo soddisfano già i requisiti per l'interoperabilità sintattica, grazie alla codifica in formato JSON. Saranno successivamente adeguati a livello semantico, mediante l'uso di vocabolari controllati, di standard e di modelli di dati aperti e condivisi tra gli appartenenti alla comunità di ricerca, in particolare il modello semantico CRMpe, sviluppato all'interno di PARTHENOS (12).

Questo consentirà al DMP di avvicinarsi ulteriormente alle caratteristiche dei machine-actionable DMP e offrirà ai ricercatori la possibilità di trarre beneficio dalla condivisione delle informazioni.

3. Conclusioni

Il modello DMP di PARTHENOS si basa su quello di H2020, con l'aggiunta di un utile supporto nella compilazione del documento grazie alle linee guida e ai riferimenti incrociati creati con i contenuti degli altri strumenti sviluppati dal progetto (ad es. Wizard e SSK).

Il lavoro condotto da PARTHENOS anticipa quelle che sono le raccomandazioni e le azioni promosse dal "European Commission Expert Group on FAIR data" (13) fornendo uno strumento di supporto alla compilazione del documento.

Uno strumento online realizzato dal PIN all'interno del progetto è attualmente in fase di sviluppo per essere incorporato nel sito web di PARTHENOS per fornire ai ricercatori un tool semplice per creare il proprio DMP. Lo strumento attualmente consente la compilazione del modulo, grazie a un'interfaccia intuitiva e la possibilità di scaricare una copia del documento in PDF e in JSON.

Altre attività in previsione per il prossimo futuro includono la traduzione del modello DMP di PARTHENOS in diverse lingue per fornire versioni nazionali a quei paesi che non hanno ancora sviluppato un proprio modello (ad es. italiano, spagnolo, greco, tedesco, ecc.).

Al momento, nella comunità di PARTHENOS solo il modello archeologico è stato sviluppato e testato. In futuro, altre comunità di ricercatori delle discipline umanistiche, che sono disposte a contribuire con le loro pratiche quotidiane, saranno coinvolte nella

definizione di linee guida mirate.

Al fine di promuovere e diffondere il modello DMP di PARTHENOS e le sue linee guida alle comunità di ricerca archeologiche, si prevede l'organizzazione di seminari di formazione e servizi di consulenza per garantire una coerente diffusione dei risultati di PARTHENOS, allo scopo di aumentare la consapevolezza sulla ricerca aperta in archeologia.

Riferimenti bibliografici

(1) European Commission, Research and Innovation, Open Science, https://ec.europa.eu/research/openscience/index.cfm (ultima consultazione 8/12/2018)

(2) FORCE 11, Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing Version B1.0, https://www.force11.org/fairprinciples (ultima consultazione 8/12/2018)

(3) M. D. Wilkinson, M. Dumontier et al., 2016, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3:160018, https://doi.org/10.1038/sdata.2016.18 (ultima consultazione 8/12/2018)

(4) Data Management Template in Horizon 2020

http://ec.europa.eu/research/participants/docs/h2020-fundingguide/cross-cutting-issues/open-access-data-management/datamanagement_en.htm#A1-template (ultima consultazione 8/12/2018)

(5) S. Bassett, S. Di Giorgio, et al., 2017, A DMP template for Digital Humanities: the PARTHENOS model, pp. 74-77, Proceedings of GARR Conference "The Data Way to Science", Venezia,15-17 Novembre

(6) Data Curation Centre, https://dmponline.dcc.ac.uk/ (ultima consultazione 8/12/2018)

(7) PARTHENOS (Pooling Activities, Resources and Tools for HeritageE-research Networking, Optimization and Synergies) https://www.parthenos-project.eu (ultima consultazione 8/12/2018)

(8) ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) www.ariadne-infrastructure.eu (ultima consultazione 8/12/2018)

(9) ARIADNE Portal, http://portal.ariadne-infrastructure.eu/

(10) M. Grootveld, E. Leenarts, et al., 2018, OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (Version 1.0.0 Data set). Zenodo. http://doi.org/10.5281/zenodo.1120245 (ultima consultazione 8/12/2018)

(11) T. Miksa, S. Simms, et al., Ten simple rules for machine-actionable data management plans, https://zenodo.org/record/1172673#.XBOIOGhKiUk (ultima consultazione 8/12/2018)

(12) G. Bruseker, M. Doerr, et al., Report on the common semantic framework, PARTHENOS project deliverable 5.1, 2017 http://www.parthenosproject.eu/Download/Deliverables/D5.1_Common_Semantic_Framework_Appendices.pdf (ultima consultazione 8/12/2018)

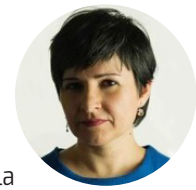
(13) S. Hodson, et al., 2018, FAIR Data Action Plan: Interim recommendations and

actions from the European Commission Expert Group on FAIR data (Version Interim draft). http://doi.org/10.5281/zenodo.1285290 (ultima consultazione 8/12/2018)

Autori

Sara Di Giorgio - sara.digiorgio@beniculturali.it

Collabora dal 2004 con l'Istituto centrale per il Catalogo Unico delle biblioteche italiane del Ministero dei beni e delle attività culturali, per lo sviluppo di CulturaItalia, il portale della cultura italiana del MiBACT, e di numerosi progetti europei per la digitalizzazione, la fruizione e la conservazione a lungo termine del patrimonio culturale digitale. Partecipa al progetto Parthenos, coordinando le attività per la definizione dei requisiti per lo sviluppo di policy sul data lifecycle. Coordina i Task 'Managing FAIRness of archaeological data and IPR' e ' Innovating services with industries' nell'ambito del progetto ARIADNE Plus. E' membro del Council di Europeana, la digital library dell'Unione Europea.



Paola Ronzino - p.ronzino@gmail.com

Paola Ronzino, è una ricercatrice post-doc presso il PIN. Ha conseguito un Master in Conservazione dei Beni Culturali e un Dottorato in Science and Technology for Cultural Heritage. I suoi interessi di ricerca riguardano lo sviluppo di ontologie e standard di metadati per la documentazione archeologica e il patrimonio culturale, con interesse specifico per la documentazione degli edifici archeologici. Ha partecipato a diversi progetti europei coordinati dal PIN sul patrimonio culturale digitale, ed è stata attivamente coinvolta nelle attività di ricerca del progetto ARIADNE. Attualmente, partecipa alle attività del progetto PARTHENOS, in particolare nella definizione delle esigenze degli utenti e l'elaborazione del Data Management Plan per le discipline umanistiche. E' Project Manager del progetto ARIADNEplus all'interno del quale si occupa delle attività relative all'integrazione di dataset geografici.

DOI: 10.26314/GARR-Conf18-proceedings-08

Beni archivistici e big data. Il progetto GAWS: Garzoni, Apprenticeship, Work, Society

Andrea Erboso

Archivio di Stato di Venezia

Abstract. Gli archivi rappresentano uno dei più ricchi e importanti "database" di informazioni del passato la cui consultazione, tuttavia, è resa difficoltosa dalla varietà di supporti, tipologie e grafie che caratterizza la documentazione manoscritta. L'Archivio di Stato di Venezia ha collaborato con importanti istituti universitari europei per un progetto dedicato alla valorizzazione del proprio patrimonio archivistico. Il progetto GAWS ha consentito, dopo un lungo lavoro di digitalizzazione integrale del materiale, la realizzazione di un database di vastissime dimensioni, unico nel panorama degli studi storici. Sul database, sul processo di inserimento dei dati e sulle possibilità di interrogazione degli stessi sono stati testati nuovi approcci offerti dalle tecnologie informatiche.

Keywords. Archivi, Garzoni, Big data, Open data, Machine learning

Introduzione

Il dibattito sempre più attuale sull'uso di big data a supporto della ricerca sta interessando, ormai da diversi anni, il mondo degli archivi: non deve stupire che anche un settore principalmente e tradizionalmente rivolto agli studi umanistici si interessi dei risultati più recenti offerti dalle computer sciences e dalle tecnologie di intelligenza artificiale nello sviluppo di applicazioni per la gestione, l'organizzazione e la manipolazione di grandi quantità di dati. Gli archivi, infatti, sono costituiti da complessi documentali già di per sé ricchissimi di contenuti informativi, a loro volta ulteriormente ampliati dai legami e dalle interrelazioni che in maniera organica intercorrono tra le carte: nel loro complesso gli archivi rappresentano un potenziale di conoscenza senza eguali. La stessa disciplina archivistica, quasi anticipando certi sviluppi dell'attuale ricerca sui big data, ha da sempre cercato di affinare le metodologie che consentono di descrivere i complessi archivistici ad un livello di definizione tale da permettere ad un ricercatore di orientarsi in un patrimonio che per vastità esclude a priori la possibilità dell'approccio diretto e analitico. Le metodologie di ordinamento e descrizione degli archivi hanno proprio l'obiettivo di fornire gli elementi generali necessari a strutturare gerarchicamente l'informazione, consentendo con relativa rapidità di identificare le unità che possono contenere i dati ricercati con precisione.

L'esperienza degli archivi è fondamentale anche per il tema, strettamente collegato, degli open data: a condivisione della conoscenza acquisita dalle esperienze di valorizzazione degli archivi è uno degli obiettivi dell'amministrazione archivistica. I documenti conservati presso gli Archivi di Stato, infatti, sono beni culturali a tutti gli effetti e, in quanto tali, disciplinati dal Codice dei beni culturali (D.Lgs 42/2004), che ne sancisce la natura di bene

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

pubblico. La predisposizione di strumenti di ricerca innovativi e aggiornati, necessari alla fruizione e valorizzazione degli archivi stessi, è dunque uno dei principali compiti degli Archivi di Stato che ne devono garantire la completa pubblicità e condivisibilità.

1. Il progetto GAWS

Un esempio di applicazione di tecnologie tipiche dei big data al contesto archivistico è il progetto GAWS, Garzoni Apprenticeship Work and Society, finanziato dall'Agence Nationale de la Recherche ANR e dal Fonds National Suisse FNS e frutto della collaborazione tra l'Università di Lille 3, l'Università di Rouen, l'EPFL di Losanna e l'Archivio di Stato di Venezia. Il progetto è in via di conclusione e ha già prodotto un complesso database che sarà pubblicato nell'autunno del 2019 (https://garzoni.hypotheses.org/).

Il progetto ha preso avvio da una fonte documentaria conservata presso l'Archivio di Stato: la serie degli Accordi dei garzoni del fondo della Giustizia Vecchia, una raccolta sistematica di 53.890 contratti di apprendistato tra garzone e maestro, relativi a oltre 1.300 professioni e raccolti in 32 registri (per un totale di 14.236 carte manoscritte).

La particolarità che rende unica al mondo questa serie è che, a differenza di quanto accade in tutta l'Europa dell'età moderna dove il contratto di apprendistato è considerato di natura privatistica e quindi sottoscritto dalle parti di fronte al notaio, a Venezia questa pratica era espressamente vietata. Nella Serenissima era lo Stato - la magistratura della Giustizia Vecchia appunto - ad occuparsi della registrazione di questo tipo di accordi (Bellavitis et al., 2017). Questa particolare disposizione ha garantito alla documentazione una omogeneità sorprendente con la registrazione costante della stessa tipologia di informazioni per tutto l'arco di tempo per cui si conserva la fonte (1575-1772 con lacune).

L'obiettivo del progetto è stato, dunque, sperimentare un metodo per strutturare le informazioni contenute nella fonte e renderle esplorabili in modalità altrimenti impossibili.

2. Descrizione e inserimento dei dati

In una prima fase si è proceduto alla digitalizzazione sistematica di tutto il materiale che è stato poi caricato in un'interfaccia di navigazione e lavoro compatibile con gli standard IIIF (Ehrmann et al., 2016).

Contemporaneamente un team di storici e archivisti si è occupato di effettuare uno studio preliminare sulla fonte ed estrarre da essa il modello delle informazioni che un gruppo di informatici ha adoperato per sviluppare l'ontologia e l'architettura informatica. La tipologia di contenuti presenti nei singoli accordi e inseriti nel database sono:

- il nome del garzone con alcuni elementi identificativi (patronimico, provenienza);
- la professione che il garzone vuole imparare;
- il nome del maestro che insegnerà il mestiere, anch'esso con alcuni elementi identificativi (localizzazione della bottega);
- il nome di una persona che funge da garante e che molto spesso è un parente del garzone stesso;
- le condizioni materiali del contratto;

- la durata dell'apprendistato in anni (regolamentata da ogni singola corporazione);
- le condizioni salariali.

L'inserimento dei dati è avvenuto direttamente a partire dalla riproduzione digitale del

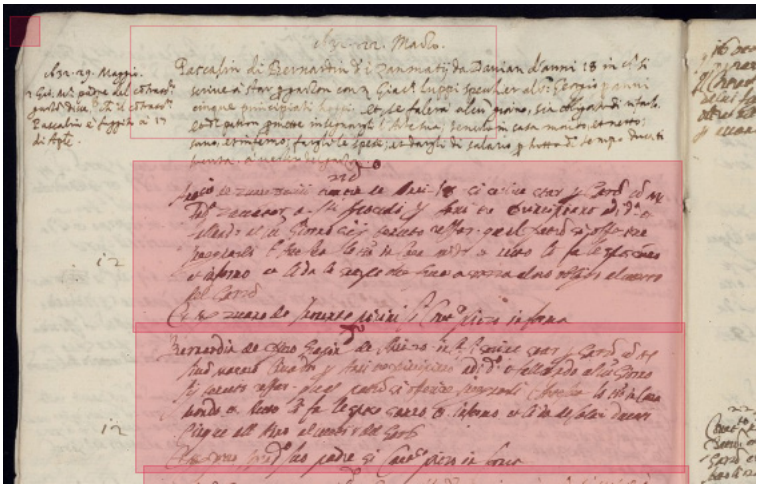


Fig. 1
Esempi di segmentazione della riproduzione digitale del documento

documento che è stata segmentata in riquadri, uno per contratto.
Direttamente dal segmento di immagine è possibile aprire la maschera di inserimento dei dati per la trascrizione della informazioni. Questa attività, condotta da un team di specialisti, ha prodotto più di 470.000 trascrizioni, alle quali sono state associate più di 420.000 tag semantiche individuate nello studio preliminare sulla fonte. Ogni trascrizione con le relative tag è stata annotata direttamente sull'immagine come fosse un suo metadato: in tutto il database, dunque, ogni informazione rimane costantemente associata alla riproduzione del documento stesso.

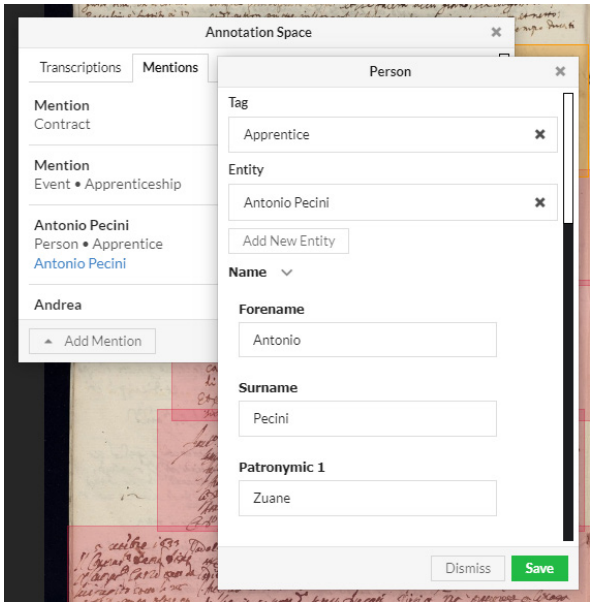


Fig. 2
Esempio della maschera di trascrizione e inserimento dati

2.1 Tecniche di machine learning per il data entry

L'aspetto più critico della compilazione è stato quello riguardante i nomi di persona: tra i 187.266 nominativi, i casi di omonimia e similarità sono stati frequentissimi. Le modalità di compilazione hanno previsto la trascrizione letterale del nome come presentato nella fonte, la cosiddetta person mention, e l'attribuzione di una person id normalizzata.
I casi problematici emersi dalla compilazione sono riconducibili a due fattispecie:

- diverse mention riferite alla stessa persona (stesso individuo citato più volte);
- mention identiche riferite però a persone diverse (omonimia).

La possibilità di disambiguare tra i possibili omonimi e i casi di variante grafica dello stesso nome è stata offerta dalle tecniche di machine learning messe a punto appositamente per il progetto. All'atto di attribuzione del person id normalizzata infatti, il software di compilazione proponeva al compilatore i nomi simili già inseriti nel database, offrendo in aggiunta una sintesi dei dati correlati al nominativo: in questo modo al compilatore venivano forniti tutti gli elementi disponibili (e il link diretto alla fonte per la verifica puntuale) per scegliere criticamente se assegnare un nuovo id (omonimia) o attribuendone uno già esistente (variante grafica). Man mano che il database si è popolato di informazioni, maggiore è stata la precisione del tool nell'indicare al compilatore la rosa di casi tra i quali operare la disambiguazione, eliminando sempre più il rumore causato dai risultati non pertinenti.

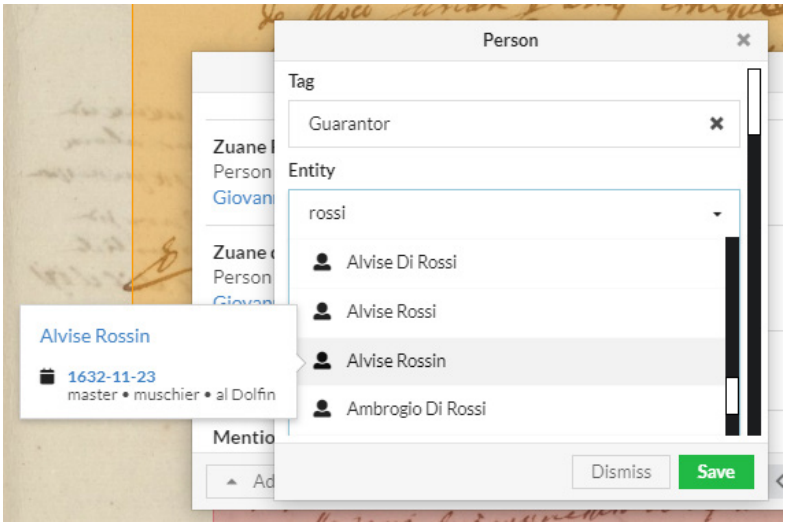


Fig. 3
Esempio di funzionamento del tool di suggerimento per la disambiguazione

3. Modalità di interrogazione

L'interfaccia di esplorazione dei dati, attualmente in fase di test, consente di interrogare il database con ricerche a testo libero, filtri e query sparql oltre che, naturalmente, per segnatura archivistica, navigando tra le immagini. Soprattutto il sistema di filtri si sta rivelando particolarmente efficace e consente di estrarre liste di contratti sempre più precise

che lo studioso potrà consultare online, scegliendo quali informazioni visualizzare tra le molte contenute nei contratti, oppure potrà scaricare nei formati xlsx, ods e json per elaborarli con software più complessi, per esempio per analisi statistiche su lunghi periodi.

In tutti i casi, sia nella consultazione online, che nei dati scaricati, sarà presente il collegamento con l'immagine del documento, evidenziata nella parte che contiene le informazioni. Questa possibilità è frutto di una precisa scelta metodologica, relativamente nuova negli studi umanistici, che consentirà agli studiosi di verificare agevolmente la fonte citata: la trascrizione, con i suoi inevitabili errori, diventa pertanto un punto di accesso all'informazione, senza sostituirla.

4. Conclusioni

L'esperienza del progetto GAWS è stata estremamente fruttuosa per sperimentare nuovi metodi di valorizzazione del patrimonio archivistico: da un lato, la digitalizzazione del materiale ha consentito di rendere disponibili agli studiosi i registri degli Accordi dei garzoni, preservando per il futuro gli originali dall'usura dovuta alla consultazione, dall'altro la banca dati offre opportunità di ricerca prima impensabili. La possibilità di interrogare il database a partire da un nome o da una professione, per fare un esempio, consente di far emergere aspetti che la consultazione tradizionale, limitata alla sola lettura degli accordi registrati in successione cronologica, non permetterebbe, come le reti di relazioni sociali tra maestri, garzoni e garanti o i flussi migratori legati all'apprendistato.

Il progetto ha poi offerto un terreno di prova per la creazione di un team multidisciplinare che ha visto collaborare fianco a fianco archivisti, storici, storici dell'arte, informatici, scienziati del linguaggio, in un clima collaborativo e di reciproca contaminazione, costituendo un incoraggiante precedente per la progettazione di nuovi interventi di valorizzazione del patrimonio archivistico.

Riferimenti bibliografici

Bellavitis A., Frank M., Sapienza V., a cura di (2017), Garzoni Apprendistato e formazione tra Venezia e l'Europa in età moderna, Mantova, Universitas Studiorum.

Ehrmann M., Colavizza G., Topalov O., Cella R., Drago D., Erboso A., Zugno F., Bellavitis A., Sapienza V., Kaplan F. (2016), From Documents to Structured Data: First Milestones of the Garzoni Project, DH Commons Journal, (2), pp. 1-18.

Autore

Andrea Erboso - andrea.erboso@beniculturali.it

Laureato in Storia dell'Arte Moderna presso l'Università Ca' Foscari di Venezia, ha conseguito il diploma in Archivistica, Paleografia e Diplomatica presso l'Archivio di Stato di Venezia. Dal 2018 è funzionario archivista di Stato presso l'Archivio di Stato di Venezia. Si è occupato principalmente di arte veneziana del secondo Cinquecento, di botteghe e apprendistato nelle arti. Ha collaborato a progetti di digitalizzazione documentaria e descrizione archivistica con importanti istituzioni internazionali.

DOI: 10.26314/GARR-Conf18-proceedings-09

Abylia, un ambiente digitale per la scuola in trasformazione

Davide Zucchetti[1], Filippo Chesi[2], Mirko Marras[3], Silvio Barra[1]

[1]Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica, [2]Abylia Srl, [3]Lattanzio Advisory

Abstract. In risposta alla crescente complessità dei modelli e dei processi educativi, ai docenti è richiesto di acquisire nuove competenze e di sviluppare una professionalità più ampia e trasversale per rafforzare l'efficacia del loro ruolo educativo e formativo verso "nativi digitali". In questa cornice prende forma il progetto di ricerca co-finanziato dal MIUR "ILearn TV Anytime Anywhere" con l'obiettivo di ridisegnare i modelli di aggiornamento professionale dei docenti valorizzando la funzione abilitante delle tecnologie digitali per la collaborazione e la cooperazione nella comunità scolastica, a supporto dell'innovazione sociale e territoriale. Accessibilità, leggibilità e interoperabilità di contenuti e oggetti formativi digitali sono requisiti essenziali per raggiungere questi risultati. I processi di produzione multi-autorale sono stati definiti per rispondere alle esigenze della knowledge school e al suo ruolo di centro di formazione permanente, anche al difuori delle attività didattiche e della relazione docente-discente. Abylia, questo il nome dell'ambiente digitale sviluppato e oggi in fasi di sperimentazione, valorizza sia la dimensione umana sia quella tecnologica dell'innovazione rafforzando il valore strategico della scuola e dell'apprendimento continuo come fattore di crescita e sviluppo individuale e collettivo.

Keywords. Scuola digitale, competenze chiave, trasformazione digitale, aggiornamento docenti, formazione continua

1. La scuola: un ecosistema basato sulla conoscenza

La scuola è, per sua natura un crocevia di conoscenza: il luogo dove convergono persone e organizzazioni con interessi e obiettivi plurali, ma appartenenti alla stessa comunità. Sono studenti, docenti, genitori, imprese, dirigenti e personale amministrativo, organizzazioni del terzo settore, istituzioni. Rivalutare il ruolo sociale e culturale della scuola è quindi una premessa indispensabile per esplorare paradigmi e modelli innovativi nel settore strategico dell'istruzione, capaci di affrontare le sfide poste dalla trasformazione digitale e di coglierne le opportunità [1].

In questa auspicabile accezione, la scuola è un luogo aperto, popolato da svariati attori, saperi e interessi interconnessi, tutti coinvolti nei processi di produzione e fruizione della conoscenza, che la rendono – come indicato dall'art.1 della legge 107/2015 c.d. "buona scuola" (una "comunità attiva, aperta al territorio e in grado di sviluppare e aumentare l'interazione con le famiglie e la comunità locale) [2]. È questa comunità il perimetro entro cui circola la conoscenza veicolata dai diversi attori, si produce e si acquisisce sapere.
Ogni interazione nell'ecosistema produce un "impatto formativo" e ogni attore coinvolto, allo stesso tempo, trae beneficio e contribuisce all'aumento incrementale del capitale di

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

conoscenza dell'ecosistema: una risorsa intangibile di grande valore strategico nella società interconnessa [3]. Risulta evidente quanto l''interdipendenza e l'interazione tra attori "interni" ed "esterni" alla scuola generino complessità – requisito indispensabile per l'innovazione – e implichino l'urgenza di superare i vecchi modelli lineari e cumulativi nell'organizzazione dei saperi, andando oltre le logiche di reclusione delle conoscenze, rileggendole in una moderna prospettiva di condivisione, cooperazione e collaborazione [4].
Innovare la formazione in questa direzione significa quindi abilitare processi di interconnessione, senza semplificare il sistema, ma preservando la sua complessità rendendola invisibile e funzionale al raggiungimento degli obiettivi degli attori in una dinamica win-win. Gestire tale complessità è un obiettivo strategico dipendente in buona parte dalla capacità di elaborare e condividere conoscenze, organizzandole sistematicamente e funzionalmente [5].

In questa nuova visione ecosistemica dove la conoscenza è la risorsa circolante che genera valore attraverso gli scambi, diventando così una forma di capitale, la scuola come istituzione e comunità rappresenta il punto nevralgico della rete di interconnessioni, driver naturale di ogni processo di innovazione basato sulla conoscenza [6]. I benefici per ciascun attore dell'ecosistema derivano dall'ottimizzazione di tali processi di condivisione e produzione della conoscenza che metta a sistema il capitale conoscitivo frutto di ogni esperienza formativa formale, non formale, informale, trasformando un insieme di portatori di interesse in una comunità collaborativa.

2. Un SaaS per abilitare la produzione collaborativa di oggetti formativi evoluti

Ogni processo di produzione di conoscenza all'interna della comunità prevede la presenza di un insegnante che assume ruoli diversi: educatore e docente con gli studenti, tutor per l'alternanza scuola-lavoro, referente delle organizzazioni del territorio, interlocutore delle famiglie e altro ancora. Questa figura è, quindi, determinante per la finalizzazione di nuovi modelli di apprendimento, per il passaggio al nuovo paradigma e per la crescita personale e professionale dei discenti, sempre più individui motivati e capaci di auto-costruire i loro percorsi formativi nel corso di tutta la loro vita [7].

In supporto all'attuale fase di transizione vissuta dagli oltre 700.000 docenti in Italia, ILEARNTV interviene definendo nuovi modelli per la loro formazione, sostenuti da Abylia: un ambiente digitale progettato appositamente e da nuove modalità di veicolazione multicanale di contenuti formativi digitali (Enhanced Learning Object - ELO) mediante pc, tablet e smartphone. La piattaforma consiste di un avanzato sistema di produzione e gestione di contenuti formativi di tipo Software-as-a-Service in cui convergono un'infrastruttura Cloud per la gestione delle risorse hardware e una soluzione software attraverso cui le funzionalità applicative della piattaforma sono implementate e rese fruibili agli utenti finali.

Un digital repository funge da livello sottostante di gestione dei contenuti entro cui il materiale di apprendimento viene indicizzato e reso facilmente disponibile per il riutilizzo. Sopra questo, è costruito un livello applicativo pensato per una creazione collaborativa

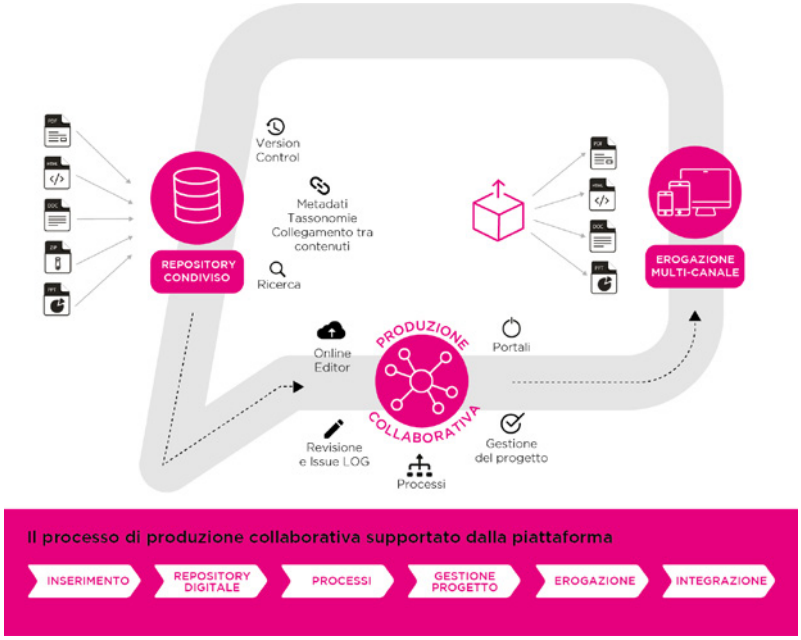


Fig. 1
Processi di produzione collaborativa di Learning Object in Abylia

di contenuti formativi basata su modelli e processi predefiniti e personalizzabili che favoriscano una produzione facile, veloce e coerente di contenuti multi-formato e multi-canale da parte dei docenti. Il ciclo produttivo dei contenuti, ovvero i software necessari, i modelli da applicare e la generazione di contenuti, così come i processi di gestione dei contenuti, dalla progettazione fino al montaggio e verifica, e le figure professionali che ne concorrono alla realizzazione sono state definite e perfezionate con l'obiettivo di sperimentare e diffondere i risultati nelle istituzioni scolastiche.

Conclusioni

In questo modo, il progetto ILEARNTV offre soluzioni cruciali per vincere le sfide dell'innovazione in una realtà complessa e interconnessa come quella scolastica, supportando gli insegnanti nella gestione dei mutamenti impressi dalla trasformazione digitale nella scuola, rendendo uniforme il loro grado di preparazione professionale, abbattendo le spese e i tempi generalmente richiesti dalla formazione tradizionale.

Un repository centralizzato e condiviso (Shared Repository) permette di archiviare, indicizzare, condividere, cercare e riutilizzare con semplicità e rapidità i contenuti didattici inseriti nella piattaforma dagli utenti, opportunamente memorizzati e arricchiti di metadati estratti in maniera automatica nella fase di caricamento (Ingestion). La piattaforma è in grado di manipolare contenuti redatti con i programmi più comunemente utilizzati come presentazioni preparate in Microsoft PowerPoint, documenti scritti su Microsoft Word, file salvati in formato PDF, contenuti didattici salvati nel formato standard SCORM e contenuti multimediali come immagini, audio e video.

