UPGRADE
The European Online Magazine for the IT Professional
http://www.upgrade-cepis.org
Edizione italiana a cura di ALSI e Tecnoteca
http://upgrade.tecnoteca.it

Applicazione dell'XML al Giornalismo

di
Luis Sánchez-Fernández, Carlos Delgado-Kloos, Vicente Luque-Centeno,
María del Carmen Fernández-Panadero, e Laura Martínez-Bermejo

Traduzione italiana di Anna Fino, ALSI

Il Giornalismo è un classico esempio di dominio applicativo in cui l'XML sta producendo dei cambimanti nel modo di lavorare. Oggi gli articoli sono conservati e scambiati usando formati proprietari. Questi formati stanno diventando obsoleti anche a causa dello sviluppo di Internet e delle tecnologie multimediali. Per questo il Consiglio Internazionale di Telecomunicazioni della Stampa (IPTC), composto da organizzazioni e aziende che lavorano nel campo dell'immagazzinamento, distribuzione e pubblicazione delle informazioni a livello mondiale sta standardizzando nuovi formati basati su XML per migliorare le correnti procedure di lavoro in questo campo. In questo articolo presenteremo il lavoro che noi e alcuni ricercatori del Dipartimento di Scienze dell'Informazione della nostra università hanno effettuato in questo campo. Il nostro lavoro si basa su un modello che abbiamo sviluppato per i processi di lavoro in campo giornalistico usando l'XML.

Parole chiave: XML, Giornalismo, Supporto Multipiattaforma, Personalizzazione, E-commerce

1. Introduzione

Negli anni settanta i progressi nel campo dei computer e delle telecomunicazioni per lo scambio e l'archiviazione dei documenti ha portato all'automazione dei processi di lavoro in molti campi, incluso il giornalismo. In quegli anni erano stati sviluppati dei formati di descrizione degli articoli che permettevano lo scambio e l'archiviazione elettronica dei documenti, per esempio ANPA 1312 [Wire 1979] e IPTC 7901 [IPTC 1979].

I formati sono stati elaborati da applicazioni sviluppate appositamente per questo scopo.

Un articolo codificato in uno di questi formati contiene il testo dell'articolo (sono supportati solo articoli in formato testo), i campi, le categorie e gli indici aggiuntivi come per esempio l'identificativo numerico, la data, la sezione, la priorità.

L'evoluzione della tecnologia (Internet, multimedialità etc.) e l'aumentata attività giornalistica sta rendendo questi formati sempre più antiquati.

Di seguito un elenco delle carenze segnalate per questi formati:

- Non supportano la descrizione di contenuti multimediali
- Non hanno metadati per agevolare la classificazione e la ricerca degli articoli (ricerche per persone, luoghi etc)
- Le categorie di indici disponibili mancano di dettaglio
- Non supportano l'intero ciclo di vita di un articolo, per esempio non e' possibile identificare l'autore di un articolo, includere quale versione e' attualmente disponibile chi l'ha ricevuta.
- Non permettono collegamenti ad altri articoli correlati
- Non permettono la composizione degli articoli (per esempio associare una fotografia al testo dell'articolo).

Per rispondere a queste necessità le organizzazioni per gli standard giornalistici hanno deciso di sviluppare nuovi standard basati su applicazioni XML.

In questo articolo viene presentato un modello di processo che noi abbiamo proposto per la futura attività giornalistica.

Questo modello è stato sviluppato nel corso di due progetti di ricerca "El Periotrònico" e "Infomedia". Nella sezione due presentiamo le applicazioni XML per il giornalismo che sono state sviluppate dalla IPTC [IPTC].

Nella sezione 3 e 4 sono descritti il modello e la sua implementazione. E, infine, la sezione 5 contiene le conclusioni e identifica le aree per un ulteriore lavoro

2. NITF a NewsML

Consapevole delle limitazioni degli attuali formati di descrizione delle notizie, l'IPTC ha sviluppato nuovi formati basati sull'XML [XML 2000].

NITF [NITF] è stato il primo standard XML sviluppato dall'IPTC negli anni novanta. NITF è una applicazione XML per la descrizione di articoli in formato testuale, sebbene possa contenere oggetti multimediali. Questo linguaggio mira a superare alcune delle caremze che sono state identificate negli standard definiti in precedenza in quanto incorpora molti elementi di metadati.

