Trattamento degli allegati (Proposta di rinnovamento)

Introduzione

Il trattamento degli allegati da parte del server eXtraWay ha due aspetti.
Il primo consiste nella loro gestione: collocazione su *file system*, spostamento da una posizione ad un'altra, rimozione, destinazione a *storage area* di diverso livello e così via.
Il secondo consiste nel compierne l'opportuno utilizzo in fase di indicizzazione dei contenuti e/o nel riconoscere gli allegati effettivi da quelli derivati[1)].

Quanto detto si è sino ad ora attuato avvalendosi di un particolare *name space* e di elementi ed attributi riferiti a tale namespace aventi caratteristiche fisse su cui il server potesse fare affidamento. Lo stesso risultato si può ottenere avvalendosi di *Processing Instruction* che assolvono allo stesso compito ed hanno il vantaggio di non avere effetti diretti o indiretti sull'eventuale DTD che regola il contenuto di un'Unità Informativa XML.
In ultimo, era stata introdotta tempo fa la possibilità di indicare gli allegati in elementi/attributi liberi limitandosi a dichiarare quali essi fossero[2)]. Questo sistema ha però una carenza: impedisce di definire che tipo di derivazioni compiere dall'allegato in esame e come trattare il risultato.

Proposta di revisione della gestione allegati e loro conversioni

Per completare lo scenario descritto si suggerisce l'implementazione di quanto segue:

- Gli allegati saranno comunque identificati in una delle modalità sino ad ora supportate: con un percorso assoluto o relativo ovvero tramite un identificativo logico (un *ID* rappresentato da un numero seguito da un'estensione).
- La trasformazione che si desidera effettuare per questi allegati deve sempre fare riferimento all'identificatore di cui al punto precedente ed altrettanto varrà per l'esito della conversione, così da costituire un legame che possa essere mantenuto e gestito nel tempo (ad esempio rimuovendo gli allegati derivati di un allegato all'atto dalla sua rimozione).
- Per ottenere la conversione di un file è necessario esplicitare che tipo di conversione si desidera ottenere e che tipo di trattamento il server deve compiere ad operazione completata. In pratica, l'unico dato realmente utile al server è sapere se l'esito della conversione sia da indicizzare o meno.
- A conversione avvenuta, la richiesta di conversione va rimossa ed il suo posto va altresì preso dall'esito della stessa, a sua volta rappresentato da una *Processing Instruction*

Condizione necessaria e sufficiente perché il server eXtraWay abbia un comportamento corretto nei confronti di quanto presente nella *Processing Instruction* è che essa sia conformata in modo equivalente ad un elemento XML auto-chiuso caratterizzato da uno o più attributi. Questa caratteristica vale sia per la *P.I.* di richiesta conversione che per quella che identifica la conversione già avvenuta.

Il procedimento avrà quindi il seguente comportamento:

- L'applicazione appone nel record XML, nella posizione ritenuta più consona ai propri scopi, una *P.I.* conformata come indicato avente come label la stringa 'xw-toderive'.
  In tale *P.I.* devono essere presenti uno o pià *pseudo-attributi*. Essi sono:
  - name: al pari dell'attributo `name` dell'elemento `xw:file` e del corrispondente pseudo-attributo della *P.I.* `xw-file`, esso identifica l'allegato. **Obbligatorio**;
  - index: anche in questo caso il senso di questo pseudo-attributo è analogo a quello degli elementi descritti. Se presente e con valore pari a 'yes', indica al server che l'esito della conversione dev'essere sottoposto ad indicizzazione. Facoltativo;
  - Uno o più ulteriori pseudo-attributi necessari al sistema di conversione per sapere come operare. Tra essi possono apparire (valori suggeriti)
    - content_type (Questo potremmo quasi considerarlo obbligatorio)
    - label
    - …
- Il server riconosce la richiesta di conversione e passa l'intero contenuto della medesima ad un processo “near on line” del server stesso che provvederà a richiedere ad un servizio opportuno la conversione richiesta passando ad esso tutti i parametri trovati nella *P.I.*;
- Ottenuto il risultato della conversione, il server modifica il documento come segue:
  - Identifica la *P.I.* `xw-toderive` cui corrisponde la conversione effettuata.
  - Rimuove la *P.I.* appena identificata e la sostituisce[3)] con un'equivalente *P.I.* che conterrà:
    - name: identificatore del nuovo allegato derivato da quello orignario. **Obbligatorio**;
    - der_from: identificatore dell'allegato originario. **Obbligatorio**;
    - index: ereditato dalla *P.I.* `xw-toderive`. Indicherà se compiere o meno l'indicizzazione. Facoltativo;
    - Tutti gli altri pseudo-attributi sottoposti alla richiesta di conversione. In questo modo si saprà sempre come questo file è stato derivato (con quali parametri etc. etc.) e si avrà ad esempio cognizione del sul `content_type` e così via.

Non sono stati imposti limiti o vincoli sugli pseudo-attributi proprio per garantire il massimo grado di libertà nella loro definizione. Su suppone però che una parte di essi sia necessaria al processo di conversione per sapere in quali modalità operare ed una parte

avrà invece scopi applicativi tesi ad identificare il ruolo di ogni allegato derivato, il perché esso sia stato elaborato e quindi come farne uso applicativamente.

Visto che al momento in cui si scrive il servizio di conversione che possa operare secondo le modalità indicate ancora non esiste, i dettagli sugli pseudo-attributi necessari alla conversione verranno concordati in seguito, se pure è ragionevole ritenere che il `content_type` (o altro equivalente) dovrà necessariamente esistere.