Il layout responsive inoltre consente l'accesso ai contenuti in ogni momento e da qualsiasi dispositivo grazie all'interfaccia grafica studiata e sviluppata considerando la variabilità delle esperienze formative digitali per garantire la migliore esperienza d'uso su dispo-

sitivi fissi e mobili in ottica di "ubiquitous learning". [8]

Riferimenti bibliografici

1. Association for Smart Learning Ecosystems and Regional Development ASLERD (2016), Timisoara Declaration: Better Learning for a Better World through People Centred Smart Learning Ecosystems. Url: http://www.mifav.uniroma2.it/inevent/events/aslerd/ docs/TIMISOARA_DECLARATION_F.pdf.

2. Articolo 1, Legge 13 luglio 2015, n. 107: Riforma del sistema nazionale di istruzione e formazione e delega per il riordino delle disposizioni legislative vigenti della legge 107/2015. Url: http://www.gazzettaufficiale.it/eli/id/2015/07/15/15G00122/sg.

3. Charband Y., Navimipour N. J. (2016), Article in Information Systems Frontiers, 18(6), 1131-1151, Online knowledge sharing mechanisms: a systematic review of the state of the art literature and recommendations for future research, Springer, Berlino.

4. Dominici P. (2016), Scuola e istruzione: prerequisiti fondamentali di una cittadinanza matura, attiva e non eterodiretta. Url:https://www.statigeneralinnovazione.it/online/scuola-e-istru-zione-pre-requisiti-fondamentali-di-una-cittadinanza-matura-attiva-e-non-eterodiretta/.

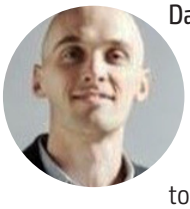
5. Chang, M., Chen, Huang R., Kinshuk, Kravcik M., Li Y., Popescu E., N.-S. (Eds.) (2016), State-of-the-art and Future Directions of Smart Learning, Springer, Berlino.

6. Divitini M., Mealha O., Rehm M. (2017), Citizens, Territory and Technologies: Smart Learning, Springer, Berlino.

7. Asghar H. M., Rashid, T. (2016), Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations, Article in Computers in Human Behavior, 63, 604 – 612, Elsevier, Amsterdam.

8. Gröhbiel U., Mateescu M., Pimmer, C. (2016) Mobile and ubiquitous learning in higher education settings. A systematic review of empirical studies. Article in Computers in Human Behavior, 63, 490-501, Elsevier, Amsterdam.

Autori



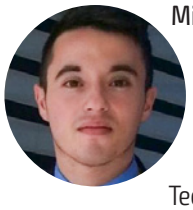
Davide Zucchetti - zucchetti@lattanziokibs.it

Sociologo di formazione, è consulente in processi di innovazione nella PA e nelle PMI. In questa cornice adotta un approccio human-oriented alla progettazione di sistemi complessi e alle trasformazioni di natura organizzativa, tecnologica e di business. Per Abylia Srl è Project Manager del progetto di ricerca ILearnTV. In collaborazione con il Dipartimento di Matematica e Informatica dell'Università di Cagliari, LATTANZIO Advisory Spa, ENEA e NEXERA Spa, partner del progetto, sta lavorando alla sperimentazione di un ambiente digitale che abilita la produzione collaborativa di contenuti e oggettivi formativi, a supporto di processi di apprendimento "aperti" nel paradigma della knowledge school.

Filippo Chesi - chesi@lattanziokibs.com

Da sempre opera nella consulenza di direzione alla pubblica amministrazione e alle im-

prese. Esperto di modelli organizzativi e sistemi di governance, di pianificazione e controllo e di reporting direzionale, nonché di analisi e progettazione di sistemi informativi. In Italia e all'estero, coordina progetti complessi di innovazione gestionale, monitoraggio e valutazione delle politiche pubbliche, in particolare nel settore della giustizia, dello sviluppo di impresa, dell'industria creativa e della pianificazione territoriale. Co-fondatore di LATTANZIO E ASSOCIATI e partner di LATTANZIO Group, ricopre l'incarico di amministratore di diverse società del gruppo.



Mirko Marras - mirko.marras@unica.it

Dottorando in Informatica presso il Dipartimento di Matematica e Informatica dell'Università degli Studi di Cagliari (Italia). Ha ricevuto la Laurea Magistrale in Informatica dalla stessa Università nel 2016. I suoi interessi di ricerca sono focalizzati su Deep Learning, Biometric Systems, Semantic-Aware Systems, Recommender Systems e E-Learning Technologies.

Silvio Barra - silvio.barra@unica.it

Nato nel 1985 a Battipaglia (SA). Riceve nel 2009 la Laurea Triennale (cum laude) e successivamente nel 2012 la Laurea Magistrale (cum laude) in Computer Science presso l'Università degli studi del Salento. Nel 2017 ottiene il Dottorato all'Università degli studi di Cagliari. Attualmente è Professore ricercatore presso la stessa università. E' membro del CVPL (ex GIRDR). I suoi interessi di ricerca riguardano principalmente Pattern Recognition, Video e Biometrics analysis, Analytics.

DOI: 10.26314/GARR-Conf18-proceedings-10

Modelli e strumenti semantici per la documentazione dell'Heritage Science

Lisa Castelli[1], Achille Felicetti[2]

[1]INFN, Istituto Nazionale di Fisica Nucleare, [2]PIN, Università degli Studi di Firenze

Abstract. L'Heritage Science è una branca multidisciplinare delle scienze applicate che mira allo studio, la valorizzazione e la conservazione dei beni culturali. Le analisi di laboratorio condotte da questa disciplina producono informazioni digitali che necessitano di un'opportuna sistematizzazione, nell'ottica di una loro migliore reperibilità e di un loro efficiente riutilizzo. Questa comunicazione illustra un modello di codifica inteso a documentare tali informazioni per mezzo di ontologie, vocabolari e sofisticati strumenti di metadatazione. Il fine ultimo è la creazione di archivi semantici di conoscenza digitali che, uniti ad opportuni strumenti di ricerca, rendano i dati scientifici facilmente reperibili, accessibili e integrabili, in modo da consentirne il potenziale riutilizzo in ambiti diversi.

Keywords. Heritage Science, Ontologie, Modelli di metadati, Conservazione e restauro, Beni culturali

Introduzione

Che si tratti di stabilire la provenienza o la datazione di un reperto archeologico, di pianificare un restauro o la strategia di conservazione di un dipinto o di un monumento, la scienza è ormai divenuta indispensabile per lo studio dei beni culturali. Strumentazioni, tecniche e metodologie nate per la fisica, la chimica, la biologia, sono oggi utilizzate nei modi più svariati per espandere la conoscenza di oggetti e beni artistici (per comprenderne ad esempio la fattura o la provenienza) e per garantire la qualità della loro vita presente e futura. L'Heritage Science è ormai divenuta una branca interdisciplinare della ricerca, che mette in campo l'analisi scientifica fornendo un contributo insostituibile a supporto dello studio dei beni.

Esistono tuttavia problematiche legate alla gestione dei dati prodotti da questo tipo di ricerca che richiedono particolare attenzione. L'uso di strumenti digitali per le analisi scientifiche, ad esempio, determina la creazione di una mole notevole di informazioni digitali estremamente eterogenee (documenti, immagini, modelli 3D, ecc.) che, una volta utilizzate per il fine immediato della ricerca in atto, finiscono spesso per essere archiviate in modo poco accurato, tanto da risultare in seguito difficilmente reperibili, specialmente da ricercatori esterni all'istituzione che le ha prodotte.

Alcune iniziative europee stanno tentando di fornire possibili soluzioni al problema. In particolare, fra quelle finanziate dall'Unione Europea, ARIADNE (www.ariadne-infrastructure.eu), PARTHENOS (www.parthenos-project.eu) e E-RIHS (www.e-rihs.eu) hanno proposto diversi approcci al tema dell'organizzazione organica di dataset digitali e per la loro indicizzazione, soprattutto per mezzo dell'adozione dei principi FAIR, stabiliti per

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

promuovere un utilizzo proficuo dei dati: l'utilizzo di standard internazionali e di tecnologie semantiche sono infatti gli unici strumenti in grado di garantirne l'interoperabilità e la riusabilità. L'utilizzo di ontologie, vocabolari e altri strumenti per la creazione di metadati descrittivi è fondamentale in tal senso per conferire coerenza a informazioni per loro natura estremamente variegate, e per descriverne gli elementi essenziali in modo uniforme, tale da renderle poi facilmente reperibili.

1. Il modello semantico

Il modello qui proposto, sviluppato dalla collaborazione fra PIN, INFN e CNR, tutti partner di ARIADNEplus e E-RIHS, ha lo scopo di creare descrizioni standardizzate dei dataset prodotti dall'Heritage Science, indipendentemente dal loro formato digitale, al fine di indicizzarne e archiviarne i contenuti in modo razionale ed efficiente per renderli accessibili e rintracciabili (anche online) con strumenti di query complessi. Si tratta, in realtà, di un meta-modello concettuale, ovvero un formato di scambio indipendente da ogni piattaforma o software, progettato per descrivere il contenuto dei dataset e le informazioni relative alle persone, le istituzioni, gli eventi e gli strumenti coinvolti nella loro produzione in un formato standard basato su ontologie e vocabolari controllati.

Da un punto di vista tecnico, ciò che viene generato per ogni dataset (per mezzo di interfacce utente o di procedure semi automatiche in grado di attingere a metadati esistenti) è una serie di moduli XML definiti secondo uno schema XSD appositamente predisposto. Lo schema implementa un meta-formato, lineare ed intuitivo sul lato utente, dal quale è poi possibile generare ulteriori contenuti, in formati relazionali anche complessi (Fig. 1), da utilizzare con qualunque archivio o base di conoscenza digitale, inclusi i database SQL e quelli semantici di ultima generazione, quali il DIGILAB di E-RIHS. Il livello più articolato, quello semantico, è implementabile attraverso un meccanismo di mappatura e conversione (già definito nel modello stesso) per la trasformazione dei metadati in formato CIDOC CRM (RDF), l'ontologia di alto profilo scelta per questa sperimentazione (http://cidoc-crm.org). I metadati descrittivi così generati sono consultabili con sistemi di query semantiche e perfettamente integrabili con metadati simili provenienti da ambiti legati all'Heritage Science (archeologia, storia dell'arte, conservazione beni culturali, ecc.). Questo consente la creazione di grafi semantici complessi e interdisciplinari, un risultato che è sempre stato fra gli obiettivi di PARTHENOS e che è alla base della creazione del DIGILAB di E-RIHS.

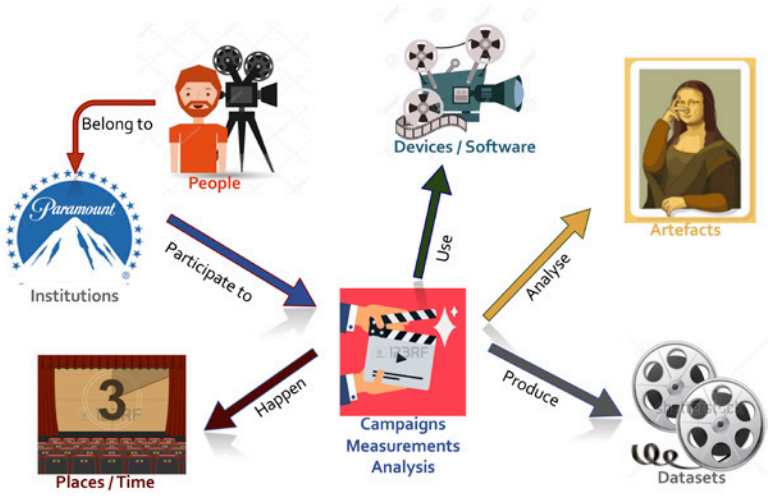


Fig. 1
Moduli e relazioni del modello semantico

2. Analisi e codifica di dati scientifici

Il modello è stato creato dopo un'attenta analisi di sistemi di metadati scientifici esistenti nel panorama internazionale, quali CERIF (http://www.eurocris.org/cerif/main-features-cerif) e OBOE (http://semtools.ecoinformatics.org/oboe), e una vasta ricognizione dei dataset contenenti risultati analitici di alcune importanti istituzioni scientifiche, impegnate nell'Heritage Science ed interessate alla definizione di un sistema di metadati funzionale. In particolare, per la sua creazione sono stati investigati dataset di laboratorio del CNR, dell'INFN, dell'Opificio delle Pietre Dure, dell'ISCR, ponendo attenzione al loro formato di codifica e al metodo di archiviazione utilizzato.

Le entità della prima release del nostro modello sono state utilizzate dai ricercatori INFN in stretta collaborazione con gli esperti informatici del PIN. INFN è attivo nell'ambito della Heritage Science tramite la sua rete per i beni culturali, INFN-CHNet, costituita da circa 20 laboratori diffusi sul territorio italiano, con competenze molto eterogenee nell'ambito della diagnostica sui beni culturali: fra le tecniche applicate e perfezionate ogni giorno dai laboratori di rete possono essere citate per esempio tecniche a raggi X, quali radiografia, tomografia e XRF, datazioni tramite radiocarbonio o termoluminescenza e analisi con fasci di particelle accelerate. La varietà di indagini diagnostiche disponibili presso i laboratori di rete rende i dati acquisiti da CHNet un buon banco di prova per testare l'efficacia del modello proposto.

Test preliminari sono stati eseguiti applicando il modello ad una serie di dataset di CHNet derivanti da analisi diagnostiche di tipo diverso, provando a individuare, per ogni tipo di analisi, quali fossero i metadati rilevanti ai fini di una ricerca successiva nel database. L'esito dei primi test ha dato esito decisamente positivo, tanto da incoraggiare il prosieguo della sperimentazione.

3. Conclusioni

Il modello, pur se ancora in fase di sviluppo, ha dimostrato di essere uno strumento versatile e sufficientemente capace di descrivere dataset scientifici già nella sua forma attuale. I test effettuati hanno reso inoltre possibile una valutazione del suo potenziale e degli sviluppi necessari per una sua ulteriore evoluzione. Nel prossimo periodo, una nuova serie di test verrà effettuata su nuove tecnologie diagnostiche, sia di INFN che di altri laboratori, al fine di consolidare il modello e di ampliarlo con la definizione di nuove entità e relazioni.

Il fine ultimo è la creazione di uno strumento flessibile in grado di catturare l'essenza di ogni tipo di dataset scientifico (di ambito Heritage Science ma non solo) e di fornirne una completa descrizione formale. L'ontologia CIDOC CRM e la sua estensione scientifica CRMsci (http://cidoc-crm.org/crmsci), unita all'utilizzo di strumenti terminologici utilizzati per la codifica dei metadati, consentirà una perfetta integrazione con informazioni dello stesso tipo, generate in altri domini. Particolare attenzione sarà dedicata all'implementazione di interfacce utente user friendly, in grado di garantire un'efficiente raccolta dei dati iniziali e una rapida consultazione degli archivi integrati. L'ultimo importante step riguarderà lo sviluppo della piattaforma digitale e del repository per l'archiviazione dei dati stessi, oltre che dei metadati: un processo che si concretizzerà nel DIGILAB, attualmente in fase di progettazione in E-RIHS.

Riferimenti bibliografici

Niccolucci F. et al. (2017), Condivisione dei dati sui beni culturali: DIGILAB, l'esperienza di ARIADNE e di E-RIHS. Proceedings of GARR Conference, pp.129–132, Venice (2017). DOI: 10.26314/GARR-Conf17-proceedings-24

Pezzati L., Felicetti A. (2017), DIGILAB: A New Infrastructure for Heritage Science, ERCIM NEWS 111, 2017

https://ercim-news.ercim.eu/en111/special/digilab-a-new-infrastructure-for-heritage-science

CHNet – La rete INFN dedicate ai beni culturali: http://chnet.infn.it

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

Niccolucci F., Felicetti A. (2018), A CIDOC CRM-based Model for the Documentation of Heritage Sciences, Paper accepted at Digital Heritage Congress 2018, 26-30 October 2018, San Francisco, USA

Autori

Lisa Castelli - lisa.castelli@fi.infn.it

Lisa Castelli è una fisica impegnata dal 2013 nell'ambito delle tecnologie applicate ai beni culturali, in particolare nello sviluppo e applicazione di strumentazione portatile basata sui raggi X per la caratterizzazione elementale dei materiali nelle opere d'arte. Dal 2015 si occupa della gestione e organizzazione delle attività della rete per i beni culturali dell'Istituto Nazionale di Fisica Nucleare, INFN-CHNet, che raccoglie una ventina di strutture dell'Ente con competenze di vario tipo nell'ambito dell'archeometria.

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 delle estensioni CIDOC CRM CRMarchaeo, per la codifica di dati di scavo, e CRMtex, per l'integrazione di informazioni epigrafiche.

DOI: 10.26314/GARR-Conf18-proceedings-11

Governance dei dati e conformità in materia di privacy: GDPR e ricerca scientifica

Nadina Foggetti

Università degli Studi di Bari A. Moro

Abstract. Con la diffusione di Big data è necessario il ricorso all'uso di approcci di tipo Data lake, ovvero di architetture in grado di avere un "contenitore" distribuito di dati che sostituisca i tradizionali repository.
Si è passati dal modello tradizionale basato sui silos di immagazzinamento di dati ad un approccio di tipo integrato secondo il quale si ricorre all'uso di data lake e di data warehouse che operano in modalità integrata. Al fine di rispondere alle differenti esigenze di storage, gestione e analisi di qualsiasi tipo di dato si rischia di memorizzare i dati in modo non legal-aware.
Obiettivo del lavoro è esaminare gli effetti che le recenti disposizioni fondate sull'accountability, comportano nel quadro della raccolta, analisi e gestione dei dati in ambito scientifico con particolare riguardo ai dati sanitari. Si analizzerà la disciplina vigente per esaminare le problematiche connesse alla governace dei dati, in ambito scientifico anche alla luce della garanzia del principio dell'Open Access.

Keywords. GDPR, Open Access, anonimizzazione, pseudonimizzazione

1. GDPR e ricerca scientifica

Alla luce del GDPR Reg. 2016/679 e del D.Lgs. 101/2018, occorre innanzitutto esaminare cosa debba intendersi per ricerca scientifica, considerato che solo nei confronti di questa particolare categoria di dati la golden rule del consenso può essere derogata. In particolare il GDPR stabilisce che il trattamento di dati personali per finalità di ricerca scientifica dovrebbe essere interpretato in senso lato e includere ai sensi del Cons. 159 sviluppo tecnologico, ricerca fondamentale, ricerca applicata e ricerca finanziata da privati, oltre a tenere conto dell'obiettivo dell'UE di istituire uno spazio europeo della ricerca ai sensi dell'art. 179 TFUE; nonché gli studi svolti nell'interesse pubblico nel settore della sanità pubblica. Il GDPR richiama l'uso aperto dei data set, nel rispetto del principio di libera circolazione, stabilendo che "dovrebbero applicarsi condizioni specifiche, in particolare per quanto riguarda la pubblicazione o la diffusione in altra forma di dati personali nel contesto delle finalità di ricerca scientifica, se il risultato della ricerca scientifica, in particolare nel contesto sanitario, costituisce motivo per ulteriori misure nell'interesse dell'interessato".

Il Regolamento n. 1290 del 2013 avente ad oggetto la condivisione dei risultati della ricerca, effettua una tripartizione rispetto agli obblighi relativi alla diffusione dei risultati della ricerca, tra Comunicazione, Dissemination e exploitation dei dati. La prima è volta a dimostrare alla società come i fondi europei contribuiscano ad affrontare certe sfide. La seconda riguarda un'attività relativa l'accesso e l'uso dei dati raccolti. L'ultima ha l'obiettivo di consentire lo sfruttamento economico della conoscenza sviluppata nel quadro della

Data Sharing Policy per i progetti H2020.

Quando parliamo di sfruttamento dei dati della ricerca scientifica occorre segnare una dicotomia tra i dati pseudonimizzati e quelli anonimizzati. Un principio fondamentale al centro del GDPR è il l'anonimizzazione. Dati anonimizzati sono quei dati privati di tutti gli elementi identificativi e quindi non sono soggetti alle norme a tutela dei dati personali (G. Ziccardi, 2018).

I dati pseudonimi, invece, sono quei dati personali nei quali gli elementi identificativi sono stati sostituiti da elementi diversi, quali stringhe di caratteri o numeri (hash), oppure sostituendo al nome un nickname, purché sia tale da rendere estremamente difficoltosa l'identificazione dell'interessato. I dati pseudonimizzati, a differenza di quelli anonimizzati, sono comunque dati personali. Questa distinzione ha un impatto significativo in relazione al trattamento dei dati aventi natura scientifica.

I primi riguardano le attività di ricerca che rientrano pienamente nel quadro delle deroghe previste dall'articolo 9 GDPR. L'articolo in parola, infatti introduce un divieto di trattamento di particolari categorie di dati personali (tra cui i dati genetici, dati sanitari), disponendo al contempo deroghe specifiche, tra cui quella relativa alla ricerca scientifica, purché siano rispettati il principio di necessità e proporzionalità e siano introdotte misure per tutelare i diritti dell'interessato.

Nell'ipotesi relativa allo sfruttamento dei risultati consistenti in dati anonimizzati, il GDPR non si applica. L'ente destinatario dei dataset dovrà, tuttavia, adottare misure organizzative per tenere distinte le attività di ricerca da quelle commerciali, e quindi anche dataset "fisicamente" separati, per fare in modo che il regime agevolato si applichi ai trattamenti effettivamente svolti per scopo di ricerca.

In questi casi, il titolare del trattamento adotta misure appropriate per tutelare i diritti, le libertà' e i legittimi interessi dell'interessato. Occorre precisare che in queste specifiche ipotesi il programma di ricerca dovrà essere oggetto di motivato parere favorevole del competente comitato etico a livello territoriale e dovrà essere sottoposto a preventiva consultazione del Garante ai sensi dell'articolo 36 del GDPR. La normativa risolve altresì l'annoso problema relativo all'esercizio dei diritti dell'interessato ai sensi dell'articolo 16 GPRD, nelle ipotesi dei trattamenti in parola, stabilendo che la rettifica e l'integrazione dei dati sono annotate senza apportare modifiche tali da pregiudicare il risultato della ricerca. Occorre precisare che l'art.11 e il cons. 57 GDPR si occupano di definire l'impatto sull'analisi dei dati a fini di ricerca scientifica, specificando che "se le finalità per cui un titolare del trattamento tratta i dati personali non richiedono o non richiedono più l'identificazione dell'interessato, il titolare del trattamento non è obbligato a conservare, acquisire o trattare ulteriori informazioni per identificare l'interessato ai soli fini di rispettare il regolamento". Qualora invece, nelle ipotesi disciplinate dal paragrafano 1 dell'art. 11, il titolare del trattamento possa dimostrare di non essere in grado di identificare l'interessato, ne informa l'interessato, se possibile. In tali casi, gli articoli da 15 a 20 non si applicano tranne quando l'interessato, al fine di esercitare i diritti di cui ai suddetti articoli, fornisce ulteriori informazioni che ne consentano l'identificazione". Nel Cons. 57 si precisa che se i dati che tratta non gli consentono di identificare una persona fisica, il titolare del trattamento non

dovrebbe essere obbligato ad acquisire ulteriori informazioni per identificare l'interessato. Tuttavia il titolare del trattamento non dovrebbe rifiutare le ulteriori informazioni fornite dall'interessato al fine di sostenere l'esercizio dei suoi diritti.

L'art. 11, tuttavia segna un terza via tra anonimizzazione e pseudonimizzazione, poichè consente al titolare di adottare, modalità di trattamento di dati "deindicizzati" o comunque privi di modalità di identificazione delle persone, anche se questo comporta la privazione della possibilità di esercitare i loro diritti.

Ai sensi del Cons. 57, un titolare acquisisce dati che sono personali perché riferibili a persone identificate o identificabili, ma non ha interesse a raccogliere ed utilizzare anche gli elementi che possono consentire tale identificazione, in quanto inutili ai fini dei trattamenti che vuole porre in essere. Alla luce di quanto sostenuto è possibile affermare che il Cons. 57 estende l'efficacia dell'art. 11, in quanto non fa riferimento ad un'anonimizzazione, ma non si tratta neppure di una pseudonimizzazione, che richiede specifiche tutele rispetto all'interessato. Questa terza via si inserisce, grazie ad un'esegesi dell'art.11 in combinato disposto con il Cons. 57, tra il concetto di pseudonimizzazione e anonimizzazione dei dati. I dati non sono pseudonimi perchè il titolare non è mai venuto in possesso degli identificativi (Cons, 57) o ne è venuto in possesso ma poi li ha staccati in modo irreversibile, neppure anonimi perchè inizialmente sono stati raccolti come dati personali. L'articolo in parola consetirebbe ai titolari di porre in essere attività di Big Data e Data Analysis utilizzando dati strutturalmente personali che possono consentire trattamenti di norma non sottoposti ad una rigida tutela degli interessati, perché raccolti senza identificativi o perché trattati cancellando ogni indicatore relativo a persone.

2. The right to know

Alla luce del Regolamento 1290 del 2013, nonchè del nuovo articolo 53 del codice dell'amministrazione digitale, che sancisce il principio dell'Open Data by default, il diritto all'acceso ai dati e all'uso dei dati stessi assume un'importanza cruciale. Già il Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (Frank La Rue, 2011) evidenziava che "Internet has become an indispensable tool for realizing a range of human rights, combating inequality, and accelerating development and human progress, ensuring universal access to the Internet should be a priority for all States" e richiedeva agli Stati di adoperarsi al fine di " to support initiatives to ensure that online information can be accessed in a meaningful way by all sectors of the population". L'affermazione del diritto all'accesso (open science) afferente al diritto internazionale pone il problema dell'individuazione della natura giuridica della pretesa informativa e delle posizioni giuridiche ad essa annessa.

La Corte EDU è intervenuta in materia di "right to know" nell'ambito del diritto all'accesso ai dati relativi alla ricerca scientifica nella sentenza Magyar Helsinki Bizottság (Magyar Helsinki Bizottság c. Ungheria). La Corte EDU riconosce definitivamente il diritto di accesso alle informazioni di pubblico interesse, poiché connesso al diritto di ricevere e comunicare informazioni ex art. 10 in quanto 'mezzo' per tutelare altri diritti (M. MCDONAGH, 2013), consolidando la propria giurisprudenza (Bubon c. Russia). A fronte

di questo diritto la Corte non evidenzia l'esistenza di un obbligo giuridico di diffondere informazioni; secondo la Corte, infatti, "the fact that the information requested is ready and available ought to constitute an important criterion in the overall assessment". Sebbene l'approccio esegetico della Corte EDU sia fondamentale sotto il profilo di una ricostruzione della libertà di espressione in termini non meramente negativi, l'interpretazione risulta comunque restrittiva (R. PELED e Y. RABBIN, 2008).

Riferimenti bibliografici

La Rue F. (2011), Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf

McDonagh A. (2013), The Right to Information in International Human Rights Law, Human Rights Law Review', 13, pp. 25-55.

Peled R. - Rabbin Y. 2008, The Constitutional Right to Information, Columbia Human Rights Law Review, 42(2), pp. 357-401.

Sent. Corte EDU, Bubon c. Russia, ric. n. 63898/09, del 7 febbraio 2017.

Sent. Corte EDU Magyar Helsinki Bizottság c. Ungheria, ric. n. 18030/11, dell'8 novembre 2016.

Ziccardi G., (2018) Ciberspazio e diritto. Il GDPR: una nuova era per la protezione dei dati?, Mucchi, Milano.

Autrice

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. Esperto per le PA dei processi di digitalizzazione e Open Data. 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-Conf18-proceedings-12

Soluzioni di Deep Learning per la Cyber Security

Nicola Cannistrà, Fabio Cordaro, Francesco La Rosa, Umberto Ruggeri, Riccardo Uccello

Università degli Studi di Messina

Abstract. Un componente fondamentale di un'infrastruttura di cyber security è il sistema di Network Intrusion Detection (NIDS). Un NIDS viene usato per identificare, analizzando il traffico di rete su nodi chiave, attività malevole volte a violare la confidenzialità, l'integrità e la disponibilità dei dati e dei sistemi. Molti tra i NIDS più moderni fanno uso di tecniche di Machine Learning (ML) o Deep Learning (DL) per la loro capacità di adattamento ad attacchi di rete sconosciuti (zero-day). In questo articolo proponiamo un NIDS basato su una deep neural network già adottata con risultati notevoli in ambiti come quelli della Computer Vision e del Natural Language Processing.

Keywords. Cybersecurity, Machine Learning, Deep Learning, AI, network

Introduzione

Creare difese efficaci contro vari tipi di attacchi di rete e garantire la sicurezza delle apparecchiature di rete e delle informazioni è diventato un problema di particolare importanza. I Network Intrusion Detection System identificano attacchi dannosi analizzando il traffico di rete su nodi chiave e sono diventati una parte importante dell'architettura di cyber security.

Esistono tre tipi di analisi del traffico di rete in base alle quali è possibile classificare i NIDS: misuse-based, anomaly-based e hybrid. Nel caso dell'analisi misuse-based il rilevamento dell'abuso è basato sul confronto di eventi registrati con schemi predefiniti di attacco, che, ad oggi, costituisce l'approccio più utilizzato. Questo sistema di rilevamento delle intrusioni è lento nel generare uno schema, rendendo difficile rilevare efficacemente nuovi tipi di attacco emergenti su Internet. Di contro, essi vengono utilizzati per tipi noti di attacchi senza generare un numero elevato di falsi allarmi.

Le soluzioni anomaly-based si basano su un modello del normale comportamento della rete e del sistema e identificano le anomalie come deviazioni da esso. Sono interessanti per la loro capacità di rilevare attacchi zero-day. Il principale svantaggio delle tecniche anomaly-based è il potenziale elevato tasso di falsi allarmi.

La detection ibrida combina l'approccio misuse-based con quello anomaly-based. Diverse sono le soluzioni proposte (Buczak and Guven, 2016) in cui tecniche di ML sono state applicate al problema dell'intrusion detection (Ford and Siraj, 2014). Il ML si concentra principalmente sulla classificazione e la discriminazione in base a caratteristiche note precedentemente apprese per mezzo di dati raccolti per l'addestramento (learning). Il Deep Learning (LeCun et al., 2015) è un nuovo campo di ricerca che ricade nell'ambito del ML.

La sua motivazione sta nella creazione di una rete neurale che simula il comportamento del cervello umano nell'apprendimento analitico. I metodi di DL si possono classificare in metodi di apprendimento supervisionato e apprendimento non supervisionato.

Il vantaggio del DL, rispetto alle soluzioni tradizionali di ML (Meshram and Haas, 2017), è la possibilità di un apprendimento non supervisionato o semi-supervisionato che consente l'estrazione automatica (ed efficiente) di feature (Fiore et al., 2013). Gli algoritmi di DL (Alrawashdeh and Purdy, 2016) comunemente utilizzati includono ad esempio gli Autoencoders, le Belief Networks (DBMs), le Convolutional Neural Networks (CNNs) e le Long Short-Term Memory networks (LSTMs). Di recente sono stati proposti diversi NIDS basati su Convolutional Neural Networks (Wang et al., 2018) che hanno mostrato una precisione nel rilevamento (detection rate) degli attacchi superiore a quella degli approcci precedenti. In questo articolo proponiamo un anomaly-based NIDS che fa uso di una Convolutional Neural Network. I risultati sperimentali mostrano che la nostra soluzione supera approcci analoghi, presenti in letteratura, in termini di accuratezza, Detection Rate (DR) e False Acceptance Rate (FAR).