Il bisogno di supportare le tecnologie multimediali ha permesso di produrre un nuovo standard, NewsML [NewsML]. NewsML per superare le carenze degli formati correnti, identificati nell'introduzione.

NewsML utilizza linguaggi complementari, per esempio NITF piò essere usato per la descrizione di notizie testuali all'interno di una descrizione NewsML.

Le caratteristiche principali di NewsML sono le seguenti:

- Introduce il concetto di Topicset che permette di definire un insieme di parole chiave che possono essere usate per descrivere le sezioni di un articolo, per esempio città, persone etc. Essendo un formato flessibile NewsML può essere ulteriormente sviluppato.
- Supporta tutti i tipi di media e formati digitali. Per ottenere questo formato e mezzo sono definiti attraverso Topicsets; questi possono essere estesi per supportare i formati sviluppati in futuro. Il contenuto multimediale stesso può essere incluso in un documento NewsML o collegato con un "Uniform Resource Locator" (URL)
- Permette l'aggiunta di un metadato a un articolo multimediale, per esempio altri dettagli di persone coinvolte o dettagli del luogo a cui fa riferimento la notizia. Poiché non fissa il linguaggio in cui devono essere scritti i contenuti NewsMl permette che si usino altri linguaggi per descrivere i contenuti per esempio NITF e' raccomandato per i contenuti testuali.
- Permette di controllare la versione di un articolo.
- Permette di identificare l'autore, il modificatore e il destinatario di un articolo.
- Permette di definire le relazioni tra articoli
- Permette di definire insiemi di articoli
- Permette la definizione di articoli che sono essi stessi composti da altri articoli (per esempio testo e immagine).
- Supporta versioni alternative dello stesso articolo (per esempio versioni in lingue diverse).

NITF e NewsML sono aggiornati periodicamente. Attualmente sono disponibili la versione 3.0 di NITF e la versione 1.0.1 di NewsML.

3. Modello di attività giornalistica basato su XML.

Le attività giornalistiche possono essere così suddivise

1. Creazione di un articolo fatta da un cronista (nella maggior parte dei casi per conto di una agenzia di informazione).
2. Selezione degli articoli da pubblicare (effettuata di solito, da una redazione).
3. Impaginatura degli articoli diversa per un giornale cartaceo o per uno elettronico.

Attualmente i contenuti delle notizie prodotte in una agenzia sono principalmente in formato testuale, di solito disponibile ai media attraverso dei database. Gli articoli sono creati usando un formato di descrizione delle notizie (per esempio ANPA 1312) che, tra le altre limitazioni, non ammette i contenuti multimediali.

Una redazione, dopo aver selezionato quali articoli pubblicare tra quelli resi disponibili dalle agenzie e dalle altre risorse, stabilisce l'impaginazione del giornale sia per la versione cartacea che per quella elettronica.

Il modello che noi proponiamo e' descritto nella figura 1. In questo modello l'agenzia di informazioni diverrà una agenzia di informazioni multimediali in grado di produrre notizie con contenuti in tutti i formati digitali (testo, immagine, audio, video, grafica e animazione computerizzata).

Il contenuto giornalistico sarà contrassegnato usando l'applicazione standard dell'XML e archiviato in un database XML che conterrà i contenuti di notizie passate e presenti.

I News Media saranno in grado di fare ricerche simultanee sulle basi dati di differenti agenzie, sfruttando il fato che tutti saranno in grado di usare lo stesso formato per le notizie. Il supporto di metadati offerto da NewsML permetterà ricerche molto più intelligenti ed efficienti.

Il giornale sarà creato usando le notizie in formato NewsML delle altre agenzie o preparate dalla redazione stessa. Usando differenti fogli di stile sarà possibile creare in modo automatico diverse versioni del giornale per le distinte piattaforme: cartacea per la distribuzione tradizionale, HTML [HTML] per chi accede dal Web, WML [WML 2002] per i telefoni cellulari e la televisione digitale.

