Introduzione a Watch Doc

Introduzione

La libraria Watch Doc ha lo scopo di dare la possibilità di alimentare una base dati sottoponendole dei files XML contenenti unità informative. Ci si attende che queste unità informative siano riconoscibili per mezzo di un *element* e che l'archivio sia stato disegnato per ospitare simili tipi di documento. Dati questi presupposti che non si ritiene necessario approfondire in questa sede, vediamo cosa comporta l'uso di questo strumento e che vantaggi comporta.

Sottoporre alla libreria Watch Doc un file XML consiste nel produrne una copia consumabile in una delle directory che vengono monitare dalla liberia. La libreria, infatti, sulla base del contenuto di un file di configurazione, prende visione del contenuto di una o più directory. Se all'interno di una di esse viene rilevato un file con estensione .xml su di esso viene iniziata una lavorazione tesa alla scoperta delle unità informative in esso conenute ed ognuna di esse viene salvata singolarmente nell'archivio associato al file XML in esame. Perché questo avvenga, come detto precedentemente, è necessario che la libreria sia opportunamente configurata ma è anche necessario che il nome del file sia parlante, cosi da poter indentificare l'archivio di destinazione e l'operatore che ha sottoposto il file cosi che esso venga preso in esame nella fase di salvataggio.

Al termine della lavorazione del file esso viene rimosso dalla directory che lo ospitava. Se non tutte l'elaborazione è andata a buon fine, rimarrà in sua vece, nello stesso posto, un file con estensione .xml.fail

Trattamento del nome del file

In primo luogo bisogna fare la massima attenzione alla fase di copia del file nella directory predestinata. La copia, infatti, non deve ASSOLUTAMENTE avvenire per files aventi estensione .xml.
La condizione prevista è che il file non abbia una simile estensione o che esso venga rinominato subito prima della copia e, solo al termine della copia, modificato nel proprio nome (o nella sola estensione) perché possa essere riconosciuto come file .xml. Se la copia avviene con il file avente già tale estensione, la LIbreria può tentare di iniziarne la lavorazione quando esso è ancora presente in modo parziale portando a potenziali errori.

Quello che si suggerisce, ad esempio, è di cambiare l'estensione standard in .xxx, effettuare la copia e poi modificare nuovamente l'estensione in .xml

Veniamo ora al nome vero e proprio del file. Il file deve avere una denominazione riconducibile alla seguente regola:

```
nomearchivio<separatore>componente variabile.estensione
```

Nota: Il nome del file deve essere composto come indicato. A seconda della vetustà della libreria e del server utilizzati, se il nome non prevede il separatore e la componente variabile, la procedura potrebbe non avviarsi. Prevedere un file il cui nome sia composto esclusivamente dal *nomearchivio.estensione* potrebbe non rappresentare un nome valido. Ciò impatta anche nella scelta del separatore, profilabile a partire dalla versione 19.5.11.* del server eXtraWay.

Il *nomearchivio* deve rappresentare l'identificatore logico dell'archivio, per intendersi lo stesso identificatore che si deve dichiarare nel file di configurazione del server per dichiarare il legame tra tale identificatore e l'archivio fisicamente presente su disco fisso

Il separatore è rappresentato dal carattere '-' o dal carattere '_' salvo diversa configurazione. La possibilità di introdurre una diversa configurazione si ha dalla versione 19.5.11.* del server eXtraWay. La possibilità che non ci sia parte variabile e quindi che il file sia composto solo dal nome archivio e dall'estensione è stata introdotta con una versione precedente e quindi certamente disponibile da quella indicata.

Indipendentemente dall'uso del carattere '_' o '-' come separatore tra il nome logico dell'archivio e la parte variabile, anche il carattere '.' ricopre un ruolo particolare. Per non avere impatti con il comportamento di Watch Doc, il server inibisce la realizzazione di nomi logici d'archivio contenenti i caratteri '_' e '-', come precedentemente detto, salvo diverse indicazioni che coinvolgono il file di configurazine xwwd.conf.xml. Oltre ad essi, però, anche il carattere '.' non può essere utilizzato nei nomi logici d'archivio, pena un errato comporamento su alcune funzionalità. Esso verrà presto inibito.