1. Il Sistema

1.1 Convolutional Neural Networks

Le Convolutional Neural Networks (CNNs) sono un tipo specializzato di rete neurale adatta al processamento di dati sotto forma di matrice, come serie temporali e immagini. La tipica architettura di una CNN (LeCun et al., 2015) consiste di un layer di input ed uno di output completati da diversi layer nascosti.

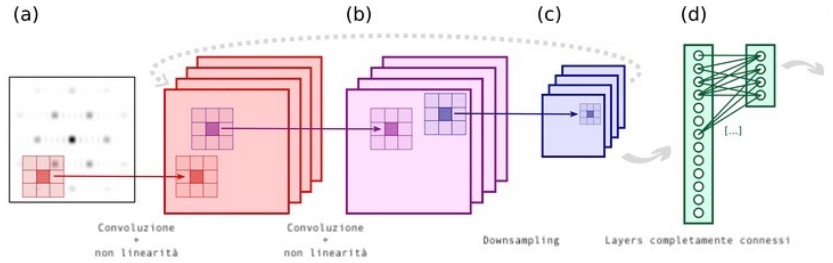


Fig. 1
L'architettura di una Convolutional Neural Network

I layer nascosti sono sia dei convolutional layer che dei pooling layer o dei layer completamenti connessi. Un'architettura tipica delle CNNs è mostrata in Fig. 1. Il convolutional layer costituisce il blocco principale di una CNN e i suoi parametri coincidono con i coefficienti dei filtri (o kernel) che verranno applicati sull'immagine d'ingresso (o matrice). Un altro elemento tipico delle CNNs è il pooling, che è una forma di down-sampling non lineare. Lo strato di pooling serve a ridurre le dimensioni spaziali della rappresentazione, a ridurre il numero di parametri e quindi a limitarne l'overfitting. Strati completamente connessi, infine, connettono ogni neurone di uno strato a ciascun neurone nello strato successivo, come in una rete neurale multi-strato tradizionale (MLP).

1.2 L'architettura proposta

Il classificatore, che ha come core la CNN, è stato implementato usando Tensorflow (Abadi et al., 2016). La soluzione proposta è caratterizzata da 24 canali generati dall'applicazione di altrettanti kernel di dimensione 3*3 e con un passo pari a 1. La matrice, da cui si ottengono i canali, contiene i dati raw acquisiti dalla rete ed ha una dimensione fissa di 28*28 elementi (784 bytes). La CNN è composta da due convolutional layer, 2 layer di pooling (con un 2*2 max-pooling) ed una rete completamente connessa con 3 strati nascosti e 5 uscite. Per il training si è usato l'Adam optimizer su batch di 60 esempi con una funzione di costo cross-entropy. Si sono adottati, inoltre, un learning rate ed un training time rispettivamente pari a 0.002 e 50 epoche.

2. Misure Sperimentali

Per favorire il confronto tra la soluzione proposta e altre presenti in letteratura, si è realizzato un benchmark basato sul dataset DARPA98 (Graf et al., 1998). Nel dataset DARPA98, al traffico normale, sono sovrapposti attacchi di rete appartenenti a 4 famiglie diverse: DoS, Probe, R2L e U2R (vedi tab.1). Per valutare le prestazioni della soluzione proposta, abbiamo adottato tre metriche, accuratezza, DR e FAR, comunemente utilizzate nel campo dell'intrusion detection. L'accuratezza è utilizzata per ottenere una stima complessiva delle prestazioni del NIDS. La DR è usata per stimare le prestazioni del sistema rispetto all'identificazione degli attacchi. Il FAR, invece, permette di quantificare gli errori di classificazione in caso di traffico normale. La definizione di ciascuna delle metriche adottate è riportata in eq. 1. Con riferimento alla eq. 1, dato X come evento attacco e non-X come evento traffico normale, potremo indicare con TP (true positive) il numero di istanze correttamente classificate come X, con TN (true negative) il numero di istanze classificate correttamente come non-X, con FP (false positive) il numero di istanze classificate erroneamente come X e con FN (false negative) il numero di istanze classificate erroneamente come non-X.

$$ACC = \frac{TP+TN}{TP+FP+FN+TN} \quad DR = \frac{TP}{TP+FN} \quad FAR = \frac{FP}{FP+TN} \quad (1)$$

Categoria	Pattern di attacco
Dos	back, land, neptune, pod, smurf, teardrop
R2L	ftp-write, guess-passwd, imap, multihop, phf, spy, warezclient, warezmaster
U2R	buffer-overflow, loadmodeule, perl, rootkit
Probe	ipsweep, nmap, portsweep, satan

Tab. 1
Categorie
degli attacchi

Dataset	Acc%	DR%	FAR
Dos	99.62	99.23	0.03
Probe	99.30	83.43	0.02
R2L	99.75	75.1	0.03
U2R	99.98	67.2	0.03
Totale	99.72	97.82	0.08

Tab. 2
Prestazioni
del NIDS

In tab.2 sono riportati accuratezza, DR e FAR riscontrati per ciascuna tipologia di attacco. Dai risultati ottenuti è emersa una "buona" classificazione per quasi tutte le categorie di attacchi oggetto di questa sperimentazione. Da un'analisi più approfondita dei risultati ottenuti è emerso che una parte degli attacchi di DDoS sono stati "confusi" con del traffico normale. La ragione di tale risultato sta nel fatto che alcuni flussi di rete relativi a DDoS sono molto simili a quelli del traffico normale.

3. Conclusioni

In questo articolo abbiamo proposto un Network Intrusion Detection System basato su una Convolutional Neural Network. I risultati sperimentali evidenziano le prestazioni del sistema in termini di accuratezza, DR e FAR. Due sono i problemi che richiedono un ulteriore approfondimento. Il primo problema consiste nel migliorare le prestazioni della rete su dataset sbilanciati (il traffico dovuto ad attacchi è piccolo rispetto al traffico normale). Il secondo consiste nella possibilità di migliorare le prestazione del NIDS proposto combinando i dati (raw) raccolti con feature tradizionali (Wang et al., 2018).

Riferimenti bibliografici

Alrawashdeh K., Purdy C. (2016), Toward an online anomaly intrusion detection system based on deep learning, 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 195–200.

Buczak A., Guven E. (2016), A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys Tutorials, (18), pp. 1153–1176.

Abadi M. et al. (2016), Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Fiore U., Palmieri F., Castiglione A., De Santis A. (2013), Network anomaly detection with the restricted boltzmann machine, Neurocomput., (122), pp. 13–23.

Ford V., Siraj A. (2014), Applications of machine learning in cyber security, (10).

Graf I., Lippmann R., Cunningham R., Fried D., Kendall K., Webster S., Zissman, M., (1998). Results of DARPA 1998 offline intrusion detection evaluation. https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-data-set.

LeCun Y., Bengio Y., Hinton G. (2015), Deep learning, (521), pp. 436–444.

Meshram A., Haas C. (2017), Anomaly detection in industrial networks using machine learning: A roadmap. Machine Learning for Cyber Physical Systems, pp. 65–72.

Wang W., Sheng Y., Wang J., Zeng X., Ye X., Huang Y., Zhu M. (2018), Hast-ids: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection, IEEE Access, (6), pp. 1792–1806.

Autori

Nicola Cannistrà - nicola.cannistra@unime.it
APM-GARR per l'Università di Messina. Responsabile dell'U.Op. "Infrastrutture ICT" dell'Università degli Studi di Messina, ha sviluppato competenze in ambito sistemistico, progettazione e gestione reti in ambito MAN, sistemi di WiFi centralizzato, sistemi VoiP, sicurezza e sistemi di autenticazione.

Fabio Cordaro - fabio.cordaro@unime.it
Vice responsabile dell'Unità Operativa "Infrastrutture ICT" presso l'Università degli Studi di Messina, Technical Contact per il dominio della stessa università, ha sviluppato competenze nell'amministrazione di sistemi Unix-like, progettazione e gestione reti, amministrazione di servizi di rete, implementazione di applicativi web.

Francesco La Rosa - francesco.larosa@unime.it
Ha conseguito un dottorato di ricerca in Computer Science c/o l'Università degli Studi di Messina. E' coautore di decine di articoli pubblicati su atti di convegno e riviste internazionali. Negli ultimi anni ha ricoperto ruoli di responsabilità c/o il CIAM (Centro Informatico Ateneo di Messina), Università degli Studi di Messina.

Umberto Ruggeri - umberto.ruggeri@unime.it
Laureato in Ingegneria Elettronica presso l'Universita degli Studi di Messina. Responsabile dell'Unità Organizzativa Sistemi ed Infrastrutture ICT presso l'Universita degli Studi di Messina. Ha sviluppato competenze in ambito sistemistico, virtualizzazione di sistemi, progettazione di reti, sicurezza e sistemi di autenticazione.

Riccardo Uccello - riccardo.uccello@unime.it
Laureato in Fisica presso l'Universita degli Studi di Messina. Ha ricoperto negli ultimi anni ruoli di responsabilità nel settore dell'ICT, prima presso l'Universita Mediterranea ultimamente presso il Centro Informatico dell'Università degli Studi di Messina. Ricopre il ruolo di APA-GARR per l'Ateneo messinese.

DOI: 10.26314/GARR-Conf18-proceedings -13

Un Approccio di gestione e prevenzione per l'agricoltura di precisione

Francesca Maridina Malloci

Università degli Studi di Cagliari

Abstract. Negli ultimi anni, i Big Data ricoprono un ruolo sempre più importante nell'agricoltura di precisione. L'analisi delle immagini satellitari, dei dati meteorologici e dei dati agronomici attraverso un modello di previsione, aiuta gli agricoltori a migliorare i raccolti e a ridurre i costi. I Big Data rappresentano un'opportunità significativa per migliorare il processo agricolo. La difficoltà è riuscire a gestirli e analizzarli al fine di fornire linee guida che aiutino gli agricoltori a prendere decisioni accurate. Il presente lavoro, si propone di trattare un approccio di gestione e prevenzione che tramite diversi modelli matematici, analizza una massiccia quantità di dati per fornire un adeguato supporto alle operazioni in situ.

Keywords. Big Data, Big Data Analytics, Precision Farming

Introduzione

L'Agricoltura è stata da sempre l'attività più importante per gli esseri umani sin dai tempi antichi. Recentemente, abbiamo assistito all'impiego di pratiche agricole intensive che hanno portato gradualmente a un degrado ambientale. L'uso massiccio di fertilizzanti chimici coadiuvato dall'impiego di pesticidi, a lungo andare hanno degradato il suolo e portato all'insorgenza di nuovi parassiti e malattie che ancora oggi influenzano la produzione agricola. Negli ultimi anni, per far fronte alle suddette emergenze, stiamo assistendo ad una forte spinta innovativa nel settore agricolo verso un'Agricoltura più sostenibile, biologica e al contempo più produttiva. Tale esigenza non riguarda solo la produttività, la competitività e le prestazioni ambientali, ma secondo le stime della FAO la popolazione mondiale entro il 2050 raggiungerà i 9,2 miliardi di persone e ciò comporterà un conseguente aumento della richiesta alimentare del 70%. Si può facilmente dedurre che l'Agricoltura si trova a dover affrontare un cambiamento radicale nel processo produttivo che ha adottato sino ad oggi.

In tale contesto, l'integrazione tecnologica nel settore agricolo fornisce una soluzione alle necessità di innovazione, in cui l'Agricoltura di Precisione (AP) è oggi lo strumento che certamente risponde alle esigenze emerse negli ultimi decenni.

L'AP è stata definita negli anni Novanta come approccio alla gestione del processo produttivo agricolo che consente di "fare la cosa giusta, nel posto giusto, al momento giusto, con la giusta quantità". Pertanto, la AP basa la sua applicabilità sull'uso delle tecnologie per rilevare e decidere cosa è "giusto". Quest'ultima è una delle tante definizioni che sono state proposte sino ad oggi.

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

I metodi dell'AP si basano principalmente sull'uso congiunto di diverse tecnologie: telerilevamento (RS), sistema di informazione geografica (GIS), sistemi di posizionamento (GPS) e il controllo dei processi.

Finora gli agricoltori hanno preso le decisioni sulla base delle loro esperienze e intuizioni. Tuttavia, quest'ultime sono insufficienti per prevedere un processo decisionale a lungo termine, che potrebbe migliorare la produttività ed evitare costi inutili legati alla raccolta, all'uso di pesticidi e fertilizzanti. Tali obiettivi possono essere realizzati tramite applicazioni di Big Data Analytics. Un'analisi corretta dei dati provenienti da varie fonti consente di misurare le variazioni inter ed intra campo al fine di adoperare interventi selettivi e sistematici nei processi agricoli.

Importanti investimenti da parte di numerose agenzie di ricerca e governi di tutto il mondo stanno incentivando i ricercatori e le aziende nello sviluppo di sistemi di supporto decisionale che facciano uso dei Big Data. Nel 2018, sono state pubblicate le linee guida per lo sviluppo dell'AP in Italia in linea con il Piano strategico per l'innovazione e la ricerca nel settore agricolo alimentare e forestale 2014-2020 (PSIR). A seguito dei fondi stanziati, sono nati numerosi progetti a livello nazionale, quali Life-Agricare, Life-HelpSoil, ERMES e MED-GOLD che hanno coinvolto diverse regioni italiane. Tra le regioni italiane maggiormente impegnate nella digitalizzazione agricola vi è la Sardegna.
Il presente lavoro descrive un sistema predittivo di integrazione dati multidimensionali al fine di prevenire e mantenere la salute delle colture, aumentare la produttività e ridurre i costi delle operazioni agricole.

La sezione I fornisce una panoramica dello stato dell'arte dei Big Data nell'ambito dell'AP. La sezione II descrive l'architettura e le funzionalità del sistema.

1. Big Data nell'Agricoltura di Precisione

I Big Data rappresentano un patrimonio informativo così elevato in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici specifici per la loro trasformazione in valore. In generale, il termine si riferisce a un'ampia varietà di dati strutturati, semi strutturati e non strutturati con un potenziale nascosto esplorabile attraverso algoritmi dinamici, che consentono di conoscere, comprendere e valutare la relazione, i modelli e le tendenze rilevati dai dati al fine di rendere le decisioni più accurate.

Il settore dei Big Data Analytics rappresenta uno strumento di lavoro sempre più importante nell'agricoltura moderna. Esso, a partire dai dati è in grado di estrarre informazioni e conoscenze utili atti a risolvere e semplificare i processi agricoli. I dati nell'AP sono generati dall'agricoltura in loco, dal telerilevamento o dall'agricoltura satellitare. Sono raccolti sotto forma di tipo strutturato e non strutturato. I set di dati dell'AP si riferiscono a crop patterns, crop rotations, weather parameters, environmental conditions, soil types, soil nutrients, Geographic Information System (GIS) data, Global Positioning System (GPS) data, farmer records, agriculture machinery data, yield monitoring e Variable Rate Fertilizers (VRT). In generale, i set di dati riguardanti l'AP sono classificati in: Historical Data, Agricultural Equipment's and Sensor Data, Social and Web Based Data, Publications, Streamed Data, Business Industries and External Data.

Un'accurata integrazione dei dati consente di prevedere e consigliare le decisioni future agli agricoltori e alle aziende agricole.

2. Architettura

Il sistema di integrazione dati multidimensionali si occupa di raccogliere, elaborare e trasferire, con un flusso continuo, diversi tipi di dati e informazioni, i quali vengono interpretati tramite diversi modelli matematici per fornire un adeguato supporto alle operazioni in situ. Esso, è stato sviluppato per l'Agenzia Laore Sardegna. In particolare, l'Agenzia Laore si occupa di fornire servizi di consulenza, informazione, formazione e assistenza nel settore agricolo regionale. L'Agenzia, cura la pubblicazione di notiziari fitosanitari finalizzati alla segnalazione delle avversità a carico delle principali colture e all'adozione delle più idonee strategie di difesa. Per fornire un'adeguata assistenza tecnica, essi avevano la necessità di supportare le loro attività tramite l'ausilio di uno specifico sistema informatico che fosse in grado di analizzare differenti tipi di dato (climatici, colturali, ambientali e geolocalizzati) e suggerire delle linee guida precise da adoperare in campo.
Per conseguire il suddetto obiettivo e garantire un servizio ottimale, in linea con le direttive del Mipaff, il processo è articolato in tre fasi: Data Collection, Data Analysis, Data Evaluation (Figura 1).

Fig. 1
Le fasi che costituiscono l'Agricoltura di Precisione

Il sistema è stato sviluppato seguendo il modello architetturale REST di tipo Client-Server. Il Server REST rappresenta il fulcro funzionale dell'intera piattaforma. Si occupa di gestire i dati ricevuti dall'applicativo web e dai database esterni ed eseguire e gestire il carico computazionale degli algoritmi matematici atti all'elaborazione dei dati. Esso, effettua con un flusso continuo, l'integrazione di dati multidimensionali provenienti da diverse fonti, Nello specifico, l'acquisizione dati avviene dalle seguenti tecnologie: sistemi satellitari, sistemi GPS, sistemi GIS, stazioni agro-meteorologiche e data services online. Ogni dato raccolto da ciascuna delle suddette fonti presenta il proprio formato e/o semantica. Una volta terminata la fase di acquisizione, il sistema memorizza i dati ricevuti e tramite diver-

si modelli matematici effettua un'analisi incrociata e dinamica di quest'ultimi. La Figura 2 fornisce graficamente il work flow del sistema.

Fig. 2
Work flow
del sistema

La base di partenza dell'intero sistema è il monitoraggio della produzione. La prima raccolta dati è rappresentata dalla tecnica del telerilevamento, la quale impiega speciali sensori in grado di misurare e registrare l'energia elettromagnetica riflessa o emessa dalla superficie dell'areale indagato evitando di effettuare delle operazioni invasive. Tale operazione, è svolta dalla fotocamera multispettrale posta sul satellite (1), la quale invia i dati acquisiti al Web GIS dell'Agenzia Laore (2).

Il Web GIS dell'Agenzia Laore, analizza le immagini tramite differenti e molteplici indici di vegetazione (e.g. NVDI, Normalized Difference Vegetation Index) e in generale, funzioni matematiche che consentono di ricavare delle informazioni riguardanti la variabilità spaziale, nello specifico le proprietà del terreno e lo stato colturale delle superfici indagate. Il prodotto finale dell'elaborazione è rappresentato dalle mappe tematiche. La sovrapposizione di più mappe congiunte ai dati derivanti dall'analisi del suolo permette di ottenere delle mappe di prescrizione che individuano delle zone omogenee fra loro per caratteristiche e proprietà produttive. Tuttavia, esse non forniscono un aiuto preciso per la formulazione degli interventi colturali da adoperare in situ. Per tale motivo, il sistema (3) integra i suddetti dati con i dati agro-metereologici provenienti dalle stazioni meteorologiche e i dati agro-fenologici inseriti dai tecnici Laore tramite l'applicativo web (4). Dall'applicativo web arrivano i dati scaturiti dai rilievi in campo delle diverse colture costantemente monitorate dall'Agenzia. I dati inseriti riguardano la fenologia, le catture e l'infestazione.
Una volta acquisiti tutti i dati sopracitati, il passo successivo consiste nel tradurre in termini operativi e gestionali le differenze individuate nelle varie zone omogenee tramite l'utilizzo di modelli previsionali. Nel sistema sono stati mutuati diversi modelli previsionali che per ogni singola coltura permettono di:

- prevedere il rischio infettivo,

- prevedere le fasi fenologiche,
- prevedere lo stato di salute della coltura,
- fornire il grado di salute e vigoria della coltura,
- calcolare la dose ottimale di prodotto fitosanitario da distribuire in dose variabile.

Terminata la fase di elaborazione e interpretazione della variabilità in campo, i risultati ottenuti vengono inviati all'applicativo web. I tecnici Laore, dopo aver valutato gli interventi e le strategie da seguire, inviano quest'ultimi agli operatori in campo che tramite delle macchine agricole dotate di una centralina a bordo e sensori GPS e macchinari con tecnologia TAV applicano in modo diversificato gli input chimici e biologici.

Conclusioni

Nel presente lavoro, è stato proposto lo sviluppo di un sistema, a carattere predittivo e strategico, che grazie all'analisi incrociata di fattori ambientali, climatici e colturali consente di monitorare gli areali di competenza dell'Agenzia Laore Sardegna. Per conseguire tale obiettivo e garantire un servizio ottimale, il sistema per il monitoraggio e la previsione agro-metereologica è stato realizzato seguendo il modello architetturale REST di tipo Client-Server, il cui utilizzo ha permesso di fornire un servizio strutturato e maggiormente manutenibile. L'approccio proposto consente agli agricoltori e ai tecnici Laore di identificare le migliori strategie e interventi da adoperare in situ. Il suo impiego permette di beneficiare dei vantaggi derivanti dall'uso di tecnologie che impiegano i Big Data per trasformare le informazioni in linee guida accurate che consentono di procedere con un'agricoltura più precisa, sostenibile ed economica.

Bibliografia

P. Steduto, T. Hsiao, E. Fereres e D. Raes, «Crop Yield Response to Water» FAO Irrigation and Drainage, p. 500, 2012.

R. D. Grisso, M. M. Alley, P. McClellan, D. E. Brann, and S. J. Donohue, Precision Farming. A Comprehensive Approach, 2009. doi: 10919/51373

Precision agriculture – An opportunity for EU farmers - Potential support with the CAP 2014-2020.

Pierce, F.J., Nowak, P., 1999. Aspects of Precision Agriculture. Adv. Agron. 67, 1–86. doi: 10.1016/S0065-2113(08)60513-1

Zhang, N., Wang, M., Wang, N., 2002. Precision agriculture—a worldwide overview. Comput. Electron. Agric. 36, 113–132. doi:10.1016/S0168-1699(02)00096-0.

Lamrhari, Soumaya et al. "A profile-based Big data architecture for agricultural context." 2016 International Conference on Electrical and Information Technologies (ICEIT) (2016): 22-27. doi: 10.1109/EITech.2016.7519585.

S. K. Seelan, S. Laguette, G. M. Casady, and G. A. Seielstad, Remote sensing applications for precision agriculture: A learning community approach, Remote Sens. Environ., vol. 88, no. 1–2, pp. 157–169, Nov. 2003. doi: 10.1016/j.rse.2003.04.007

M. Neményi, P. á. Mesterházi, Z. Pecze, and Z. Stépán, The role of GIS and GPS in precision

farming, Comput. Electron. Agric., vol. 40, no. 1–3, pp. 45–55, Oct. 2003. doi: 10.1016/S0168-1699(03)00010-3

R. D. Grisso, M. M. Alley, and G. E. Groover, Precision Farming Tools. GPS Navigation, 2009. doi: 442/442-501/

Hirafuji, Masayuki. “A Strategy to Create Agricultural Big Data.” 2014 Annual SRII Global Conference (2014): 249-250. doi: 10.1109/SRII.2014.43

De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016). "A Formal definition of Big Data based on its essential Features". Library Review. 65: 122–135. doi:10.1108/LR-06-2015-0061

B. Brisco, R. Brown, T. Hirose, H. McNairn, and K. Staenz, “Precision agriculture and the role of remote sensing: a review,” Canadian Journal of Remote Sensing, vol. 24, no. 3, pp. 315–327, 1998. doi:10.1080/07038992.1998.10855254

S. W. Searcy, Precision farming: A new approach to crop management. Texas Agricultural Extension Service, Texas A & M University System, 1997.

M. R. Bendre, R.C. Thool, V. R. Thool, Big Data in Precision Agriculture: Weather Forecasting for Future Farming, 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4-5 September 2015. doi: 10.1109/NGCT.2015.7375220

Autrice

Francesca Maridina Malloci - francescam.malloci@unica.it

Francesca Maridina Malloci è dottoranda in Matematica e Informatica presso l'Università degli Studi di Cagliari. Si interessa dei settori di Data Analysis, Prediction Algorithms e Networking. È inserita nel progetto di prototipizzazione di un modello di supporto alla Precision Farming in collaborazione con l'Agenzia LAORE Sardegna.

DOI: 10.26314/GARR-Conf18-proceedings-14

Qualità dell'aria: algoritmi di machine learning applicati a dati simulati dal sistema modellistico AMS-MINNI

Angelo Mariano[1], Mario Adani[2], Gino Briganti[3], Mihaela Mircea[2]
in rappresentanza del gruppo MINNI[4]

[1]ENEA laboratorio DTE-ICT-IGEST, Bari, [2]ENEA laboratorio SSPT-MET-INAT, Bologna, [3]ENEA laboratorio SSPT-MET-INAT, Pisa, [4]www.minni.org

Abstract. L'applicazione di algortmi di machine learning nell'ambito della qualità dell'aria richiede la disponibilità di una grande mole di dati per individuare trend stagionali e realizzare algoritmi predittivi che consentano di simulare l'effetto delle molteplici fonti di inquinamento atmosferico. In questo studio sono stati utilizzati il dataset ufficiale nazionale relativo alla simulazione per l'anno 2010, prodotto mediante il sistema modellistico della qualità dell'aria AMS-MINNI e raccolto negli anni sui sistemi di storage distribuito dell'ENEA, e osservazioni estratte dal database BRACE. A questi dati sono stati applicati modelli di deep learning per predire le concentrazioni orarie di alcuni inquinanti. I risultati di questa sperimentazione sono oggetto del presente contributo e suggeriscono una nuova via di ricerca nell'integrazione dei sistemi modellistici della qualità dell'aria con modelli di machine learning.

Keywords. Machine Learning, Deep Learning, Modellistica computazionale, Qualità dell'aria

Introduzione

L'obiettivo di questo studio è verificare quanto il machine learning e le reti neurali "deep" siano in grado di estrarre le caratteristiche ("features") più rilevanti e di dedurre le correlazioni che permettano di predire le concentrazioni di determinati inquinanti per un set di dati simulati con il sistema modellistico AMS-MINNI e le osservazioni estratte dal database BRACE, relativi all'anno 2010. Il sistema modellistico atmosferico (AMS) del modello nazionale MINNI (Modello Integrato Nazionale a supporto della Negoziazione Internazionale sui temi dell'inquinamento atmosferico [1]; http://www.minni.org) è stato sviluppato nell'ambito dell'omonimo progetto avviato da ENEA nel 2002 e finanziato dal Ministero dell'Ambiente e della Tutela del Territorio e del Mare (MATTM). AMS-MINNI è utilizzato per fare simulazioni modellistiche della qualità dell'aria sull'Italia in adempimento del D.Lgs. 155/2010 (art. 22, comma 5) che ha recepito la Direttiva Europea 2008/50/EC.

I dati simulati con AMS-MINNI per l'anno 2010 [2] ad una risoluzione spaziale di 4 km, occupano attualmente circa 20TB di storage. La simulazione della qualità dell'aria richiede una quantità rilevante di risorse computazionali per coprire l'intero territorio nazionale in alta risoluzione ed è stata eseguita con l'ausilio dell'infrastruttura di supercalcolo ENEA CRESCO di Portici [3], inclusa nella griglia computazionale ENEA Grid, distribuita su 6 centri di ricerca connessi dalla rete GARR.

Le reti neurali deep (DNN) e gli algoritmi che le caratterizzano non sono un concetto com-

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

pletamente nuovo, in quanto già teorizzate circa 20 anni fa, ma solo recentemente sono diventate uno strumento molto efficace e in via di forte espansione grazie all'estrema abbondanza di dati digitali, di tecnologie evolute per la gestione dello storage e per il computing, oltre che per ottimizzazioni sopravvenute sugli algoritmi stessi [4]. Ultimamente, il deep learning è la tecnologia alla base di moltissime applicazioni di intelligenza artificiale, come le auto a guida autonoma, la riproduzione del parlato e dei suoni, il riconoscimento facciale e gli assistenti virtuali. Sicuramente si può ritenere che una delle potenzialità più sorprendenti delle reti DNN sia quella di trovare una rappresentazione semplificata delle features iniziali e di stabilire le relazioni tra le grandezze di input tabula rasa da un'ampia mole di dati, grazie ad un'elevata potenza computazionale e alla presenza di big data che anche esperti del dominio sono in grado di interpretare solo parzialmente. In particolare, questi algoritmi con le loro caratteristiche di predittori possono trovare un ampio impiego nel settore relativo all'analisi della qualità dell'aria, grazie alla possibilità di integrare grosse moli di dati provenienti da fonti diverse e di evidenziare le correlazioni tra le condizioni meteo, le condizioni orografiche, le emissioni di inquinanti.

1. Metodi e strumenti

1.1 Descrizione del modello AMS-MINNI

Il sistema modellistico AMS-MINNI è composto principalmente da un modello meteorologico RAMS, un post-processore diagnostico SURFPRO che stima le variabili micrometeorologiche usando come input i dati meteorologici prodotti da RAMS, un processore per le emissioni antropiche (Emission Manager) e un modello di trasporto chimico FARM.

Le componenti del sistema modellistico (Figura 1) e la simulazione eseguita per il 2010 è descritta in dettaglio in [2].

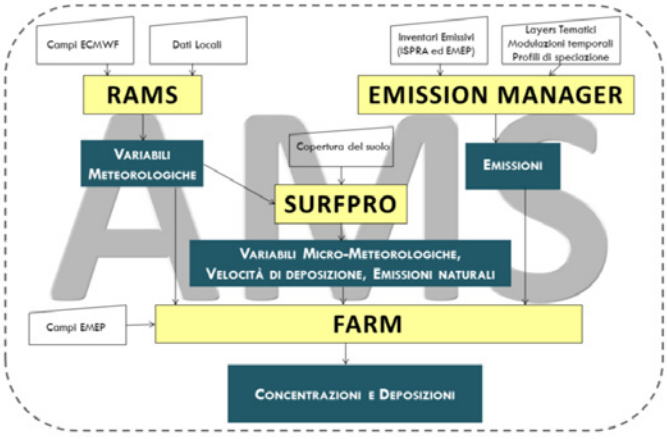


Fig. 1
Rappresentazione schematica di AMS-MINNI

1.2 Descrizione del modello predittivo Machine Learning (ML)

Il modello Machine Learning utilizzato si basa sull'uso estensivo delle celle Long Short-Term Memory (LSTM) [5]. Si tratta di reti neurali "deep" ricorrenti, cioè in grado di analizzare in particolare le dipendenze delle serie temporali o delle sequenze di segnali. Una semplice unità LSTM è composta da una cella, un gate di input, un gate di output ed un

gate di aggiornamento della memoria. La cella è attraversata da una sorta di nastro trasportatore che rappresenta la memoria a breve termine, nella quale le sequenze di segnali vengono scritte o cancellate nel tempo. I compiti nei quali le celle LSTM eccellono sono legati alla classificazione, al processamento e alla formulazione di predizioni nel caso delle serie temporali, dal momento che ci possono essere intervalli di lunghezza non specificata nella ricorrenza di eventi importanti. Le unità LSTM sono state sviluppate per affrontare il problema del gradiente che si azzera durante la correzione a ritroso della rete neurale ricorrente tradizionale (RNN). L'insensibilità rispetto all'intervallo temporale nella ripetizione è uno dei vantaggi delle celle LSTM rispetto alle RNN e più in generale agli altri algoritmi di machine learning applicati alle serie temporali.