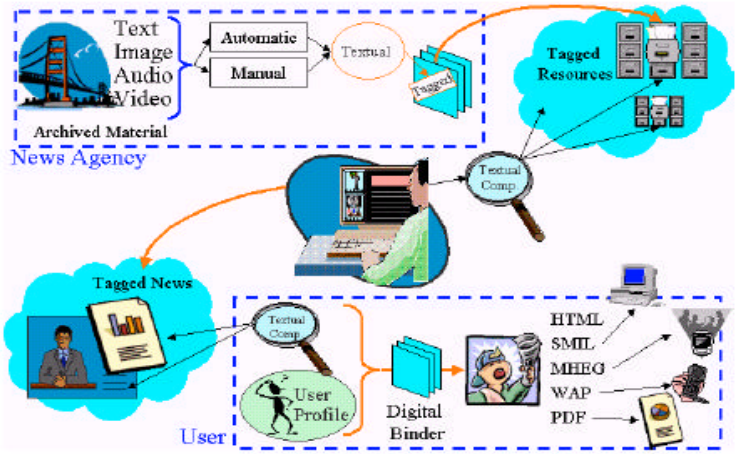

Figure 1: Model of Journalistic Activity in the Future

Come NewsML permette che un giornale sia pubblicato in formati differenti così potrebbe permettere ulteriori personalizzazioni. Il giornale potrebbe offrire all'utente finale un database di notizie giornaliere che potrebbero avere una copertura molto più ampia di quella tradizionale.
L'utente finale potrebbe registrare le sue preferenze sul server del giornale e accedendo al database delle notizie il server potrebbe scegliere quelle che più si avvicinano alle preferenze dell'utente finale.

4. Implementazione del Modello.

Il modello presentato nel paragrafo precedente è stato parzialmente implementato attraverso due progetti di ricerca finanziati dallo stato: "El Periotronico" e "Infomedia". In "El Periotronico" abbiamo lavorato all'implementazione di questo modello prima che comparisse NewsML. Dopo che NewsML è divenuto uno standard, lo abbiamo adottato come il nostro formato di descrizione delle notizie nel progetto "Infomedia", sostituendolo all'applicazione dell'XML al Linguaggio di Modellazione del Giornalismo (JML) sviluppato per il progetto "El Periotronico" come spiegato di seguito.

4.1 "El Periotronico".

Nel progetto "El Periotronico" è stato creato il prototipo di un giornale elettronico [Fernandez et al . 1999]. In questo prototipo si chiedeva all'utente finale di completare un modulo da un browser Web che offriva l'opportunità di registrare alcune preferenze per i contenuti (sezioni del giornale, parole chiave, autori) e per la presentazione (colori, font).
Il prototipo accede a un database di notizie. Usando moduli software appositamente sviluppati per il progetto sono eseguite le seguenti funzionalità:

- Un modulo software periodicamente scarica i contenuti più recenti del database su files pronti a essere inviati al web server.
- Un CGI (Common Gateway Interface) è usato per costruire il giornale personalizzato. Per fare ciò, ogni giorno sono selezionati articoli tra quelli "freschi" adatti alle preferenze dell'utente.

In questa versione prototipale gli articoli sono conservati in formato HTML.
Eravamo consapevoli dell'interesse per la descrizione delle notizie in formato XML, anche se quando il progetto "El Periotronico" ebbe inizio non esistevano né NewsML né NITF. E' stato sviluppato uno specifico DTD (Document Type Definition) in JML per trattare le notizie in formato testo. Inoltre, era stato sviluppato un semplice editor che permetteva ai giornalisti di creare gli articoli in formato JML.

4.2 "Infomedia".

Il giornale elettronico creato con il progetto "El Periotronico" è stato poi usato nel progetto "Infomedia" per implementare tre aspetti del modello: supporto multi piattaforma, personalizzazione dinamica, ed e-commerce.

Supporto multipiattaforma.

L'obiettivo di questo aspetto del lavoro era di permettere la creazione di versioni del giornale per le differenti piattaforme, partendo da un insieme di notizie in formato NewsML, per questo è necessario considerare che la generazione di versioni differenti cambia a seconda delle seguenti variabili:

- *La Presentazione.* La considerazione principale è che la quantità di contenuto che può essere presentato in simultanea potrebbe cambiare in relazione alla piattaforma da usare. Le differenti piattaforme sono state

ordinate basandosi sulla quantità di contenuto che può essere presentato su di esse. Carta, computer, televisione, cellulare. La presentazione, quindi, deve essere adattata alle caratteristiche della piattaforma. Bisogna tenere conto, sopratutto, della qualità della presentazione, per esempio bisogna evitare di lasciare parole isolate sull'ultima linea di un paragrafo.

