Guida rapida all'installazione del software TheMatrix in una ASL

Rosa Gini, Emanuele Carlini
Settembre, 2014

Release 1.45 e superiore

TheMatrix è un software che esegue i comandi contenuti in appositi file di testo, chiamati *script*, per trasformare le banche dati sanitarie di un'azienda ASL o regione e produrre da essi dei nuovi dataset. Il software è in corso di sviluppo nell'ambito del progetto MATRICE, ma versioni prototipo sono già disponibili e funzionanti e possono essere utilizzate. L'utilizzo del software è finalizzato a facilitare la procedura di produzione di dati derivati omogenei in ASL o regioni diverse, per obbiettivi interni o per la partecipazione a progetti in rete con altre ASL o regioni.

1. Premessa: formato IAD

Per eseguire le proprie trasformazioni il software ha bisogno di accedere ai dati in un formato standard. Il formato che è stato elaborato si chiama IAD (Italian Administrative Database) e ricalca i tracciati record contenuti negli allegati tecnici alla normativa nazionale che ha istituito i diversi flussi informativi. Il formato contiene ad oggi otto tabelle

PERSON - l'anagrafe
EXE - le esenzioni
HOSP - le schede di dimissione ospedaliera
DRUGS - la farmaceutica territoriale
DDRUG - i farmaci a erogazione diretta
OUTPAT - le prestazioni ambulatoriali
HOME - l'assistenza domiciliare
RESIDENT - l'assistenza residenziale

il cui tracciato record è descritto nei file disponibili a questo link

https://zimbra.ars.toscana.it:443/home/rosa.gini@ars.toscana.it/Briefcase/MATRICE/documentazione_IAD_per_aziende.zip

Ogni script per funzionare deve accedere a una o più di queste fonti di dati. Una procedura semplice per mettere a disposizione di TheMatrix i dati viene descritta nella sezione 2.2 di questo documento. Nel manuale è descritta una procedura più complessa, che in questo documento non viene presa in considerazion

2. Procedura di installazione

Questo documento propone una procedura rapida di installazione di TheMatrix su un computer con sistema operativo Windows. Tuttavia è possibile installare ed eseguire TheMatrix anche su sistemi operativi OS X e Linux (fare riferimento alla Sezione 4: Note per utenti Linux e OSX).

La procedura è divisa in tre passi

1) Installazione del software ed esecuzione di test su dati falsi
2) Scaricamento dei dati veri della ASL
3) Lancio di script di qualità su dati veri

2.1. Installazione del software ed esecuzione di test su dati falsi

L'archivio di installazione si ottiene ad oggi da Agenas. Una volta ottenuto l'archivio di installazione la procedura è la seguente

a) Scompattare l'archivio nella directory C
b) Rinominare la directory 'TheMatrix'
c) Scaricare i file reperibili a questo link

https://zimbra.ars.toscana.it/home/rosa.gini@ars.toscana.it/Briefcase/MATRICE/script_installazione

nella cartella C:\TheMatrix\scripts. I file possono essere aperti con un qualsiasi editor di testo ed esaminati. Si tratta di alcuni script di test per verificare la correttezza e qualità dei dati, più uno script che genera un dataset più complesso. Ogni file contiene sequenze di comandi nel linguaggio di scripting di TheMatrix. Il linguaggio di scripting è documentato nella distribuzione di TheMatrix.

d) Scaricare i file reperibili a questo link

https://zimbra.ars.toscana.it/home/rosa.gini@ars.toscana.it/Briefcase/MATRICE/archivi_falsi_IAD

nella cartella C:\TheMatrix\iad. Si tratta di dataset di dati falsi, che però hanno il formato IAD. E' come se i dati fossero quelli di una finta ASL, con circa 5mila assistiti. I file possono essere aperti con un editor di testo ed esaminati.

e) Aprire una shell

f) Lanciare la seguente sequenza di comandi. Questi comandi leggono ciascuno il file csv corrispondente e generano un file xml che rende visibile a TheMatrix il dataset.

cd C:\TheMatrix
MetaXMLCreatorWIN PERSON.csv
MetaXMLCreatorWIN HOSP.csv
MetaXMLCreatorWIN EXE.csv
MetaXMLCreatorWIN OUTPAT.csv
MetaXMLCreatorWIN DDRUG.csv
MetaXMLCreatorWIN DRUGS.csv

g) Lanciare il seguente comando (Questo è uno script di prova che genera il dataset dei 4760 soggetti che al 1 gennaio 2011 risultano assistiti dalla finta ASL, marca chi di essi risulta diabetico da farmaci, esenzioni o ricoveri precedenti, e indica chi ha svolto almeno un test dell'emoglobina glicata nel corso del 2011)

