

Linked Heritage: multilinguismo e terminologie

Resoconto dal seminario di Parigi

Maria Teresa Natale

ICCU

Si presentano qui in sintesi i risultati del seminario sul multilinguismo e le terminologie, tenuto a Parigi lo scorso 18 aprile 2013 e organizzato nell'ambito del progetto europeo Linked Heritage.

L'obiettivo di questo seminario era dare una panoramica dello stato dell'arte su questi temi nell'ambito di alcuni progetti europei attivi nell'ecosistema di Europeaana e di presentare alcune iniziative realizzate in Francia.

Per chi volesse approfondire ulteriormente, la maggior parte delle presentazioni è disponibile sul sito di Linked Heritage¹, mentre i video sono presenti su VIMEO².

La prima sessione, intitolata "Terminologie e multilinguismo: una sfida su scala europea", è stata introdotta da Rossella Caffo, direttore dell'ICCU e coordinatore del progetto Linked Heritage, che ha sintetizzato le prime esperienze in tema di multilinguismo portate avanti da alcuni progetti europei conclusi o ancora in corso. Nel 2006, nell'ambito di MinervaPlus, venne condotta un'indagine che aveva lo scopo di verificare lo stato dell'arte del multilinguismo nei siti web, se in questi ultimi si utilizzassero vocabolari controllati e thesauri, se fossero disponibili servizi multilingue³. Il portale MICHAEL, attivo ormai da diversi anni e attualmente in corso di reingegnerizzazione, attraverso il quale gli utenti sono in grado di trovare informazioni ed esplorare le collezioni digitali culturali di migliaia di istituzioni europee, utilizza terminologie multilingue condivise⁴. Successivamente, anche i progetti ATHENA⁵ e Linked Heritage⁶ hanno focalizzato l'attenzione sul tema degli strumenti terminologici e della semantica, elaborando delle linee guida e il prototipo di un registro per la gestione di terminologie multilingue, che verrà stabilizzato nel corso del progetto AthenaPlus e di cui si tratterà più diffusamente in questo contributo.

Per sfruttare al meglio i risultati a livello europeo, è molto importante cooperare e

¹ <http://www.linkedheritage.org/index.php?en/146/events/85/paris-seminar-on-multilingualism-and-terminology>.

² <http://vimeo.com/user14659022>.

³ *Report on inventories and multilingualism issues: Multilingualism and Thesaurus*, <<http://www.minervaeurope.org/publications/multilingualismandthesaurus.htm>>.

⁴ <http://www.michael-culture.org/>.

⁵ <http://www.athenaeurope.org>.

⁶ <http://www.linkedheritage.eu>.

condividere le proprie conoscenze nell'ambito di altre iniziative e con progetti attivi nel settore. Europeana, portale di accesso multilingue al patrimonio culturale europeo, riflette la diversità culturale europea. La gestione delle terminologie e del multilinguismo rappresenta però una delle sfide più significative per migliorare l'accesso ai milioni di contenuti digitali presenti nella piattaforma.

Vivien Petras, responsabile delle tematiche sull'accesso multilingue nella rete tematica Europea v2.0 e partner nel progetto PROMESSE, laboratorio di ricerca partecipativa per la valutazione di sistemi informativi multimediali e multilingue, ha parlato del multilinguismo nell'ambito di Europeana e dei diversi approcci utilizzati per l'allineamento delle terminologie multilingue sviluppate dai diversi progetti, nonché del modello seguito da Europeana, che si basa sui principi dei Linked Data.

Nell'ambito del progetto EuropeanaConnect si è proceduto all'allineamento con alcuni vocabolari pivot, come UDC, DDC, VIAF, TGN, Geonames, Wordnets, dpPedia. I metodi per l'allineamento sono stati diversi (accoppiamento lessicale, accoppiamento basato sulla struttura ecc.) e i diversi allineamenti sono stati combinati.



