DigItalia

Anno VII, Numero 2 - **2012**

ISSN 1972-6201

Rivista del digitale nei beni culturali

ICCU-ROMA

Interoperabilità e armonizzazione nelle biblioteche digitali: l'esperienza del progetto European Film Gateway[1]

Alessia Bardi - *Dipartimento di ingegneria dell'informazione, Università di Pisa*
Paolo Manghi - *Consiglio nazionale delle ricerche - Istituto di scienza e tecnologie dell'informazione "A. Faedo"*
Franco Zoppi - *Consiglio nazionale delle ricerche - Istituto di scienza e tecnologie dell'informazione "A. Faedo"*

Introduzione

Grazie a un rilevante sforzo speso nella digitalizzazione di collezioni di materiale multimediale (documenti audio, video, fotografie, locandine, manifesti, disegni, testi ecc.), oggigiorno sono presenti in Europa molti archivi digitali di materiale filmografico. Essi forniscono l'accesso alle loro collezioni tipicamente attraverso portali web che, supportati da opportune tecnologie, permettono la ricerca e la visualizzazione di oggetti digitali, ovvero i dati descrittivi, o metadati, e le versioni digitalizzate del materiale archiviato. Sebbene il servizio che questi archivi offrono sia senz'altro utile, la loro parcellizzazione rappresenta un limite alla domanda di accesso globale espressa ormai dalla comunità degli utenti del settore.
La Best Practice Network European Film Gateway[2] (EFG), finanziata dal 2008 al 2011 dalla Commissione europea nell'ambito del programma eContentplus[3], fornisce agli utenti un singolo punto di accesso attraverso il quale è possibile interrogare in modo uniforme il materiale filmografico eterogeneo conservato da alcuni fra i maggiori archivi cinematografici d'Europa quali Istituto Luce, Das Deutsches Filminstitut, Cinemateca Portuguesa. Per rendere possibile la costituzione di un tale spazio informativo omogeneo e accessibile sia da singoli utenti sia da terze parti autorizzate (ad esempio la European Digital Library – Europeana[4]), è necessario aggregare ed armonizzare i contenuti degli archivi sorgenti. A tal fine, EFG fornisce un'infrastruttura dati per la collezione, armonizzazione ed esportazione di metadati in formato XML.

[1] Questo articolo costituisce la sintesi delle attività svolte dal progetto EFG, del quale l'istituto CNR-ISTI è stato partner tecnologico, <http://www.europeanfilmgateway.eu/>.
[2] Progetto European Film Gateway, <http://www.europeanfilmgateway.eu>.
[3] Programma eContentPlus, <http://ec.europa.eu/information_society/activities/econtentplus>.
[4] Europeana, <http://www.europeana.eu>.

Tale infrastruttura è stata realizzata per superare le problematiche di interoperabilità legate al processo di aggregazione di metadati eterogenei. Infatti, sebbene gli archivi cinematografici contengano oggetti simili, i loro modelli dei dati (e i relativi schemi di metadati) possono differire molto sia nella struttura sia nella semantica, così come il loro contenuto può essere soggetto a errori o duplicazioni. I principali requisiti che l'infrastruttura di EFG è tenuta a soddisfare sono perciò (i) la capacità di processare metadati eterogenei, cioè conformi a diversi formati e relativi a diverse tipologie di entità, al fine di creare uno spazio informativo integrato su cui servizi automatici e persone possono operare; (ii) la capacità di fornire servizi ai curatori per migliorare la qualità dei dati nello spazio informativo creato; (iii) fornire un supporto completo per la ricerca avanzata nello spazio informativo; e (iv) supportare protocolli standard per l'esportazione dei metadati verso servizi di terze parti.

Per quanto riguarda i problemi di interoperabilità, EFG ha adottato un approccio a due fasi, iniziando con la definizione di un modello dei dati comune, del relativo schema dei metadati e di vocabolari controllati specifici per il dominio applicativo e concludendo con l'implementazione della tecnologia per raccogliere dagli archivi i record di metadati, trasformarli nello schema comune e armonizzarli.