- *La Struttura*. E' necessario organizzare le informazioni in livelli per poter presentare queste informazioni su piattaforme differenti. Ciò e' conseguenza del punto precedente in quanto per poter presentare meno informazioni sono necessari più livelli, per le piattaforme che consentono una minore quantità simultanea di informazioni è necessario ridurre la quantità totale di informazioni che il giornale contiene. E' necessaria, quindi, una struttura che renda l'uso del giornale più semplice e che permetta di riunire le notizie per soggetto e di ridurre il numero di livelli.
- *Il Contenuto*. Oltre alla struttura e alla presentazione il contenuto stesso delle notizie può cambiare secondo la piattaforma e non solo per la quantità di informazione fornita. Si sa che le caratteristiche degli utenti dei diversi canali sono distinte, per esempio il tipico utente Internet e' una persona giovane che ha qualche conoscenza del computer, mentre l'utente di servizi di televisione digitale può avere caratteristiche molto differenziate, nella maggior parte dei casi con necessità culturali differenti. Così i contenuti non devono solo soddisfare la quantità di informazione richiesta, ma anche il modo in cui l'informazione è presentata.

Il lavoro che si sta facendo è quello di disegnare un insieme di fogli di stile XSL[XSL 2001] che permettano la trasformazione del NewsML in diversi formati di presentazione. XSL è uno degli standard che il consorzio per il World Wide Web sta sviluppando attorno all'XML. Un documento XSL è esso stesso un documento XML che definisce le regole di trasformazione basate su pattern che consentano la trasformazione di un documento XML in un documento in un altro formato, per esempio HTML o WML.

Per ottenere una presentazione di qualità è necessario che il progettista grafico di un giornale guidi il processo di trasformazione. Stiamo sviluppando uno strumento che permetta di definire dei modelli in cui siano stati identificati i campi (per esempio i titoli, il testo, l'immagine etc.). Questi campi possono essere riempiti con i contenuti estratti da documenti in formato NewsML. Sarà possibile ottenere il giornale nei formati richiesti attraverso questo insieme di modelli sommati ai documenti in formato NewsML,.

I modelli che sono stati costruiti devono permettere la selezione dei paragrafi da un articolo, la scelta del mezzo (testo, video, immagine, audio, etc.), la fusione dei contenuti nei modelli e la produzione del documento finale. I modelli possono includere anche le preferenze del lettore (personalizzazione) e i contenuti possono essere personalizzati in quanto NewsML è in grado di mantenere differenti versioni dello stesso articolo.

Personalizzazione Dinamica

La personalizzazione può essere ottenuta in due modi distinti: personalizzazione dinamica e personalizzazione statica. Nel caso della personalizzazione statica, l'utente finale riempie e invia un modulo per indicare chiaramente le preferenze per i contenuti e per la presentazione. In "El Periotronico" è stata implementata questo tipo di personalizzazione.

Per costruire un profilo dell'utente, in questo caso, si raccolgono le informazioni relative al comportamento dell'utente finale. Conoscendo gli accessi effettuati in passato da un utente finale è ragionevole supporre che a quali argomenti sarà ancora interessato in futuro. Se un giorno un utente legge una parte di notizia circa un determinato argomento (per esempio un pezzo sul Mercato all'Ingrosso), é molto probabile che mostrerà lo stesso interessamento per quell'argomento il giorno successivo. L'implementazione di questo tipo di personalizzazione è uno degli scopi di "Infomedia".

Le tecniche di estrazione di testo (test mining) sono usate per implementare il modulo di personalizzazione dinamica. L'algoritmo usato è composto di tre fasi. La prima fase comprende una pre-elaborazione del testo per estrarre le parole chiave dagli articoli che l'utente finale ha visto. La seconda fase sviluppa un processo di estrazione per costruire un profilo utente usando un algoritmo generico - una funzione logica booleana -. In una terza fase, ogni giorno gli articoli che potrebbero essere interessanti per l'utente finale sono selezionati usando il profilo dell'utente.

Per estrarre le parole chiave si possono usare diverse strategie. Noi abbiamo impostato la selezione delle parole chiave sulle parole che sono presenti piu' frequentemente escluse le parole "vuote" (articolo, pronomi, etc.), assegnando un peso speciale a quelle che compaiono nei titoli degli articoli o nei sommari.