La serie temporale di riferimento è la sequenza delle concentrazioni orarie di NO2 misurate da rilevatori disposti in tre punti geografici in Emilia Romagna: sito A non urbano nel comune di Besenzone (PC), sito B urbano nel comune di Piacenza, Giordani-Farnese, sito C urbano nel comune di Piacenza, Parco Montecucco, 24 ore su 24 nell'arco del 2010. Per l'analisi di base del comportamento della serie temporale si è fatto riferimento alle procedure di forecast per le serie temporali Prophet [6], che si basa su un modello additivo in cui si identificano i trend non lineari su base annuale, mensile, settimanale, giornaliera ed oraria.

Le variabili di input considerate sono state le features non nulle che di solito il modello AMS-MINNI utilizza in input, rilevate in un range di 2 ore indietro nel tempo, in accoppiata con la concentrazione di NO2 misurata dalla centralina nella stessa ora. La topologia del modello ML per la predizione delle concentrazioni orarie di NO2 ha una struttura composta da un layer di 50 celle LSTM e un layer denso che fornisce la predizione sulla concentrazione di biossido di azoto in un determinato sito. Le celle LSTM quindi processano e identificano al loro interno le sequenze temporali estese fino a 2 ore indietro nel tempo e le correlazioni tra le features del modello AMS-MINNI e la concentrazione di biossido di azoto a parità di ora. Per evitare l'overfitting dei dati di input, all'interno della rete neurale deep si è inserita la funzionalità del dropout, ossia la rottura random dei collegamenti fra i nodi ad ogni livello della rete nel 20% dei casi. Il dataset di input, su questi siti specifici è composto da 1GB di dati ed è stato così suddiviso: training il primo 70%, test il secondo 20%, validazione il restante 10%.

Una prima sperimentazione è stata effettuata prendendo in considerazione un singolo sito A (non urbano) corrispondente alla presenza di una centralina e addestrando la rete neurale a replicare i risultati reali. La seconda sperimentazione ha avuto l'obiettivo di replicare con la rete neurale deep i risultati della simulazione di AMS-MINNI, per ottenere uno strumento veloce ed affidabile per verificare l'incidenza delle perturbazioni dei parametri di input sull'output del modello. In quest'ultimo caso le concentrazioni di NO2 a parità di fascia oraria sono state non quelle misurate dalla centralina, ma quelle simulate con il modello.

2. Risultati e discussioni

Fig. 2 e 3 mostrano i "pattern" caratteristici ottenuti dal modello additivo per le concentrazioni di biossido d'azoto su una scala temporale settimanale e, rispettivamente, su scala giornaliera nel sito A. Si evidenziano i tipici fenomeni di accumulo di inquinanti in cor-

rispondenza del traffico veicolare intenso di mattina e pomeriggio e nei giorni lavorativi
La Fig.4 mostra, nell'ambito del set di validazione, il valore atteso della concentrazione di NO2 predetto dal modello ML, il valore misurato dalla centralina ed il valore "airq"

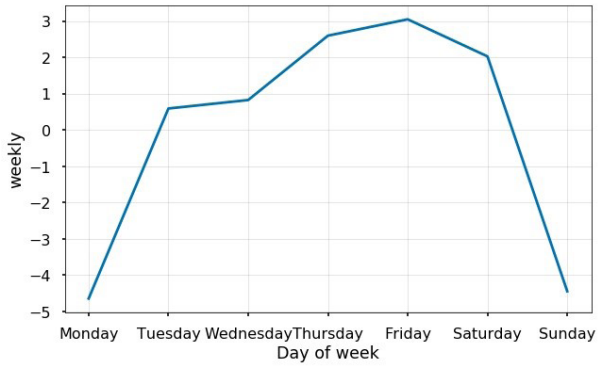


Fig. 2
Illustrazione
del trend a base
settimanale

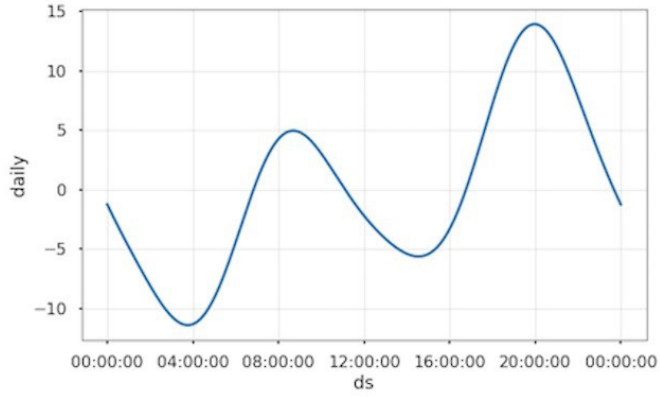


Fig. 3
Illustrazione
del trend a base
giornaliera

corrispondente al valore simulato dal modello AMS-MINNI per il sito A. Come si vede dal grafico, il modello ML prevede con buona approssimazione i valori di concentrazione dell'inquinante nel sito considerato. Il "relative squared error", definito rispetto al valore atteso $\mathbf{y}$, al valore predetto $\hat{\mathbf{y}}$ e alla deviazione standard $\boldsymbol{\sigma}$ come

$$RSE = \frac{\sqrt{(y - \hat{y})^2}}{\sigma}$$

passa da 1.76 nel caso del modello AMS-MINNI a 0.76 nel caso del modello ML, guadagnando quindi una deviazione standard rispetto ai dati presi in esame anche a causa dell'utilizzo preponderante dei dati misurati.
La capacità di questo modello di riprodurre il dato della centralina è però limitato dalla disponibilità di dati di concentrazioni in real time: ML richiede i dati di concentrazione nelle N ore precedenti. Fino a quando questa limitazione di tipo organizzativo non verrà rimossa, sarà necessario far riferimento a modelli di machine learning che prescindano dalle concentrazione dello stesso inquinante nelle ore precedenti, che però al momento danno risultati meno promettenti, con un RSE di 1.13 sullo stesso set di dati di validazione.
Il secondo esperimento utilizzata una rete neurale deep analoga a quella precedente, in

cui però si è cercato di prevedere il dato di concentrazione simulato nel modello AMS-MINNI. In questo caso il training è effettuato su un solo sito B (urbano) e il modello ML "addestrato" è in grado di predire i dati di concentrazione di NO2 sia nel sito originario B

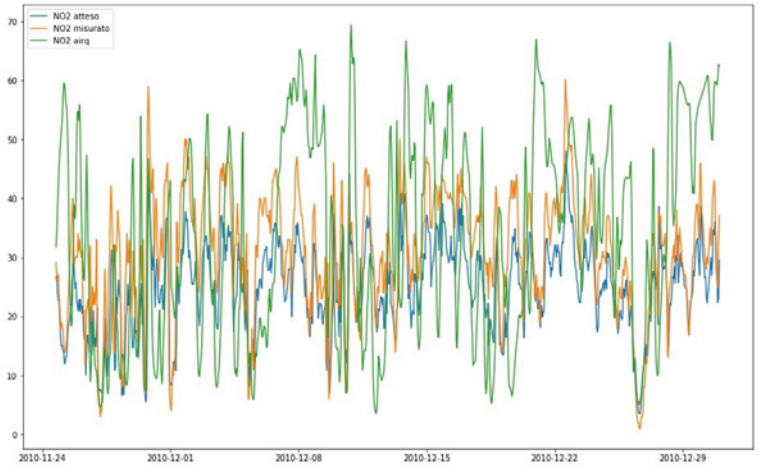


Fig. 4
Confronto tra le concentrazioni orarie di NO2 nel sito A predette dal modello ML (in blu), misurate dalla centralina (in arancio) e simulate dal modello AMS-MINNI (in verde)

(Fig.5) con RSE pari a 0.08 che in altri siti diversi sia urbani (sito C) (Fig.6) che non (sito A) (Fig.7), in quest'ultimo caso con RSE pari a 0.10.
I tempi di calcolo in entrambe le sperimentazioni sono stati dell'ordine di mezzora su un computer portatile. Da questi risultati si evince come l'algoritmo di machine learning sia stato in grado di estrarre da questo grande dataset informazioni utili per stimare in maniera efficace e rapida le concentrazioni a breve termine dell'inquinante considerato.

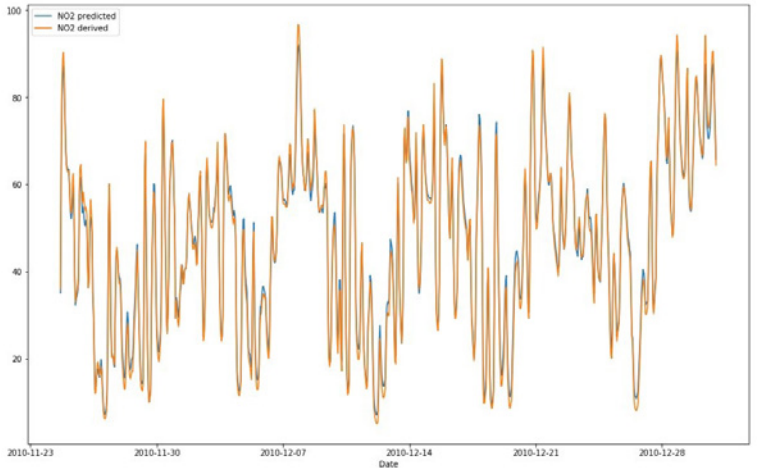


Fig. 5
Confronto tra le concentrazioni orarie di NO2 nel sito B predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio)

3. Conclusioni

Questo studio dimostra le potenzialità di questo grande dataset distribuito ospitato sull'infrastruttura ENEAGrid – CRESCO e degli algoritmi di machine learning nella costruzione di predittori che supportino la ricerca nel campo della qualità dell'aria. I risultati

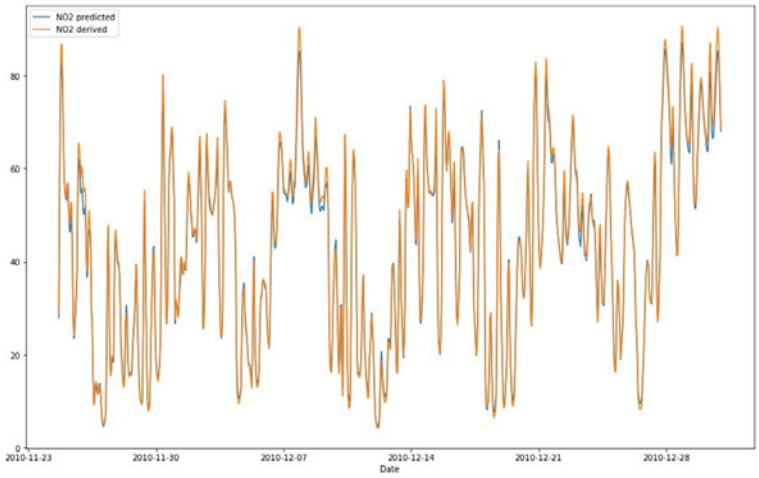


Fig. 6
Confronto tra le concentrazioni orarie di NO_2 nel sito C predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio) con training sul sito B.

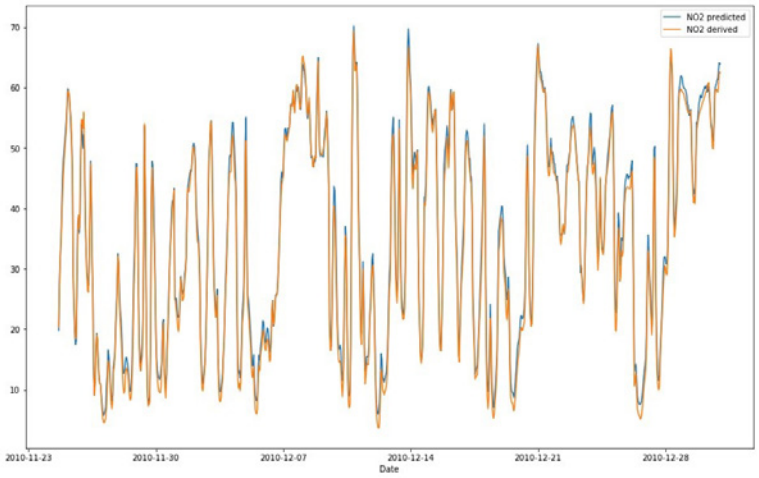


Fig. 7
Confronto tra le concentrazioni orarie di NO_2 nel sito A predette dal modello ML (in blu) e simulate da AMS-MINNI (in arancio) con training sul sito B.

aiuteranno a sviluppare approfondite analisi su altri inquinanti dannosi per la salute pubblica, come l'ozono e il particolato PM10 e PM2.5.
Si ringrazia per il supporto il gruppo di lavoro ENEAGrid [7].

Riferimenti bibliografici

[1] Mircea, M., Ciancarella, L., Briganti, G., Calori, G., Cappelletti, A., Cionni, I., Costa, M., Cremona, G., D'Isidoro, M., Finardi, S., Pace, G., Piersanti, A., Righini, G., Silibello, C., Vitali, L., Zanini, G., (2014). "Assessment of the AMS-MINNI system capabilities to predict air quality over Italy for the calendar year 2005. Atmospheric Environment", 84, 178–188, ISSN 1352-2310, http://dx.doi.org/10.1016/j.atmosenv.2013.11.006.

[2] Ciancarella L. et al. (2016), "La simulazione nazionale di AMS-MINNI relativa all'anno 2010 - Simulazione annuale del SistemaModellistico Atmosferico di MINNI e validazione dei risultati tramite confronto con i dati osservati", Rapporto Tecnico ENEA, RT/2016/12/ENEA

[3] Ponti G. 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" In Proceedings of the International Conference on High Performance Computing and Simulation, HPCS 2014, Bologna, Italy, 21-25 July, 2014

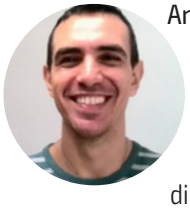
[4] Bengio Y. (2009) "Learning Deep Architectures for AI", Foundations and Trends in Machine Learning, 2 (1), pp. 1-127.

[5] Hochreiter S., Schmidhuber J. (1997). "Long short-term memory". Neural Computation. 9 (8): 1735– 1780.

[6] Taylor SJ, Letham B. (2017). "Forecasting at scale." PeerJ Preprints 5:e3190v2

[7] http://www.eneagrid.enea.it/people/2018EneaGridPeople.html

Autori



Angelo Mariano - angelo.mariano@enea.it

Dottore di ricerca in fisica teorica dal 2004, con una tesi nel campo della decoerenza in meccanica quantistica, ricercatore ICT in ENEA dal 2010, con competenze nel campo della progettazione informatica, dell'integrazione tra i sistemi ICT, del calcolo scientifico ad alte prestazioni, del cloud computing, dell'analisi di sistemi complessi. Attualmente si occupa di informatica gestionale, gestione dei big data, machine learning e intelligenza artificiale.

Mario Adani - mario.adani@enea.it

Ha lavorato nel campo della modellistica e assimilazione dati in oceanografia, partecipando allo sviluppo del modello NEMO. Ha contribuito a diversi progetti europei finalizzati allo sviluppo della capacità previsionali e di reanalisi. Attualmente, sta lavorando con modelli di dispersione inquinanti in atmosfera a fini previsionali e alle relazioni tra il cambiamento climatico e l'inquinamento atmosferico.

Gino Briganti - gino.briganti@enea.it

Gino Briganti si è laureato in fisica nel 1990, con una tesi concernente simulazioni su CRAY della QCD non perturbativa. Si occupa di sviluppo di codici per la simulazione della micrometeorologia dello strato limite e della dispersione e trasformazione chimica degli inquinanti in atmosfera. Una considerevole parte di lavoro riguarda set-up e run di modelli di simulazione su GRID ENEA. La sua attività ha come obiettivo la valutazione di impatto sulla qualità dell'aria di scenari emissivi, a supporto della negoziazione internazionale sull'inquinamento atmosferico.

Mihaela Mircea - mihaela.mircea@enea.it

Mihaela Mircea lavora come ricercatrice nel campo delle scienze atmosferiche, essendo particolarmente interessata alla conoscenza della composizione chimica della troposfera e ai suoi effetti sulla meteorologia, clima, ambiente e salute. I suoi studi sono dedicati allo sviluppo di modelli in modo integrato con le osservazioni per riprodurre in modo realistico i processi atmosferici e la loro interazione con il resto dell'ambiente con lo scopo di poter prevedere con precisione l'impatto delle future attività antropiche sulla qualità dell'aria e sul clima.

DOI: 10.26314/GARR-Conf18-proceedings-15

Infrastrutture, terminologie e policy per la ricerca umanistica: note per un confronto interdisciplinare

Annarita Liburdi, Cristina Marras, Ada Russo

CNR-ILIESI, Istituto per il Lessico Intellettuale Europeo e Storia delle Idee

Abstract. L'intervento riguarda le infrastrutture digitali e la Policy dei dati della ricerca. Descrive l'esperienza del gruppo di ricerca del CNR-ILIESI nel progetto PARTHENOS. Si concentra sull'open access e sull'open science nella preparazione, pubblicazione, gestione, accesso e riuso di documenti/testi/informazioni su piattaforme interoperabili. Guarda con attenzione alla terminologia e alla necessità di un Glossario dedicato nella stesura di linee guida per la gestione e regolamentazione di questa interoperabilità. Presenta le specificità della ricerca storico filosofica e umanistica digitale e alcuni problemi e prospettive aperti dal confronto con la costruzione di infrastrutture digitali e di workflows per la ricerca, facendo emergere come inevitabile un confronto con le pratiche e i protocolli delle altre discipline.

Keywords. Interdisciplinarità, Open access, Open Science, Scienze Umane, Glossari

Introduzione

Il programma scientifico dell'Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (ILIESI) del Consiglio Nazionale delle Ricerche (CNR) fa dello studio della terminologia di cultura la chiave d'accesso alla storia del pensiero filosofico e al costituirsi dell'identità culturale europea. I risultati delle attività di ricerca sono rappresentati da studi critici, archivi digitali, metodologie digitali per l'analisi testuale. Obiettivo dell'ILIESI è integrare i risultati di 50 anni di ricerca alle nuove tecnologie, confrontandosi con esperienze interdisciplinari nazionali e internazionali. In particolare vengono testati strumenti e processi di standardizzazione per la creazione di ambienti integrati e interoperabili; strumenti di annotazione semantica a supporto dello scambio e creazione di nuova conoscenza; modelli per l'organizzazione della conoscenza e la ricerca interdisciplinare; strumenti per la gestione e la valorizzazione del multilinguismo. In questo contesto si innesta la partnership con il progetto Horizon 2020 Parthenos, Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies e la specificità dei tasks assegnati al gruppo di ricerca ILIESI (di cui fanno parte oltre alle autrici di questo intervento Giancarlo Fedeli e Hansmichael Hohenegger).

Parthenos vede insieme 16 partners al fine di realizzare una infrastruttura di ricerca europea che si ponga come cluster capace di rafforzare e promuovere la collaborazione tra i diversi settori delle scienze umane. Per questo uno dei suoi obiettivi principali è la definizione di standard comuni e di azioni condivise, oltre a strumenti e applicazioni per una armonizzazione delle policies.

In particolare, le attività si sono concentrate sulla stesura di "Guidelines for Common

Policies Implementation", che costituiscono un ponte tra differenti campi e stakeholders nelle scienze umane. La ricerca ha previsto a tal fine una mappatura delle policies riguardanti il data management e la qualità dei dati, metadati, repository e IPR (Intellectual Property Rights), open data e open access e ha individuato linee di integrazione, i principi FAIR, per la qualità dei dati. L'idea di fondo è "make your data as FAIR as possible".

1. Infrastrutture digitali e problemi di terminologia

Uno degli aspetti della riflessione condotta dal gruppo ILIESI è legato alla terminologia condivisa in Parthenos, progetto che pone tra i suoi obiettivi la condivisione di dati. Nella redazione dei documenti di progetto è emersa infatti la difficoltà a convergere su una terminologia di riferimento univoca e condivisa. Da ciò la necessità di creare un glossario per definire i termini, sciogliere gli acronimi e documentare standard e tecnologie.

Nella piattaforma SSK "Standardization Survival Kit", una piattaforma dedicata alla promozione di un uso diffuso degli standard nelle scienze umane, è stata realizzata una tassonomia ordinata gerarchicamente con un approccio sintetico per macro temi, all'interno dei quali sono contenuti i termini. Ma tale approccio non facilita l'accesso ai termini, non valorizza la molteplicità di domini di utilizzo, e costringe a percorsi di ricerca predefiniti non sempre esaustivi. Inoltre, pur se tassonomie e glossari sono vocabolari controllati, nelle tassonomie la classificazione delle parole è intesa principalmente per la ricerca, mentre nei glossari è orientata ai contenuti e al loro significato.

Da qui la nostra riflessione sull'opportunità di un glossario collaborativo, strutturato non per macro temi ma per termini che funzionano come chiavi d'accesso e che permettono la costruzione dinamica di raggruppamenti. Per esempio l'acronimo TEI (Text Encoding Initiative) è nello stesso tempo un acronimo che va sciolto, uno standard che va documentato e l'istituzione' che gestisce lo standard. Non deve dunque comparire in modo ridondante in macro temi diversi, ma può assumere il ruolo di chiave di accesso per questa molteplicità. È importante poi ricostruire il contesto di ogni termine per renderlo il più completo possibile: definizione, fonte della definizione, e, là dove opportuno, sitografia e

Fig. 1
Problemi di terminologia

istituzioni di riferimento. Essenziale è, inoltre, assicurare la correttezza dell'informazione, attraverso ad esempio la cronologia delle modifiche che contestualizzano l'entrata del glossario. Non ultimo è il problema del multilinguismo: la terminologia di riferimento per la comunità scientifica è l'inglese, ma il glossario deve essere capace di poter parlare a una comunità plurilingue stabilendo tuttavia una 'nomenclatura' univoca e condivisa (Fig. 1). A differenza di un thesaurus, il glossario non prevede gerarchie. In questo modo è possibile far emergere la molteplicità tematica di appartenenza di un singolo termine, di mettere in evidenza le possibili associazioni e intersezioni tra ambiti senza essere incanalati in percorsi precostituiti, e si favorisce la dimensione euristica della ricerca terminologica.

2. Open science: teorie e pratiche

Gli archivi digitali dell'ILIESI costituiscono un esempio di infrastruttura ad accesso aperto che raccoglie materiali eterogenei, ma coerenti rispetto a tematiche di carattere filosofico e/o terminologico. Gli archivi si basano su programmi open-source e codifiche standard testuali e per i metadati, usano strumenti per l'arricchimento semantico e tecnologie per la rappresentazione della conoscenza. La loro finalità è favorire e supportare l'attività di ricerca, disseminare i risultati, consentirne il riuso. Le linee guida seguite dall'ILIESI per la realizzazione degli archivi digitali rispecchiano i requisiti definiti a livello nazionale e internazionale per le collezioni digitali e riguardano soprattutto la necessità di garantirne l'affidabilità e la sostenibilità nel tempo; di condividere standard per dati e metadati per favorire l'uso e il ri-uso; di assicurare l'accesso aperto agli archivi (Fig. 2, Liburdi et al. 2015).

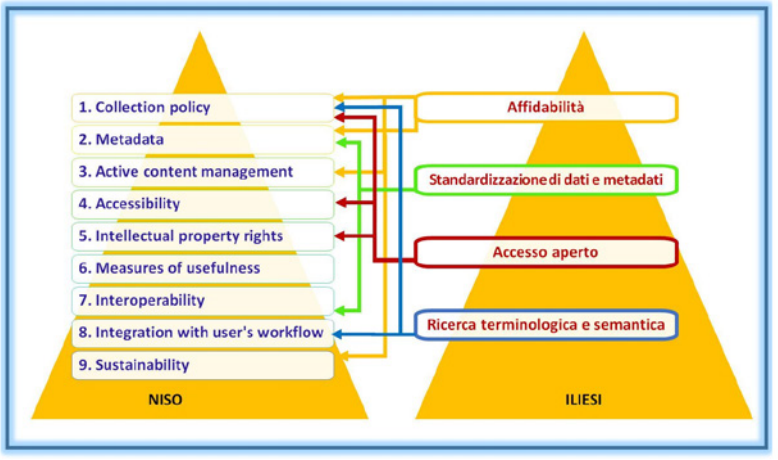


Fig. 2
Linee guida ILIESI per una collezione digitale

Nella realizzazione degli archivi digitali, sono stati centrali gli aspetti legati alla condivisione dei dati della ricerca e all'accesso aperto alle pubblicazioni, perché indispensabili per la realizzazione della scienza aperta. É evidente come anche in un ambito di ricerca umanistica specifico il problema della policy e del rapporto tra teoria e pratica è cruciale. Il CNR, ente pubblico di ricerca, per esempio è impegnato a sostenere i principi della libera circolazione della conoscenza, ha aderito alla "Berlin declaration" e ha firmato il "Position

Statement" sull'accesso aperto ai risultati della ricerca in Italia. Tuttavia, non si è dato una policy che guidi in modo chiaro e trasparente le pratiche dei ricercatori dell'Ente negli aspetti legali degli open data e dell'open access. In qualche modo anche nell'ambito della nostra esperienza di Istituto sentiamo l'esigenza di avere delle best practices condivise che pur tenendo conto delle differenze disciplinari - come per esempio la differenza del periodo di embargo per la pubblicazione - contemplate dalla legge italiana (Legge 7/12/2013, n. 112), tutelino il ricercatore di ogni ambito disciplinare nell'attuazione dell''open access e dell'open science.

3. Conclusioni

Se l'accesso aperto alle pubblicazioni scientifiche cresce e diventa una pratica sempre più consolidata e regolamentata, quello relativo ai dati della ricerca è diventato centrale solo di recente. Per lo più si esita a renderli pubblici per il timore di vederli riportati impropriamente, di non vederli attribuire correttamente o di avvantaggiare eventuali competitors.

Una protezione eccessiva dei dati si scontra con il modello collaborativo della scienza aperta, inibendo una possibile dialettica virtuosa tra la salvaguardia/protezione della paternità di un dato/idea/manufatto e della sua creazione, e la libertà di riutilizzarlo e migliorarlo. L'accesso aperto può inoltre entrare in conflitto con l'IRP, o meglio con una interpretazione della proprietà intellettuale della ricerca piuttosto restrittiva.

Attualmente la ricerca pone l'accento su collaborazione e condivisione anziché sui tradizionali concetti di 'proprietà intellettuale' e pubblicazione. Indubbiamente l'accesso aperto ai dati è uno dei pilastri della moderna metodologia di ricerca che si basa sempre più sulla cooperazione nel lavoro e su nuove modalità di distribuzione e diffusione della conoscenza attraverso le tecnologie digitali.

Su queste basi lo scambio dinamico di open data innesca processi di innovazione e riproducibilità creando contesti scientifici sempre più interattivi e integrati; in tali contesti, e nelle infrastrutture scientifiche che ad essi fanno riferimento, il linguaggio scientifico e tecnico è ampiamente condiviso e la terminologia di dominio diventa un indicatore sensibile dello stato e sostenibilità delle infrastrutture stesse. In questa prospettiva la policy assume un ruolo fondamentale di guida per i ricercatori, nell'attuazione di open access e open data e per l'identificazione di perimetri comuni (terminologia, standard, etc..) nell'ambito dei quali esercitare la scienza aperta.

Riferimenti bibliografici

Archivi Digitali ILIESI: http://www.iliesi.cnr.it/risorsedigitali

Decreto Legge 8 agosto 2013, n. 91: http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legge:2013;91

FAIR data principles: https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

Istituto per il Lessico Intellettuale Europeo e Storia delle Idee: http://www.iliesi.cnr.it

Liburdi, A., Russo, A., Tardella, M. (2015): Ricerca filosofica e archivi digitali: quale im-

patto sulla diffusione della cultura scientifica e sui beni culturali?, Poster. IV Convegno AIUCD, Digital Humanities e beni culturali: quale relazione? Torino, 17-19/12/ 2015.

Position Statement CNR: https://www.cnr.it/it/position-statement

Progetto PARTHENOS: http://www.parthenos-project.eu

Autrici

Annarita Liburdi - annarita.liburdi@cnr.it

Dal 1982 è nei ruoli del Consiglio Nazionale delle Ricerche e dal 2006 è Primo Tecnologo assegnato all'Istituto per il Lessico Intellettuale Europeo e Storia delle idee. Ha lavorato presso l'Istituto di Analisi dei Sistemi ed Informatica, la Biblioteca Centrale del CNR, il Comitato Nazionale di consulenza per le Scienze Matematiche e presso l'Istituto di Scienze e Tecnologie della Cognizione. Le sue aree di interesse sono: biblioteche e archivi digitali; accesso aperto; ipertesto.

Cristina Marras - cristina.marras@cnr.it

Ricercatrice presso l'Istituto per il Lessico Intellettuale Europeo e Storia delle Idee del Consiglio Nazionale delle Ricerche, affianca la sua ricerca in filosofia, filosofia del linguaggio e umanistica digitale con attività di valorizzazione del dialogo interdisciplinare attraverso l'esplorazione dei diversi linguaggi e delle tecnologie che favoriscono la condivisione di metodi, pratiche e risultati della ricerca.

Ada Russo - ada.russo@cnr.it

Tecnologo presso l'Istituto per il Lessico Intellettuale Europeo e Storia delle Idee del Consiglio Nazionale delle Ricerche, si occupa di trattamento automatico di dati documentali e testuali; realizzazione di strumenti specifici per l'elaborazione, l'organizzazione e reperimento di informazioni di tipo linguistico-lessicografico; uso di thesauri terminologici strutturati come chiave di accesso a materiali lessicografici; analisi di sistemi di gestione di banche di dati documentali e testuali.

DOI: 10.26314/GARR-Conf18-proceedings-16

Analisi emozionale di recensioni di corsi online basata su Deep Learning e Word Embedding

Danilo Dessì[1], Mauro Dragoni[2], Mirko Marras[1], Diego Reforgiato Recupero[1]

[1]Università degli Studi di Cagliari, [2]Fondazione Bruno Kessler

Abstract. La moltitudine di dati generata dagli utenti, distribuita su infrastrutture e servizi di rete mediante i social media, è un elemento essenziale per la crescita delle odierne comunità online. In questo articolo, analizziamo le recensioni lasciate dagli studenti dopo aver seguito corsi online. L'obiettivo è mostrare come le reti neurali profonde, abbinate a rappresentazioni semantiche e sintattiche di informazioni del contesto e-learning, siano più efficaci, rispetto alle tradizionali tecniche applicate su risorse te-stuali di carattere generico, nel predire la polarità del sentimento veicolato dalle re-censioni nello stesso contesto e-learning.

Keywords. Deep Learning, Sentiment Analysis, Word Embedding

1. Introduzione

Una interessante ed emergente area di ricerca riguarda l'analisi del sentimento e delle emozioni veicolate da un testo. Tale campo prende il nome di Sentiment Analysis. I word embedding, una rappresentazione vettoriale e distribuita delle parole in grado di memorizzare informazioni sia semantiche che sintattiche, sono impiegati con successo in questo settore. Tuttavia, gran parte dei modelli per la loro creazione non considera la distribuzione delle parole in un contesto specifico e le rappresentazioni risultanti perdono così informazioni rilevanti. Per mitigare il problema, Li et al., 2017 ha integrato conoscenze preliminari e indagato l'influenza di ogni parola sul sentimento, mentre Tang et al., 2016 ha usato il sentimento del testo per creare word embedding contestualizzati.

