

NICOLA FERRO*

La gestione e conservazione dei dati scientifici prodotti dalla valutazione sperimentale dei motori di ricerca

MOTIVAZIONI

Con una popolazione di più di 500 milioni di abitanti nei suoi 27 stati membri in cui vengono parlate più di 80 lingue in 3 alfabeti, senza contare quelle derivanti dall'immigrazione, i cittadini e le aziende dell'Unione Europea necessitano di sistemi informativi in grado di consentire loro di interagire con contenuti culturalmente, politicamente e linguisticamente diversi e spesso disponibili in svariati media.¹

In questo contesto, i sistemi di gestione e accesso all'informazione, tra cui i motori di ricerca e le biblioteche digitali, divengono ogni giorno più complessi in quanto si trovano a dover soddisfare tipologie di utenti e supportare attività sempre più variegate e articolate. Infatti, gli utenti non percepiscono più come "inviolabili" i confini tra lingue e media diversi ma considerano l'accesso all'informazione come intrinseco al processo di comunicazione, rivestendo non solo il ruolo tradizionale di consumatori di informazione ma anche e sempre più spesso di produttori di informazione.

Nel caso del patrimonio culturale, gli utenti desiderano sistemi di biblioteca digitale che permettano loro di interrogare simultaneamente cataloghi di biblioteche, archivi e musei per accedere a libri, scansioni di manoscritti, riproduzioni di quadri, opere musicali, documenti audio-visivi, e così via. Tali sistemi dovrebbero fornire agli utenti la possibilità di "contestualizzare" l'informazione recuperata, collegandola con altre risorse pertinenti disponibili su Web, arricchendola con annotazioni e recensioni fatte da altri utenti, e personalizzandola in base alle preferenze dell'utente.

(*) Dipartimento di Ingegneria dell'Informazione - Università degli Studi di Padova.

(¹) COMMISSION OF THE EUROPEAN COMMUNITIES. *Multilingualism: an asset for Europe and a shared commitment*. COM(2008) 566 final, September 2008.



FIG. 1 - Esempio di oggetto del patrimonio culturale in media e lingue diverse.

Questo è il caso, ad esempio, di Europeana,² un progetto che mira alla creazione della biblioteca digitale europea in cui confluiscono i materiali provenienti da biblioteche, musei, archivi e archivi audiovisivi al fine di valorizzare e rendere accessibile il patrimonio culturale europeo attraverso un portale unitario. Per conseguire questo ambizioso obiettivo diviene necessario ideare e sviluppare modelli di dati avanzati, capaci di mediare tra le diverse modalità di descrizione del patrimonio culturale che caratterizzano le varie istituzioni che sono demandate a mantenerlo, così come progettare e realizzare sistemi in grado di implementare tale modello di dati e offrire funzionalità di ricerca basate su esso in modo efficiente e scalabile.

Di conseguenza, la progettazione e lo sviluppo dei sistemi di gestione e accesso all'informazione multilingua e multimediale diventano un'attività sempre più complessa e che richiede una costante verifica per assicurare che essi siano conformi ai requisiti specificati e garantiscano le prestazioni attese dagli utenti. In questo processo di progettazione e sviluppo svolge un ruolo centrale la *valutazione sperimentale di tali sistemi*, cioè l'attività volta a verificare e quantificare le effettive prestazioni di tali sistemi rispetto a dei "campioni" di riferimento.

(²) <http://www.europeana.eu/>

LA VALUTAZIONE SPERIMENTALE

La valutazione sperimentale è un'attività fondamentale per supportare e guidare lo sviluppo dei sistemi di accesso all'informazione multilingua e multimediale. Essa è una parte essenziale del procedimento scientifico in quanto, grazie all'utilizzo di collezioni di dati e scenari di valutazione condivisi, è possibile confrontare i sistemi di accesso all'informazione, comprendere e analizzarne le prestazioni e perseguire e dimostrare i progressi ottenuti.

La valutazione sperimentale viene usualmente condotta nel contesto di campagne internazionali di valutazione su larga scala al fine di favorire la condivisione delle risorse, la comparabilità e il confronto tra i risultati sperimentali, le tecniche e gli algoritmi, e la formazione di una comunità con le competenze multidisciplinari necessarie per sviluppare i sistemi di accesso all'informazione multilingua e multimediale.