Per garantire sostenibilità, scalabilità e robustezza, l'infrastruttura EFG è stata realizzata utilizzando il software open-source D-NET[5]. D-NET è un software abilitante per la realizzazione e manutenzione di infrastrutture dati. Esso fornisce un insieme ricco e personalizzabile di servizi di gestione dei dati per la raccolta, memorizzazione, indicizzazione, trasformazione, armonizzazione ed esportazione di metadati XML. Inoltre, D-NET permette di configurare repliche distribuite dei servizi e backup programmatici dello spazio informativo. D-NET rende inoltre disponibili servizi per lo sviluppo di portali che possono essere configurati in base ai requisiti applicativi della comunità di utenti che li dovrà utilizzare, oltre a supportare vari formati di scambio e servizi di mediazione per sistemi esterni che dovessero accedere dallo spazio informativo, quali OAI-PMH[6] e SRW/CQL[7].

Per quanto riguarda l'armonizzazione e cura dei dati, D-NET è stato esteso con servizi per la modifica e per il controllo della qualità dei dati, oltre che per la gestione di file autoritativi (ad esempio per il problema della de-duplicazione). Nel seguito, la Sezione 2 presenta il workflow di aggregazione dello spazio informativo implementato dall'infrastruttura EFG. La Sezione 3 presenta brevemente il software abilitante per infrastrutture dati D-NET. La Sezione 4 descrive come D-NET sia stato configurato ed esteso con servizi e strumenti per la cura dei dati in EFG. Si conclude infine, nella Sezione 5, riportando alcuni temi di sviluppo futuri.

[5] D-NET Software Toolkit, <http://www.d-net.research-infrastructures.eu>.

[6] Carl Lagoze, Herbert Van de Sompel, *The making of the open archives initiative protocol for metadata harvesting*, «Library Hi Tech», 21 (2003), n. 2, p. 118-128.

[7] SRU Protocol Family, <http://www.loc.gov/standards/sru/>.

Il workflow per l'aggregazione di archivi cinematografici

Come anticipato nell'introduzione, un workflow di aggregazione deve risolvere problematiche di interoperabilità dei dati derivanti dalla natura eterogenea degli archivi. Di fatto, il contenuto di archivi diversi è in genere organizzato secondo modelli dei dati e schemi diversi, la cui struttura può variare da gerarchie o grafi complessi a semplici insiemi "piatti" di elementi. Inoltre, il contenuto descritto può riferirsi sia a entità diverse come pure a entità uguali, ma con diversa semantica (ad esempio vocabolari di termini o formati standard diversi per la rappresentazione di nomi, date, periodi).

Per risolvere questo problema in EFG è stato da un lato definito il modello dati comune e il relativo schema XML, sul quale sono mappati i record di metadati degli archivi; dall'altro è stata realizzata un'infrastruttura di gestione dei dati, i cui servizi permettono di raccogliere i record XML dagli archivi e trasformarli in record conformi allo schema XML comune, e di "curare" i record così ottenuti identificando e correggendo gli errori semantici e i duplicati.

Il *workflow* di aggregazione dello spazio informativo (schematizzato in fig. 1) è costituito da più fasi e richiede l'interazione tra gli esperti del dominio applicativo e gli amministratori dell'infrastruttura, adeguatamente supportati dai servizi dell'infrastruttura stessa. Questi diversi attori sono guidati da una precisa metodologia operativa, il cui scopo è la realizzazione di un ciclo di alimentazione controllato e iterato che porti incrementalmente alla formazione di uno spazio informativo di alta qualità. Il *workflow* si articola sulle seguenti quattro fasi:

- Fase 1: definizione del mapping dei metadati. Gli esperti del dominio applicativo presso gli archivi cinematografici analizzano i metadati che intendono fornire per determinare le corrispondenze semantiche e strutturali con lo schema di EFG. Tali corrispondenze sono formalizzate in regole descritte in documenti di testo e tabelle di corrispondenza e fornite agli amministratori dell'infrastruttura che le codificano in script eseguibili.
- Fase 2: trasformazione e armonizzazione dei metadati. I record di metadati forniti dagli archivi sono raccolti dall'infrastruttura utilizzando protocolli standard (OAI-PMH, FTP ecc.) e processati dagli script prodotti in Fase 1 per generare i record EFG corrispondenti. Questi sono memorizzati in uno spazio informativo di "pre-produzione" per essere ulteriormente elaborati nella successiva Fase 3. Le Fasi 1 e 2 possono essere ripetute più volte per raffinare le regole di mapping e raggiungere la migliore qualità possibile dei metadati.
- Fase 3: controllo della qualità e arricchimento dei metadati. I record nello spazio informativo di pre-produzione – prima di essere definitivamente validati – vengono esaminati con diversi strumenti web per identificare eventuali refusi, errori di mapping e record duplicati. In particolare, gli strumenti *Content* e *Vocabulary Checker* permettono di verificare la correttezza delle regole strutturali e semantiche.

Lo strumento *Authority File Manager* identifica invece possibili record duplicati, ovvero record diversi relativi allo stesso oggetto del mondo reale. Questo processo di controllo della qualità può portare alla ridefinizione delle regole di mapping (Fase 1), alla modifica degli script di mapping (Fase 2), o al processo di arricchimento dei dati mediante lo strumento *Metadata Editor*, il quale consente di modificare, aggiungere o rimuovere record nello spazio informativo.

– Fase 4: pubblicazione dei metadati. I record validati nella Fase 3 vengono replicati in uno spazio informativo di "produzione" dove vengono resi disponibili al portale della comunità EFG e possono anche essere esportati verso terze-parti, come Europeana.

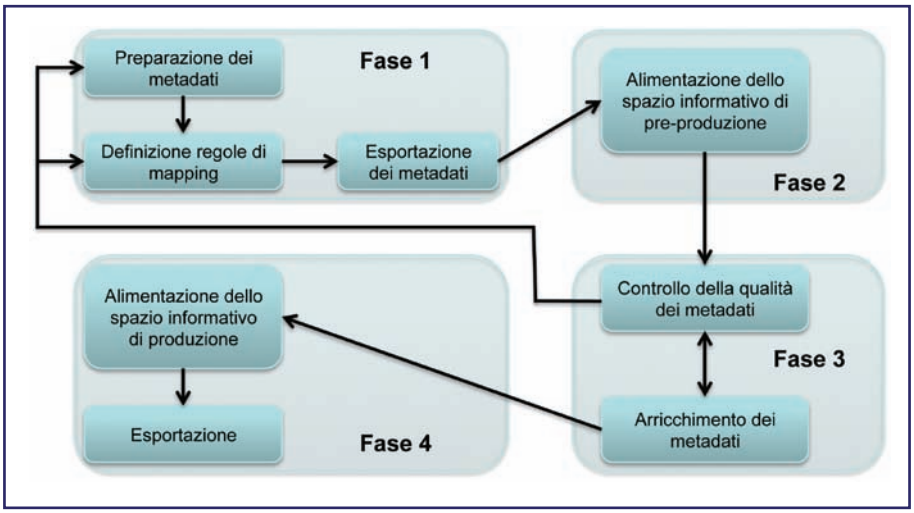

Figura 1. *Le fasi del workflow di alimentazione di EFG*

D-NET: software abilitante per infrastrutture dati

Nell'ultima decade, come testimoniato da svariate iniziative nazionali (ad esempio BASE[8], DAREnet[9], OAIster[10]) e progetti finanziati dalla Comunità Europea (ad esempio Europeana[11], Bricks[12], ScholNet[13], DILIGENT[14], D4Science[15], DRIVER[16],

[8] BASE: Bielefeld Academic Search Engine, <http://www.base-search.net>.
[9] DAREnet: Digital Academic Repositories, <http://www.darenet.nl/>.
[10] AIster Official Site, <http://www.oaister.org>.
[11] Europeana, <http://www.europeana.eu>.
[12] Bricks Project, <http://www.brickscommunity.org/>.
[13] ScholNet Project, <ftp://ftp.cordis.europa.eu/pub/ist/docs/rn/scholnet.pdf>.
[14] DILIGENT Project, <http://diligent.ercim.eu/>.
[15] D4Science Project, <http://www.d4science.eu/>.
[16] DRIVER Project, <http://www.driver-community.eu/>.