In questo articolo, il nostro obiettivo è dimostrare come gli approcci capaci di catturare la semantica del contesto nel quale vengono poi applicati siano più efficaci rispetto a quelli progettati e allenati senza tener conto del contesto di applicazione. Così, abbiamo prima creato dei word embedding specifici per il contesto e-learning, usando una vasta raccolta di recensioni di corsi online e approcci allo stato dell'arte per la loro generazione. Successivamente, abbiamo progettato una rete neurale basata su deep learning per inferire la polarità del sentimento di ogni recensione. Infine, abbiamo valutato l'efficacia dei nostri word embedding specifici del contesto rispetto a quelli generici e confrontato l'efficacia di diverse strategie per il rilevamento di polarità nelle recensioni dei corsi online. I risultati hanno mostrato come il nostro approccio basato su reti neurali e word embedding specifici del contesto superi diverse baseline.

2. Il Nostro Contributo

Il nostro approccio e le sue componenti sono mostrati in Fig. 1.

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

Il Review Splitter riceve in ingresso un insieme di recensioni di corsi, ognuna contenente un commento testuale e la relativa valutazione. Inoltre, accetta (i) un intero N che definisce quante recensioni sono scelte complessivamente per ogni classe per il training e il test della predizione della polarità e (ii) un intero M < N che definisce il numero di recensioni per classe utilizzate specificamente per la fase di training. Il modulo restituisce un corpus testuale impiegato per la creazione dei word embedding contestuali e due sottoinsiemi di recensioni per il rilevamento della polarità, uno per il training e uno per il test.

Il Neural Word Embedding Generator riceve un corpus testuale. Il suo output è un insieme di vettori, ciascuno rappresentante il word embedding associato ad una parola del corpus. Inoltre, il modulo accetta parametri per indicare la dimensione e l'algoritmo di creazione degli embedding: Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), FastText (Joulin et al., 2016).

I Review Pre-Processor di Training e Test accettano un insieme di recensioni, ognuna identificante un commento testuale e una valutazione. Restituiscono una serie di recensioni pre-elaborate, dove ogni elemento è composto da un vettore ordinato di indici numerici interi associati alle parole del commento originario e un'etichetta numerica che rappresenta la valutazione originaria lasciata dall'utente per quel commento. Ogni differente parola del commento originario è mappata su un indice intero univoco. Le recensioni pre-processate vengono passate al Deep Neural Network Trainer (DNNT) e Regressor (DNNR).

Il Trainer accetta un insieme di word embedding e un insieme di recensioni pre-elaborate. Con essi, allena una rete neurale. Il Regressor prende tale rete e un commento, e restituisce la polarità del sentimento per quel commento. Il modello di rete neurale è ispirato a quello usato da Atzeni e Reforgiato, 2018. Diversamente da loro, abbiamo adottato solo un livello di tipo Bidirectional LSTM (Long Short-Term Memory) e abbiamo configurato l'ultimo livello per restituire un valore continuo che rappresenta il valore associato alla polarità.

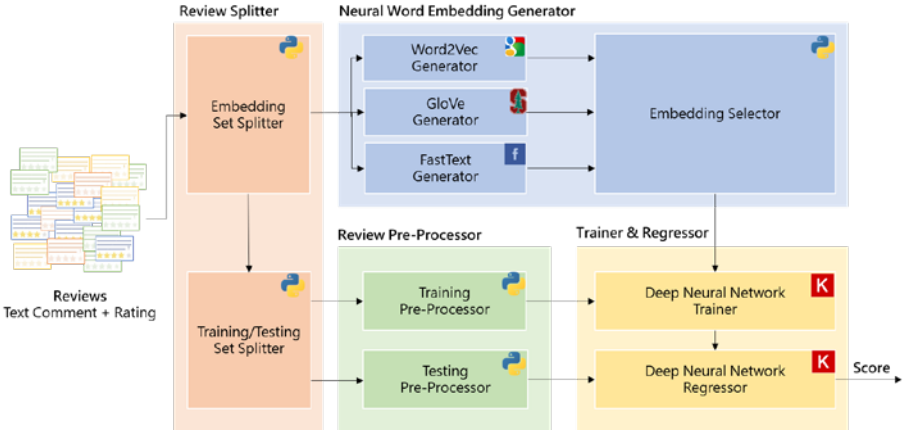


Fig. 1
Il nostro approccio basato su reti neurali e word embedding contestuali.

3. Valutazione Sperimentale

Gli esperimenti sono stati eseguiti sul dataset COCO (Dessì et al., 2018). Da esso, il Review Splitter ha considerato 1.396.312 recensioni in Inglese, con N = 6500 e M = 650.

Quindi, 1.396.312 - 6.500 * 10 sono state usate per creare i word embedding, 6500 * 10 per il training e 650 * 10 per il test. Per ogni test, è stato impiegato un protocollo basato su 10-fold stratified cross-validation e sono stati misurati il MSE (Mean Squared Error) e il MAE (Mean Absolute Error).

La rete neurale proposta nel nostro approccio è stata comparata con regressori tradizionali: il Support Vector Machine (SVR), il Random Forests (RF) e il Multi-Layer Perceptron (MLP). Tutti sono stati allenati con i nostri word embedding contestuali. Per i regressori tradizionali, la media dei word embedding delle parole del commento è stata calcolata per rappresentare ciascun commento con un unico vettore. Osservando Tab. 1, la baseline con le migliori prestazioni è MLP. Tuttavia, come è possibile notare, MLP conduce a valori di errore superiori rispetto alla nostra rete neurale con ogni tipo di word embedding.

Tab. 1
Comparazione tra la nostra rete neurale e i regressori tradizionali

Regressore	Generatore	MSE	MAE
RF	Word2Vec	5.25	1.85
	GloVe	5.47	1.90
	FastText	5.17	1.84
SVR	Word2Vec	4.17	1.63
	GloVe	5.38	1.91
	FastText	5.35	1.92
MLP	Word2Vec	4.06	1.59
	GloVe	4.27	1.63
	FastText	4.00	1.53
DNNR (ours)	Word2Vec	**3.35**	**1.41**
	GloVe	3.85	1.54
	FastText	3.82	1.55

Successivamente, i word embedding generati dalle recensioni in COCO sono stati confrontati con quelli creati sulla base di dati testuali generici: Word2Vec, 2018; GloVe, 2018; FastText, 2018. Tutti i word embedding sono stati passati in input alla nostra rete neurale. Da Tab. 2, si evince come la nostra rete neurale, allenata con word embedding contestuali, fornisca valori inferiori di MSE e di MAE rispetto alla baseline proposte. Le migliori prestazioni sono state ottenute allenando la rete neurale con word embedding contestuali creati da Word2Vec.

Tab. 2
Comparazione tra i word embedding contestuali e i word embedding generici

Tipo	Generatore	MSE	MAE
Contestuale	Word2Vec	**3.35**	**1.41**
Generico		4.58	1.73
Contestuale	GloVe	3.85	1.54
Generico		**3.79**	**1.54**
Contestuale	FastText	**3.82**	**1.55**
Generico		4.71	1.73

4. Conclusioni

In questo articolo, abbiamo generato word embedding dipendenti dal contesto delle recensioni di corsi online e li abbiamo incorporati in una rete neurale progettata per predire la

polarità del sentimento in recensioni provenienti dallo stesso contesto. I word embedding contestuali sono stati comparati con i word embedding generici esistenti in letteratura e la nostra rete neurale è stata comparata con regressori tradizionali. I risultati hanno mostrato come i word embedding contestuali siano adatti per l'analisi della polarità del sentimento e la rete neurale proposta superi le baseline considerate nel predire tale polarità.

Riferimenti bibliografici

Atzeni, M., Reforgiato, D. 2018. Deep Learning and Sentiment Analysis for Human-Robot Interaction. In European Semantic Web Conference 2018, 14-18. Springer.

Dessì, D., Fenu, G., Marras, M., Reforgiato, D. 2018. COCO: Semantic-Enriched Collection of Online Courses at Scale with Experimental Use Cases. In World Conference on Information Systems and Technologies, 1386-1396. Springer.

FastText, Facebook, 01/12/2018, https://fasttext.cc

GloVe, 01/12/2018, https://nlp.stanford.edu/projects/glove

Word2Vec, 01/12/2018, https://code.google.com/archive/p/word2vec

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T. 2016. FastText. zip: Compressing text classification models. arXiv:1612.03651.

Li, Y., Pan, Q., Yang, T., Wang, S., Tang, J., Cambria, E. (2017). Learning word representations for sentiment analysis. Cognitive Computation, 9(6), 843-851.

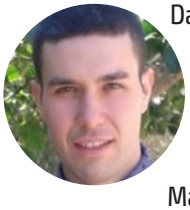
Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Pennington, J., Socher, R., Manning, C. 2014. Glove: Global vectors for word representation. Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532-1543.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., Zhou, M. 2016. Sentiment embeddings with applications to sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(2), 496-509.

Udemy, 01/12/2018, https://www.udemy.com

Autori

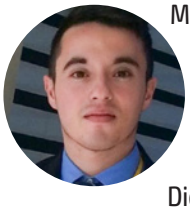


Danilo Dessì - danilo_dessi@unica.it

Danilo è Dottorando in Informatica presso il Dipartimento di Matematica e Informatica dell'Università degli Studi di Cagliari. I suoi interessi di ricerca sono focalizzati su Natural Language Processing, Semantic Web e Deep Learning.

Mauro Dragoni - dragoni@fbk.eu

Mauro è Ricercatore presso l'Information Technology Center della Fondazione Bruno Kessler. I suoi interessi di ricerca sono focalizzati su Information Retrieval, Artificial Intelligence e Sentiment Analysis.



Mirko Marras - mirko.marras@unica.it

Mirko è Dottorando in Informatica presso il Dipartimento di Matematica e Informatica dell'Università degli Studi di Cagliari. I suoi interessi di ricerca sono focalizzati su Semantic-aware Systems, E-Learning Technologies e Deep Learning.

Diego Reforgiato Recupero - diego.reforgiato@unica.it

Diego è Professore Associato presso il Dipartimento di Matematica e Informatica dell'Università degli Studi di Cagliari. I suoi interessi di ricerca sono focalizzati su Semantic Web.

DOI: 10.26314/GARR-Conf18-proceedings-17

Cultura e educazione ai tempi del digitale

Vittorio Midoro

Libero professionista, già Istituto Tecnologie Didattiche, CNR

Abstract. La rivoluzione digitale origina nuovi contenitori del sapere, gli oggetti digitali che sono multimediali, aperti, computabili, interattivi, immateriali e pervasivi. Nuovi paradigmi subentrano a quelli della cultura scritta su cui poggia la scuola: si passa così dalla monomedialità alla multimedialità, dalla chiusura all'apertura, dalla linearità alla reticolarità, dal consumo al consumo-produzione, dalla unidirezionalità all'interattività, dalla competizione alla collaborazione, dal materiale all'immateriale. A fronte di questi cambiamenti, è necessario ideare una scuola diversa considerando che quello educativo è un sistema complesso, costituito da entità strettamente correlate, che bisogna esplicitare, analizzare, comprendere e cambiare, individuando modi e processi del cambiamento.

Keywords. Oggetti digitali, scuola, conoscenza, innovazione, nuovi paradigmi

1. La Rivoluzione Digitale

Quella digitale è una delle più grandi rivoluzioni nelle tecnologie della conoscenza dall'origine del linguaggio. Iniziata intorno alla metà del '900 e sviluppatasi agli inizi degli anni settanta, ha originato nuovi contenitori del sapere: gli oggetti digitali (dati e programmi). Se le lettere, le parole e le frasi erano il supporto della cultura scritta, oggi gli oggetti digitali vanno oltre gli scritti, fornendo il supporto alla nascita di una cultura digitale, che ingloba quella scritta. Ma quali sono le loro caratteristiche?

La multimedialità. Un oggetto digitale può avere diversi formati, a cui corrispondono canali di comunicazione diversi. Può così presentarsi sotto forma di testo, di immagini, di video, di suoni e di ologrammi, in tutte le combinazioni.

L'apertura. Un oggetto digitale può essere legato a qualsiasi altro oggetto digitale. Gli oggetti digitali vivono nel contesto sociale della rete e sono facilmente producibili, riproducibili, modificabili, immagazzinabili e condivisibili.

La computabilità e l'interattività. I programmi eseguibili dai processori sono oggetti digitali. Per esempio, le APP sui dispositivi mobili sono programmi che amplificano le possibilità di interazione con l'ambiente biofisico, sociale e individuale.

L'immaterialità. Gli oggetti digitali risiedono in memorie digitali, disponibili su nuvole, di cui pochissimi conoscono la collocazione. Così, per un utente comune, gli oggetti digitali fluttuano nel cielo, catturabili solo con dispositivi personali, divenuti protesi del corpo.

La pervasività. Per determinarne la funzione, gli oggetti digitali sono presenti dovunque ci sia un processore: nelle lavatrici, nei sistemi di riscaldamento, ecc.

L'avvento delle tecnologie digitali ha avuto un forte impatto su tutti gli aspetti della vita sociale. Per esempio un'enorme quantità di informazione è prodotta a livello mondiale a

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

una velocità impressionante. Si stima che oggi la conoscenza raddoppi ogni anno e che nel 2020 raddoppierà all'incirca ogni 3 mesi. La conoscenza cresce dunque esponenzialmente e contemporaneamente si sviluppano gli strumenti per produrla, immagazzinarla e condividerla, in qualsiasi momento e in qualsiasi luogo.

La rivoluzione digitale ha investito anche l'economia. Scrive Rullani "La nostra economia è diventata un'economia in cui la conoscenza è messa al lavoro." La conoscenza è diventata dunque il motore che movimenta industria, commercio e servizi, favorendo la globalizzazione dei mercati. In questa economia, la finanza, quasi completamente informatizzata, è divenuta il cuore. Si stanno diffondendo forme di transazione finanziaria, come le blockchain, che originano monete virtuali, come i bitcoin.

Al di là dell'economia, la rivoluzione digitale investe tutti gli aspetti della vita, modificando profondamente la società e gli individui che la compongono, cambiando modi di lavorare, di informarsi, di divertirsi, di apprendere, di relazionarsi con gli altri, di partecipare alla vita pubblica e con essi abitudini, atteggiamenti, comportamenti e stili di vita.

2. Nuovi Paradigmi

Riguardo alla conoscenza, nuovi paradigmi stanno sostituendosi a quelli legati alla cultura scritta. Dalla monomedialità alla multimedialità. Fino all'avvento del digitale, la cultura era inscindibilmente legata ai libri e agli scritti. Infatti, a scuola la sua trasmissione è principalmente basata sui libri e nell'accademia, i risultati delle ricerche sono pubblicati su riviste specializzate. In tutti i campi, oggi è possibile acquisire e condividere conoscenza in modi diversi. La rivoluzione digitale slega il sapere dagli scritti, che diventano una risorsa tra le tante, e anch'essi spesso in formato digitale.

Dalla chiusura all'apertura, dalla linearità alla reticolarità. Il sapere, necessariamente sequenziale nei libri, riacquista nel web la sua reale dimensione di rete ipermediale, dove tutto si lega, consentendo al fruitore di navigare in un grafo potenzialmente infinito e di lasciare traccia dei suoi pensieri. Il digital literate può navigare gran parte della conoscenza con il suo dispositivo personale, standosene in poltrona. Questa caratteristica delle tecnologie digitali riconfigura il modo di intendere il conoscere, non più come processo lineare e chiuso, ma come attività reticolare e aperta, che richiede grande capacità di autoregolazione.

Dal consumo al consumo-produzione. In una cultura scritta, il lettore è un fruitore dei testi, su cui può tutt'al più apporre commenti. In un ambiente digitale, l'utente può fruire di un oggetto digitale, ma può anche modificarlo, riusarlo, produrre nuovi oggetti digitali assemblando quelli esistenti. In altri termini, non è solo un consumatore ma anche un produttore, che rende disponibili per gli altri le sue creazioni, è un prosumer.

Dalla unidirezionalità all'interattività. Spesso la trasmissione del sapere è identificato con un flusso unidirezionale che va da chi sa a chi non sa. Questa concezione si concretizza nella disposizione delle aule con una cattedra di fronte a tanti banchi, ma anche nello studio individuale basato su testi. Gli oggetti digitali restituiscono alla conoscenza l'interattività, perché da un lato possono essere concepiti per reagire agli input dell'utente, dall'altro consentono di dialogare con chiunque sia in rete. Assistiamo così alla nascita di comunità

che dialogano tra di loro, aiutandosi a risolvere problemi comuni, o condividendo prodotti, come quella dell'open software o quella dei makers.

Dalla competizione alla collaborazione. La nascita di comunità di pratica virtuali indica una tendenza indotta dalla rivoluzione digitale, quella di valorizzare la collaborazione tra individui e istituzioni nella soluzione di problemi e nello svolgimento di attività. Un esempio è la collaborazione di alcune delle più prestigiose università nell'offerta di corsi denominati MOOC (Massive Open Online Course), basati sulla collaborazione, sia tra chi offre i corsi, sia tra gli studenti che li fruiscono.

Dal materiale all'immateriale. La rivoluzione digitale sta determinando la sparizione di molti oggetti concreti, sostituiti da informazione in rete. Affidiamo alla banca i risparmi e li gestiamo online o con carte di credito, senza toccare banconote o monete. Prenotiamo online il biglietto dell'aereo o del treno e al controllore indichiamo solo una lettera. Per ascoltare il nostro brano preferito, ci colleghiamo a Spotify, senza alcun CD. Per leggere un libro, lo scarichiamo dalla rete, pagandolo con la carta di credito. In sintesi, la rivoluzione digitale induce una dematerializzazione di molti oggetti e soprattutto degli oggetti che trattano l'informazione. I dispositivi personali con i quali accediamo ad Internet sono indispensabili per gestire la dimensione immateriale.

3. Impatto sulla scuola

Questi cambiamenti di paradigma rendono superata la scuola tradizionale, le cui finalità sono descritte efficacemente da Baldacci; "Una scuola con una struttura chiusa e selettiva, capace di separare la formazione delle classi dirigenti da quella delle classi subalterne".

La struttura del sistema scolastico riflette sia queste finalità sia l'organizzazione del lavoro del secolo scorso: la scuola primaria e secondaria di primo grado mirano ad alfabetizzare tutti; i licei a fornire una cultura prevalentemente umanistica ai futuri dirigenti, professionisti e artisti, destinati a proseguire gli studi; gli istituti tecnici a formare quadri aziendali e impiegati e le scuole professionali operai, artigiani e agricoltori. I paradigmi su cui è basata questa scuola sono quelli di una cultura scritta:

Monomedialità. La lezione e i libri sono gli strumenti principali per la trasmissione della conoscenza e il supporto principale della didattica.

Chiusura. L'apprendimento ha luogo in un'aula chiusa, con l'insegnante che fa lezione e assegna i compiti a casa, basati sui libri di testo, anch'essi oggetti "chiusi".

Linearità. La conoscenza è rigidamente suddivisa in discipline che non si parlano. Il programma ministeriale, inglobato nei libri di testo, stabilisce l'ordine con cui gli argomenti devono essere appresi, spesso legato all'evoluzione temporale della disciplina.

Consumo. Le materie, trattate nei libri di testo, forniscono un corpo di nozioni che devono essere assimilate. La produzione di conoscenza da parte degli studenti ha poca importanza.

Unidirezionalità. L'apprendimento è visto principalmente come un flusso di informazione unidirezionale tra chi sa (il docente e il libro) e chi non sa (lo studente), che da solo deve rielaborare le nozioni ricevute.

Competizione. La competizione è spesso considerata un valore. Intelligenze diverse da

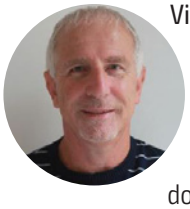
quelle linguistiche e logico-matematiche sono ignorate. La collaborazione tra pari è ostacolata dalla disposizione della classe, pensata per ascoltare e non per lavorare con gli altri. Materialità. La scuola fa principalmente uso di oggetti materiali, come libri, quaderni, penne, astucci ecc. Qualche raro dispositivo digitale ha un ruolo marginale nelle attività didattiche ed è spesso malvisto.

La scuola tradizionale crolla se togliamo i pilastri su cui è fondata, cioè la cultura scritta, e non bastano pochi ritocchi digitali per tenerla in piedi. Allora bisogna immaginare una scuola nuova, fondata sui nuovi paradigmi che caratterizzano l'era del digitale.
Per fare questo è necessario costruire una visione di scuola diversa, partendo dall'idea che la scuola sia un sistema complesso, costituito da entità strettamente correlate, che bisogna esplicitare, analizzare, comprendere e cambiare, individuando modi e processi del cambiamento. Le entità principali da considerare sono:

- studenti;
- finalità, struttura, curriculum e modi di apprendere, organizzazione;
- spazi e tecnologie;
- docenti.

Per costruire una scuola nuova è necessario comprendere a fondo queste entità e tenerne conto nella creazione di una idea nuova di scuola adeguata alle sfide poste dalla rivoluzione digitale.

Autore



Vittorio Midoro - vittorio.midoro@gmail.com

È uno dei fondatori del settore delle Tecnologie Didattiche in Italia. Dal 1974 al 2008 è stato ricercatore dell'Istituto Tecnologie Didattiche del CNR. Attualmente svolge attività di ricerca su metodi di lettura in età prescolare e di consulenza a livello nazionale e internazionale. Nel 1993, ha fondato la rivista TD TECNOLOGIE DIDATTICHE. Nel 2018, è stato docente a contratto presso l'Università Ca' Foscari di Venezia per una serie di corsi su "Tecnologie Educative e Didattiche".

DOI: 10.26314/GARR-Conf18-proceedings-18

Interoperabilità e aggregazione di dati eterogenei per il monitoraggio dei risultati scientifici all'Istituto Italiano di Tecnologia

Ugo Moschini, Elisa Molinari

Istituto Italiano di Tecnologia

Abstract. Istituti di ricerca e università si trovano spesso a dover generare reportistica relativa alla loro produzione scientifica e allo staff, per uso interno o esterno. Tali dati sono di solito presenti su sistemi eterogenei e molto tempo viene speso nel raccogliere le informazioni dai vari uffici, con metodologie poco efficaci e soggette a errori. L'integrazione di servizi messi a disposizione dai sistemi dei diversi uffici si pone come unica soluzione per migliorare sia la qualità sia l'automatizzazione con cui le informazioni sono gestite e estratte. In questo articolo, viene descritta l'esperienza all'Istituto Italiano di Tecnologia nel raccogliere in modo automatico i dati dai diversi sistemi e nel generare automaticamente reportistica. In particolare, l'applicazione web Scientilla, sviluppata nell'istituto, ha il potenziale di diventare un collettore unico per tutti i dati relativi all'attività di ricerca e fornire una panoramica sempre aggiornata di informazioni eterogenee.

Keywords. Interoperabilità, web-services, reportistica, valutazione, R

Introduzione

In ambienti accademici o istituti di ricerca, succede frequentemente che siano richiesti statistiche e reportistica relativa al personale o ai gruppi di ricerca. I motivi sono molteplici: benchmark nazionali o internazionali a scopi valutativi richiesti da comitati interni o esterni, promozioni di carriera di ricercatori, informazioni da visualizzare su pagine web istituzionali e molti altri.

Nella maggioranza dei casi, l'informazione che permette di rispondere a tali richieste è già presente nei vari sistemi IT di cui un istituto è provvisto. La parte che spesso richiede maggior tempo è il mettere insieme tutti i dati proprio da questi sistemi, che sono molto differenti fra di loro. Solitamente, questi variano da file Microsoft Office archiviati su qualche server a web-services più strutturati. A causa delle molteplici tecnologie usate, la mole di lavoro più grande mentre viene compilata la reportistica è rappresentata dal raccogliere e strutturare nuovamente informazioni diversificate, spesso a mano. Per esempio, tale attività consiste solitamente nel contattare (e attendere e elaborare la risposta) gli uffici deputati alla gestione dei progetti e fondi vinti, brevetti, informazioni dalle Risorse Umane, ..., che possono anche essere dislocati geograficamente in sedi molto distanti, e raccogliere tutti i dati con frequenti operazioni di "copia-incolla", soggette a errori. Inoltre, è sempre presente la preoccupazione che, se una richiesta cambia anche leggermente (ad esempio, si vogliono analizzare dati su un intervallo temporale diverso da quello previsto), tutto il

lavoro di contattare gli uffici, sistemare nuovamente i dati, e così via, andrebbe ripetuto nuovamente.

L'Istituto Italiano di Tecnologia (IIT) sta sviluppando soluzioni e tecnologie per far fronte alle problematiche menzionate sopra. È in atto al momento uno sforzo collettivo fra i vari uffici amministrativi e l'ufficio ICT per fornire sistemi e servizi con cui questi possano interagire e comunicare fra loro. Una applicazione web chiamata Scientilla, sviluppata completamente in IIT, prevede di diventare un collettore unico di dati dei diversi uffici, per fornire allo staff dell'istituto una visione d'insieme delle informazioni.

1. I sistemi all'Istituto Italiano di Tecnologia

In IIT, la produzione scientifica che è maggiormente oggetto di reportistica riguarda pubblicazioni scientifiche e interventi a convegno, la capacità di attirare fondi e vincere grant nazionali e internazionali, l'impatto sulla società, misurato come numero di brevetti e aziende spin-off, e i dati sullo staff. I principali sistemi che gestiscono tutte queste informazioni sono:

- pubblicazioni scientifiche: Scientilla – al momento è la piattaforma IIT per gestire pubblicazioni scientifiche, includendo sia riferimenti bibliografici sia dati bibliometrici. Scientilla si può definire come un sistema CRIS (Current Research Information System), un sistema per la gestione dei dati di ricerca. È una applicazione web sviluppata internamente all'Istituto. I ricercatori stessi sono responsabili dell'inserimento dei loro dati in Scientilla, sia per il loro profilo personale sia per il profilo del gruppo di ricerca di cui sono eventualmente responsabili.
- Progetti e fondi: MONIIT – l'informazione è mantenuta e accessibile tramite software ideato sia per progetti dal settore pubblico sia da quello privato. È sviluppato all'interno dell'istituto e si appoggia ad applicazioni SAP.
- Brevetti: Patents Management – i dati riguardati le applicazioni brevettuali sono ottenute tramite un servizio web collegato a un database, mantenuto dal Patents Office. Anche questa soluzione è sviluppata all'interno dell'istituto.
- Risorse Umane: i dati relativi allo staff e i dati contrattuali sono gestiti da un sistema SAP.

In generale, ciascuno di questi sistemi ha una interfaccia tramite la quale servizi esterni sono in grado di connettersi sulla rete e accedere ai dati.

2. Interoperabilità e integrazione

2.1 Scientilla: un esempio di interoperabilità fra servizi eterogenei

In IIT, l'applicazione Scientilla rappresenta un ottimo esempio di interoperabilità fra servizi. Gli articoli scientifici, con la loro informazione bibliografica e i dati citazionali, sono aggiornati in modo automatico raccogliendo le informazioni da database online esterni o servizi a pagamento, come Elsevier Scopus/Scival o Web of Science (numero di citazioni, impact factor, ...). Una routine notturna mantiene ogni informazione aggiornata per le migliaia di articoli in Scientilla. In tal modo, statistiche e indicatori sono mostrati aggior-

nati ogni giorno, senza alcun intervento manuale. Grazie all'interfaccia di cui Scientilla dispone per permettere a servizi esterni di interfacciarsi con il suo database, il sito web dell'istituto e i siti web dei gruppi di ricerca sono in grado di caricare automaticamente l'informazione aggiornata sulle loro pagine web.

In Scientilla sono presenti cruscotti tramite i quali i ricercatori visualizzano indicatori bibliometrici relativi alla loro attività di ricerca. I dati bibliografici possono essere esportati facilmente per esser poi riutilizzati (per esempio, nella stesura di un curriculum o di una reserch proposal). Inoltre, i responsabili di un gruppo di ricerca possono vedere la pagina relativa al loro gruppo, con le pubblicazioni prodotte dal loro staff e informazioni sul personale, così come è generato dai servizi messi a disposizione dall'ufficio Risorse Umane.

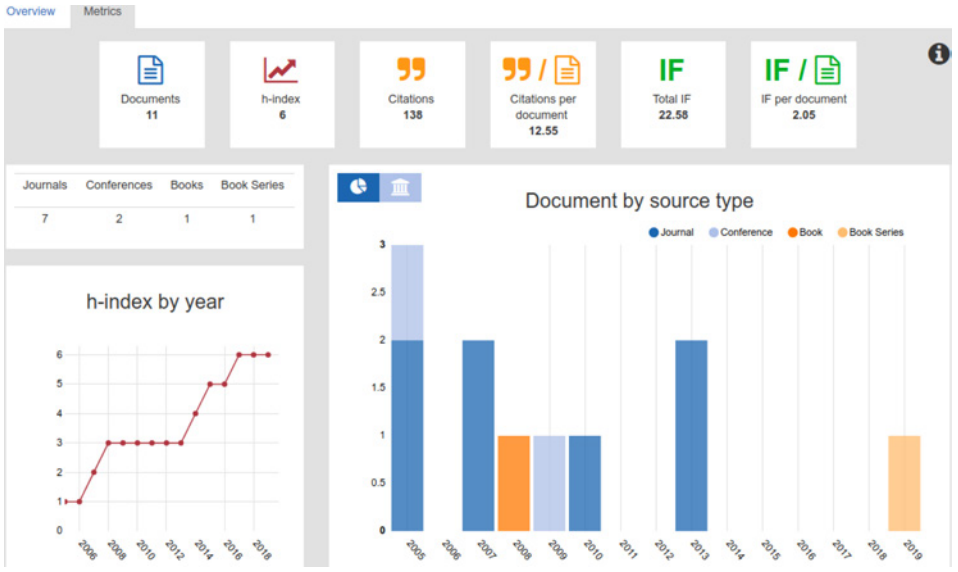


Fig. 1
Un esempio di alcuni cruscotti e report presenti in Scientilla, che integrano informazione bibliografica con informazione bibliometrica (da Scopus e Web of Science).

2.2 Automatizzazione della reportistica

In caso sia necessario creare reportistica riguardante la produzione scientifica, è auspicabile avere completa integrazione fra ogni sistema. Si può definire come un problema di estrazione e visualizzazione di dati provenienti da sistemi eterogenei. Per leggere e aggregare tutti questi tipi di informazione, è stato adottato R, un linguaggio per l'analisi statistica: è molto versatile, facile da usare e provvisto di moltissime librerie free e open source. Va evidenziato che uno script R è capace di generare con facilità tabelle e grafici in formato pdf o PowerPoint, facilitando la distribuzione e la leggibilità della reportistica con l'utilizzo di formati largamente diffusi.

I sistemi elencati nella Sezione 1 sono accessibili via web-services: con R, sono recuperati i dati delle pubblicazioni scientifiche, progetti e grant, brevetti e informazioni contrattuali sul personale. In seguito, l'informazione è analizzata e vengono prodotti statistiche e indicatori. Per esempio, grafici e tabelle possono descrivere la distribuzione di documenti o grant per tipo e per anno, la percentuale di documenti più citati, e così via. Dal momen-

to che tutto è generato automaticamente tramite un programma R, i report e i grafici possono essere creati velocemente ogni volta che ce ne sia la necessità o l'informazione venga aggiornata nei sistemi. Inoltre, impostazioni come il nome di un gruppo o l'intervallo temporale da analizzare possono essere dati come input al programma R, rendendo il report adattabile a parametri personalizzabili: non c'è bisogno di nessun lavoro extra, semplicemente il report può essere generato di nuovo eseguendo il solito identico script R. In generale, in un istituto, l'ufficio deputato alla reportistica può quindi concentrarsi sul contenuto vero e proprio, cioè cosa visualizzare e quale sia la visualizzazione più efficace, invece di concentrare gli sforzi sulla semplice attività di raccogliere i dati dai diversi uffici.