Le maggiori campagne di valutazione sperimentale a livello internazionale sono la *Text REtrieval Conference* (TREC)³ organizzata dal *National Institute for Standards and Technology* (NIST) negli Stati Uniti, la *The CLEF Initiative (Conference and Labs of the Evaluation Forum)*⁴ in Europa, e il *NII-NACSIS Test Collection for IR Systems* (NTCIR)⁵ organizzato dal *National Institute of Informatics* (NII) in Giappone e Asia. Inoltre, stanno sorgendo nuove iniziative di questo genere per dare voce e supporto a nuove comunità e per indirizzare esigenze specifiche, come ad esempio il *Forum for Information Retrieval Evaluation* (FIRE)⁶ in India.

Tutte le campagne di valutazione seguono un paradigma comune, chiamato paradigma di Cranfield e sviluppato nella seconda metà degli anni Sessanta del secolo scorso, che fa ricorso all'uso di *collezioni sperimentali* per quantificare l'efficacia di un motore di ricerca. Una collezione sperimentale consiste di una collezione di documenti rappresentativi, per argomento e numerosità, del dominio in cui il motore di ricerca dovrà operare, di una collezione di esigenze informative, da cui si derivano le interrogazioni da porre al motore di ricerca, e di una serie di giudizi di rilevanza che, per ogni interrogazione, indicano quali documenti sono rilevanti e quali non lo sono. In questo modo è possibile determinare quali dovrebbero essere i risultati "ideali" forniti

(³) TREC, <http://trec.nist.gov/>

(⁴) CLEF, <http://www.clef-initiative.eu/>

(⁵) NTCIR, <http://research.nii.ac.jp/ntcir/>

(⁶) FIRE, <http://www.isical.ac.in/~fire/>

da un motore di ricerca in risposta a una data interrogazione e si ha quindi la possibilità di confrontare i risultati effettivi prodotti da un dato sistema con essi e determinare così le prestazioni del sistema.

In questo contesto, le campagne di valutazione sperimentale forniscono un contributo significativo all'avanzamento dello stato dell'arte della ricerca nel settore e all'innovazione industriale. Inoltre, come riportato in uno studio condotto dal NIST riguardante l'impatto economico di TREC, per ogni dollaro investito dal NIST e dai suoi associati in TREC, sono stati prodotti da 3,35 a 5,07 dollari in benefici per la comunità di ricerca e il settore in generale.⁷

LA GESTIONE E CONSERVAZIONE DEI DATI PRODOTTI DALLA VALUTAZIONE

Durante la loro vita, le campagne di valutazione sperimentale su larga scala hanno prodotto e producono una grande quantità di dati scientifici di estremo valore. Non solo essi rappresentano il prodotto di sistemi che hanno richiesto molti anni di lavoro per la loro ideazione, progettazione e sviluppo, ma anche i dati sono spesso difficili da riprodurre, sia per la complessità dei sistemi stessi sia per il costo delle infrastrutture hardware su cui essi funzionano, che non sono alla portata di tutti i gruppi di ricerca o organizzazione. Questi dati sperimentali forniscono poi le basi per tutta la produzione scientifica successiva e per lo sviluppo dei sistemi ed è estremamente importante che essi siano facilmente accessibili e che si possa verificarli e confrontarli quando necessario. Infine, questi dati hanno valore anche da un punto di vista economico se si considera che, sempre nello stesso studio menzionato in precedenza, il NIST stima in 30 milioni di dollari l'investimento complessivo in TREC e di cui i dati scientifici prodotti sono una parte essenziale.

Tuttavia, minore attenzione è stata posta fino a oggi alla modellazione, gestione, cura, accesso, preservazione e riuso di tali dati scientifici, anche se l'importanza generale dei dati scientifici è stata evidenzia-

(⁷) B. R. ROWE, D. W. WOOD, A. L. LINK, D. A. SIMONI. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. <http://trec.nist.gov/pubs/2010.economic.impact.pdf>, 2010.

(⁸) COMMISSION OF THE EUROPEAN COMMUNITIES. *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation*, COMM(2008) 56 Final, February 2007.

ta da diverse istituzioni come, ad esempio, la Commissione Europea,⁸ il *National Scientific Board* (NSB)⁹ negli Stati Uniti, e l'*Australian Working Group on Data for Science*¹⁰ in Australia.