OpenAIRE[17], CLARIN[18], HOPE[19]), la diffusione delle Digital Library (DL) nell'ambito di singole comunità, è stata seguita da una nuova necessità di aggregare e integrare contenuti da DL diverse, per renderli poi disponibili attraverso un unico punto di accesso e andare incontro al carattere sempre più multidisciplinare della ricerca moderna.
Nell'ambito dei progetti sopra menzionati sono state ideate diverse soluzioni tecnologiche[20] per la realizzazione di infrastrutture per l'aggregazione di metadati, in grado di offrire funzionalità per raccogliere dati da sorgenti eterogenee (ad esempio repository, archivi, database), elaborare tali dati per creare spazi informativi omogenei, e sviluppare servizi per portali personalizzati per poter operare su tali spazi informativi (ad esempio ricerche, inferenza di riferimenti tra pubblicazioni, calcolo delle citazioni).
Tra questi software abilitanti, di particolare interesse è il software D-NET[21] sviluppato nell'ambito dei progetti europei DRIVER e OpenAIRE. D-NET è una soluzione open-source specificamente progettata per la realizzazione e la gestione di infrastrutture dati personalizzate sostenibili.

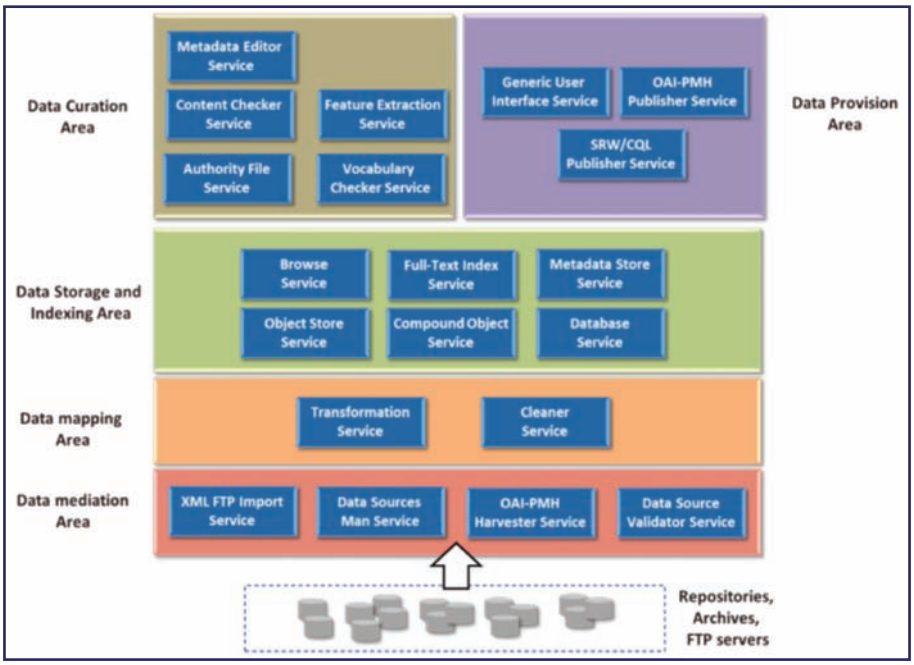

Figura 2. *Architettura dei servizi del software D-NET*

[17] OpenAIRE Project, <http://www.openaire.eu/>.
[18] CLARIN Project, <http://www.clarin.eu/>.
[19] HOPE Project, <http://www.peoplesheritage.eu>.
[20] Paolo Manghi - Marko Mikulicic - Leonardo Candela - Michele Artini - Alessia Bardi, *General-Purpose Digital Library Content Laboratory Systems. ECDi Proceedings of the 14th European conference on Research and advanced technology for digital libraries, Berlin - Heidelberg: Springer-Verlag, 2010, p. 14-21.*
[21] D-NET Software Toolkit, <http://www.d-net.research-infrastructures.eu>.