Il processo di selezione delle notizie e' basato sull'applicazione di una assegnata funzione booleana in cui ogni argomento della funzione indica se una parola chiave è presente (valore logico VERO) o no (valore logico FALSO) nell'articolo che si sta analizzando.
Se la funzione logica ha valore uguale a Vero, l'articolo e' selezionato. Per effettuare un effettivo processo di personalizzazione e' necessario avere un esteso insieme di articoli che l'utente ha già letto. Noi stimiamo che il numero di articoli necessari per ottenere risultati utili sia di circa 100 articoli, sebbene sia necessaria una ulteriore sperimentazione. Questo insieme di articoli è stato ottenuto in giorni consecutivi, con le parole chiave e le funzioni booleane aggiornate mediante l'aggiunta di articoli che sono stati letti più di recente e la rimozione di quelli più vecchi.

La personalizzazione dinamica sarà usata insieme alla personalizzazione statica già implementata nel primo prototipo.

E-commerce

E' risaputo che la maggioranza dei giornali elettronici perde denaro. Noi riteniamo che i giornali elettronici nel medio e lungo periodo non potranno più fondare i redditi esclusivamente sulla pubblicità. Stiamo sviluppando un modello in cui l'utente finale paghi sia per ogni articolo a cui accede sia per ogni accesso ad articoli specializzati.

5 Conclusioni e Aree di ulteriori Ricerche.

La tecnologia XML è stata usata con successo in molti campi in cui è necessario un formato per catalogare, immagazzinare e scambiare le informazioni.

Nel caso del giornalismo l'uso dell'XML, e gli strumenti standard che si stanno sviluppando attorno ad esso, in futuro permetteranno un accesso più efficiente e flessibile alle informazioni giornalistiche. Questo offrirà all'utente finale nuovi servizi e nuovi modi per accedere alle informazioni giornalistiche.

Abbiamo in programma di proseguire il lavoro, continuando a lavorare utilizzando i risultati ottenuti, per far si che le redazioni lavorino con l'XML.
Questo nuovo lavoro includerà aspetti quali il supporto del flusso di informazioni, il ciclo di vita dei contenuti (creazione, modifica, etc.) e ricerche efficienti.

Ringraziamenti.

Il lavoro presentato in questo articolo e' stato parzialmente finanziato dal Centro di Investigazione Scientifica e Tecnologica (CICYT) della Spagna attraverso i progetti TEL97-0788 "El Periotronico" e TEL99-0207 "Infomedia". Si ringraziano per la loro collaborazione fruttuosa Antonio Hernández-Pérez, Tomás Nogales-Flores e David Rodríguez-Mateos.

Referenze

[Fernández et al. 1999]
M. Carmen Fernández Panadero, Vicente Luque Centeno, Carlos Delgado Kloos, Andrs Marín López, Carlos García Rubio, Luis Sánchez Fernández y Antonio Hernández Pérez. Mass-Customizing
Electronic Journals. In Electronic Publishing'99. Redefining the Information Chain-New Ways and Voices, May 1999.
[HTML]
HyperText Markup Language (HTML) Home Page. World Wide Web Consortium. <http://www.w3.org/MarkUp/>.
[IPTC]
International Press Telecommunications Council (IPTC). <http://www.iptc.org>.
[IPTC 1979]
The IPTC Recommended Message Format IPTC TEC-7901 (ITPC 7901). International Press Telecommunications Council, 1979.
[NewsML 2000]
NewsML Version 1.0-Functional Specification. International Press Telecommunications Council, October 2000.
<http://www.iptc.org/site/NewsML/speci.cation/NewsMLv1.0.pdf>.
[NIFT]
News Industry Text Format (NITF). <http://www.nitf.org>.
[Wire 1979]
Wire Service Transmission Guidelines (ANPA 1312). American Newspaper Publishers Association, February 1979.
<http://www.naa.org/technology/standard/89-3msw.pdf>
[WML 2002]
Wireless Markup Language (WML) version 2 speci.cation. WAP Forum. <http://www.wapforum.org/what/technical.htm>.
[XML 2000]
Extensible Markup Language (XML) 1.0 Recommendation, October 2000.
<http://www.w3.org/TR/2000/REC-xml-20001006>.
[XSL 2001]
Extensible Stylesheet Language (XSL) 1.0 Recommendation, October 2001. <http://www.w3.org/TR/xsl/>.