Occorre quindi sviluppare e fornire una infrastruttura e un ambiente unificati in cui collezionare i dati scientifici e gli strumenti a supporto della valutazione sperimentale per le comunità di utenti, individui e organizzazioni commerciali che si occupano di progettare, realizzare, migliorare, adottare e utilizzare i sistemi di accesso all'informazione multilingue e multimediale.

Il primo passo in questa direzione è rappresentato dalla realizzazione del sistema di biblioteca digitale per la gestione dei dati scientifici prodotti dalla valutazione sperimentale chiamato DIRECT, *Distributed Information Retrieval Evaluation Campaign Tool*^{11,12,13,14}, che:

- introduce uno schema concettuale per modellare e rendere chiaro quali siano le entità coinvolte nello spazio informativo della valutazione sperimentale, i loro attributi e le relazioni tra loro;
- fornisce le basi per sviluppare formati di metadati che consentano di descrivere il significato dei dati scientifici che vengono gestiti, di facilitare la loro condivisione, scambio e riuso nel tempo;
- adotta un meccanismo di identificazione univoca che consente la citazione esplicita dei dati scientifici, ad esempio negli articoli dove vengono utilizzati, e ne favorisce un accesso facilitato;

(⁹) National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (NSB-05-40). National Science Foundation (NSF). <http://www.nsf.gov/pubs/2005/nsb0540/>, 2005.

(¹⁰) Working Group on Data for Science. *FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science*. Report to Prime Minister's Science, Engineering and Innovation Council (PMSEIC), <http://www.innovation.gov.au/Science/PMSEIC/Documents/FromDataToWisdom.pdf>, 2006.

(¹¹) DIRECT, <http://direct.dei.unipd.it/>; un video con una demo del sistema è disponibile su YouTube nel canale della rete di eccellenza PROMISE all'indirizzo: <http://www.youtube.com/watch?v=fDsXDCUPkiM>

(¹²) M. AGOSTI, N. FERRO. *Towards an Evaluation Infrastructure for DL Performance Evaluation*. In G. Tsakonas, C. Papatheodorou (Eds.), *Evaluation of Digital Libraries: An insight into useful applications and methods*, Chandos Publishing, Oxford, UK, 2009, pp. 93-120.

(¹³) M. AGOSTI, G. M. DI NUNZIO, M. DUSSIN, N. FERRO. *10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go*. In T. Sakay, M. Sanderson, W. Weber (Eds.), Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010), National Institute of Informatics, Tokyo, Japan, 2010, pp. 16-24.

(¹⁴) N. FERRO, A. HANBURY, H. MÜLLER, G. SANTUCCI. Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems. *Procedia Computer Science*, 4:740-749, 2011.

- gestisce i diversi aspetti di una campagna di valutazione sperimentale come, ad esempio, la gestione delle collezioni di documenti, la sottomissione degli esperimenti, il calcolo delle misure di prestazione, e così via.

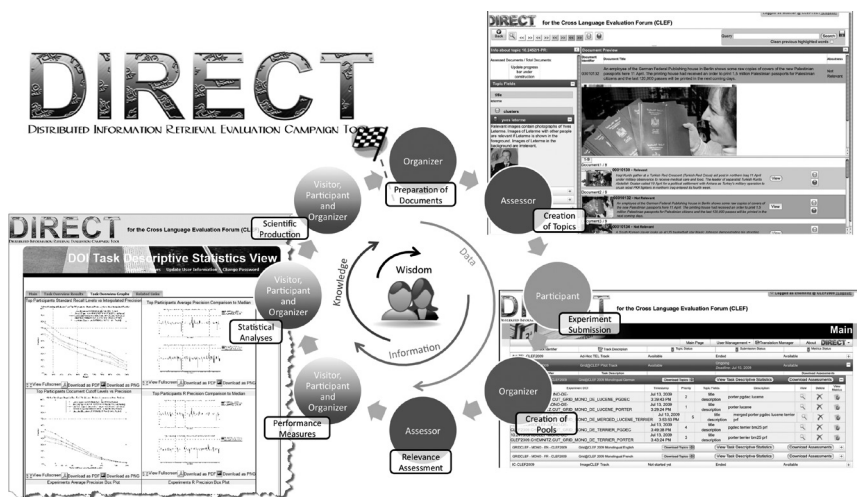


FIG. 2 - Esempio di funzionalità supportate dal sistema di biblioteca digitale per dati scientifici DIRECT.