D-NET fornisce una piattaforma *service oriented* sulla quale le infrastrutture dati possono essere sviluppate secondo un approccio modulare, scegliendo e combinando opportunamente i singoli servizi D-NET. Le infrastrutture risultanti sono personalizzabili sulla base del modello dei dati comune e di altre logiche dipendenti dal dominio, estendibili (nuovi servizi possono essere integrati per estendere le funzionalità della piattaforma) e scalabili (ad esempio repliche dei servizi di memorizzazione e indicizzazione possono essere attivati su nodi remoti per far fronte all'aumento di accessi concorrenti o della dimensione dei dati gestiti).
L'insieme dei servizi offerti da D-NET è molto ampio (fig. 2) e copre aspetti come la raccolta dei dati (*mediation area*), il mapping tra formati diversi (*mapping area*), la memorizzazione (*storage and indexing area*), l'accessibilità (*provision area*) e la cura (*curation area*) dei dati. I servizi possono anche essere personalizzati e combinati assieme per soddisfare gli specifici requisiti di una determinata comunità di utenti.

L'infrastruttura dati EFG

L'infrastruttura dati EFG è stata realizzata configurando e combinando opportunamente i servizi del framework D-NET. In particolare, i servizi dell'area di Data Curation sono il risultato delle attività del progetto. Essi sono stati concepiti per rispondere ai requisiti della comunità di EFG, ma ingegnerizzati in modo da poter essere configurati per l'uso in altri contesti.
Il cuore dell'infrastruttura è il modello dei metadati comune che è stato definito dal consorzio EFG. I servizi D-NET necessari per la realizzazione del *workflow* presentato nella Sezione 2 sono configurati in conformità a tale modello e alla sua implementazione in XML-Schema.

Il modello e lo schema dei metadati di EFG

Il modello dei metadati EFG è stato definito a seguito dell'analisi dei modelli e degli schemi adottati da tutte le maggiori organizzazioni operanti nel dominio audio/video, a partire ovviamente dai fornitori dei contenuti della comunità EFG. Questo studio ha preso in considerazione standard come FRBR[22] e Dublin Core[23], così come standard specifici del settore come il Cinematographic Works Standards EN 15907[24].

[22] IFLA: Study group on the functional requirements for bibliographic records, *Functional requirements for bibliographic records: final report*, München: K. G. Saur, 1998, (UBCIM publications. n. s.; 19).
[23] Stuart L. Weibel, *Metadata: The Foundations of Resource Description*, «Annual review of OCLC research», 1995, p. 52-56. <http://www.dlib.org/dlib/July95/07weibel.html>.
[24] CEN Technical Committe, Metadata standards for cinematographic works (2005).

Come risultato di tale studio, il modello EFG[25,26] definisce otto entità in relazione tra loro: *AVCreation* (creazione audio/video, tipicamente un film), *AVManifestation* (manifestazione fisica di un aAVCreation, ad esempio la versione di un film in una particolare lingua), *NonAVCreation* (creazione non audio/video, ad esempio foto, poster, documenti di censura), *NonAVManifestation* (manifestazione fisica di una NonAVManifestation), *Item* (file digitale), *Agent* (persone e organizzazioni), *Event* (evento che può verificarsi durante il ciclo di vita di una creazione, ad esempio la vittoria di un premio, una proiezione pubblica), *Collection* (gruppo di AVCreation e NonAVCreation).
In figura 3 è mostrato un esempio relativo al film *2001: Odissea nello spazio* diretto da Stanley Kubrick. Il record del film è collegato alle sue manifestazioni, relative locandine, regista e attori.