Datacloud sviluppato nel progetto EuropeanaConnect, 2011

Tra le prospettive di sviluppo di Europeana per il multilinguismo, sono state indicate: migliori euristiche dell'arricchimento e normalizzazione più rigida; annotazione dei metadati attraverso il contributo dell'utente (social tagging, classificazione); geoparser e gazetteer per la creazione di dati geografici basati sui nomi dei luoghi; ontologie aperte per i periodi storici; arricchimento di Agenti e Concetti basati su Dbpedia.

Marie-Veronique Leroi ed Eva Coudyzer, rispettivamente del Ministero francese della cultura e della comunicazione e dei Musei reali d'arte e storia di Bruxelles, hanno presentato i risultati ottenuti nel corso del progetto ATHENA, in particolare una serie di raccomandazioni e linee guida per la gestione della terminologie e del

multilinguismo con l'obiettivo di sensibilizzare le istituzioni culturali al principio dei linked data e alle tecnologie del web semantico. ATHENA si rivolgeva soprattutto ai musei europei, molti dei quali utilizzano propri vocabolari per descrivere le collezioni. Le linee guida sono state sintetizzate nella pubblicazione *Your terminology as a part of the semantic web: recommendations for design and management*⁷, liberamente scaricabile dal web. Le principali raccomandazioni sono lo sviluppo dell'interoperabilità tecnica e l'utilizzo di formati standard (XML, RDF, OWL, SKOS). Questo lavoro preliminare è proseguito nel progetto Linked Heritage dove l'approccio teorico di Athena è stato messo in pratica con lo sviluppo di un prototipo di piattaforma per la gestione delle terminologie (TMP).

La Terminology Management Platform è stata illustrata nella seconda sessione del seminario, dedicata agli strumenti per la creazione, armonizzazione e gestione delle terminologie multilingue. La TMP doveva rispondere ai bisogni identificati dai gruppi di lavoro dei progetti ATHENA e Linked Heritage:

- registrazione di una terminologia in un repository;
- skosificazione della terminologia;
- ricerca e navigazione in una rete di vocabolari;
- mapping della terminologia con un thesauro;
- arricchimento del thesauro;
- moderazione collaborativa per la revisione di un thesauro.



Sintesi dei bisogni della TMP

I requisiti della TMP erano i seguenti:

- essere un servizio web per un lavoro collaborativo in rete
- avere un'interfaccia amichevole, adatta a un uso da parte di utenti non esperti, provenienti da istituzioni culturali della memoria
- combinare componenti open source
- essere strutturata logicamente in un workflow intuitivo
- essere sufficientemente flessibile per essere adattata a nuovi standard.

La piattaforma verrà stabilizzata nel corso del progetto AthenaPlus, avviato a marzo 2013.

⁷ <http://www.linkedheritage.org/getFile.php?id=244>.