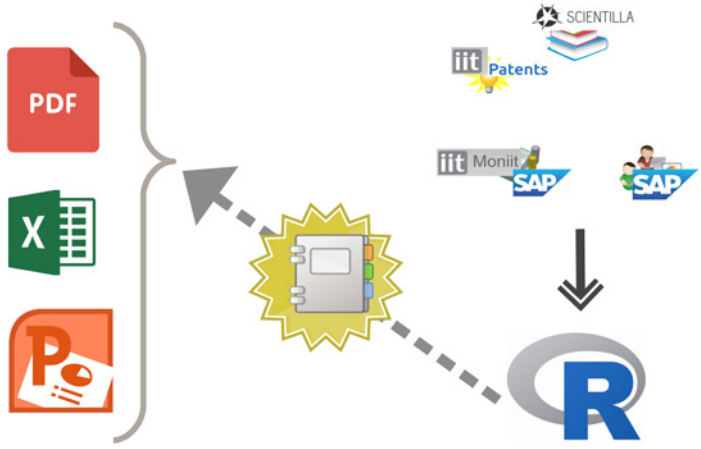


Fig. 2
Lo schema riassume come la reportistica è generata in IIT: i sistemi elencati nella Sezione 1 (in alto a destra) vengono interrogati tramite R, con cui i report sono generati in maniera automatica nei formati classici Microsoft o Pdf.

3. Conclusioni

All'Istituto Italiano di Tecnologia, la generazione di reportistica è senz'altro più immediata adesso rispetto ai mesi precedenti, sfruttando la nuova interoperabilità fra sistemi. Anche la qualità del dato è migliorata sia in dettaglio sia in qualità, grazie alla comunicazione con servizi esterni come, ad esempio, la base dati di Scopus. Una volta che un report è stato creato, gli altri uffici all'interno dell'organizzazione possono usarlo per i loro scopi, sapendo che il dato è controllato, aggiornato e certificato: tutta l'informazione è stata ottenuta dai vari sistemi in un preciso momento e poi elaborata.

Naturalmente, c'è ancora lavoro da fare. I servizi in rete possono essere espansi per fornire più informazioni. Comunque, la logica dei processi che coinvolgono la reportistica rimane la medesima: l'interfaccia dei servizi per interagire con i sistemi può rimanere invariata, ma con più informazioni utili. Al momento, nel nostro istituto ci sono ancora dati contenuti in file Excel: l'integrazione dei vari sistemi non è ancora completata, sebbene la maggior parte del lavoro sia stata compiuta. Al momento è in sviluppo anche un sistema di Research Data Managament.

Lo scopo ultimo dell'Istituto Italiano di Tecnologia è dotarsi di un singolo punto di accesso in cui ricercatori e personale amministrativo possano avere immediatamente il quadro generale dei dati nei vari sistemi e generare reportistica tramite pochi click in una applicazione web. L'applicazione Scientilla, insieme ai servizi con cui è integrata e alle interfacce che offre, è un grande passo avanti in questa direzione. In particolare, il fatto che

i ricercatori stessi osservino e confermino le informazioni loro riguardanti in Scientilla fornisce un controllo ulteriore per raggiungere una migliore qualità del dato finale e far emergere eventuali discrepanze fra i sistemi.

Ringraziamenti

Gli autori colgono l'occasione di ringraziare i colleghi del Data Analysis Office sviluppatori di Scientilla, Federico Bozzini, Dieter Casier e Federico Semprini.

Autori

Ugo Moschini - ugo.moschini@iit.it

Ugo Moschini ha conseguito un dottorato in Informatica presso l'Università di Groningen, Olanda, nell'analisi e classificazione di dati e immagini astronomiche. Ha lavorato per due anni all'Agenzia Spaziale Europea in Olanda e Germania nel supporto ad operazioni satellitari, brevettando un algoritmo di compressione dati. Al momento lavora al Data Analysis Office dell'IIT e si occupa dello sviluppo di metodi e strumenti per monitorare i risultati scientifici dell'istituto.

Elisa Molinari - elisa.molinari@iit.it

Elisa Molinari ha conseguito un dottorato in Bioingegneria e Bioelettronica al DIST di Genova, Italia, progettando e sviluppando una piattaforma hardware e software per la gestione di dati neuroscientifici. Ha poi ottenuto un postdoc all'IIT e collaborato con università e l'ospedale pediatrico Gaslini nel campo dell'analisi di immagini biomedicali. È adesso Lead Data Analyst all'IIT di Genova e si occupa della coordinazione e dello sviluppo di sistemi di monitoraggio dell'attività di ricerca.

DOI: 10.26314/GARR-Conf18-proceedings-19

The PhenoMeNal Cloud-based Virtual Data Processing and Analysis e-infrastructure

Luca Pireddu[1], Marco Enrico Piras[1], Antonio Rosato[2], Gianluigi Zanetti[1]

[1]Data-Intensive Computing Group, CRS4, [2] Magnetic Resonance Center, C.I.R.M.M.P.

Abstract. PhenoMeNal is a complete software-defined e-infrastructure for data processing and analysis on modern infrastructure-as-a-service (IaaS) cloud platforms, especially tailored for metabolomics. Through the online PhenoMeNal portal, personal instances of the PhenoMeNal Cloud Research Environment (CRE) – which includes widely used data analysis environments (such as Galaxy and Jupyter) and over 250 open source tools – can be easily deployed on-demand, through an automated process, on accessible cloud IaaS like OpenStack, Amazon Web Services or Google Compute Platform, and partner institutional providers such as ReCaS-Bari, Embassy Cloud and de.NBI Cloud. Each PhenoMeNal CRE version fixes the version of the tools it integrates, so deployments are completely reproducible. Thus, PhenoMeNal constitutes a turnkey solution for scientists to easily leverage cloud computing resources to accelerate their work on a user-friendly, workflow-driven, reproducible and shareable data analysis platform.

Keywords. Cloud computing, Infrastructure-as-Code, software containers, metabolomics, reproducibility

Introduction

PhenoMeNal (Peters K., et al. 2018) is a complete software-defined e-infrastructure for data processing and analysis on modern Infrastructure-as-a-Service (IaaS) cloud platforms, especially tailored for metabolomics. This field studies the small molecules in organisms to reveal insights into their metabolic biochemistry. Work in metabolomics typically generates large data sets that require computationally intensive analyses (Sugimoto M., et al. 2012) – e.g., terabytes of data for large cohorts with thousands of metabolite profiles (Joyce A. R., et al. 2006). Cloud computing offers a valuable resource for these computational analyses by allowing access to sufficient computing resources that can be instantiated on-demand and without capital investment.

PhenoMeNal fills the gap between cloud computing and metabolomics researchers by providing a complete and performant data analysis e-infrastructure that includes a large suite of standardised and interoperable metabolomics data processing tools and workflows. The PhenoMeNal e-infrastructure can be easily deployed onto public and private cloud environments, enabling scalable and cost-effective high-performance metabolomics data analysis while hiding the technical complexity from the user. Moreover, PhenoMeNal facilitates reproducible analyses through automated, sharable and citable workflows and through its versioned and reproducible data analysis platform. PhenoMe-

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

Nal is composed of the Cloud Research Environment, the Portal, and the integrated Scientific Tools and Workflows

1. The Cloud Research Environment

PhenoMeNal's central e-infrastructure component is the Cloud Research Environment (CRE). The CRE can be automatically deployed on most cloud provider services without any sophisticated technical input from the user. Deployments can be activated from the PhenoMeNal web-based portal (described in a later section) or from the command line. The CRE integrates:

- the Galaxy (Afgan E., et al. 2018) and Jupyter (Kluyver T., et al. 2016) data analysis platforms and the Luigi workflow engine;
- over 250 standard metabolomics software tools, packaged by PhenoMeNal;
- ten tested and validated metabolomics data analysis workflows.

Thus, once deployed, the CRE allows the user to start processing data without any further software installation. At the time of this writing, the PhenoMeNal CRE can be deployed on a number of commercial cloud providers, including Amazon Web Services, Microsoft Azure and Google Cloud Platform, and on OpenStack-based private or institutional clouds. Of this latter type, the following partner providers are preconfigured for deployments from the PhenoMeNal portal: EMBASSY Cloud, de.NBI Cloud, and ReCaS-Bari.

From a technical standpoint, the PhenoMeNal CRE is designed as a microservice architecture, using containers to provision microservices for data analysis tools and long-running services such as workflow managers. Kubernetes (k8s) is used to orchestrate containers over the computing infrastructure. PhenoMeNal implements various layers to provision k8s on top of an IaaS. To start, Terraform is used to provision the required computing resources from IaaS – e.g., storage volumes, network resources and compute instances. Terraform provides a cloud-agnostic layer which allows PhenoMeNal to more easily support different cloud providers. The virtual machines are initialised with a customized base image and a Kubernetes cluster with GlusterFS is deployed on them. This entire procedure has been automated and generalized and provided as a new k8s deployment tool called KubeNow (Capuccini M., et al. 2018), which also supports the complete management of the deployment's life-cycle – creation, maintenance, destruction. A KubeNow plugin implements the deployment of the PhenoMeNal-specific services.

This step is performed through the installation of specific Helm charts (the equivalent of a package manager for k8s) which in turn result in the deployment of Docker containers implementing the various services. To realize this solution we created suitable container images for the various services, in addition to the Helm charts themselves; relevant charts and images were contributed to the community – e.g., the Galaxy Helm chart and contributions to the galaxy-kubernetes container images (https://github.com/galaxyproject/galaxy-kubernetes). By default, secure access via https is configured via Cloudflare (http://cloudflare.com), providing dynamic DNS services and encryption for all user communication with services inside the CRE.

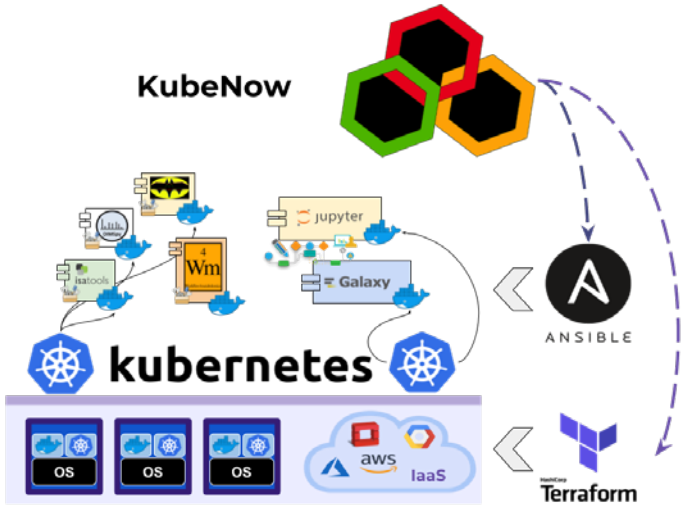


Fig. 1
PhenoMeNal Architecture. Kubernetes is used to provide a uniform platform on any IaaS. All services and tools are packaged as Docker container images and pulled on demand. Terraform and Ansible are used to provision and configure the base infrastructure.

2. The PhenoMeNal Portal

The PhenoMeNal portal (https://portal.phenomenal-h2020.eu) orients users through the features and resources provided by the e-infrastructure and, perhaps most importantly, provides the functionality to deploy and manage CREs through a web interface. Deployments to can be made using an easy-to-follow wizard and managed through a dashboard. The wizard (Fig. 2) collects all required information – e.g., credentials and parameters to access the cloud provider – and then executes the deployment procedure. Interestingly, the user can also customize the computational resources provisioned and the version of the PhenoMeNal to deploy. Once the CRE is operational, the user can follow the links on the dashboard to access the various running services – i.e., Galaxy, Jupyter, etc. – and put them to work, using all the integrated tools and workflows. In addition to this functionality, the portal also gives users access to the PhenoMeNal public instance, which is a public CRE hosted by EMBL-EBI that allows users to test-run a CRE without the need to deploy their own.

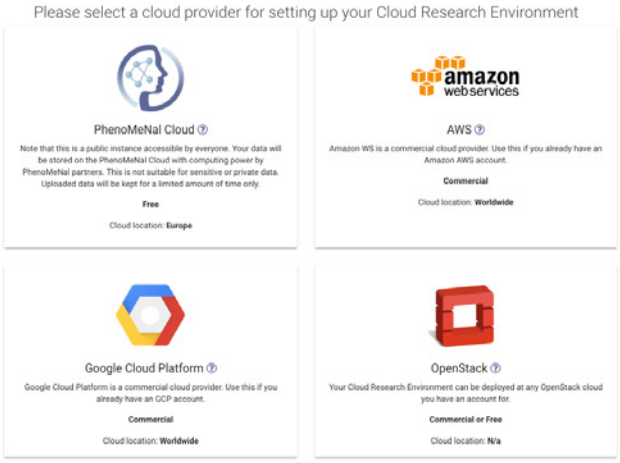


Fig. 2
CRE wizard: cloud provider selection during CRE deployment configuration on the PhenoMeNal portal

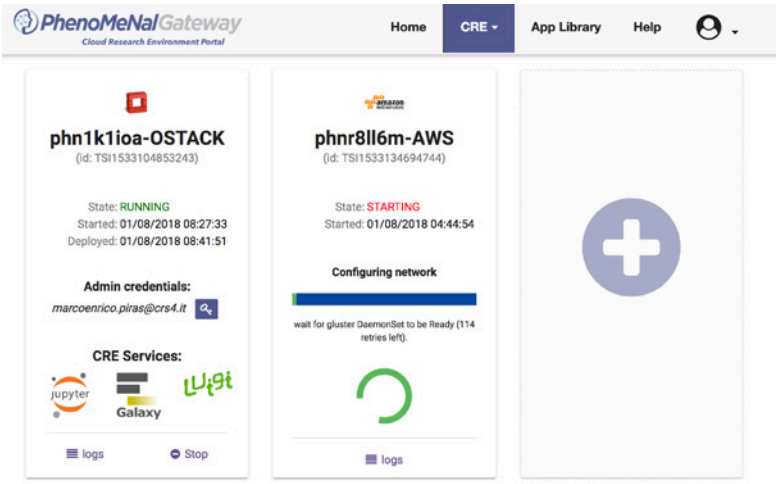


Fig. 3
CRE dashboard on the PhenoMeNal portal.

3. Integrated Scientific Workflows and Tools

PhenoMeNal integrates over 250 internally and externally developed metabolomics tools – which are mainly accessible through the CRE's Galaxy instance. These tools have been containerized by the PhenoMeNal project and integrated in the CRE. As part of the packaging process, we have defined minimal documentation requirements and implemented continuous integration testing. A number of scientific workflows built with these tools are also integrated in the CRE's Galaxy instance; many of these implement analysis procedures from published studies, making it simple to apply them to new data. To keep the CRE deployment relatively light, the container images are dynamically pulled from the PhenoMeNal Docker registry and executed on demand. Furthermore, to support interoperability PhenoMeNal has implemented native support in Galaxy for the standard ISA-Tab and ISA-JSON metabolomics data types (Rocca-Serra P., et al. 2010), coupled with tools to convert between ISA and other formats and to retrieve and store datasets in these formats to MetaboLights (Steinbeck C., et al. 2012) – one of the main metabolomics public data repositories.

4. Extensibility

The PhenoMeNal e-infrastructure can easily be repurposed to other scientific domains by simply changing the tools and workflows that are integrated. Work is already in progress to configure a CRE for genomic data analysis.

5. Conclusions

The PhenoMeNal e-infrastructure allows users to instantiate identical data analysis environments – down to the version of the individual tools – on any compatible IaaS through a simple graphical user interface. Thus, it is a turnkey solution for leveraging the potential of IaaS with very little financial and training investment. Coupled with the possibility of storing, exporting and sharing workflows, PhenoMeNal is a valuable means to ensure the reproducibility of scientific data analysis work. PhenoMeNal can be found at http://phenomenal-h2020.eu. The GitHub repository http://github.com/phnmnl hosts all source code

and the Wiki. Source code and documentation are available under the terms of the Apache 2.0 license. Integrated open source tools are available under the respective licensing terms.

6. Acknowledgements

PhenoMeNal is the work of the entire PhenoMeNal project consortium and was funded by the European Commission's Horizon2020 programme, grant agreement n. 654241.

References

Afgan E., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucl. Acids Res., 46(W1), pp W537-W544.

Capuccini M., et al. (2018), KubeNow: an On-Demand Cloud-Agnostic Platform for Microservices-Based Research Environments, arXiv, 1805.06180 (preprint).

Joyce A. R., et al. (2006), The model organism as a system: integrating 'omics' data sets, Nat. Rev. Molecular cell biology, 7(3), p 198.

Kluyver T., et al. (2016), Jupyter Notebooks-a publishing format for reproducible computational workflows, ELPUB, pp 87-90.

Peters K., et al. (2018), PhenoMeNal: Processing and analysis of Metabolomics data in the Cloud, GigaScience (to appear).

Rocca-Serra P., et al. (2010), ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinf., 26(18), pp 2354-2356.

Steinbeck C., et al. (2012), MetaboLights: towards a new COSMOS of metabolomics data management, Metab., 8(5), pp 757-760.

Sugimoto M., et al. (2012), Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis, Curr. Bioinf., 7(1), pp 96-108.

Authors

Luca Pireddu - luca.pireddu@crs4.it
Luca Pireddu coordinates the Distributed Computing Group at CRS4. His current research focuses on novel applications of distributed stream computing and IaaS, with particular interest in problems pertaining to data-intensive biology and smart infrastructures. Luca holds a Ph.D. in Biomedical Engineering from the University of Cagliari, Italy, an M.Sc. Computing Science from the University of Alberta, Canada, and a B.Sc. in Computing Science from Laurentian University, Canada.

Marco Enrico Piras - marcoenrico.piras@crs4.it
Graduated in Computer Engineering at Università "La Sapienza" of Rome, since 2014 he works as a technologist at CRS4. His current research interest deals with the design and implementation of distributed systems for the processing and storage of large volumes of scientific data (Big Data), with a particular focus on the use of microservices architectures and the automation of their deployment.
Antonio Rosato - rosato@cerm.unifi.it

Antonio Rosato is Associate Professor of Chemistry at the University of Florence. He holds a Ph.D. in Chemical Sciences from the same University. His research interests span from structural biology of metalloproteins to NMR methods in the life sciences, and all associated computational aspects. He has co-authored more than 100 articles in scientific journals.

Gianluigi Zanetti - gianluigi.zanetti@crs4.it
Gianluigi Zanetti is the director of the Data-Intensive Computing sector at CRS4. He holds a degree in Physics from the University of Bologna and a Ph. D. in Physics from The University of Chicago. His research spans many areas of computational science and is widely published in major journals and conferences (over 100 refereed publications).

DOI: 10.26314/GARR-Conf18-proceedings-20

A Sensor-based app for self-monitoring of pill intake

Daniele Riboni

Università di Cagliari, Dipartimento di Matematica e Informatica

Abstract. Accurate adherence to prescribed medications is essential for the effectiveness of therapies, but several studies show that when patients are responsible for treatment administration, poor adherence is prevalent. Existing apps to support self-administration of drugs may interfere with the normal routine of patients by providing unnecessary reminders. More sophisticated solutions, including the use of smart packaging and ingestible sensors, are currently restricted to patients involved in a few clinical studies. In this paper, we introduce a novel app to support self-administration of drugs without interfering with the patient's routines. The system relies on cheap wireless sensors attached to medicine boxes to detect medicine intake. The app uses machine learning to detect intake events, and active learning to improve recognition based on the user's feedback. The app is available on Google Play Store.

Keywords. Pill reminder app, sensor-based activity recognition, human computer interaction

Introduction

Accurate adherence to medical prescriptions is a prerequisite for the effectiveness of therapies based on pharmaceutical drugs. Unfortunately, several patients experience difficulties in complying to the medical prescriptions. This problem is particularly relevant for patients with chronic illnesses, and can lead to increased use of healthcare services and detriment of life quality (Majumdar et al. 2006).

When non-adherence is unintentional, one of the main reported reasons is forgetfulness (Bartlett 2002). Consequently, different tools have been proposed to remind the patient about the medications to take. A simple solution is the use of "medication reminder packaging", where the medicine box incorporates the prescribed date and time of drug intake. However, those solutions showed a modest improvement of adherence on the long term (Mahtani 2011). RFID tags were used in (Riboni 2016) to let the patient manually register the medicine intake. Other tools actively remind the user, using pagers, custom devices, short message service and, more recently, smartphone apps (Dayer 2013).

Those tools proved to be quite effective on the short term; however, the effects on the long term are unclear (Vervloet 2012). Moreover, existing reminders are context-oblivious: the patient is notified with a reminder at each scheduled time of intake. Hence, such reminders may interfere with the normal routine of users, since they are issued even when patients have actually taken the prescribed drug at the prescribed time. Ideally, the patient should receive a reminder only when he/she actually forgets to take the prescribed medicine.

In this paper, we address the challenging issue of devising a system to support self-ad-

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

ministration of drugs, which does not interfere with the normal routine of the patient. We introduce a novel system based on a combination of cheap wireless sensors, a smartphone app, and cloud infrastructure, to monitor medicine intake and provide reminders only when actually needed. Tiny accelerometers stuck to medicine boxes communicate motion data through a Bluetooth low energy interface to the patient's app. The app preprocesses accelerometer data to extract features of interest, and uses machine learning methods to detect intake actions. Detected intake actions are compared with the scheduled therapy. In case of missed intake, the app notifies the user with a reminder. The patient is asked to confirm of refute the detections of the app. The user's feedback is exploited by the app to fine-tune the prediction algorithm to the user's habits, using an active learning method.

1. System overview

The system includes a smartphone running the smart reminder app, tiny Bluetooth sensors (named tags) attached to medicine boxes, and communication with the cloud to acquire data about tags and to process the data at the server side. The app exploits user feedback to fine-tune medicine intake recognition to the patient's habits. A Web-based dashboard is available to clinicians for inspecting the history of medicine intakes of their patients.

As in normal pill reminder apps, the patient manually fills his/her therapies and the prescribed times of intake. However, through the smart reminder app, the patient also associates each therapy to a colored tag (Fig. 1), which is actually a tiny Bluetooth low energy (BLE) beacon with an integrated accelerometer.

The app queries the cloud to get tags information, including the color and the kind of image depicted on its surface, in order to facilitate tag identification. After the association, the patient sticks the tag to the medicine box.

When moved, the tag broadcasts packets containing its identification number and triaxial acceleration. Those packets are acquired by the app and analysed by a ML algorithm, which is in charge of classifying the movements of the medicine box in either “intake action” or “other action”.

When the app detects a medicine intake action at the prescribed time, it discreetly displays a pop-up message asking the patient to either confirm or refute the intake, and to fill a mood and pain scale.

For the sake of this project, mood and pain values are important to understand why the patient is taking or not the prescribed medicines.

If the app does not detect the intake of a prescribed medicine within a time threshold, it vibrates, emits a sound and displays a reminder (Fig. 2). The patient can either confirm that he/she forgot to take the prescribed drug, or may report a misprediction of the app, indicating the actual time of intake.

The app has a calendar function showing the prescribed medicines to be taken in the current time of the day, as well as a “performance” function displaying the rates of correct intakes and average mood and pain values in the current day, week and month.

The app is part of the DomuSafe research project (http://sites.unica.it/domusafe/), and can be used both by patients participating to the project's experimental evaluation,

and by other users.

The app data (therapies, motion data, medicine intakes, mood and pain values) are periodically communicated to a server in the cloud. The server provides a Web-based dashboard through which clinicians can evaluate the adherence to prescriptions of patients participating to the evaluation, and inspect the trends of pain and mood. Communication with the cloud is done through an encrypted channel. The data are stored on the server in an anonymous form. Each patient participating to the experiment is identified by a unique code; the association among codes and identities is known by the clinician only.

The app has been developed for the Android platform, using the Weka libraries for implementing the ML algorithms. In the current implementation, we adopt Estimote (https://estimote.com/) stickers as our tags. Stickers have an adhesive side that makes it easy to stuck them to medicine boxes. Their communication range is sufficient to cover most apartments. Stickers are disposable and have a life time of approximately one year. At the time of writing, their cost is ten US dollars each. The Web dashboard is implemented in PHP and HTML5.

Fig. 1
BLE sensor stuck
to a medicine box

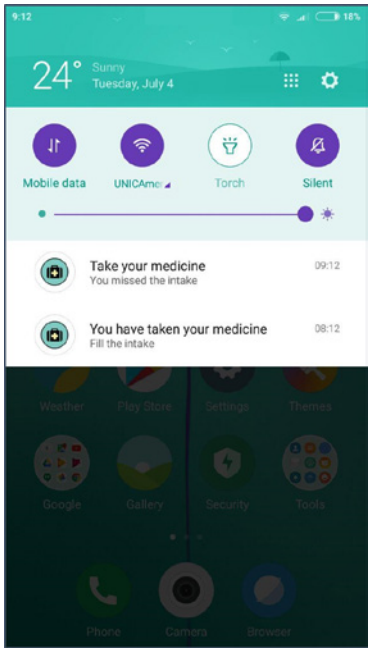


Fig. 2
Context-aware
notification

2. Conclusion and future work

Our system for self-monitoring pill intake is based on a smartphone app and cheap wireless sensors to be attached to medicine boxes. We implemented the system and conducted a preliminary experimental evaluation.

Future work includes experimenting our system with real patientsbto evaluate both the utility and usability of our system. We also plan to investigate more advanced active learning methods, in order to fine-tune the prediction model to the patient's context and habits while reducing the effort of providing feedback to the app.

References

J. A. Bartlett. (2002). Addressing the Challenges of Adherence. J Acquir Immune Defic Syndr 29, S2-S10 (2002).

L. Dayer et al. (2013). Smartphone Medication Adherence Apps: Potential Benefits to Patients and Providers. JAPhA 52 (2013), 172-181. Issue 2.

K. R. Mahtani, C. J. Heneghan, P. P. Glasziou, and R. Perera. (2011). Reminder packaging for improving adherence to self-administered long-term medications. Cochrane Database Syst Rev 9 (2011).

S. R. Majumdar, R. S. Padwal, R. T. Tsuyuki, J. Varney, and J. A. Johnson. (2006). A meta-analysis of the association between adherence to drug therapy and mortality. BMJ 333, 15 (2006).

D. Riboni, C. Bettini, G. Civitarese, Z. Haider Janjua, and Rim. Helaoui. (2016). SmartFABER: Recognizing Fine-grained Abnormal Behaviors for Early Detection of Mild Cognitive Impairment. Artificial Intelligence in Medicine 67 (2016), 57-74.

M. Vervloet, A. J Linn, J. C M van Weert, D. H de Bakker, M. L Bouvy, and L. van Dijk. (2012). The effectiveness of interventions using electronic reminders to improve adherence to chronic medication: a systematic review of the literature. J Am Med Inform Assoc 19 (2012), 696-704.

Autore

Daniele Riboni - riboni@unica.it

Daniele Riboni is associate professor at the Department of Mathematics and Computer Science of the University of Cagliari. He received his PhD in Computer Science from the University of Milan in 2007. His main research interests are in the area of knowledge management for mobile and pervasive computing, context awareness, activity recognition, and data privacy. He is the author of over 60 publications and his contributions appear in highly-ranked international journals and conference proceedings.

DOI: 10.26314/GARR-Conf18-proceedings-21

Customer Satisfaction from Booking

Maurizio Romano, Luca Frigau, Giulia Contu, Francesco Mola, Claudio Conversano

Università di Cagliari, Dipartimento di Science Economiche ed Aziendali

Abstract. Nel presente studio, analizzando i commenti dei clienti presenti su Booking, si è creato un modello affidabile al 91% che: a) supporta le strutture ricettive nell'individuazione di pregi e difetti, identificando i fattori su cui operare per ottenere un miglioramento del servizio offerto; b) consente d'identificare i punti di forza e debolezza della destinazione in cui opera la struttura ricettiva; c) può essere applicato in considerazione di differenti ambiti territoriali; d) consente di prevedere lo score di una recensione ed i relativi punti di forza/debolezza. Tale modello sfrutta l'affidabilità, la capacità e il filtraggio della Rete per effettuare le operazioni d'acquisizione periodica senza restrizioni e per rendere fruibile l'applicazione con il correlato Data Warehouse.

Keywords. Customer, Sentiment, Booking, WoM, Bayes

1. Introduzione

Questo studio è stato realizzato nell'ambito del progetto P.I.A. "Realizzazione di una piattaforma ICT a supporto del settore turistico" (RAS, 2007/2013), il cui obbiettivo è sviluppare una piattaforma ICT che consenta di valutare la Customer Satisfaction per il settore turistico sardo.

Sono state analizzate le recensioni dei clienti per ciascuna struttura ricettiva operante in Sardegna e presente su Booking. Considerando che nel settore turistico si ricorre all'applicazione automatica della sentiment analysis sia tramite machine learning che tramite metodi basati sul lessico per prevedere quando una recensione sia positiva o negativa (Schmunk et al., 2013), si è proceduto implementando un programma Python che estraesse tutte le informazioni utili fornite da Booking relative alle strutture d'interesse. Successivamente, le recensioni sono elaborate in linguaggio naturale al fine di comprendere per quali aspetti il cliente sia soddisfatto o meno e per fornire uno strumento di benchmarking delle strutture ricettive.

2. Obiettivi

La scelta della piattaforma Booking è dipesa dal fatto che le recensioni in esso presenti implicano che il recensore abbia soggiornato in una struttura. Booking utilizza una valutazione in scala [2.5; 10] e suddivide ogni recensione in due parti: un commento positivo ed uno negativo.

Gli scopi dell'analisi sono quindi molteplici:

- creare un classificatore che, dato un commento, lo classifichi in negativo/positivo;
- misurare quantitativamente l'incidenza (negativa o positiva) di una data parola all'interno del commento;

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

- sviluppare un modello di previsione per lo Score di Booking in funzione delle recensioni;
- produrre uno strumento di benchmarking.

Per raggiungere gli scopi prefissati si è reso necessario attuare due fasi: 1) data collection: i dati sono stati estrapolati da Booking e trattati in preparazione dell'analisi statistica; 2) analisi delle recensioni: è stato implementato un classificatore Naive Bayes* ad-hoc e un modello di previsione dello score di una recensione su Booking in funzione delle parole in essa contenute.

3. Data Collection

E' stato implementato un estrattore Python che, mediante web scraping, ha estrapolato tutte le possibili informazioni d'interesse offerte pubblicamente da Booking e le ha inserite in tabelle create ad hoc.

Fig. 1
Esempio di una pagina di Booking contenente i giudizi dei clienti

I dati di Booking riguardano due tipologie d'informazione (Fig. 1):

- le 619 strutture alberghiere operanti in Sardegna;
- le 66237 recensioni per ciascuna struttura, in italiano o in inglese, scisse nei due commenti (positivi e negativi) che le compongono, per un totale di 106800 commenti.