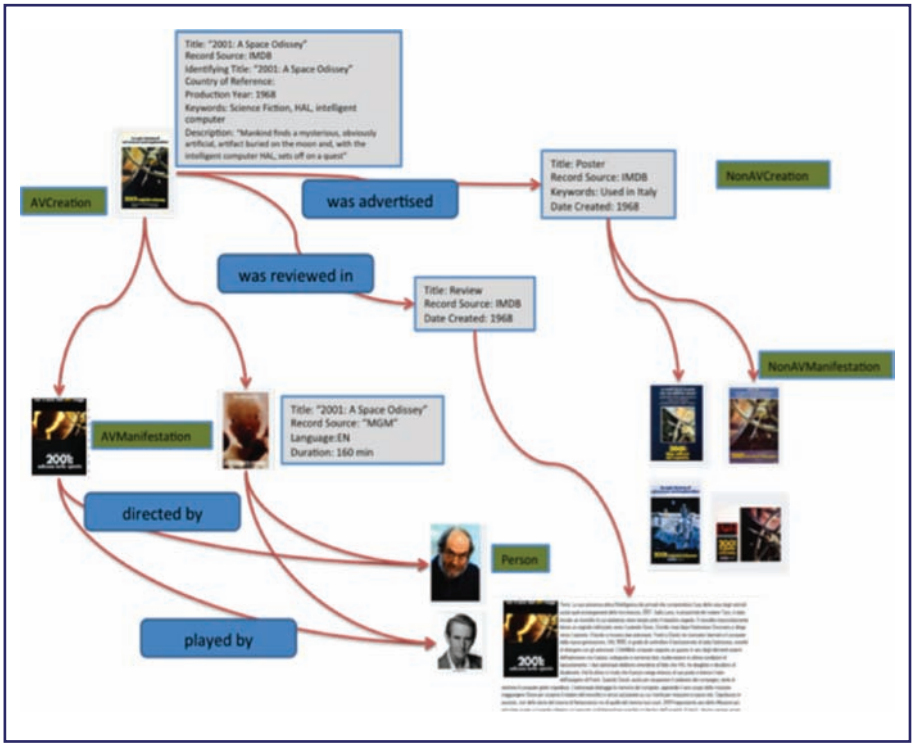

Figura 3. *Esempio di metadati associati al film 2001: A Space Odyssey*

[25] Detlev Balzer, Franca Debole, Pasquale Savino, *Common interoperability schema for archival resources and filmographic descriptions*, Deliverable D2.2, EFG Project.

[26] Pasquale Savino - Franca Debole - Georg Eckes, *Searching and browsing film archives. The European Film Gateway Approach, Cultural Heritage Cairo 2009*. 4th International Congress on Science and Technology on the Safeguard of Cultural Heritage in the Mediterranean Basin, Cairo, Egypt, 6-8 December 2009. Proceedings, Roma: Fondazione Roma Mediterraneo, 2009, p.359-364.

Lo schema[27] XML di EFG implementa il modello comune appena descritto. Esso definisce i tipi di elementi e gli attributi delle otto entità e le loro proprietà. Lo schema EFG è stato concepito come l'unione di otto distinti schemi XML (uno per ogni entità) in modo che un record XML di EFG rappresenti un'entità assieme a tutte le sue relazioni con altre entità. Inoltre esso definisce i cosiddetti "elementi controllati", cioè elementi XML i cui valori devono rispettare un dato vocabolario di termini.

Definizione dei mapping, trasformazione e armonizzazione dei metadati

Gli archivi – e gli esperti del dominio applicativo – che entrano a far parte della comunità EFG sono supportati da una metodologia di lavoro che facilita la definizione del mapping strutturale dagli schemi degli archivi allo schema comune EFG e del mapping semantico dai propri vocabolari ai vocabolari usati dalla comunità. Un mapping è un insieme di regole che gli amministratori dell'infrastruttura usano per configurare i servizi dell'area di Data Mapping. Qui, i servizi *Transformation* e *Cleaner* eseguono degli script che analizzano, validano e trasformano i record sorgenti nel formato comune di EFG secondo le regole definite.
Il servizio Transformation provvede all'applicazione di regole strutturali. Esse definiscono la corrispondenza tra elementi e attributi dello schema dell'archivio sorgente ed elementi e attributi dello schema EFG. Il mapping strutturale è tutt'altro che banale, poiché tipicamente un record sorgente è mappato su più record EFG – in relazione tra loro – che rappresentano diverse entità del modello dei dati EFG.
Il servizio *Cleaner* è invece responsabile dell'applicazione delle regole semantiche. Ciascuna regola rappresenta un vocabolario dei sinonimi, stabilendo corrispondenze fra i valori utilizzati nei record sorgenti e i termini dei vocabolari adottati in EFG.