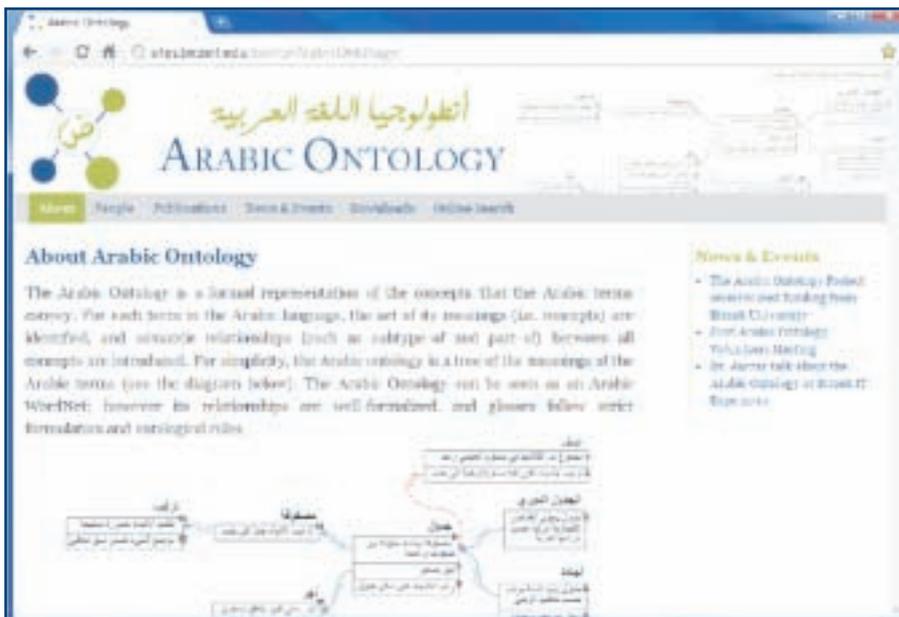


Schermata di accesso alla TMP⁸

Successivamente, Christophe Roche, docente presso l'Università della Savoia, ha presentato il progetto Integrating Sina Institute into the European Research Area (Siera). Lo sviluppo dei mercati transfrontalieri e della globalizzazione impone di facilitare la condivisione delle conoscenze tra i popoli. È quindi indispensabile procedere nella digitalizzazione dei documenti che vanno resi accessibili e comprensibili alle altre comunità. La diversità linguistica, culturale e normativa è uno dei principali ostacoli allo scambio di conoscenze. L'obiettivo generale del progetto è il rafforzamento della cooperazione scientifica tra i paesi dell'Unione europea e i ricercatori palestinesi nel campo delle tecnologie di condivisione della conoscenza multilingue e multiculturale. In particolare, nel corso del progetto verranno realizzate le seguenti iniziative:

- un'indagine sulla presenza della lingua araba all'interno di portali per la condivisione di conoscenza
- un'estensione dei thesauri e delle ontologie multilingue nei settori della cultura e dell'ecologia (localizzazione araba del thesaurus Michael, skosificazione del thesaurus Bethléem)
- la creazione di una corrispondenza tra Wordnet e l'Ontologia araba, un progetto avviato nel 2010 dall'Università palestinese di Birzeit.

⁸ <http://www.culture-terminology.org/>.



Schermata del sito della Arabic Ontology

Katell Briatte, del Ministero francese della cultura e della comunicazione, ha presentato GINCO, un software libero sviluppato dal Ministero francese per la gestione dei vocabolari, nel rispetto dei principi definiti dallo standard ISO 25964-1:2011 Information and documentation — Thesauri and interoperability with other vocabularies — Part 1 : Thesauri for information retrieval. GINCO è rilasciato con licenza CeCiLL v2 license (GNU GPL compatibile) ed è disponibile sul servizio di hosting GitHub⁹.

Le caratteristiche principali già sviluppate sono:

- l'interfaccia grafica e la gestione di liste multiple di autorità, tassonomie, thesauri;
- la gestione dei termini: status, lingua, identificatore unico di tipo URI, note di tipo diverso associate ai termini, assegnazione di termini ai concetti, relazioni di equivalenza tra termini, relazioni gerarchiche e associative tra concetti;
- la gestione di faccette e collegamenti virtuali;
- la gestione di gruppi di concetti (microthesauri, temi, dominio);
- l'importazione ed esportazione di thesauri in formato SKOS/RDF;
- la consultazione attraverso servizi web.

⁹ <http://github.com/culturecommunication/ginco>.

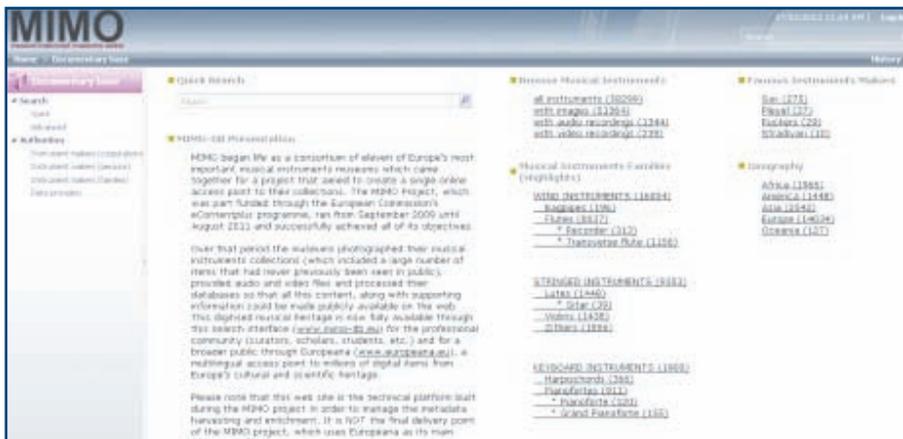
È già prevista una seconda release che prevederà:

- lo sviluppo di funzionalità collaborative
- l’allineamento tra vocabolari e risorse esterne
- l’implementazione di una terminologia di riferimento al fine di fornire agli utenti un accesso unificato a tutti i vocabolari (terminologies server).

Mark Lindeman, direttore di Picturae, azienda olandese che realizza soluzioni informatiche per i beni culturali, ha presentato le applicazioni legate al sistema OpenSkos, una piattaforma di pubblicazione che fornisce API su SKOS. Il software OpenSkos¹⁰ ha un editor per la gestione di SKOS; è un progetto open source, risultato della collaborazione tra l’Agenzia olandese per i beni culturali, Trezorix, Adlib and Picturae.

È stata quindi la volta di Nacha Van Steen che ha presentato l’evoluzione della terminologia multilingue – fonti, struttura, problematiche – nel progetto Europeaana Photography¹¹ e come essa verrà integrata all’interno di Europeaana. Allo stato attuale sono stati individuati 500 concetti tradotti in 12 lingue.

Rodolphe Bailly, responsabile del Sistema d’informazione documentaria e della digitalizzazione presso la Cité de la musique di Parigi, ha illustrato il ruolo, gli strumenti per l’elaborazione e l’utilizzo del thesaurus multilingue del progetto europeo MIMO¹², oramai concluso, che si proponeva di aggregare per Europeaana contenuti di musei europei che possiedono collezioni di strumenti musicali.



Schermata di accesso al database di MIMO¹³

¹⁰ <http://openskos.org/>.

¹¹ <http://www.europeana-photography.eu/>.

¹² <http://www.mimo-international.com/>.

¹³ http://www.mimo-db.eu/mimo/infodoc/default.aspx?_lg=EN-en.

Anche il progetto Partage Plus¹⁴, che ha come obiettivo la digitalizzazione di materiali risalenti al periodo artistico noto come Art Nouveau, sta costruendo una serie di vocabolari controllati multilingue per consentire un più facile reperimento delle relative risorse culturali digitali in Europeana. Christiane Pagel, dell'Università di Marburg, ha spiegato in dettaglio la struttura dei metadati e la definizione dei campi obbligatori per la catalogazione di base nell'ambito del progetto, i cui risultati saranno pienamente fruibili all'inizio del 2014.

La terza parte del seminario verteva sulla presentazione di casi di riuso e applicazioni basate sui Linked Data.

Thibault Grouas, responsabile della creazione di progetti tecnologici a supporto della protezione della lingua francese e delle 75 lingue parlate in Francia, ha presentato SemanticPédia, un programma a supporto dei programmi di ricerca e sviluppo applicati a progetti collaborativi culturali, utilizzando dati estratti dai progetti di Wikimedia. Il progetto nasce dalla collaborazione tra Ministère de la culture et de la communication, Institut national de recherche en informatique et en automatique (Inria) e Wikimédia France. Cosa ha stimolato l'avvio di SemanticPédia? Il fatto che Wikipedia conti quasi un milione e mezzo di articoli in francese, che di questi più del 40% sono contenuti culturali diversificati, che una parte di queste informazioni è già strutturata e potenzialmente riusabile, che il progetto iniziale Dbpedia.org non includeva la lingua francese.