La componente variabile è a scelta dell'operatore e solitamente serve a dare un ordine di elaborazione dei files. La Libraria, infatti, procede secondo l'ordine alfabetico dei files rilevati. In questa componente è altresì possibile indicare un nome utente da associare all'operazione di inserimento o modifica che deriva dall'operato della libreria. Perché ciò avvenga in questa parte del nome del file dev'essere chiaramente identificabile una sequenza

usr=

seguita dal nome dell'utente che si vuole dichiarare (tipicamente l'operatore che sottopone il file ad eXtraWay). Il nome dell'utente deve a sua volta terminare perché si incontra un punto (che isola l'estensione) o uno dei suddetti separatori. In questo modo si possono indicare files la cui composizione risulta essere

```
nome archivio + separatore + operatore + progressivo per ordinamento.estensione
nome archivio + separatore + progressivo per ordinamento + operatore.estensione
```

a seconda che si intenda privilegiare l'ordine per operatore e poi per l'ordine dato o vice versa.

Nota: Si fa presente che i caratteri separatori di default sono l'underscore '_', il trattino o segno meno '-' ed ovviamente il punto '.'. E' altresì possibile configurare nel file xwwd.conf.xml quali separatori possono essere utilizzati in alternativa a questi che sono il default[1)].

Modificazioni del nome dei files e loro significato

Quando la Libreria decide di prendere in carico un file aggiunge in coda al suo nome l'estensione .wrk per indicare che è in lavorazione (working). Dopo aver aggiunto quest'estensione supplementare, la Libreria porta alla creazione di un ulteriore file avente lo stesso nome ma con estensione .xml.fail anzichè .xml.wrk. Questo nuovo file viene creato preventivamente e conterrà tutti i frammenti XML che la procedura di interpretazione non fosse stata in grado di riconoscere o comunque di salvare nell'archivio come nuove unità informative o come modifiche di unità esistenti.

Al termine delle operazioni di acquisizione, prima di passare al file successivo la procedura deve rimuovere entrambe i files. posson quindi verificarsi i seguenti casi:

- Sono spariti tutti e due i files. L'operazione ha avuto luogo correttamente, si può passare ad elaborare un nuovo file.
- E' presente il solo file .xml.fail. Ne consegue che la lavorazione ha incontrato condizioni che hanno impedito di salvare alcune unità informative. Si richiede quindi di effettuare un intervento correttivo sul contenuto del file, anche sulla base dei files di log stilati dalla procedura, per comprendere quali siano stati gli impedimenti al salvataggio. Una volta apportate le modifiche al file, esso può essere rinominato in .xml e sottoposto nuovamente alla procedura di acquisizione.
- E' presente il solo file .xml.wrk. Questo comporta che la procedura è stata predisposta ma non è mai realmente partita. Potrebbe dipendere, ad esempio, da una carenza di licenze o da altre ragioni rilevabili nei logs.
- Sono presenti entrambe i files. Ciò comporta che il server stia normalmente elaborando tale file ma se così non fosse è presumibile che si sia verificato un crash di eXtraWay.

Alla luce di quanto detto, le condizioni normali sono quelle in cui tutti e due i files vengono correttamente rimossi ovvero quando sopravvive solo ed esclusivamente il file .xml.fail. In esso sono quindi registrati i frammenti XML malformati o non riconoscibili ovvero quelli che violano l'univocità su archivi dove non è prevista la sovrascrittura delle unità informative o che violano una molteplice univocità. Sul contenuto di questi files è quindi necessario intervenire basandosi su quanto rilevato nel file di log wd_journal.log la cui dislocazione ricalca quella di tutti i logs di eXtraWay.

Un ultima nota sul trattamento dei nomi dei files riguarda la possibilità di forzare il salvataggio di unità informative che violano l'univocità. Per ottenere questo risultato si ipotizza di avere, dopo una fase di elaborazione un file avente estensione .xml.fail. A tale file si deve accodare una nuova porzione di estensione perché venga nuovamente riconosciuto come un file XML. La nuova estensione completa sarà quindi .xml.fail.xml. Questo notifica alla Libreria l'intenzione di salvare le unità informative determinate senza curarsi della violazione di univocità (semplice o molteplice) ignorando quindi sia l'impostazione nel file xwwd.conf.xml sia l'impostazione di univocità dell'archivio. Al termine della nuova elaborazione il file deve sparire e con esso il suo ulteriore file dei fallimenti. Se per contro sopravvive all'elaborazione un file avente l'estensione composta .xml.fail.xml.fail esso contiene i frammenti XML che non è stato possibile salvare come unità informative indipendentemente dalle univocità. Salvo diversa indicazione tali frammenti dovrebbero quindi essere malformati e quindi irriconoscibili (o più semplicemente non accettabili).

