Design of PubmedFetch

Method used to ingest data from PubMed SOAP interface. Brings in data as XML selected by either queries or record ranges and returns a stream of raw RDF/XML. Method can call a variety of fetch methods that allow selecting records based on a range of different attributes such as date added, date modified, number range, affiliation, etc.

Usage

To successfully harvest from PubMed:

1. Model - vivo.xml should be configured to point to your chosen vivo.
2. Task - create a pubmedfetch.xml (2 examples are provided).(Help for the search term)
3. Datamap - pubmed-to-vivo.xsl currently maps the data to the UF implementation this will have to be adjusted.

Methods

serializeFetchRequest

Runs, sanitizes, and outputs the results of a EFetch request to the xmlWriter

1. create a buffer
2. connect to pubmed
3. run the efetch request
4. get the article set
5. create XML writer
6. output to buffer
7. dump buffer to string
8. use sanitizeXML on string

sanitizeXML

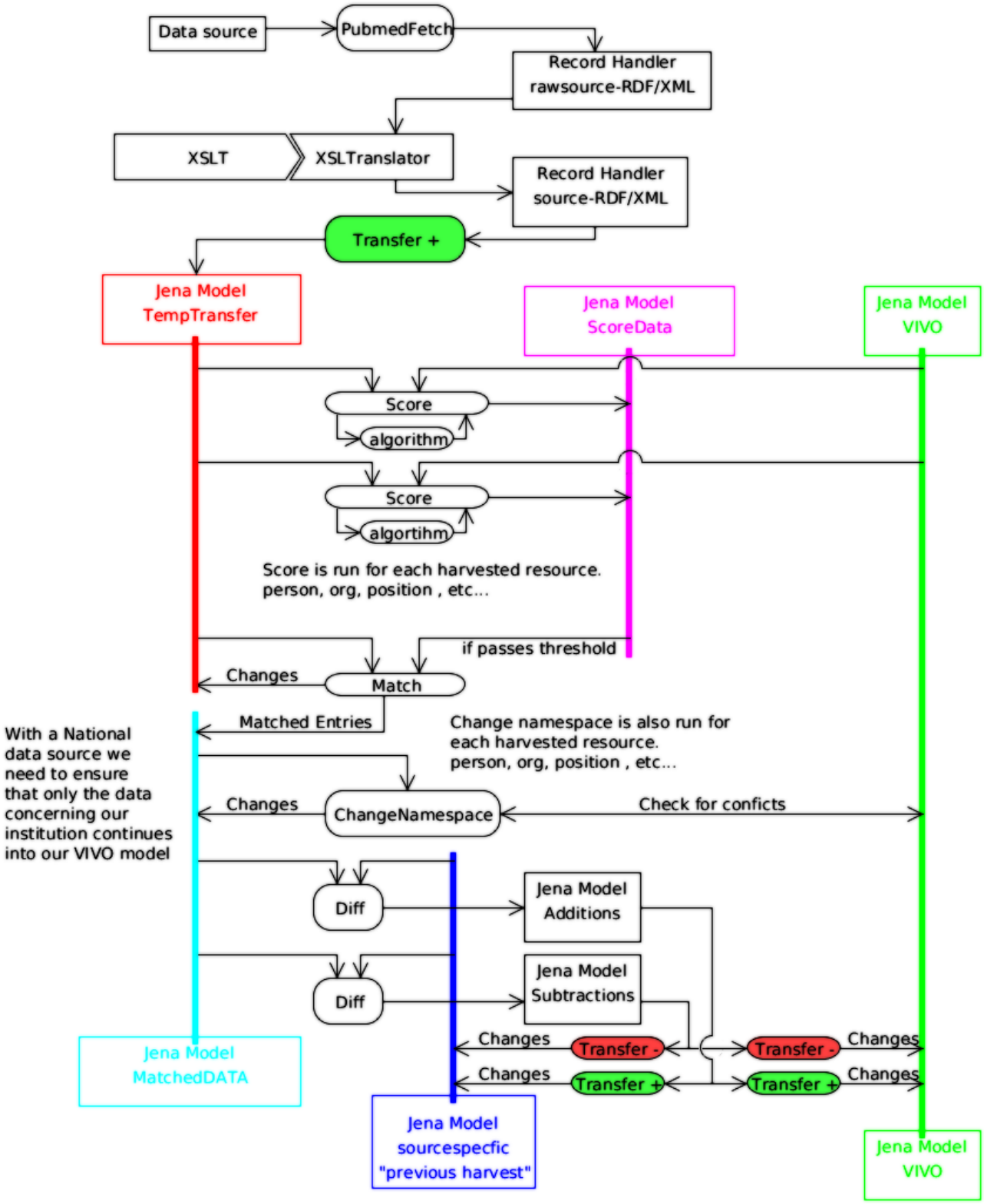
Sanitizes the XML in preparation for the output stream

1. replaces the input characters
2. writes to the output stream
    a. the OsWriter is provided in the superclass NIHFetch

Configuration file example

```
<?xml version="1.0" encoding="UTF-8"?>
<Task type="org.vivoweb.harvester.fetch.PubmedSOAPFetch">
        <Param name="email">swilliams@ichp.ufl.edu</Param>
        <Param name="output">config/recordHandlers/Pubmed-XML-h2RH.xml</Param>
        <Param name="termSearch">ufl AND edu[ad]</Param>
        <Param name="numRecords">100</Param>
        <Param name="batchSize">1000</Param>
</Task>
```

Flowchart


Data source
PubmedFetch
Record Handler
rawsource-RDF/XML
XSLT
XSLTranslator
Record Handler
source-RDF/XML
Transfer +
Jena Model
TempTransfer
Jena Model
ScoreData
Jena Model
VIVO
Score
algorithm
Score
algortihm
Score is run for each harvested resource.
person, org, position , etc...
if passes threshold
Changes
Match
Matched Entries
Change namespace is also run for
each harvested resource.
person, org, position , etc...
With a National
data source we
need to ensure
that only the data
concerning our
institution continues
into our VIVO model
Changes
ChangeNamespace
Check for conficts
Diff
Jena Model
Additions
Diff
Jena Model
Subtractions
Changes
Transfer -
Transfer -
Changes
Changes
Transfer +
Transfer +
Changes
Jena Model
MatchedDATA
Jena Model
sourcespecfic
"previous harvest"
Jena Model
VIVO