A oggi, DIRECT gestisce più di 8,6 milioni di documenti in 16 lingue diverse e scritti con 4 differenti alfabeti (latino, cirillico, arabo, e cinese); mantiene 4.200 esperimenti che ammontano a più di 300 milioni di elementi e sono stati prodotti da 430 partecipanti in 41 nazioni diverse; contiene più di 3.2 milioni di giudizi di rilevanza su più di 10.000 interrogazioni prodotte da oltre 270 valutatori in 19 nazioni diverse; infine, dà accesso a oltre 27 milioni di misurazioni e statistiche descrittive sugli esperimenti gestiti e a più di 50.000 grafici e analisi statistiche.

DIRECT si fonda su un'architettura flessibile e modulare, basata su servizi Web, che consente un accesso e una manipolazione dei dati gestiti indipendenti dalla specifica applicazione che si desidera realizzare. In questo modo, sullo stesso sistema, si possono realizzare diverse applicazioni, adatte a diverse esigenze e categorie di utenza. Ad esempio, è stata realizzata l'applicazione Web descritta in precedenza e mostrata in Figura 2 per la gestione delle campagne di valutazione

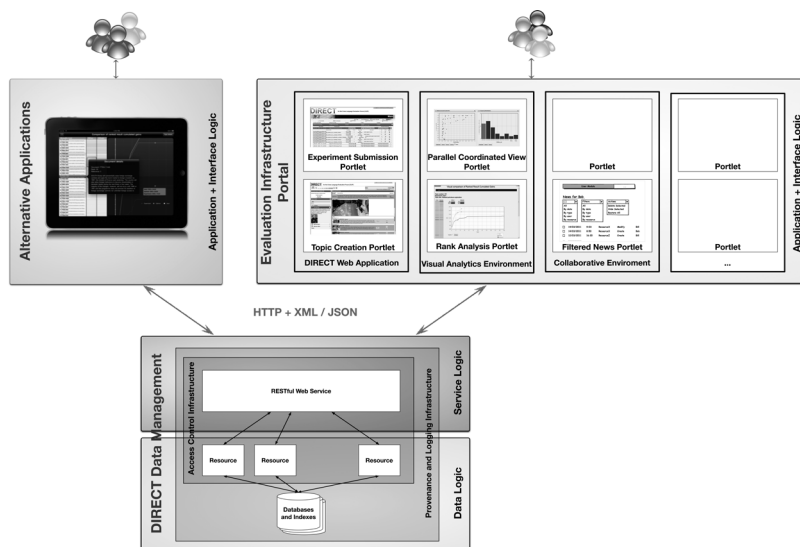


Fig. 3 - Architettura del sistema DIRECT.

sperimentale ma è anche in corso di sviluppo un'applicazione per iPad, descritta nel seguito, per l'esplorazione dei risultati sperimentali.

FUNZIONALITÀ AVANZATE PER L'ACCESSO AI DATI SPERIMENTALI

Avere a disposizione un'infrastruttura per la gestione e l'accesso ai dati sperimentali offre l'opportunità di sviluppare applicazioni avanzate su di essa, che favoriscano la collaborazione, il mantenimento e l'interazione con i dati sperimentali.

Una possibilità interessante è quella di consentire l'aggiunta da parte degli utenti di annotazioni¹⁵ sui dati gestiti, come mostrato in Figura 4.

Ad esempio, le annotazioni possono consentire agli utenti sia di collaborare, utilizzando i dati sperimentali, ad esempio fornendo spiegazioni e discutendo le prestazioni di un dato sistema, sia di migliorare la qualità dei dati mantenuti, segnalando possibili errori o inconsistenze. Tutto ciò favorisce il mantenimento dei dati nel tempo e la possibilità di poterli interpretare e comprendere correttamente anche a distanza di anni.

⁽¹⁵⁾ M. AGOSTI, N. FERRO. (2008). A Formal Model of Annotations of Digital Content. *ACM Transactions on Information Systems (TOIS)*, 26(1):3:1-3:57.