I dati delle strutture alberghiere sono organizzati in una tabella (Fig. 2) ove ciascuna riga rappresenta una struttura e le cui colonne sono il risultato dell'accorpamento delle informazioni reperibili sulle stesse su Booking (tipologia di struttura, CAP, comune, valutazioni sulla pulizia, comfort, ecc.)

Nome Struttura	Tipologia	Cap	Comune/Località	...
Struttura 1	Extralberghiere	09044	Sant' Isidoro	...
Struttura 2	3 Stelle	09049	Villasimius	...
Struttura 3	3 Stelle	07013	Mores	...
Struttura 4	4 Stelle	09123	Cagliari	...
...	...	...	...	...

Fig. 2
Estratto della struttura tabellare contenente i dati delle strutture alberghiere

I dati di ciascuna recensione successivamente vengono organizzati in un'altra tabella (Fig. 3) contenente, oltre al commento, le informazioni utili per risalire alla struttura e alla recensione originale, nonché i dati sul recensore e sulla tipologia di cliente (viaggio di lavoro, viaggio di piacere ecc.).

Nome Struttura	ID Commento	ID Review	Commento	Neg-Pos	Score	...
Struttura 1	1	1	christina was the best...	Pos	10.0	...
Struttura 1	2	2	we booked an appartment...	Pos	10.0	...
Struttura 1	3	3	we travelled into cagliari...	Pos	9.2	...
Struttura 1	4	4	it was fantastic	Pos	10.0	...
...	...	...	...	...	...	...

Fig. 3
Estratto della struttura tabellare contenente i dati dei commenti/recensioni

4. Riorganizzazione e Data Cleaning

I dati sono stati riorganizzati raggruppando le parole con il medesimo significato e ripulendoli da congiunzioni, punteggiatura, numeri e altre stopwords.
Affinché sia possibile trattare nello stesso modo parole dal medesimo significato e, in funzione di questo, calcolare opportunamente le frequenze di ciascuna di esse, si è reso necessario procedere a un accorpamento. Il modo più semplice per implementarlo consiste nel sostituire nei commenti originali ogni parola accorpabile con una "più comune" per significato.
Successivamente, tutte le parole di ogni commento sono state estrapolate e raggruppate in un insieme unico (Bag Of Words – BOW, Fig. 4). Questo raggruppamento ha consentito di calcolare le frequenze per ciascuna parola.

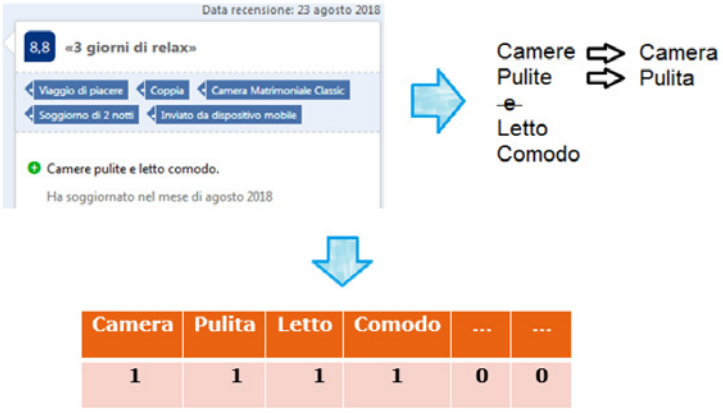


Fig. 4
Schema riassuntivo della logica di produzione del BOW

L'analisi delle frequenze delle parole ha consentito di definire delle macro-categorie che andassero a raggrupparle (Fig. 5). Nello specifico, sono state individuate le seguenti categorie: bar, cleaning, comfort, food, hotel, position, pricequalityrate, room, services, sleepquality, staff, wifi, other.

La categorizzazione è stata utilizzata esclusivamente per trattare i dati in maniera aggregata. Ciò consente di analizzare i pregi e i difetti relativi ai servizi offerti per le singole categorie. Il classificatore Naive Bayes* non le sfrutta poiché un tale accorpamento ridur-

rebbe sensibilmente la variabilità determinando un peggioramento generale della capacità predittiva del modello. Tuttavia, l'accorpamento viene utilizzato per sostituire temporaneamente le parole presenti in un commento con la categoria rilevata, al fine di comprendere in maniera immediata a quali categorie lo stesso appartenga o meno.

Fig. 5
Estratto della struttura tabellare a supporto per la categorizzazione

Word	Category
Colazione	Food
Ristorante	Food
Bread	Food
Cakes	Food
Mangiare	Food
Conto	Price-quality rate
Caro	Price-quality rate
Pagamento	Price-quality rate
Pay	Price-quality rate
Gestore	Staff
Stintino	Position
Orosei	Position
...	...

5. Analisi delle recensioni

L'analisi delle recensioni consiste nel creare un classificatore in grado di prevedere accuratamente se un dato commento sia positivo o negativo in funzione delle parole che lo compongono. Sono stati applicati i classificatori più comuni suggeriti dalla letteratura e, osservando la qualità dei risultati da loro prodotti (Fig. 8), il modello Naive Bayes* è risultato il migliore in base alla capacità di generalizzazione.
Successivamente, considerando l'unione del commento positivo e negativo di ogni recensione, si è applicato il classificatore anche alle recensioni nel loro insieme, e quindi in contesti in cui la separazione fra aspetti positivi e negativi non è netta.

*5.1 Il modello Naiva Bayes**

Sfruttando un approccio frequentista, si calcola la probabilità che un commento sia negativo in rapporto alla probabilità che esso non lo sia (Rischio Relativo o Likelihood Ratio, LR). Il teorema di Bayes semplifica i calcoli e consente di stimare il LR in funzione di numero e tipologia di parola, o del commento stesso. Si descrive di seguito formalmente il modello.
Sia W l' insieme di tutte le parole comparse nelle recensioni.
Siano $C_{pos} \subseteq W$, $C_{neg} \subseteq W$ due commenti.
Sia $R = C_{pos} \cup C_{neg}$ una recensione.
Il Likelihood Ratio è dato da con $LR_C = \dfrac{P(neg|C)}{P(pos|C)}$ $con\ C \subseteq W$

Per il teorema di Bayes, $\log(LR_C) \cong pres_C + abs_C + \log\left(\dfrac{P(neg)}{P(pos)}\right)$.
in cui:

$$pres_C = \sum_{w \in C} \log\left(\frac{P(w|neg)}{P(w|pos)}\right)$$

$$abs_C = \sum_{w \in W \setminus C} \log\left(\frac{P(\overline{w}|neg)}{P(\overline{w}|pos)}\right)$$

$pres_C$ e abs_C rappresentano, rispettivamente, le intensità delle parole presenti e non presenti nel commento.

Il *LR* può essere calcolato sia per un commento, sia per una recensione che per una singola parola od un insieme ben preciso delle stesse (ad es. una categoria). Ogni parola ha dunque un valore di *log(LR)* nel caso in cui essa sia presente nel commento ed un altro valore di *log(LR)* in cui non sia presente. Il *log(LR)* del commento sarà dato dalla somma di $pres_C$ e abs_C.

Tale valore viene calcolato per ogni singola parola presente nel BOW. I risultati vengono inseriti in una tabella (rappresentata in Fig. 6): nelle colonne sono indicate le parole e, nelle righe, i valori di *log(LR)* nel caso in cui la parola sia presente o assente e la proporzione di commenti negativi/positivi che contengono la parola stessa.

	"Assolutamente"	*"Mare"*	...
P(neg)	0,011	0,026	...
P(pos)	0,007	0,075	...
Log(LR) Parola presente	0,411	-1,077	...
Log(LR) Parola assente	-0,004	0,052	...

Fig. 6
Estratto della tabella in output dopo il calcolo dei log(LR)

Ad ogni commento (positivo o negativo) è dunque assegnato un valore di *log(LR)*. Si ricorre successivamente alla Cross Validation per stimare il valore soglia (τ) tale che:

se *log(LR)* > τ → il commento è classificato come "Negativo",

se *log(LR)* ≤ τ → il commento è classificato come "Positivo".

La scelta di τ, così come si osserva in Fig. 7, è stata effettuata minimizzando l'indice di errata classificazione (Misclassification Error) per entrambi gli errori che il classificatore Naive Bayes* può compiere al variare dei valori assumibili da τ.

Il valore stimato di τ è: τ=1,138.

La somma dei *log(LR)* dei commenti positivi e negativi di una recensione indica, in funzione

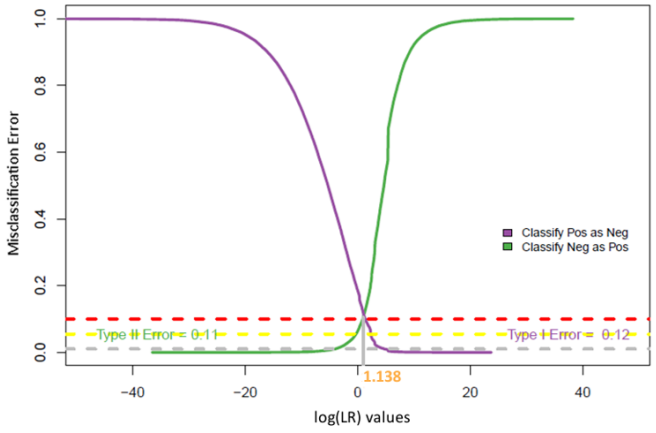


Fig. 7
Le due tipologie di errore verificabili nella classificazione dei commenti e gli andamenti del valore soglia τ

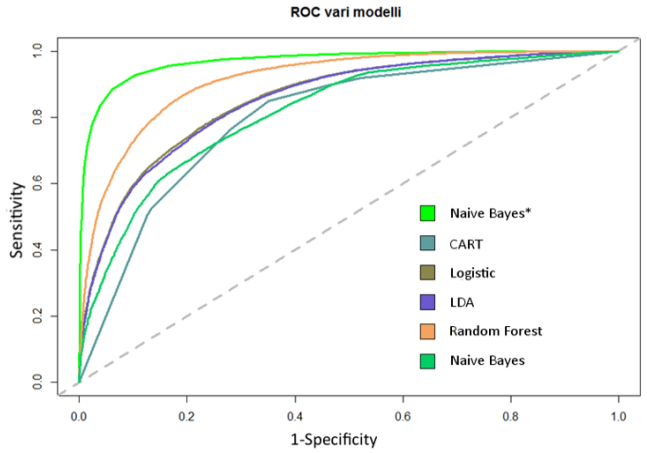


Fig. 8
Curve ROC del classificatore Naive Bayes* e degli altri classificatori utilizzati nell'analisi

del τ, quale sia l'andamento della stessa, ovvero se la recensione è da considerarsi, nel suo insieme, come negativa/positiva e di quanto questa lo sia (a seconda della distanza da τ).

In Fig. 8 e nella Tab. 1 è possibile confrontare le performance del modello Naive Bayes* e degli altri classificatori. Il modello qui proposto performa meglio rispetto a ciascuna misura di benchmarking.

Tab. 1
Tabella delle più comuni misure di benchmarking per i vari classificatori

Model	Misclassification Error	Accuracy	Sensitivity	Fall-out	F1 score	Matthews Correlation Coefficient
Naive Bayes*	**0,089**	**0,911**	**0,929**	**0,117**	**0,926**	**0,813**
Logistic	0,150	0,850	0,884	0,532	0,877	0,361
Random Forest	0,189	0,811	0,873	0,591	0,849	0,303
Naive Bayes (e1071)	0,194	0,806	0,804	0,389	0,834	0,390
Naive Bayes (klaR)	0,194	0,806	0,804	0,389	0,834	0,390
CART	0,232	0,768	0,842	0,587	0,815	0,272
LDA	0,236	0,764	0,860	0,641	0,816	0,246

5.2 Previsione score di Booking

Al fine di produrre una previsione dello score di una recensione su Booking in funzione delle parole in essa contenute, sono stati addestrati diversi modelli. Il miglior modello di previsione dello score di una recensione è risultato essere il random forest (errore quadratico medio: 0,6704). Come predittori sono state utilizzate le seguenti variabili create mediante l'uso di Naive Bayes*:

- review_position, review_bar, review_cleaning, ..., review_services, che rappresentano i valori di *log(LR)* per ciascuna delle 13 categorie definite;
- polarità, ossia la classificazione in pos/neg del *log(LR)* generale della recensione tramite τ.

6. Conclusioni

Il presente lavoro ha consentito di definire un modello (affidabile al 91%) che, sulla base

del *LR* calcolato sulle recensioni dei clienti su Booking, è capace di:

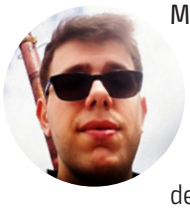
- supportare le strutture ricettive nell'individuazione dei propri punti di forza e debolezza identificando gli aspetti sui quali operare per ottenere un miglioramento del servizio offerto;
- consentire l'identificazione di pregi e difetti della destinazione in cui opera la struttura ricettiva;
- essere applicato in differenti ambiti territoriali;
- prevedere lo score di una recensione su Booking e, per ciascuna classe definita, i punti di forza/debolezza. In questo modo è possibile definire gli aspetti su cui intervenire per migliorare il risultato ottenuto.

Cardine del Progetto è l'impiego della rete scientifica GARR in quanto consente di effettuare sia le operazioni di acquisizione periodica senza restrizioni, sia di rendere fruibile l'applicazione del modello secondo la modalità nativa as-a-Service per il relativo Data Warehouse (DW) offrendo al contempo ampi spazi di sviluppo in settori scientifici limitrofi.

Riferimenti bibliografici

Schmunk Sergej, Wolfram Höpken, Matthias Fuchs, and Maria Lexhagen. (2013). Sentiment Analysis: Extracting Decision Relevant Knowledge from UGC. In Information and Communication Technologies in Tourism 2014, Zheng Xiang and Iis Tussyadiah, Cham.

Autori

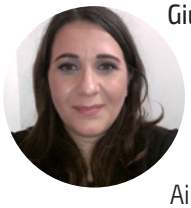


Maurizio Romano - romano.maurizio@unica.it

Maurizio Romano è dottorando in Scienze Economiche ed Aziendali presso l'Università degli Studi di Cagliari. Laureatosi in Informatica con una tesi afferente al Semantic Web, ha vinto una borsa di ricerca per la "Creazione di algoritmi per l'analisi di dati del settore turistisco". Tutor didattico nel corso "Metodi di Apprendimento Statistico per il Data Science" della Laurea Magistrale in Data Science, Business Analytics e Innovazione, incentra la sua ricerca su Statistical Learning e Big Data.

Luca Frigau - frigau@unica.it

Luca Frigau è ricercatore presso il dipartimento di Scienze Economiche ed Aziendali dell'Università degli Studi di Cagliari, dove insegna Statistica e Analisi di Mercato. Ha conseguito il dottorato di ricerca in Probability and Mathematical Statistics presso la Charles University (Praga) e i suoi principali interessi di ricerca sono il Machine Learning, il Classification assessment, il Data Mining e il Data Science.



Giulia Contu - giulia.contu@unica.it

Giulia Contu è un dottore di ricerca in "Scienze del turismo: metodi, modelli e politiche", Università di Palermo, Italia), è attualmente dottoranda presso il Dipartimento di Scienze economiche e Aziendali dell'Università degli Studi di Cagliari. I suoi interessi includono la statistica del turismo, il revenue management, il turismo sommerso e il fenomeno di Airbnb. È membro della Società Italiana di Scienze del Turismo (SISTUR).

Francesco Mola - mola@unica.it

Francesco Mola è professore ordinario di Statistica Metodologica, si occupa di Statistica Computazionale e Data Science. È Prorettore Vicario dell'Università degli Studi di Cagliari, dove tiene corsi di Statistica e Data Science ed è stato Direttore del Dipartimento di Scienze Economiche ed Aziendali. I principali temi di ricerca riguardano la statistica multivariata, computazionale e il Data Mining. È stato membro del board dell'International Association for Statistical Computing.

Claudio Conversano - conversa@unica.it

Claudio Conversano è professore associato di Statistica metodologica e docente di Statistica, Modelli statistici per l'asset allocation, Analisi di big data e Quantitative Methods for Management presso l'Università degli Studi di Cagliari. I suoi principali interessi di ricerca riguardano i metodi di apprendimento statistico (statistical learning), i Big Data, la finanza computazionale, l'inferenza causale e la qualità dei dati statistici.

DOI: 10.26314/GARR-Conf18-proceedings-22

"Open Monuments Engine", proposta per un database aperto a supporto dei GIS, dei sistemi di navigazione satellitare e della Storia

Luigi Serra

CNR - ISEM, Istituto di Storia dell'Europa Mediterranea

Abstract. Il presente contributo scaturisce, come sua naturale evoluzione, dalla proposta presentata alla conferenza internazionale della Japanese Association for Digital Humanities del 2018 (Serra L. 2018), anche se in realtà si configura come uno dei requisiti essenziali per la sua attuazione. Salutata con entusiasmo dagli informatici e umanisti nipponici sposa i punti-chiave suggeriti dalla conferenza GARR 2018. La proposta progettuale consiste nell'avvio dello studio, implementazione e popolamento collaborativo di un database aperto, liberamente consultabile, per il censimento e la schedatura di monumenti secondo strutture aggregate non solo su base geografica, ma soprattutto storica, temporale e istituzionale, finalizzato all'impiego su sistemi di navigazione satellitare.

Keywords. Database, Digital Humanities, Humanities Transfer, Open Data, Open Science

1. Introduzione

Aggregare criticamente i monumenti desiderati secondo logiche diverse e per tutte ottenere comunque un percorso guidato dai navigatori satellitari, è una funzionalità che attualmente mi è difficile trovare sui dispositivi in circolazione. Cimentarsi nel pianificare un viaggio con un VNS (Vehicle Navigation System) o anche con web App su smart device nel tentativo di inserire specifici monumenti appartenenti a determinate epoche storiche, a seconda dei soggetti di interesse, è cosa tutt'altro che semplice: i navigatori sono eccellenti con i "dove" nel districare percorsi (Quddus M.A. et al. 2007), e ottimizzare le rotte (Dijkstra E.W. 1959), ma poco o nulla è presente sul "quando" o sul "perché" del monumento.

Quando si inseriscono i PoI (Point of Interest), vengono proposti in genere sulla traccia calcolata e in prossimità del percorso seguito (ad es. Google propone una cosa simile a "cerca nelle vicinanze" o "cerca in questa zona"). I Point of Interest scaturiscono dalle query secondo categorizzazioni gerarchiche di alto livello in cui la tassonomia ha una granularità di due o tre ordini, prevalentemente nell'intorno del punto geografico, del tracciato o della ristretta zona selezionata.

I navigatori GPS, o in generale GNSS (Global Navigation Satellite System), sono giustamente focalizzati sui "luoghi", ma ben poco raccontano della loro "storia" perché questa non è contemplata nei DB GIS (Geographic Information System) di riferimento dei navigatori e relativi PoI.

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

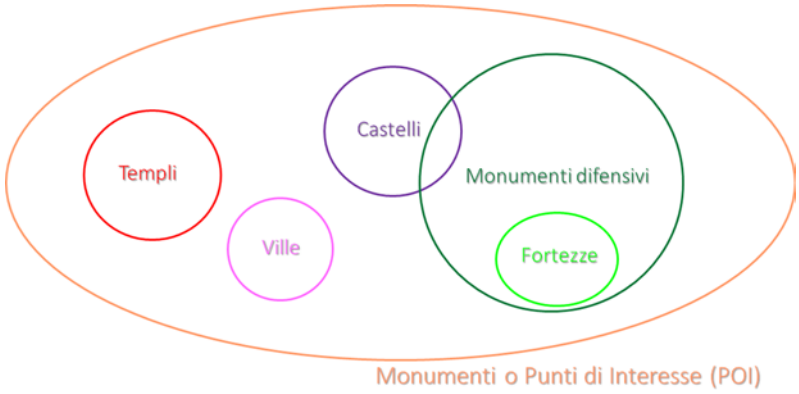


Fig. 1
Esempio possibile di tassonomia dei monumenti sui VNS

Questa lacuna nei record dei database in uso sui VNS, è una delle cause della mancata completa fruibilità trasversale della tecnologia, dedicata all'orientamento e alla navigazione affidabile, ma non alla pianificazione critica degli itinerari di qualche utilità agli studiosi e ai ricercatori (Carta M. et al. 2010), (Serreli G. 2015). Vorrei che il percorso venisse calcolato non solo su destinazioni e punti GIS, ma anche su monumenti, periodi, temi e argomenti di interesse (Serra L. 2017), epoca di appartenenza, Istituzione che li ha costruiti, Stato o appartenenza statuale (Casula F.C. 1997).

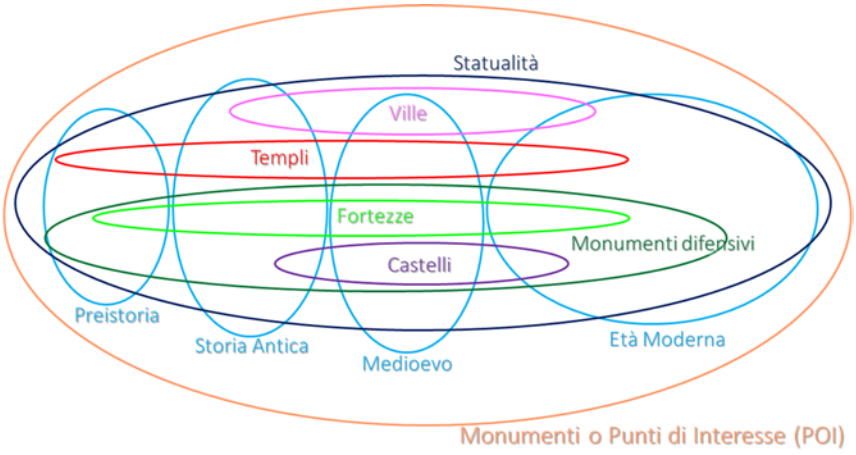


Fig. 2
Esempio di tassonomia desiderata da implementare sui VNS

Attuando una siffatta categorizzazione, assisteremmo a una dinamicità che sistemi opportunamente istruiti o progressivamente dotati di Expert Systems e Artificial Intelligence, potrebbero proporci sulla base dei nostri interessi. Da questo punto di vista osserveremmo infatti una "migrazione statica" del monumento che si evidenzierebbe solamente con l'aggiunta della variabile temporale, salvo rari casi di spostamento fisico causato da esigenze cogenti (ad es. Chiesa romanica di Zuri (OR-Italia), Igrexa de San Xoán de Portomarín (Galizia-Spagna) Abu Simbel (Egitto)). L'informazione puramente geografica sul monumento è per sua natura statica in quanto legata al luogo della sua collocazione, ma con l'aggiunta di quella storica diverrebbe dinamica se vista sotto gli occhi della sua appartenenza alle diverse Istituzioni che si sono succedute nei secoli.

Ci sono innumerevoli e utili programmi per la navigazione su PC e App per smart device, ciascuno con differenti funzioni e peculiarità spesso disgiunte perché alimentate da database proprietari, non comunicanti tra loro, che ci impongono di utilizzarle separatamente a seconda delle esigenze e dell'obbiettivo prefissato: geo-localizzazione, tracking, posizionamento sentieri, monumenti e attrazioni. La ricchezza di contenuti in ciascuna di esse è il valore aggiunto del produttore che decreta il successo di una piuttosto che un'altra soluzione; disporre di un'unica fonte autorevole da cui attingere liberamente dati, favorirebbe il perfezionamento delle implementazioni software attuali e future.

2. Gruppo di lavoro e linee guida

Per creare un database che consenta la catalogazione dei monumenti con record funzionali alle group by per tema o periodo storico, è necessaria un'accurata progettazione che coinvolga esperti, ricercatori e accademici ad integrazione degli studi sulle esigenze di fruibilità dei monumenti. Una loro classificazione e precisa collocazione storica, sono funzionali all'aggregazione tempo-correlata a prescindere dalla tipologia: ad esempio non "chiese" tout-court, ma chiese "romaniche", chiese "bizantine", chiese "medievali", chiese di "età moderna" o "contemporanee". Oggi i navigatori indicano solo la categoria del monumento, non ulteriori raggruppamenti storico-relati. Se fossi interessato a visitare monumenti da questa prospettiva, sarebbe utile che tali informazioni fossero presenti a sistema. La predisposizione di linee guida per la creazione di questo contenitore orientato all'utilizzo storico dei dati GIS per la navigazione, fornirebbe un disciplinare descrittivo, ma anche le basi per un manuale operativo di buona condotta dell'iniziativa.

3. Popolamento collaborativo distribuito e autenticazione federata

I paradigmi generalisti come Wikipedia o delle principali piattaforme basate su sistemi GIS come Wikimapia, consentono un libero contributo da parte degli autori della comunità, ma chi garantisce per l'attendibilità, la correttezza e il rigore scientifico delle informazioni inserite?

Inoltre le due piattaforme web-based, come anche altre, seppur valide possono essere consultate solo separatamente. Collimare i contenuti tra loro diventa arduo perché indipendenti e le piattaforme non integrate. Inoltre se da un lato il popolamento distribuito facilita l'implementazione iniziale e a regime, dall'altro produce contenuti inesatti, non omogenei e non normalizzati per un loro riutilizzo autorevole.

Pertanto credo che una delle soluzioni percorribili sia offerta dall'accesso autenticato al database secondo il modello utilizzato con successo nell'Autenticazione Federata. Quest'ultima agevolerebbe la profilazione dell'utente e l'applicazione delle policies di accesso alla piattaforma da popolare. Inoltre essendo già in uso e ben accolta dalla community, apprezzata per la semplicità e comodità di utilizzo trasversale multiservizio, consentirebbe l'accesso in scrittura al database da parte di chi abbia effettivamente, per area tematica, le autorizzazioni e le competenze per introdurre informazioni scientifiche sui soggetti censiti.

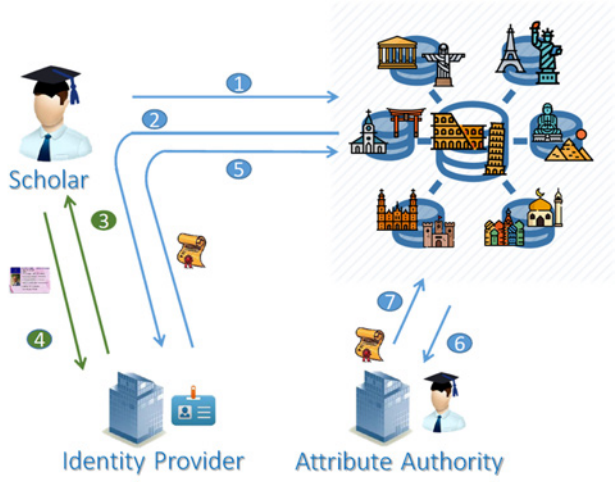


Fig. 3
Schema esemplificativo
di autenticazione per
l'accesso all'open Database

4. Visione

I database dei navigatori satellitari delle soluzioni commerciali proprietarie sono chiusi e le logiche algoritmiche non del tutto note. Queste soluzioni sono ottimizzate e indubbiamente molto performanti, ma dal punto di vista culturale, "chiudere i trovati", protetti da privative e brevetti va a detrimento della collettività. Aprire le soluzioni significa aumentare le possibilità di fruizione, ulteriore sviluppo e miglioramento progressivo.

Se l'idea fosse accolta dai big dei servizi di navigazione satellitare come ad esempio Garmin, TomTom, Google per inglobare queste informazioni Open nei loro sistemi embedded, si troverebbe un punto di incontro tra industria e accademia disponendo tutti di una risorsa in più per la pianificazione dei nostri viaggi tematici grazie all'attendibilità dei contenuti e il collaudato know-how dei principali vendor.

5. Soluzione proposta

Pianificare viaggi e itinerari che abbiano un senso anche per gli studiosi e i turisti più esigenti, è lo scopo finale di un database aperto dei monumenti. Spesso siamo confinati nel nostro piccolo e nella tipologia monumentale eredità della nostra storia e cultura. Pensiamo invece alla possibilità di estendere il paradigma a luoghi totalmente sconosciuti, lontani e inesplorati, appartenenti a culture completamente diverse dalle nostre, con le loro unicità e particolarità. Tale strumento si presta anche ad un suo utilizzo per scopi di ricerca. Generalmente la correlazione degli elementi nasce nell'individuo, a posteriori, grazie alle nozioni acquisite, ma potrebbe scaturire ed essere integrata da una aggregazione intelligente operata da sistemi informatici, per una nuova visione del panorama studiato.

6. Prospettive e conclusioni

L'Open Monuments Engine è una proposta per la costituzione di un collettore di informazioni storico-geografiche di tutti i monumenti, con aggregazioni tematiche direttamente importabili su qualsiasi navigatore satellitare per un'esperienza di navigazione personalizzata. Questa idea, che cambia la prospettiva e l'utilizzo dei GPS NAV, è un mio contributo

alla Data (R)evolution. Gli Open Data a disposizione della ricerca hanno assunto un'indiscussa valenza: essendo fruibili da tutti, incentivano ulteriore indotto per un continuo perfezionamento degli strumenti a beneficio degli utenti.

Questo contributo si propone come appello per una condivisione collaborativa delle competenze volta a delineare delle linee guida, ex novo o a integrazione di altre esistenti, che indirizzino sia strategie sia politiche di sviluppo delle infrastrutture digitali al servizio del patrimonio storico, culturale e monumentale auspicando la collaborazione del MiBACT e del mondo accademico. Si candida a essere uno strumento interoperabile che favorisca l'accesso a dati aperti sui monumenti e incoraggi l'evoluzione di servizi a contorno in modo che tutti possano trarre giovamento dalla libera fruizione dei contenuti.

È uno strumento per la trasmissione del sapere e della formazione sia individuale che collettiva, soprattutto se proposto come patrimonio trasversale per il Trasferimento Umanistico: un'estensione della conoscenza per comprendere la bellezza delle scienze umane rese semplici mediante la tecnologia e l'informatica (Serra L. 2017).

Riferimenti bibliografici

Carta M., Spagnoli L. (eds.) (2010), La ricerca e le istituzioni tra interpretazione e valorizzazione della documentazione cartografica, Gangemi Editore, Roma.

Casula F.C. (1997), La terza via della storia: il caso Italia, ETS, Pisa

Dijkstra E.W. (1959), A note on two problem in connexion with graphs, Numerische Mathematik, 1, pp 269–271.

Quddus M.A., Ochieng W.Y., Noland R.B. (2007), Current map-matching algorithms for transport applications: State-of-the art and future research directions, Transportation Research Part C, Elsevier Ltd, pp 312–328.

Serra L. (2018), "Cicerone", a monuments' guide plug-in for navigators: a proposal for a software application to increase the value of cultural heritage historically with GIS and GPS open data, Proceedings of the 8th Conference of Japanese Association for Digital Humanities, Japanese Association for Digital Humanities, Tokyo, pp 54-57.

Serra L. (2017), Relational and conceptual models to study the Mediterranean defensive networks: an experimental open database for content management systems, in Ángel Beningo González Avilés (ed.), Defensive Architecture of the Mediterranean XV to XVIII Centuries Vol. VI, Proceedings of the International Conference on Modern Age Fortifications of the Mediterranean Coast FORTMED 2017, Editorial Publicacions Universitat D'Alacant, Alicante, Spain, October 26-28, 2017, pp 369-376.

