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.

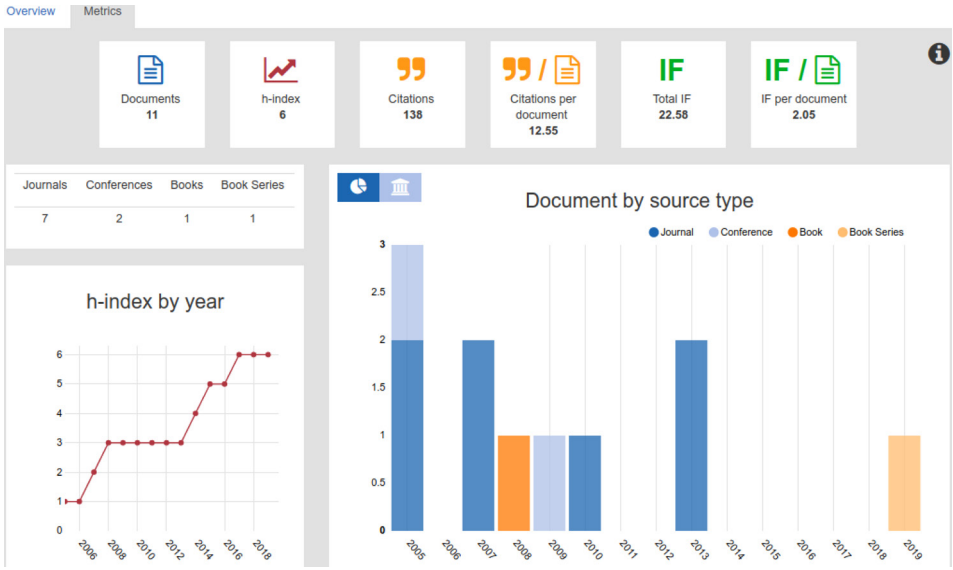


Fig. 1
Un esempio di alcuni cruscotti e report presenti in Scientilla, che integrano informazione bibliografica con informazione bibliometrica (da Scopus e Web of Science).

## 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.

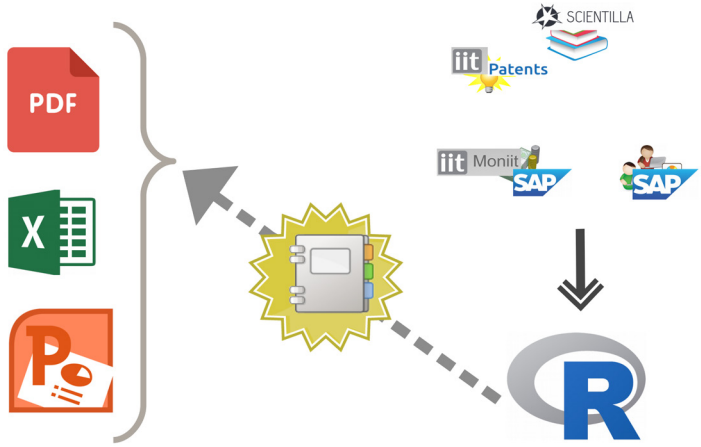


Fig. 2
Lo schema riassume come la reportistica è generata in IIT: i sistemi elencati nella Sezione 1 (in alto a destra) vengono interrogati tramite R, con cui i report sono generati in maniera automatica nei formati classici Microsoft o Pdf.

## 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.