TheMatrixWIN --logLevel 0 scriptTest.txt

h) Dopo qualche minuto il programma termina l'esecuzione. Il file

C:\TheMatrix\iad\RisultatoTest.csv

risulta generato. I file è un csv separato da virgole e può essere aperto con un editor di testo o con un foglio di calcolo ed esaminato. Risulta generato anche un file di nome RisultatoTest.xml, che contiene dei metadati, e può anch'esso essere aperto ed esaminato. Infine risultano generati dei file di nome PERSONFile.csv, HOSPFile.csv,… con i corrispondenti xml. Questi contengono i dati originali, parzialmente 'ripuliti' per accelerare le esecuzioni di script successivi.

Al termine del test rinominare la directory C:\TheMatrix\iad come C:\TheMatrix\iad_falsi e creare una nuova cartella vuota di nome C:\TheMatrix\iad.

2.2 Scaricamento dei dati veri della ASL

Nella procedura di installazione semplificata proposta in questo documento i file di dati vengono creati manualmente dall'incaricato dei sistemi informativi. I dati devono aderire al formato IAD descritto nella sezione 1 di questo documento ed essere salvati in formato csv **separato da virgole** nella cartella C:\TheMatrix\iad. Tra i campi non devono comparire spazi. Il separatore decimale è il **punto**. Nelle tabelle che descrivono il tracciato record di ciascun file è indicato in ogni riga il riferimento normativo che descrive quel dato, reperibile nell'allegato tecnico, anch'esso disponibile nell'archivio. Per esempio il riferimento normativo di un campo di OUTPAT è il campo contenuto nella terza colonna del file OUTPAT_aziende_v3.xls e si riferisce al documento 20090625_DT_Comma_5.pdf reperibile nell'archivio.

Si consiglia di scaricare per primo il file PERSON, successivamente EXE e HOSP e a seguire gli altri.

Attenzione. Nel caso di DRUGS e DDRUG (che hanno lo stesso formato, documentato nel file DRUG.xls) le variabili ATC e DURATION lasciate vuote, e sarà TheMatrix a riempirle. TheMatrix utilizza il file PRODUCT_CODE_AIFA.csv che si trova nella directory

C:\TheMatrix\lookups. Per calcolare DURATION applica la formula indicata nella riga corrispondente del file DRUG_aziende_v3.xls.

Nota. La scelta di quali file scaricare e riferiti a quanti anni dipende dallo script che si desidera eseguire: per esempio se si desidera misurare la qualità dell'assistenza ai malati cronici a partire dal 2007 si dovrà scaricare HOSP ed EXE per tutti gli anni disponibili, DRUGS e DDRUG a partire dal 2005, OUTPAT, HOME e RESIDENT a partire dal 2007 e PERSON deve essere storicizzata a partire dal 1 gennaio 2007. Se si partecipa a uno studio in rete con altri soggetti, gli anni da scaricare dipendono dallo studio e vanno concordati con i ricercatori responsabili dello studio.

2.3. Lancio di script di qualità su dati veri

Appena scaricato il file PERSON.csv eseguire la seguente sequenza di istruzioni
a) Aprire una shell
b) Lanciare la seguente sequenza di comandi

cd C:\TheMatrix
MetaXMLCreatorWIN PERSON.csv
TheMatrixWIN --logLevel 0 qualitaPERSON2011.txt

c) Dopo un tempo variabile a seconda della dimensione della ASL (fino a diverse ore) il programma termina l'esecuzione. Il file

C:\TheMatrix\iad\QualitaPERSON2011.csv

risulta generato. I file è un csv separato da virgole di una sola riga, che conta le righe del file PERSON e il numero di soggetti che risultano assistiti nel 2005, 2006,... fino al 2011. Inoltre risulta generato un file di nome PERSONFile.csv, con il corrispondente xml, che verrà riutilizzato nelle elaborazioni successive.

d) Confrontare questi numeri con quelli ottenuti dalla ASL con le sue procedure consuete.

Quando si scarica un altro file XXXX.csv eseguire la seguente sequenza di istruzioni
a) Aprire una shell
b) Lanciare la seguente sequenza di comandi

cd C:\TheMatrix
MetaXMLCreatorWIN XXXX.csv
TheMatrixWIN --logLevel 0 qualita XXXX2011 .txt

c) Dopo un tempo variabile a seconda della dimensione della ASL (fino a diverse ore) il programma termina l'esecuzione. Il file

C:\TheMatrix\iad\QualitaXXXX2011.csv

risulta generato. I file è un csv separato da virgole di una sola riga, che conta il numero di assistiti al 1 gennaio 2011, il numero di righe del file XXXX riferite all'anno 2011 e a un soggetto assistito al 1 gennaio 2011, e il numero medio di righe per soggetto. Risulta

generato anche un file di nome XXXXFile.csv, con il corrispondente xml. Nel caso di DRUGSFile.csv e DDRUGFile,csv, le varibili ATC e DURATION sono correttamente compilate.

d) Confrontare questi dati con quelli prodotti dalla ASL con le sue procedure consuete.