Serra L. (2017), Trasferimento Tecnologico e Trasferimento Umanistico: due facce della stessa medaglia, in Maria Grazia Rosaria Mele (ed.), Mediterraneo e città, Discipline a confronto, Franco Angeli, Milano, pp.259-272.

Serreli G. (2015), Il sistema difensivo del Regno di Arboréa tra il X e il XV secolo, in Giorgio Verdiani (ed.), Defensive Architecture of the Mediterranean XV to XVIII Centuries Vol. IV, Proceedings of the International Conference on Modern Age Fortifications of the Mediterranean Coast FORTMED 2016, Florence, Italy, November 10-12, 2016, pp 433-440.

Autore

Luigi Serra - luigi.serra@isem.cnr.it

Luigi Serra, Ingegnere Informatico, responsabile ICT presso l'Istituto di Storia dell'Europa Mediterranea del CNR a Cagliari, dal 2015. Referente tecnico con gli enti di ricerca nazionali ed internazionali e dei progetti ad alto contenuto tecnologico, gestisce l'infrastruttura informatica e cura diversi aspetti di valorizzazione della ricerca e dei beni culturali dedicandosi alle Digital Humanities e Digital Heritage nell'ambito dell'ADHO, proponendo soluzioni informatiche e tecnologiche innovative al servizio della ricerca nelle scienze umane e del "Trasferimento Umanistico".

DOI: 10.26314/GARR-Conf18-proceedings-23

Interoperabilità e accesso a dati e servizi: sistema di protezione di documenti elettronici mediante accesso biometrico

Pierluigi Tuveri, Marco Micheletto, Giulia Orrù, Luca Ghiani, Gian Luca Marcialis

Università degli studi di Cagliari, Dipartimento di Ingegneria Elettrica ed Elettronica

Abstract. Al giorno d'oggi, milioni di documenti attraversano ogni secondo le reti di telecomunicazione. In questo contesto la protezione dei dati e in particolare la crittografia rappresenta uno strumento fondamentale per la difesa di informazioni sensibili e/o personali.In questo articolo verrà presentato un prototipo di sistema di protezione di documenti digitali basato sull'utilizzo di tecniche crittografiche tradizionali rafforzate da un sistema di verifica personale biometrica. Il prototipo è stato sviluppato nell'ambito del progetto "Protezione di documenti elettronici mediante accesso biometrico", finanziato dalla Regione Lombardia tramite il bando InnoDriver3 per attività di trasferimento tecnologico con l'azienda CSAMed-Net4Market srl.

Keywords. impronte digitali, biometria, autenticazione

Introduzione

Oggigiorno, gran parte della popolazione mondiale ha accesso a Internet e miliardi di informazioni sensibili o personali sono quotidianamente esposte al rischio di essere intercettate, manipolate, divulgate o utilizzate per scopi diversi da quelli per cui sono state concepite. Se consideriamo anche tutte le informazioni memorizzate localmente che necessitano anch'esse di essere preservate con un certo grado di riservatezza, abbiamo un quadro chiaro di quanto sia importante e necessario sviluppare algoritmi sofisticati in grado di garantire un elevato livello di sicurezza dei dati.

Negli ultimi decenni, numerose ricerche si sono basate sullo studio dell'interazione tra crittografia e biometria, due tecnologie di sicurezza potenzialmente complementari. La crittografia è la scienza che si occupa di trovare metodi e algoritmi per rendere dei simboli o del testo, in un primo momento chiari e leggibili, del tutto incomprensibili a chiunque non fosse a conoscenza dei procedimenti utilizzati per codificarli (Stallings 2006). Attualmente gli algoritmi di criptaggio, basano l'identificazione della persona che deve accedere alla risorsa criptata, sui classici paradigmi basati su "qualcosa che si possiede" come smart card, o su "qualcosa che si conosce", come pin o password. Questi metodi hanno però il grande svantaggio di essere facili da violare in quanto una smart card può essere rubata o smarrita, pin o password dimenticati.Per questo motivo negli ultimi anni si sta sempre più diffondendo il paradigma che permette di riconoscere un individuo sulla basa di "ciò che si è", utilizzando quindi parametri biometrici, come le impronte digitali o il volto. La

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

biometria, infatti, è la scienza che studia e analizza i metodi per riconoscere in maniera univoca l'identità di un individuo sulla base delle sue caratteristiche fisiche o comportamentali (Jain et al. 2007). L'uso di metodi di autenticazione biometrica in sistemi di protezione dei dati tramite crittografia aumenta considerevolmente il livello di sicurezza sfruttando i vantaggi dei tratti biometrici, che non possono essere dimenticati, rubati o persi. In questo articolo presentiamo un prototipo di sistema di protezione di documenti elettronici mediante accesso biometrico, in particolare tramite l'utilizzo dell'impronta digitale. Il progetto è nato dalla collaborazione tra il Dipartimento di Ingegneria Elettrica ed Elettronica dell'Università di Cagliari e l'azienda Net4Market-CSAMed s.r.l.. Il suo scopo è quello di indagare la possibilità di sfruttare la tecnologia relativa ai tratti biometrici, nello specifico relativa alle impronte digitali, per incrementare il livello di protezione nei documenti elettronici dato dal semplice utilizzo di metodi crittografici. Il prototipo è stato sviluppato nell'ambito del progetto "Protezione di documenti elettronici mediante accesso biometrico", finanziato dalla Regione Lombardia tramite il bando InnoDriver3.

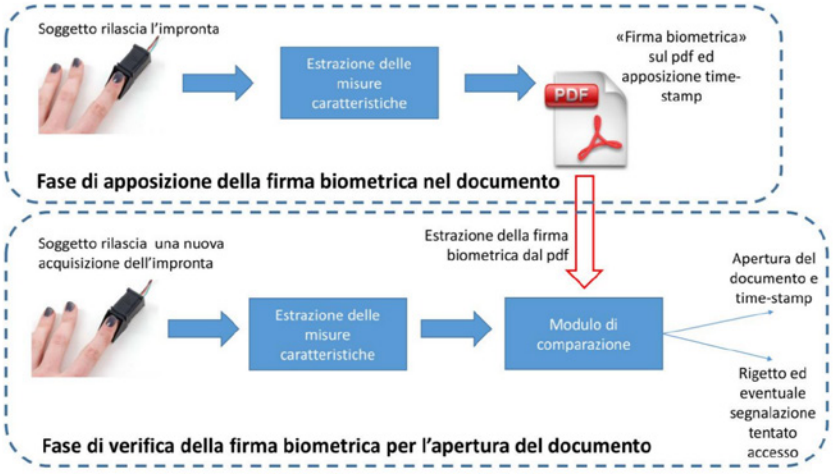


Fig. 1
Fasi funzionali del sistema di protezione di file pdf mediante accesso tramite impronta digitale

1. Il prototipo

Il sistema realizzato (Figura 1) utilizza congiuntamente un tratto biometrico, in particolare l'impronta digitale, e un metodo di sicurezza tradizionale, la password, per proteggere un documento elettronico. Usando questa due informazioni il sistema è in grado di criptare il file e di salvarlo in un altro formato con estensione proprietaria (.glm). In questo modo il nuovo file risulta illeggibile e può essere decriptato solo dalla stessa persona che l'ha criptato, cioè da chi conosce la password e possiede l'impronta digitale. In particolare, il sistema schematizzato in Figura 2 è suddiviso in tre moduli:

- modulo di estrazione delle minuzie, ossia le caratteristiche distintive dell'impronta digitale;
- modulo di protezione del template, necessario al fine di garantire la privatezza e la rinnovabilità della biometria;
- modulo di criptaggio, all'interno del quale vengono criptati sia il template che il file.

Nel modulo di estrazione delle minuzie, l'applicazione acquisisce l'impronta digitale e e-

strae i punti salienti (biforcazioni, termini, uncini, laghi, isole delle creste papillari). L'insieme delle coordinate e degli angoli di allineamento delle minuzie costituiscono il template dell'impronta digitale.

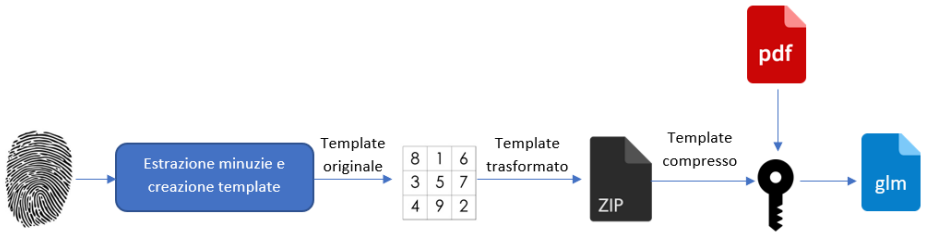


Fig. 2
Schema di funzionamento del prototipo

Il template descrive in maniera puntuale un'impronta digitale e poiché da esso è possibile ricreare l'impronta digitale da cui è stato estratto (Cao, Jain 2015) , esso deve essere protetto. Come primo passo del modulo di protezione del template viene applicata una trasformazione cartesiana al template che permette di modificare la posizione delle minuzie (Nalini et al. 2007). Per implementare la trasformazione si utilizza una matrice che cambia periodicamente (aggiornando di conseguenza i file coinvolti), come si può vedere nella Figura 3. Il template trasformato viene successivamente compresso e inserito all'interno del file pdf.

Il documento risultante viene a questo punto criptato ed esportato in formato (.glm). Alcuni dettagli delle procedure adottate non possono essere inseriti per via dell'accordo di riservatezza vigente tra il DIEE e l'azienda. Per poter accedere al file è necessario conoscere la password e superare il sistema di verifica biometrica.
Infatti, l'inserimento della password permette di decriptare ed estrarre il template trasformato.

Il template ottenuto verrà quindi decompresso e verrà applicata la trasformazione inversa tramite la stessa matrice utilizzata in fase di criptaggio ottenendo il template relativo all'impronta digitale memorizzata.

L'accesso effettivo al pdf sarà consentito solo in caso di corrispondenza (matching) tra l'impronta digitale memorizzata e quella in input al sistema.

Fig. 3
L'impronta a sinistra riporta le minuzie trovate, al centro la matrice M che permette di trasformare le minuzie, a destra le minuzie trasformate sovrapposte all'immagine originale. Come si può vedere le minuzie trasformate sono scorrelate dall'impronta.

Nel caso preso in considerazione il matching è basato sulle minuzie (Maltoni et al. 2009): le minuzie estratte dalle immagini da confrontare sotto forma di coordinate di un punto sul piano e di orientamento corrispondente.

In questo tipo di matching si cerca di raggiungere il migliore allineamento tra le due immagini in modo da ottenere una corrispondenza tra il maggior numero possibile di minuzie. Per compensare le variazioni causate del rumore e della distorsione è definita una regione di tolleranza attorno alla posizione di ogni minuzia.

2. Conclusioni

Il prototipo sviluppato nell'ambito del progetto "Protezione di documenti elettronici mediante accesso biometrico", finanziato dalla Regione Lombardia tramite il bando Inno-Driver3 per attività di trasferimento tecnologico con l'azienda CSAMed-Net4Market srl, utilizza un sistema di verifica biometrica tramite impronta digitale per proteggere ulteriormente un file pdf criptato tramite password.

Un futuro miglioramento del prototipo presentato si baserà sull'eliminazione della necessità di richiedere una password per accedere al documento, utilizzando il tratto biometrico stesso come chiave di criptaggio.

Riferimenti bibliografici

Cao K. and Jain A. K., (January 2015), Learning Fingerprint Reconstruction: From Minutiae to Image, EEE Trans. on Information Forensics and Security, vol.10, no.1, pp.104-117.

Jain A. K., Flynn P., Ross A. A., (2007), Handbook of biometrics. Springer Science & Business Media.

Maltoni D., Maio D., Jain A. K., Prabhakar S., (2009), Handbook of fingerprint recognition. Springer Science & Business Media.

Nalini K. R, Sharat C., Jonathan H. Connell, and R. M. Bolle, (April 2007), Generating Cancelable Fingerprint Templates. IEEE Trans. PAMI 29, 4 , pp 561-572.

Stallings W, (2006). Cryptography and Network Security, 4/E. Pearson Education India.

Autori

Pierluigi Tuveri - pierluigi.tuveri@diee.unica.it
Pierluigi Tuveri si è laureato nel marzo del 2014 in ingegneria elettronica presso l'Università degli Studi di Cagliari discutendo la tesi dal titolo "A user-specific approach to fingerprint liveness detecton". Dal marzo 2014 collabora con il PRA Lab, su tematiche biometriche come sviluppatore software.

Marco Micheletto - michelettomarco@gmail.com
Ha conseguito la laurea in Ingegneria Biomedica nel luglio 2017 presso l'università degli Studi di Cagliari discutendo la tesi intitolata: "Riconoscimento personale mediante le vene del palmo". Attualmente laureando in ingegneria elettronica, collabora col PRA Lab su tematiche inerenti le impronte digitali.

Giulia Orrù - giulia.orru@diee.unica.it
Dottoranda in Ingegneria Elettronica e Informatica dell'Università di Cagliari, dal 2014 collabora

con il PRA Lab nell'ambito del riconoscimento di volti, fingerprint liveness detection e sistemi biometrici adattivi.

Luca Ghiani - luca.ghiani@diee.unica.it
Luca Ghiani si è laureato nel Dicembre del 2011 in Ingegneria Elettronica con votazione 110/110 presso l'Università degli studi di Cagliari, discutendo una tesi dal titolo "Multimodal Fingerprint Liveness Detection by Local Phase Quantization and Wavelet Transforms".
Da Marzo 2012 collabora col PRA Lab, prima come dottorando di ricerca, attualmente come ricercatore post-doc.

Gian Luca Marcialis - marcialis@diee.unica.it
Dirige le attività relative alla Divisione "Biometria" del PRA Lab. Le sue attività di ricerca sono incentrate sulla biometria. Sulla base di queste attività è responsabile delle attività svolte in diversi progetti internazionali e nazionali (MAVEN, FP7 Tabula Rasa, PRIN su “Guardie biometriche”, RAS Legge 7 su “Sistemi biometrici adattativi”), commesse di aziende nazionali ed internazionali (Crossmatch, Greenbit, CSAMed) ed è organizzatore della “International Fingerprint Liveness Detection Competition” (LivDet) giunta alla sua sesta edizione.

DOI: 10.26314/GARR-Conf18-proceeding-24

Architettura della conoscenza: il sistema ICCD come modello per la descrizione dei beni culturali

Maria Letizia Mancinelli, Chiara Veninata

Istituto centrale per il catalogo e la documentazione (ICCD)

Abstract. Il paper illustra in maniera sintetica le attività dell'ICCD indirizzate ad una maggiore condivisione e valorizzazione sia dei modelli di strutturazione della conoscenza sul patrimonio culturale sia dei dati prodotti nelle campagne di catalogazione. Tali attività sono volte in particolare all'analisi e all'applicazione delle potenzialità offerte dal semantic web e dai suoi strumenti.

Keywords. semantic web, patrimonio culturale, catalogazione, ontologie, linked data

Per la realizzazione del catalogo del patrimonio archeologico, architettonico, paesaggistico, storico artistico e demoetnoantropologico, l'Istituto centrale per il catalogo e la documentazione ha elaborato un articolato "sistema di conoscenza": principi di metodo, strumenti e regole per acquisire le informazioni sui beni secondo procedure e criteri omogenei in tutto il territorio nazionale. La componente fondamentale di questo sistema è rappresentata dalle schede di catalogo: modelli descrittivi costituiti da una sequenza predefinita di voci, che raccolgono in modo formalizzato le notizie sui beni, seguendo un percorso conoscitivo che guida il catalogatore e al tempo stesso controlla e codifica i dati sulla base di precisi parametri. Nelle schede viene registrato il codice univoco nazionale assegnato a ciascun bene culturale, punto di riferimento di tutto il processo di catalogazione.

In relazione ai vari tipi di beni - mobili, immobili e immateriali - l'ICCD ha elaborato ad oggi trenta diverse schede di catalogo, che afferiscono a nove settori disciplinari (beni archeologici, beni architettonici e paesaggistici, beni storico artistici, ecc.): nel quadro del sistema degli standard, impostato con una logica fortemente unitaria e coerente, ogni tipo di scheda è stato definito con lo scopo di porre in evidenza la valenza culturale di una certa tipologia di bene, descrivendone le caratteristiche peculiari. Nel tracciato si snoda la sequenza delle informazioni: quelle identificative; quelle descrittive, tecnico scientifiche e storico critiche; quelle geografiche e di contesto, per relazionare il bene al territorio e agli altri beni, secondo una prospettiva spazio temporale; quelle di carattere amministrativo, che qualificano e certificano i contenuti; il tutto corredato da una o più immagini e da altri eventuali documenti (grafici, video, fonti, bibliografia, ecc.). I dati sono registrati in sezioni omogenee chiamate paragrafi, a loro volta organizzati in campi e sottocampi, per una scomposizione capillare delle informazioni, funzionale alla gestione informatizzata e particolarmente "adatta" alla condivisione mediante formati digitali aperti. Nel corso

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

del tempo, l'applicazione dei modelli descrittivi ICCD ha portato ad un'evoluzione e ad un progressivo raffinamento, fino all'apparato di normative più aggiornato e all'elaborazione della c.d. Normativa trasversale, ovvero la "normativa quadro" predisposta dall'Istituto per la definizione degli standard di ultima generazione 4.00, che riguarda contenuti, proprietà e caratteristiche comuni alle diverse tipologie di schede per la catalogazione dei beni culturali, per qualsiasi settore disciplinare: intorno alle sezioni "trasversali" (l'ossatura portante dei diversi modelli, che garantisce il dialogo e l'interazione con le altre componenti del sistema) vengono organizzati i nuclei specialistici propri di ciascuna tipologia di bene. L'applicazione degli standard ICCD ha favorito negli anni la produzione di dati catalografici "granulari" e di qualità, costituendo un serbatoio di informazioni dalle enormi potenzialità e tale da consentire relazioni concettuali e ontologiche particolarmente ricche.

Fig 1
Le schede di catalogo: criteri di ordinamento

SETTORI DISCIPLINARI	SCHEDE DI CATALOGO ICCD		CATEGORIA	SCHEDE IN USO (versioni 3.00 e 3.01)	SCHEDE 4.00
beni archeologici	AT	Reperti antropologici	BENI MOBILI	versione 3.01 - anno 2007	in corso costituzione GdL
	CA	Complessi archeologici	BENI IMMOBILI	versione 3.00 - anno 2003	
	MA	Monumenti archeologici	BENI IMMOBILI	versione 3.00 - anno 2003	
	RA	Reperti archeologici	BENI MOBILI	versione 3.00 - anno 2003	
	SAS	Saggi stratigrafici	BENI IMMOBILI	versione 3.00 - anno 2003	
	SI	Siti archeologici	BENI IMMOBILI	versione 3.00 - anno 2003	
	TMA	Tabella materiali archeologici	BENI MOBILI	versione 3.00 - anno 2003	
beni architettonici e paesaggistici	A	Architettura	BENI IMMOBILI	versione 3.00 - anno 2003	
	CNS	Centri/nuclei storici	BENI IMMOBILI		in elaborazione
	PG	Parchi/giardini	BENI IMMOBILI	versione 3.00 - anno 2003	
beni demoetnoantropologici	BDI	Beni demoetnoantropologici immateriali	BENI IMMATERIALI	versione 3.01 - anno 2006	rilasciata - 2016
	BDM	Beni demoetnoantropologici materiali	BENI MOBILI	versione 2.00 - anno 2000	rilasciata - 2016
beni fotografici	F	Fotografia	BENI MOBILI	versione 3.00 - anno 2003	rilasciata - 2016
	FF	Fondi fotografici	BENI MOBILI		rilasciata - 2016
beni musicali	SM	Strumenti musicali	BENI MOBILI		rilasciata - 2016
	SMO	Strumenti musicali-Organo	BENI MOBILI	versione 3.01 - anno 2008	in elaborazione
beni naturalistici	BNB	Beni naturalistici-Botanica	BENI MOBILI	versione 3.01 - anno 2007	
	BNM	Beni naturalistici-Mineralogia	BENI MOBILI	versione 3.01 - anno 2007	
	BNP	Beni naturalistici-Paleontologia	BENI MOBILI	versione 3.01 - anno 2008	
	BNPE	Beni naturalistici-Petrologia	BENI MOBILI	versione 3.01 - anno 2007	
	BNPL	Beni naturalistici-Planetologia	BENI MOBILI	versione 3.01 - anno 2007	
	BNZ	Beni naturalistici-Zoologia	BENI MOBILI	versione 3.01 - anno 2007	
beni numismatici	NU	Beni numismatici	BENI MOBILI	versione 3.00 - anno 2004	
beni scientifici e tecnologici	PST	Patrimonio scientifico e tecnologico	BENI MOBILI	versione 3.01 - anno 2005	rilasciata - 2018
beni storici e artistici	D	Disegni	BENI MOBILI	versione 3.00 - anno 2003	in elaborazione
	MI	Matrici incise	BENI MOBILI	versione 3.00 - anno 2003	
	OA	Opere/oggetti d'arte	BENI MOBILI	versione 3.00 - anno 2003	in elaborazione
	OAC	Opere/oggetti d'arte contemporanea	BENI MOBILI	versione 3.00 - anno 2004	in elaborazione
	S	Stampe	BENI MOBILI	versione 3.00 - anno 2003	
	VeAC	Vestimenti antichi/contemporanei	BENI MOBILI	versione 3.01 - anno 2010	

Ciò ha favorito già dal 2014 l'avvicinamento dell'ICCD agli strumenti e agli standard del semantic web, con la partecipazione dell'Istituto alla modellazione e all'uso dell'ontologia Cultural-ON (disponibile all'indirizzo http://dati.beniculturali.it/cultural_on/) per la pubblicazione in formato linked open data (LOD) dei Contenitori fisici e giuridici dei beni culturali. I collegamenti tra dataset rendono possibile l'arricchimento del dato di partenza grazie all'opportunità di incrociare informazioni provenienti da varie fonti eventualmente gestite da diverse amministrazioni o da enti o da privati. Il rilascio di open data implica l'instaurarsi di un rapporto nuovo con utenti potenzialmente sconosciuti a chi pubblica

i dati. Nel caso dei linked open data questi utenti possono anche non essere degli esseri umani ma degli agenti software che interrogano e interagiscono attraverso API o query SPARQL per realizzare applicazioni, visualizzazioni o altre elaborazioni. A luglio del 2017 è stato inaugurato il portale dei linked open data del Mibac, che offre l'accesso a dati sui luoghi della cultura e sugli eventi (dati aggiornati in tempo reale), anagrafiche di biblioteche, di archivi di Stato, di "contenitori" di beni culturali, e dati sul patrimonio fotografico e archivistico dell'ICCD; il portale entro l'anno offrirà l'accesso ad oltre 800.000 schede di catalogo del patrimonio culturale italiano.

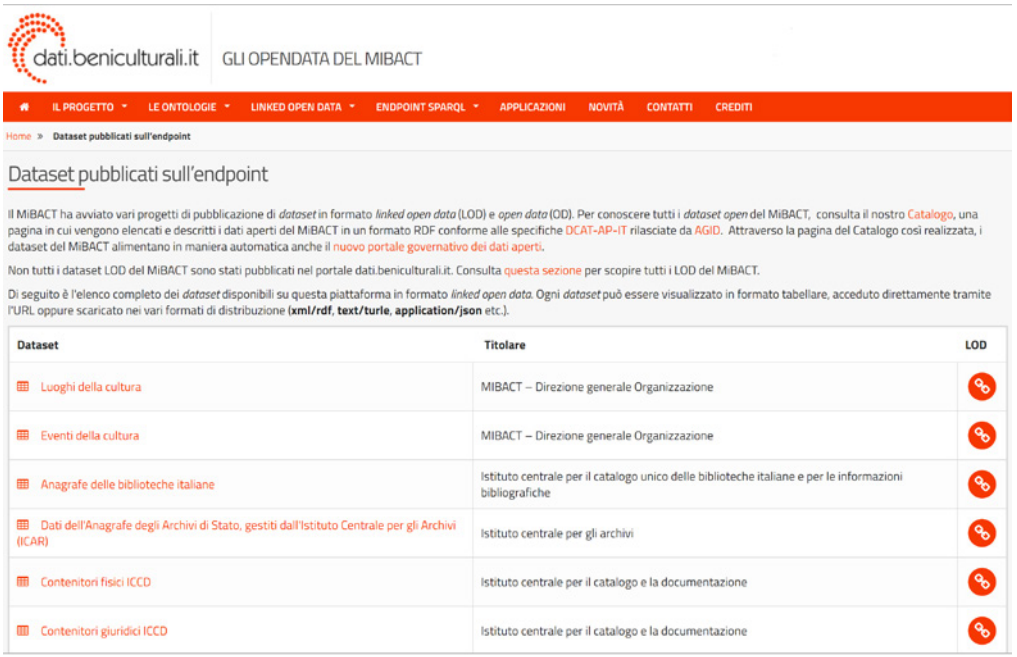


Fig. 2
Il sito
dati.beniculturali.it

Al fine di garantire una ottimale pubblicazione delle informazioni contenute nelle varie tipologie di schede catalografiche senza sacrificare nulla dell'impostazione metodologica e dei contenuti definiti nel tempo dall'ICCD, a novembre 2017 è stato avviato il progetto ArCo, con cui l'ICCD e il Semantic Technology Laboratory (STLab) dell'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del CNR (il team che lavora sul progetto ArCo è composto da Aldo Gangemi, Valentina Presutti e Luigi Asprino, Valentina Corriero e Andrea Nuzzolese) mirano a valorizzare il patrimonio artistico e culturale italiano attraverso la creazione di una rete di ontologie (1) che fornisca un'architettura della conoscenza dei beni culturali e consenta la pubblicazione di dati secondo il paradigma linked open data.

ArCo ha coinvolto fin dall'inizio numerosi potenziali sviluppatori di applicazioni che usano le ontologie e i LOD prodotti nel settore dei beni culturali, includendo un programma c.d. di early adoption. Attraverso incontri periodici svolti con lo strumento dei meetup (2) e il rilascio regolare e incrementale delle ontologie e dei dati - anche se in formato instabile - su Github (3), ICCD e STLab hanno voluto inaugurare una nuova stagione di

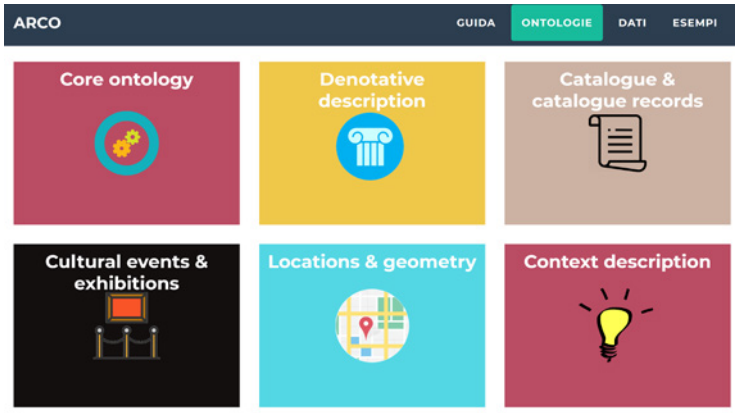


Fig. 3
Le ontologie del progetto arco

partecipazione con gli early adopter, che contribuiscono alla raccolta dei requisiti delle ontologie e allo sviluppo e miglioramento del progetto fornendo continui feedback. Nel corso delle varie release sono state finora modellate porzioni della Normativa trasversale relative alle localizzazioni dei beni, alla georeferenziazione, all'attribuzione di autore, committente e ambito culturale, allo stato di conservazione, alle informazioni relative al soggetto, al titolo, all'editore e, infine, alle indagini che hanno riguardato il bene culturale. A fine luglio è stata lanciata una call to co-creation, un contest a premi con cui l'ICCD ha inteso coinvolgere gli early adopters nella progettazione di modalità di fruizione innovative del sistema del Catalogo generale dei beni culturali. Dalle ricerche fatte all'interno della piattaforma e grazie ai LOD, l'utente potrà ottenere non solo la consultazione della risorsa digitale e dei dati descrittivi sul singolo elemento o gruppo di beni ricercato, ma anche la ricostruzione del contesto nel quale essi si collocano, evidenziando le relazioni esistenti fra gli elementi del patrimonio, i soggetti che li riguardano, i luoghi che ne sono lo scenario, le persone a cui sono legati, arricchendo le fonti di partenza con informazioni di qualità utili a fini conoscitivi, educativi, di ricerca, oltre che di valorizzazione.

Riferimenti bibliografici

Mancinelli Maria Letizia (2018), Gli standard catalografici dell'Istituto Centrale per il Catalogo e la Documentazione, in Roberta Tucci, Le voci, le opere e le cose. La catalogazione dei beni culturali demoetnoantropologici, Roma, Istituto centrale per il catalogo e la documentazione - MiBAC, pp. 279-302.

http://www.iccd.beniculturali.it

http://dati.beniculturali.it

http://www.catalogo.beniculturali.it

Note

1) Le ontologie altro non sono che tentativo di formulare uno schema concettuale esaustivo e rigoroso nell'ambito di un dominio dato. Tale schema concettuale deve essere condiviso il più possibile nell'ambito della propria comunità di interesse e deve essere formalizzato secondo un linguaggio comprensibile anche alle macchine che dovranno elaborare

automaticamente quei dati consentendo di effettuare su di essi dei ragionamenti e delle query anche complesse.
2) I meetup sono riunioni informali in cui si incontrano persone che hanno in comune un interesse culturale, tecnologico, sociale, ricreativo ecc. (dall'inglese "to meet up" che significa "incontrarsi per caso). Vedi https://www.meetup.com/it-IT/Meetup-Web-Progetto-ArCo-Architettura-della-conoscenza
3) GitHub è una piattaforma di condivisione di codice informatico e consente di gestire le modifiche al codice sorgente da parte di singoli o gruppi di lavoro anche grandi. Il software tiene traccia di ogni modifica fatta al codice permettendo di avere uno storico e di ritornare, se necessario, sui propri passi.

Autrici

Maria Letizia Mancinelli - marialetizia.mancinelli@beniculturali.it

Laurea, specializzazione e dottorato in Archeologia e Topografia medievale; attività di ricerca sul campo dal 1983 al 1999 (indagini stratigrafiche, rilievo tecnico archeologico, studio e classificazione di materiali). Dal 2000 è in servizio presso l'ICCD, attualmente come responsabile del Servizio gli standard catalografici, con funzione di coordinamento generale per la definizione e l'aggiornamento degli strumenti e delle metodologie per la catalogazione dei beni culturali e per la loro applicazione nel Sistema Informativo Generale del Catalogo (SIGECweb).

Chiara Veninata - chiara.veninata@beniculturali.it

Dopo la laurea in "Conservazione dei Beni Culturali" e il Diploma di Archivisticaa conseguito presso l'Archivio di Stato di Roma, ha ottenuto una Borsa di studio per la formazione di Ricercatori specializzati nel trattamento e nell'analisi archivistico documentale attraverso l'uso di modelli formali e tecniche informatiche, presso il Consorzio Roma Ricerche con cui ha collaborato per quasi 10 anni. Dal 2010 è di ruolo presso il Ministero per i beni e le attività culturali. Da marzo 2018 è responsabile presso l'ICCD del Servizio per la digitalizzazione del patrimonio culturale.

ISBN 978-88-905077-8-6