Interruzione flusso acquisizione

Per interrompere un flusso di acquisizione su un file, dalla versione 21.1.0.* del server, è possibile creare un file con lo stesso identico nome di quello in corso di acquisizione (riconoscibile per l'estensione .xml.wrk) con estensione .xml.wrk.stop. In presenza di un simile file il server interrompe il ciclo lasciando i documenti non acquisiti nel file .xml.fail.

Configurazione della Libreria

La configurazione, come accennato in precedenza, prevede innanzitutto di indicare una o più directory nelle quali la Libreria verificherà la presenza di files .xml. Questa è la sola configurazione realmente indispensabile senza la quale la procedura non potrà mai avere luogo. La libreria, infatti, viene caricata dal server eXtraWay alla sua partenza, in pratica dalla copia Master, ma se la configurazione è incompleta o inesatta la Libreria non può operare, scrive nei logs che l'operazione non avrà luogo, e viene scaricata dal server. Modificare la configurazione dopo la partenza del server non ha effetto e si richiede quindi di fare stop + start del server eXtraWay perché la configurazione abbia efficacia.

Oltre a quest'impostazione il file di configurazione prevede l'indicazione del tempo, in millisecondi, che deve trascorrere tra un test del contenuto delle directory da monitorare ed il successivo.

In fine è possibile dare alcune indicazioni ulteriori sull'archivio. I particolare si fa riferimento all'opportunità di sovrascrivere le unità informative esistenti con quelle sottoposte al server se in esse si rilevano violazioni di univocità. Di fatto se l'archivio è configurato per prevedere regole di univocità esse vengono verificate e l'unità informativa che ci si accinge a salvare dovrebbe naturalmente sovrascrivere quella esistente. Se alla verifica risultano esistenti più unità informative aventi le stesse caratteristiche di univocità la sovrascrittura non ha luogo ed frammento XML che rappresenta questa unità va a gonfiare le fila del file .xml.fail. Procedere alla sovrascrittura è il comportamento di default. Altrettanto si verifica quando la configurazione non preveda la sovrascrittura.

Vediamo quindi un campione del file di configurazione.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE xwwd_cfg SYSTEM "http://www.3di.it/extraway/xwwd_20030403.dtd">
<xwwd_cfg>
   <global ms_timeout="500" ms_maxtimeout="3000" arcname_separs="-"/>
   <watch dir="../wd" />
   <arc name="acque" update="off"/>
   <arc name="prova" test="wf"/>
```

```
   <arc name="testcbl" remove="yes"/> <!-- Per rimozione documenti via W.D., vedi apposito
paragrafo -->
</xwwd_cfg>
```

Nell'esempio indicato si richiede che il test sulla directory `../wd` venga effettuato ogni 500 millisecondi (quindi ogni mezzo secondo) ma che quando ci sono files in fase di acquisizione l'attesa possa crescere sino ad un massimo di 3 secondi.

Nota: Qualora questi due valori non vengano espressi essi vengono considerati pari a 15 secondi[2)]

Si stabilisce poi che il separatore dei nomi di archivi dalla restante parte del nome del file sia solo il trattino (ed ovviamente il punto che viene considerato d'ufficio). In questo modo un archivio può chiamarsi xdocwaydoc_per senza che il server tronchi il nome prima del suffisso per. Inoltre si indica che per l'archivio acque la sovrascrittura è inibita mentre per l'archivio prova si richiede che i test di sui documenti acquisiti non si compia il test di validità rispetto ad una DTD ma solo il test di Well Formedness.
I valori che possono essere assunti da questi attributi sono:

- update:[on|off]. Il valore on (default) indica che in caso di violazione di univocità il documento viene sovrascritto, off comporta il fallimento dell'inserimento, il documento rimane nel file .fail.
- test:[all|wf|off]: Il valore all (defauilt) indica che sui documenti che si intende inserire/aggiornare verrà compiuto un test di Well Formedness ed un test di rispetto di una DTD (che ovviamente viene eseguito se e solo se per l'archivio in esame esiste una DTD dichiarata, condizione non obboligatoria), il valore wf indica di compiere il solo test di Well Formedness ignorando ogni validità rispetto ad una DTD ed il valore off indica di non compiere alcun test sul contenuto dei documeni che si procede ad acquisire. Quest'ultima modalità è ampiamente deprecata e resa disponibile solo per soddisfare la necessità di compiere operazioni su dati già noti e certi con performance migliorate dall'assenza di controlli superflui.

Nota: La directory può essere espressa tanto con un percorso completo quanto con un percorso relativo alla directory degli eseguibili.

Estensione ai files Csv.

La procedura realizzata in Watch Doc per il trattamento dei files XML da utilizzarsi come strumenti di input per l'importazione/aggiornamento di documenti in una base dati stata estesa anche ai files in formato Comma Separated Value. Per tali files non si pu e non si deve parlare di importazione nel senso stretto del termine bens di predisposizione del contenuto del file csv per la successiva importazione in un archivio eXtraWay.

Attenzione:

```
  Visto che l'encoding dei files csv non  definito a priori si assume che esso sia WinLatin1.
```

Vediamo come opera la procedura. Il meccanismo sostanzialmente simile, se non identico, a quello descritto per i files xml. Il server compie monitoring delle stesse directory usate per i files xml alla ricerca di files csv e procede alla loro elaborazione. L'elaborazione ha inizio rinominando il file in csv.wrk ed avviando un'istanza del server che produrr , a partire da tale file, la forma xml dei dati contenuti nel file csv.

Per effettuare quanto detto sono necessarie queste condizioni. Il rispetto di esse condiziona direttamente il risultato ottenuto:

- La prima riga del file csv deve contenere delle etichette, ovvero diciture che consentano l'identificazione delle colonne rappresentate nel file. Tali etichette possono/devono corrispondere ai search alias impostati per l'archivio su cui si intende poi compiere l'importazione del file xml ottenuto.
- Le diverse colonne devono esere separate dal carattere ',' ovvero dal carattere ';'. Quale dei due venga utilizzato verr automaticamente stabilito analizzando la prima riga del file csv da elaborare.
- Il nome del file csv deve seguire le regole gi precedentemente indicate per i files xml. Tramite il nome del file deve quindi essere possibile identificare l'archivio con il quale si cerca di compiere un legame formale.

Se queste condizioni sono valide, vediamo come si comporta di conseguenza il server:

- Come detto, in primo luogo compie un'analisi della prima riga del file. Da essa vengono estatti tutti i search alias da confrontare con il file di configurazione d'archivio. Per le etichette che non dovessero dimostrare corrispondenza con un search alias il server assumer che il nome impostato corrisponde al nome di un elemento nella radice del frammento xml che verr generato.
- Per tutti i search alias identificati si verificher che essi appartengano tutti alla stessa primary node. Se questa condizione non verificata la procedura da errore e si interrompe.
- Se nessuna delle etichette corrisponde ad un search alias la procedura compie ugualmente la conversione e nel farlo assumer che la primary node si chiami undefined ed il suo container si chiami l_undefined. Ogni etichetta porter all'identificazione di un elemento figlio della primary node avente il nome espresso nell'etichetta stessa.

Nota:

```
  Visto l'uso che si fa delle etichette si suggerisce di evitare, nella loro stesura di usare
particolari caratteri di interpunzione o spazi, dal momento che la stessa pu  condurre al nome di
un elemento se non riconosciuta diversamente. Si ricorda inoltre che un etichetta nel formato
tabella.campo viene appunto interpretata in tal modo quindi se ad essa non si riesce ad associare
```

```
un valido search alias, la prima parte viene ignorata e solo la seconda, dopo il punto, viene
usata per identificare il nome dell'elemento che si andr@ a generare.
```

Avviata la procedura, il server produce, temporaneamente, un file csv.fail nel quale riporta la prima riga del file csv originario e dove accoder@ ogni altra riga non riconosciuta, ed un file csv.xxx nel quale verr@ stilato l'xml prodotto.

Nota:

```
  Si noti che il file prodotto, visto che viene generato nella stessa directory ove Watch Doc
cerca gli xml da importare, ha un'estensione fittizia xxx, e non xml, per impedire che il file
venga acquisito direttamente senza cotrollo da parte del richiedente.
```

Al termine della lavorazione si possono quindi presentare diversi scenari, esattamente come nel caso di Watch Doc nella lavorazione dei files xml.

- Il file csv.wrk @ sparito ed al suo posto esiste un file csv.xxx. L'operazione ha avuto luogo correttamente e non ha rilevato alcuna incongruenza, si pu@ pasare ad elaborare un nuovo file.
- Sono presenti i soli files csv.xxx ed il file csv.fail. Ne consegue che la lavorazione ha avuto luogo ma che ha incontrato almeno una condizione inesatta e queste sono state riportate nel file csv.fail mentre tutte le righe valide sono state convertite nel file csv.xxx. Solitamente questo comportamento si ha quando da una o pi@ righe del file originario non si riescono ad identificare tutte le colonne previste o se he hanno in maggior numero di quelle dichiarate con le etichiette della prima riga.
- E' presente il solo file csv.fail. Ne consegue che la lavorazione ha incontrato condizioni che hanno impedito di elaborare globalmente il file csv.wrk ed esso @ stato rifiutato in toto. Condizione tipo, ad esempio, @ che le etichette che identificano le colonne portino a riconoscere canali appartenenti ad unit@ informative diverse. Si richiede quindi di effettuare un intervento correttivo sul contenuto del file csv originario, anche sulla base dei files di log stilati dalla procedura, per comprendere quali siano stati gli impedimenti al salvataggio. Una volta apportate le modifiche al file, esso pu@ essere rinominato in csv e sottoposto nuovamente alla procedura di acquisizione.
- E' presente il solo file csv.wrk. Questo comporta che la procedura @ stata predisposta ma non @ mai realmente partita. Potrebbe dipendere, ad esempio, da una carenza di licenze o da altre ragioni rilevabili nei logs.
- Sono presenti tutti e 3 i files (csv.wrk, csv.fail e csv.xxx). Ci@ indica che il server st@ elaborando tale file in questo momento (in tal caso il file csv.xxx o il file csv.fail dovrebbero gradualmente crescere di dimensioni) ma se cos@ non fosse @ presumibile che si sia verificato un crash di eXtraWay e che l'elaborazione risulti quindi interrotta.

A partire da:

```
  Versione 2.4.0.8 o superiore e versione di eXtraWay Server 20.2.0.16 o superiore.
```

Uso di WatchDoc per la rimozione di documenti

In tempi recenti @ sorta la necessit@ di utilizzare automatismi per compiere la cancellazione di uno o pi@ documenti. Affiancata alla cancellazione documenti da selezione, esiste ora la cancellazione documenti da procedura WatchDoc.

Trattandosi di una procedura piuttosto pericolosa essa @ regolata da una serie di accorgimenti di sicurezza.

Attualmente non esiste un distinguo tra la directory che ospita i files positivi, ovvero quei files destinati all'acquisizione per inserire o modificare documenti, e quelli negativi, ovvero i files per le rimozioni dei documenti. Sono per@ pesenti altri accorgimenti e non si esclude, in futuro, di differenziare i due punti di alimentazione.

Veniamo alle caratteristiche necessarie e sufficienti per rimuovere con questo meccanismo dei documenti da un archivio.

- In primo luogo @ necessario compiere una specifica configurazione nel file xwwd.conf.xml. In esso, in corrispondenza dell'elemento arc che indica un archivio, va aggiunto un attributo remove, con valore yes, che indica che su quell'archivio @ consentita la rimozione documenti. In caso contrario il server non esegue la rimozione dandone esplicita segnalazione nei logs.
- In secondo luogo il file deve avere una specifica estensione. Se @ vero che si devono seguire gli stessi accorgimenti nell'assegnazione delle estensioni ai files, provvedendo a rinominare opportunamente il file solo dopo averne completato la copia nella directory, il file concepito per la rimozione deve avere estensione .rmxml.
- Il file deve avere una seconda sicurezza. Per essere assolutamente certi che si intenda compiere una cancellazione esso deve presentare, in un qualsiasi punto in prossimit@ dell'inizio del file stesso, la processing instruction <?wdrm?> in assenza della quale i precedenti requisiti risulteranno insufficienti.
- I documenti presenti nel file possono essere assolutamente incompleti e neppure rispettosi della DTD d'archivio. Quello che si richeide @ che l'archivio sia configurato con una regola di univocit@ precisa ed affidabile per le unit@ informative coinvolte e che il contenuto del file (rm)XML che si sottopone a WatchDoc consenta di acquisire gli estremi necessari ad identificare la violazione di univocit@ del documento da rimuovere.

Se tutte queste condizioni sono verificate il server rimuove i documenti identificati indicando quanti invece non sono stati rimossi perch@ non violano alcuna univocit@ o ne violano pi@ d'una. Va da se che condizione necessaria e sufficiente perch@ la procedura operi correttamente @ che l'archivio sia regolarmente indicizzato. In caso di fallimento viene generato un file che contiene in modo integrale o relativo i soli documenti che non sono stati riconosciuti. La creazione di questo file segue le metodiche gi@ note di

WatchDoc con l'apposizione di una ulteriore estensione .fail.

A partire da:

```
Versione 2.5.0.*. Richiede una versione 20.4.0.* del server eXtraWay.
```

[1]

A partire dalla versione 19.5.11.*

[2]

L'impostazione del tempo massimo è disponibile a partire dalla versione 2.5.1.*