Il primo progetto sviluppato nel quadro di questo programma è stato "Dbpédia en français", che ha l'obiettivo di proporre dei dati strutturati a partire dalle informazioni contenute nell'enciclopedia Wikipedia in versione francese. Esso ha già trovato un'applicazione concreta sul portale di storia dell'arte HDA-Lab¹⁵.

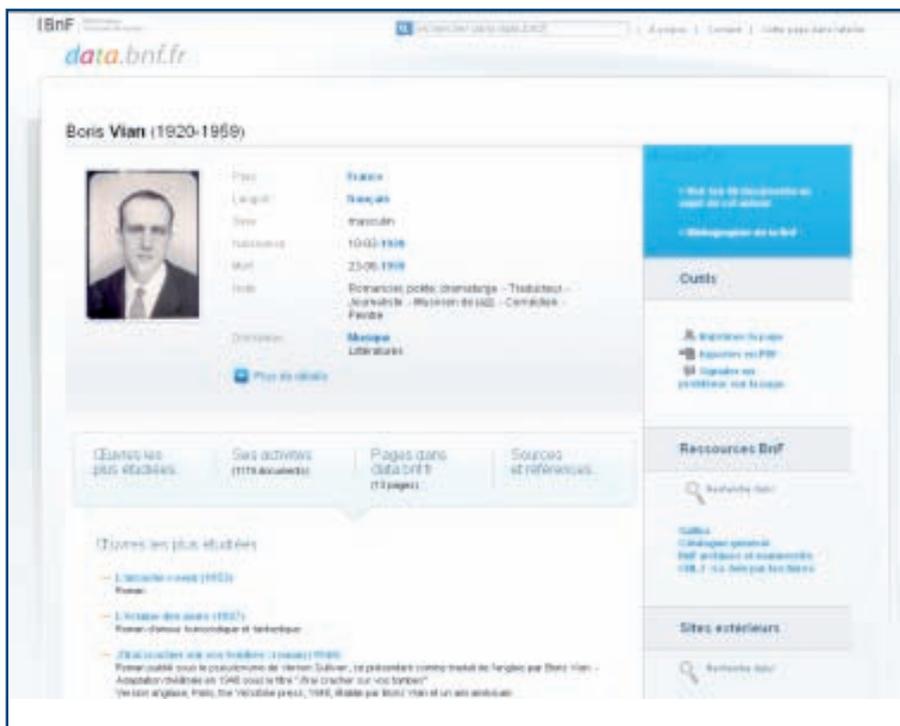


Schermata di accesso al portale di storia dell'arte HdA Lab

¹⁴ <http://www.partage-plus.eu/>.

¹⁵ <http://hdalab.iri-research.org/hdalab/>.

L'ultimo intervento è stato quello di Romain Wenz, della Biblioteca nazionale di Francia, che ha illustrato l'utilizzo di Virtual International Authority File (VIAF) in data.bnf.fr allo scopo di creare dei legami per importare delle informazioni. Il progetto della Biblioteca nazionale francese ha l'obiettivo di guidare l'utente tra le diverse risorse, raggruppando su una stessa pagina tutte le informazioni provenienti dai diversi cataloghi della biblioteca e da Gallica. Il progetto utilizza gli strumenti del web semantico e si inserisce nella strategia di fornitura di dati aperti. Volendo fare un esempio, per generare la pagina su Boris Vian visibile alla URL: <http://data.bnf.fr/13091689/boris_vian/> viene utilizzata la base di dati VIAF, che fornisce una serie di link in lingue diverse¹⁶. I link vengono integrati verso questa base di dati, ma anche verso DBpedia, grazie a dei collegamenti diretti. Questo meccanismo consente di ritrovare i diversi nomi con i quali un certo autore è stato citato nelle lingue rappresentate nel VIAF e di fornire alla macchine un link raggiungibile, quale che sia la lingua dell'utente, il quale potrà fruire dei contenuti attraverso interfacce in lingue diverse.



Pagina su Boris Vian nel sito data.bnf.fr

¹⁶ <http://viaf.org/viaf/9983585/>.

L'ultima consultazione dei siti web è avvenuta nel mese di giugno 2013.