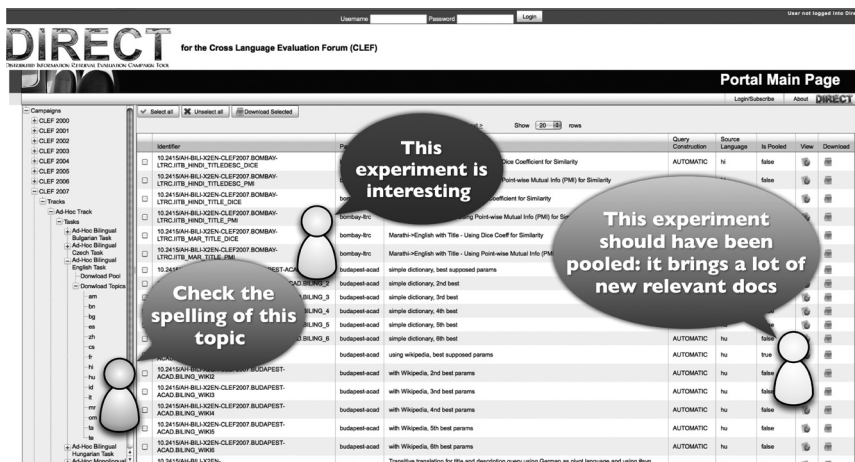


FIG. 4 - Aggiunta di annotazioni ai dati sperimentali.

L'infrastruttura consente anche di esplorare l'utilizzo di dispositivi alternativi per navigare, manipolare e interagire con i dati sperimentali, come mostrato in Figura 5 che riporta un primo prototipo di applicazione per iPad finalizzato a confrontare i risultati sperimentali con quelli ottimali e ideali. Un possibile scenario di utilizzo di una tale applicazione è quello di un gruppo di lavoro che stia sviluppando un motore di ricerca e necessita di discutere e analizzarne le prestazioni per capire meglio come sta procedendo lo sviluppo. Avere a disposizione un'applicazione su iPad che permetta di manipolare, esplorare e interagire con i dati sperimentali con una modalità di interazione naturale e basata sul tocco può rappresentare un modo efficace per condurre una riunione e far così progredire il lavoro.

Un ulteriore scenario di utilizzo di questa applicazione è lo studio di un articolo scientifico che descriva un sistema e le sue prestazioni e la possibilità di "cliccare" direttamente su un grafico e iniziare a esplorare interattivamente i dati descritti nell'articolo, come mostrato in Figura 6. In questa figura l'utente sta utilizzando l'applicazione iBook, già pre-installata su ogni iPad, e seleziona con il dito il link "topic 351" mostrato in azzurro. A questo punto si apre automaticamente l'applicazione mostrata in Figura 5 e l'utente può iniziare a esplorare i dati discussi nell'articolo che sta leggendo direttamente.

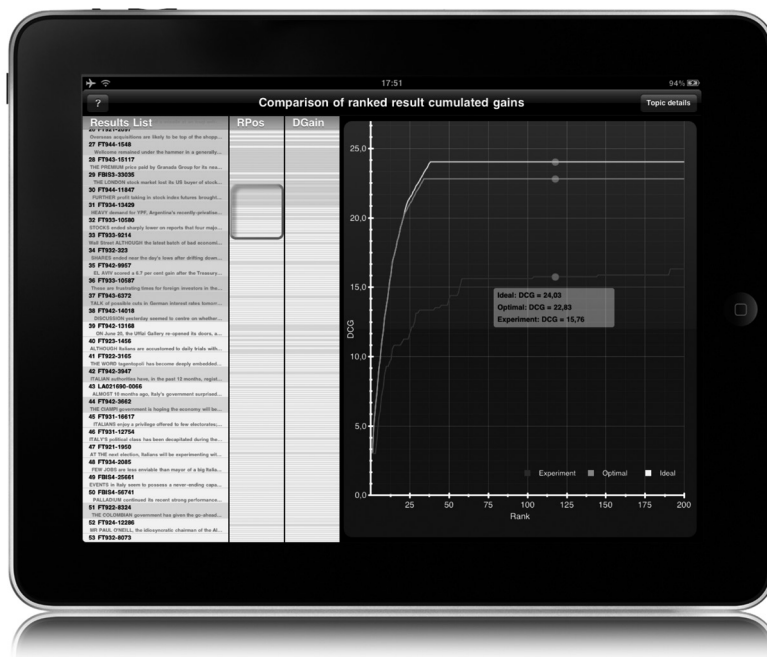


Fig. 5 - Esempio di applicazione iPad per l'esplorazione dei risultati sperimentali.