Quando queste procedure sono completate e i numeri ottenuti sono coerenti, TheMatrix è installato correttamente e può essere utilizzato. Per lanciare uno script qualsiasi (per esempio lo script scriptTest.txt) è sufficiente aprire una schell e lanciare questi comandi

cd C:\TheMatrix
TheMatrixWIN --logLevel 0 scriptTest.txt

3. Lancio di script complessi e requisiti

Gli script di TheMatrix variano da una complessità molto bassa (come quella degli script di qualità), a una intermedia (come quella dello script scriptTest.txt), fino a una complessità molto alta. A seconda della complessità dello script variano il tempo di esecuzione, lo spazio disco richiesto e più in generale i requisiti hardware necessari.

Su una macchina con almeno 2G di RAM è possibile migliorare le prestazioni di TheMatrix aprendo il file TheMatrixWIN.bat e modificando la stringa

-Xmx1024m

sostituendo al numero 1024 un numero maggiore, per esempio 2048.

Il prototipo di TheMatrix attualmente testato esegue in 15 ore uno script molto complesso su una ASL di 1,1 milioni di assistiti su una macchina con queste caratteristiche: sistema operativo a 64 bit, 4GB di RAM, CPU Pentium 1.5Ghz architettura Cedar Mill, JavaSun Java JRE 6. Lo spazio disco necessario a eseguire questo script è circa 100GB ogni 500mila assistiti.

TheMatrix è ancora in sviluppo ed è prevista una fase di ottimizzazione per ridurre l'entità di questi requisiti.

4. Note per utenti Linux e OSX

Le procedure di installazione ed esecuzione descritte in questo documento si possono applicare, con minime modifiche, anche a sistemi Linux e OSX. In particolare:

E' possibile scegliere una cartella di installazione personalizzata. E' importante che ogni comando relativo a TheMatrix sia però eseguito all'interno di questa cartella. (Esempio: se TheMatrix è installato in "/home/utente/TheMatrix" è necessario eseguire:

cd /home/utente/TheMatrix

prima di ogni sequenza di comandi.

I nomi di alcuni eseguibili sono differenti. In particolare:

- TheMatrixWIN diventa **TheMatrix**
- MetaXMLCreatorWIN diventa **MetaXMLCreator**

Quindi, una volta nella cartella di installazione, è possibile lanciare i suddetti eseguibili come:

./TheMatrix
./MetaXMLCreator

Appendice: differenze tra la release 1.45 e le precedenti

1) E' stata modificata lievemente la sintassi per lanciare gli script: ora essa è

TheMatrixWIN --logLevel 0 qualitaXXXX2011.txt

2) Gli script di Load hanno cambiato contenuto e nome: ora si chiamano loadXXXX.txt: **è necessario scaricarli nella directory –scripts- dalla cartella**

https://zimbra.ars.toscana.it/home/rosa.gini@ars.toscana.it/Briefcase/MATRICE/script_installazione

e rilanciarli. Anche se i file più lunghi (DRUGS e OUTPAT) saranno processati in un tempo abbastanza lungo, questo sarà risparmiato nelle esecuzioni successive.
3) Gli script loadDRUGS e loadDDRUG calcolano ora ATC e DURATION utilizzando la tabella di lookup che è aggiornata sia con farmaci recenti che con farmaci fuori commercio. Da un lato questo dispensa la ASL dal fare il record linkage, ma dall'altro **se una ASL ha in passato scaricato dal database solo i farmaci che si agganciavano alla vecchia tabella di lookup, deve riscaricare i dati relativi ai nuovi PRODUCT_CODE**
4) Gli script di qualità hanno cambiato nome e si chiamano ora qualitaXXXX2011.txt: **si consiglia caldamente di rilanciarli**
5) Gli output degli script sono ancora salvati nella directory iad, ma

- Hanno ora dei nomi specifici invece di riportare il nome dello script modificato
- Possono essere più d'uno
- Ciascuno viene salvato insieme a un file xml con lo stesso nome

6) I risultati di uno script possono essere riutilizzati dagli script successivi senza essere ricalcolati; questo ha come conseguenza che se si desidera che i risultati intermedi vengano ricalcolati (per esempio perché i dati sono cambiati), bisogna cancellarli manualmente.