Controllo della qualità e arricchimento dei metadati

Per la realizzazione dell'infrastruttura dati EFG, il software D-NET è stato esteso con un insieme di servizi – costituenti l'area di *Data Curation* – descritti nel seguito.

Content Checker (fig. 4) è lo strumento web di validazione che permette la ricerca e navigazione nello spazio informativo di pre-produzione allo scopo di verificare che i record di metadati siano stati raccolti e trasformati correttamente.

[27] Detlev Balzer - Franca Debole - Pasquale Savino, *Common interoperability schema for archival resources and filmographic descriptions*, Deliverable D2.2, EFG Project.

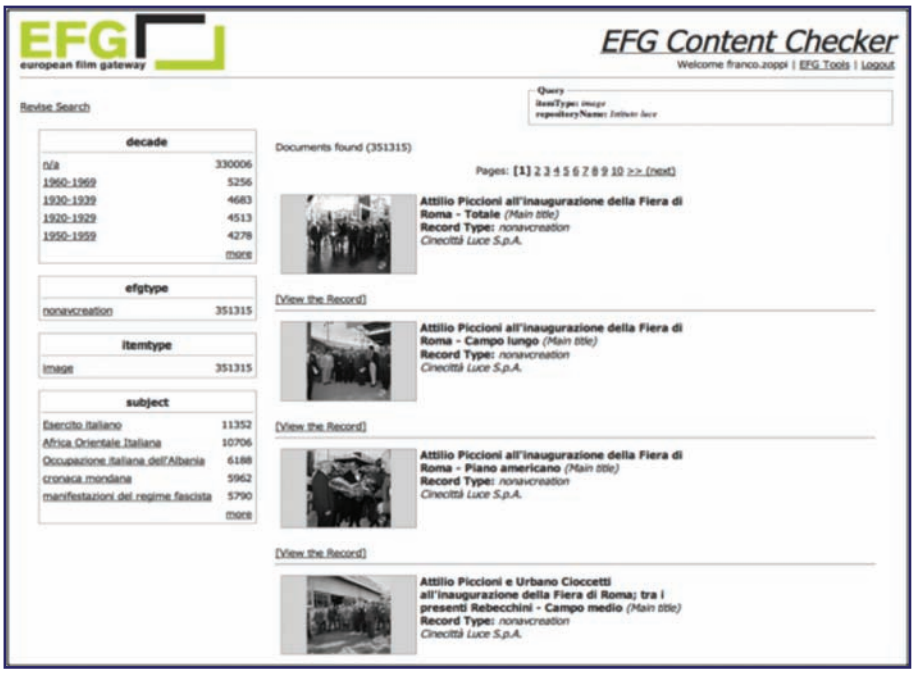

Figura 4. *EFG Content Checker*

Vocabulary Checker è lo strumento web che permette di individuare e accedere ai record di metadati che non soddisfano i vincoli imposti dallo schema e dai vocabolari controllati dopo la fase di trasformazione e armonizzazione. Il Vocabulary Checker visualizza numero, tipo e posizione degli errori nei record nello spazio informativo di pre-produzione. Grazie alla funzionalità di navigazione per tipologia di errore, i curatori possono decidere se un errore può essere corretto direttamente nello spazio informativo per mezzo del Metadata Editor, oppure nell'archivio sorgente.

Metadata Editor è lo strumento web per l'arricchimento dello spazio informativo. Esso permette ai curatori non solo di aggiungere, modificare ed eliminare record di metadati, ma anche di stabilire relazioni tra record esistenti – anche se questi provengono da archivi diversi.