IL PROGETTO PROMISE

La campagna di valutazione CLEF e il sistema DIRECT sono alcune delle attività condotte nell'ambito del progetto PROMISE,¹⁶ *Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation*. PROMISE mira a far progredire lo stato dell'arte nella valutazione sperimentale dei sistemi di gestione e accesso all'informazione multilingue e multimediale al fine di supportare individui e organizzazioni di ricerca e commerciali, ma anche comunità che progettano, sviluppano, adottano e migliorano tali sistemi.

L'obiettivo finale del progetto è fornire e rendere disponibili a livello internazionale un ambiente e un'infrastruttura software che unifichino i dati prodotti dalla valutazione sperimentale, la conoscenza da essi estratta, le metodologie di valutazione, gli strumenti per effettuare e automatizzare la valutazione e la comunità di utenti a essa interessata, accrescendo l'utilizzo di tali dati e strumenti.

(16) PROMISE, <http://www.promise-noe.eu/>

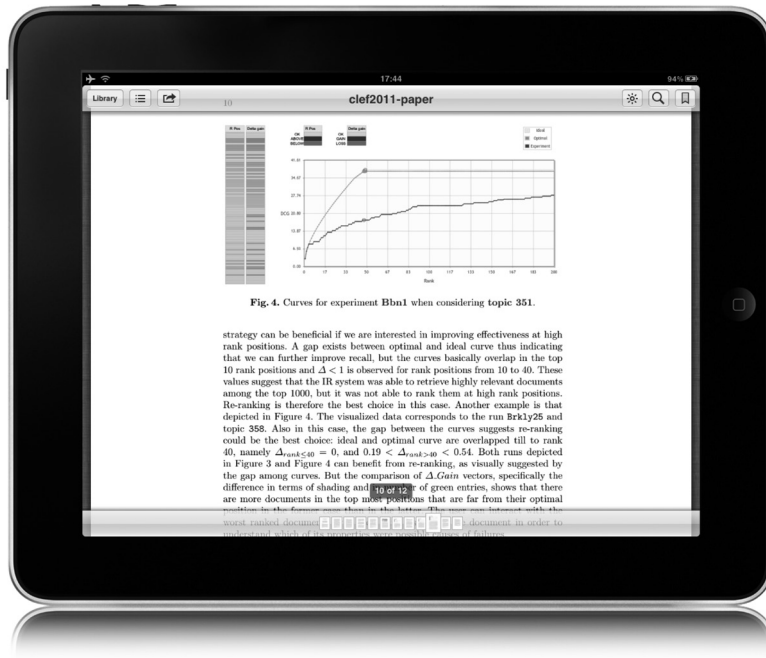


FIG. 6 - Esempio di lettura di un articolo scientifico e consultazione dei dati sperimentali corrispondenti.

PROMISE è una rete di eccellenza finanziata nell'ambito del settimo programma quadro della Commissione Europea, cui partecipano 10 partner scientifici e industriali a livello europeo, tutti leader del settore:

- Università degli Studi di Padova (coordinatore), Italia;
- Sapienza Università di Roma, Italia;
- University of Amsterdam, Olanda;
- Humboldt-Universität zu Berlin, Germania;
- Swedish Institute of Computer Science (SICS), Svezia;
- University of Applied Sciences Western Switzerland, Svizzera;
- Zurich University of Applied Sciences, Svizzera;
- Vienna University of Technology, Austria;
- Centre for the Evaluation of Language Communication Technology (CELCT), Trento, Italia;
- Evaluations and Language resources Distribution Agency (ELDA), Francia.



Il progetto ha una durata triennale a partire da settembre 2010 ad agosto 2013.

Lo strumento della rete ha l'obiettivo specifico di unire e amalgamare competenze e linee di ricerca, precedentemente separate, al fine di affrontare una problematica emergente e per la quale non sono già disponibili né soluzioni né lo spettro di competenze richiesto. La rete si propone di sviluppare le metodologie, le infrastrutture e i protocolli necessari per far progredire la valutazione sperimentale dei sistemi di gestione e accesso all'informazione multilingue e multimediale, così come svolgere attività di formazione e scambio di giovani ricercatori e organizzare eventi di disseminazione e trasferimento dei risultati raggiunti alle comunità applicative interessate e all'industria.

RINGRAZIAMENTI

Il lavoro riportato in questo articolo è stato parzialmente supportato dalla rete di eccellenza PROMISE (contratto n. 258191), parte del settimo programma quadro della Commissione Europea.

The work reported has been partially supported by the PROMISE network of excellence (contract n. 258191) project, as part of the 7th Framework Program of the European Commission.