Rispetto ai sistemi adottati sino ad ora, il legame tra *derivato* e *derivante* è mono direzionale (nel derivato si conosce l'*ID* del derivante) anziché essere bidirezionale. Le restanti informazioni sono disponibili esattamente come lo erano in passato, in quanto il fatto che una derivazione sia stata compiuta si evidenzia con la presenza del risultato (e non più della sola richiesta).

In caso di fallimento della conversione (ad esempio perché il servizio di conversione non è in grado di ottenere un file di un certo tipo partendo dal tipo di allegato originario) la *P.I.* `xw-toderive` rimane del tutto invariata (anche se il comando di conversione non viene reiterato). In questo modo, ad esempio all'atto dell'aggiornamento del servizio di conversione, si può richiedere di tentare ex-novo tutte le conversioni sino a quel momento fallite[4].

Esempio

In un documento esiste da qualche parte un allegato 1.doc.
Volendo ottenere la conversione in PDF e l'estrazione del testo per fare gli indici si indicherà:

```
<?xw-toderive name="1.doc" content_type="application/pdf"?>
<?xw-toderive name="1.doc" content_type="text/plain" index="yes"?>
```

Ad operazione completata, il server sostituisce le *P.I.* come segue:

```
<?xw-file name="2.pdf" der_from="1.doc" content_type="application/pdf"?>
<?xw-file name="3.txt" der_from="1.doc" content_type="text/plain" index="yes"?>
```

Proposta di revisione dell'Agent di Conversione/Trasforamzione Files

Alla luce della *Conference Call* tenutasi tra Roberto, Massimo, Massimiliano e Carmine, si esprime di seguito la proposta di sviluppo per sostituire l'attuale sistema di conversione files.

Premessa Attualmente avviene quel che segue:

1. Il server che salva un record valuta se in esso ci siano conversioni richieste e crea in una directory di scambio un file che viene valutato da *File Conversion Agent* (FCA);
2. L'agent FCA legge questi files, richiede al *File Conversion Service* (FCS) la conversione del caso e sull'esito fa:
    1. Crea il nuovo allegato presso il server;
    2. Carica tutti gli allegati, compreso quello appena introdotto, per il ricalcolo dell'impronta. Qui c'è una nota importante da dire. L'impronta non dovrebbe essere calcolata sugli allegati *derivati* ma solo sugli originali quindi quest'attività è sbaglita. Inoltre il server è già in grado di calcolare l'impronta di un insieme di allegati (lo fa per l'importazione da Watch Doc con allegati) e quindi sarebbe il caso di farla calcolare al server, specie se l'impronta si riferisce a tutti gli allegati. So che il nuovo documentale fa l'impronta file per file, ma se serve quella cumulativa mi pare che abbia senso.
    3. Modifica il documento aggiungendo gli allegati derivati e cambiando quindi gli attributi degli allegati originali.
3. Rimuove il file del lavoro svolto.

Questo avviene senza tener conto di altre esigenze, quali ad esempio l'opportunità di compiere un'attività modificante in un momento ben preciso.

File Transformation Service

E' un *Web Service* che dispone di funzinoalità per la trasformazione di oggetti multimediali. Originariamente concepito per convertire le immagini da un formato ad un altro, è stato esteso con librerie dinamiche per l'esecuzinoe di OCR.
Realizzato con una tecnologia che lo rende particolarmente versatile e flessibile, ha come vantaggio “derivato” che le librerie che utilizza per le conversioni possono essere usate direttamente da eXtraWay Server per attività analoghe.

Proposta Quello che is proppne è quanto segue:

1. Implementare un servizio (che potrebbe essere anche lo stesso FTS) al quale il server master richiede che vengano compiute le conversioni cin modalità analoghe a quelle attuali;
2. Il servizio tenta di compiere la conversione facendo uso delle librerie di FTS in suo possesso. Se la conversione non ha luogo in quanto nessuna delle librerie è in grado di farla, evoca FCS perché compia la conversione. Questa soluzione ci consente di sostituire gradualmetne FCS con uno strumento più potente ma ci da anche la possibilità di sfruttare un doppio fronte in materia di conversioni fruendo di codice di terze parti sia in linguaggio C che in Java.
3. Se la conversione fallisce compie le registrazioni del caso in un apposito log e mette via la conversione scartata, se invece va a buon fine non agisce direttametne sul server ma colloca il risultato del suo operato in un'ulteriore cartella di scambio dal quale il server eXtraWay provvederà ad attingere.
4. Il server eXtraWay avrà un *Thread* con il compito di interrogare la directory di scambio, acquisire il materiale e provvedere alla modifica del documento (secondo le ipotesi espresse sopra) scegliendo il momoento più adatto per farlo, vale a dire un

momento in cui l'attività degli utenti sia ridotto ai minimi termini.

In proposito all'ultimo punto vale la pena di prendere in considerazione anche il fatto che il server potrebbe modificare i documenti subito per aggiungervi gli allegati derivati che non comportano indicizzazione. Ad esempio potrebbe introdurre subito i *PDF* derivati dall'originale.
Per contro, potrebbe attendere il moment più propizio per agguingere i *TXT* per i quali si impone un indicizzazione spesso gravosa che andrebbe fatta in un momento di “scarico”.

1) Frutto di trasformazione degli allegati originari per mezzo di un servizio di conversione

2) Via XPath

3) La sostituzione avviene sul posto, se possibile o nelle più immediate prossimità, come “ultimo figlio dell'elemento parent”.

4) Il comando esiste già per gli `xw:file` e va solo esteso