Authority File Manager (PACE[28]) è uno strumento che i curatori possono utilizzare per individuare e unire record duplicati e disambiguare lo spazio informativo. Lo strumento è in grado di identificare automaticamente le coppie di record candidati

[28] Paolo Manghi - Marko Mikulicic - Claudio Atzori, *PACE: a general-purpose tool for authority control*, MTSR 2011 - Metadata and Semantic Research. 5th International Conference, Izmir, Turkey 12-14 October 2011, (Communication in computer and information science, v. 240, pt. 1), Proceedings p. 80-92, Berlin-Heidelberg: Springer, 2011.

per l'unione basandosi su una versione multi-ordinamento dell'algoritmo di *sorted neighbourhood* e su una funzione di similarità personalizzabile dai curatori (i quali possono scegliere tra un insieme di funzioni di similarità e assegnare diversi pesi ai campi dei record in base alla loro rilevanza).

Pubblicazione dei metadati

Il portale EFG[29] sfrutta i servizi D-NET per fornire agli utenti funzionalità quali la ricerca avanzata di metadati, la navigazione (per collezione, provider, data, lingua, tipologia di materiale), il raffinamento dei risultati della ricerca attraverso filtri, lo streaming video, gallerie di immagini e il servizio di news. Inoltre, l'infrastruttura EFG utilizza i servizi D-NET per l'esportazione dei metadati via OAI-PMH e SRW/CQL verso software di terze parti, in primo luogo verso Europeana di cui EFG è un fornitore diretto.

Conclusioni e sviluppi futuri

Abbiamo descritto le soluzioni adottate nella Best Practice Network EFG per raggiungere una completa integrazione di archivi cinematografici nazionali eterogenei. La soluzione è basata sulla creazione di uno schema comune di metadati con elementi caratterizzanti il dominio e capace di preservare la qualità dei metadati sorgenti. L'infrastruttura EFG colleziona i metadati degli archivi e li trasforma nel formato dello schema comune per renderli disponibili in uno spazio informativo omogeneo. Le funzionalità dell'infrastruttura sono state realizzate attraverso la configurazione, combinazione e integrazione di servizi del software D-NET, una piattaforma per lo sviluppo di infrastrutture dati personalizzabili, scalabili e sostenibili.
L'infrastruttura realizzata per il progetto EFG, iniziato nel 2008 e concluso nel 2011, ha raccolto 574.105 oggetti digitali relativi a 16 diversi fornitori di contenuti ed è stata accolta positivamente sia dalle comunità di utenti coinvolte, sia dal pubblico, tanto che il progetto ha attualmente una sua prosecuzione in EFG1914[30]. Il fine principale di EFG1914 è attrarre nuovi fornitori di contenuti per aumentare il numero degli oggetti digitali disponibili, in particolar modo quelli concernenti il periodo della Prima guerra mondiale. Gli autori si propongono di facilitare dal punto di vista tecnico l'entrata di nuovi fornitori nell'infrastruttura integrando nuovi servizi per l'importazione dei dati e per l'automazione della definizione e implementazione dei mapping.

[29] Portale European Film Gateway, <http://www.europeanfilmgateway.eu>.
[30] Progetto European Film Gateway 1914, <http://project.efg1914.eu/>.

Ringraziamenti
Questo lavoro è stato parzialmente finanziato dalla Best Practice Networks EFG – ECP 517006, FP7 EU eContentplus 2007. La sua realizzazione non sarebbe stata possibile senza la preziosa collaborazione di Marco Rendina (Istituto Luce, Italia), Georg Eckes e Francesca Schultze (Deutsches Filminstitut, Germania) per la progettazione del modello dei dati comune, e dei colleghi del CNR-ISTI Michele Artini, Federico Biagini, Franca Debole, Sandro La Bruzzo, Marko Mikulicic, Pasquale Savino che, a vario titolo, hanno contribuito all'ideazione, progettazione e realizzazione del progetto EFG e del software D-NET.

Per tutti i siti web, l'ultima consultazione è avvenuta nel mese di dicembre 2012.