semantico
delivering online publishing solutions

MEDLINE Specification

Prepared for Nature

by Semantico Limited

March 2010

Lees House, 21-23 Dyke Road
Brighton BN1 3FE UK

T +44 1273 722222
F +44 1273 723232

info@semantico.com
www.semantico.com

Table of Contents

Document Control

Version	Date	Status	Author(s)
1	31/03/10	Final	Declan Newman, Steve Mallen, Liam Sheerin

1 Introduction

1.1 MEDLINE

The MEDLINE database is described (referenced from Wikipedia) as:

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information on articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

Compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system.

1.2 Project Goals

The goal of this project is to successfully download NPG MEDLINE records from PubMed and store them within a MarkLogic XML database within Nature.

The MEDLINE records contain several elements that will be useful to current, ongoing and future projects within Nature, including their OpenSearch facility.

Whilst there has been a successful import of the MEDLINE records on previous occasions (i.e. Alf Eaton's PHP scripts), it is desirable that this functionality is wrapped-up into a Java executable and run daily to ensure that the External Metadata and NPG MEDLINE records are as up-to-date as possible.

Having discussed in some depth the nature of the data and the way it is to be stored, it has been decided that – as is described in the original MEDLINE Ingest Plan – it will comprise of the following steps:

1. Initial download of NPG records from MEDLINE and store the these verbatim in a new database (medline)
2. Iterate through all medline records, locate their corresponding article and update/insert a new external metadata element.

3. Subsequent to the initial loading of all data, further updates of new and modified MEDLINE data will update the External Metadata in their corresponding Nature article. This will run once a day.
4. Investigative work to reduce shortfall in corresponding records between MEDLINE and Nature Documents databases.

Any records that are not successfully updated will be flagged (possibly using Marklogic collections) for ease of access when these are queried in the reconciliation process.

The task in step 3 will be integral in keeping the records up-to-date. The regularity of this task will be configurable (e.g. the crontab) .

2 Method of work

The majority of the work will be carried out at Semantico's offices in Brighton. However, when required, work will be carried out at the Nature offices in London.

Wherever possible, Semantico will adhere to Nature's working practices.

During the development process Semantico will require access to some resources within Nature in order to progress with the project. These resources are:

- Jira
- Confluence
- Mercurial
- MarkLogic CQ
- Marklogic

2.1 Technologies to be used

2.1.1 Programming language

Java 1.6 – in-line with Nature's approved programming practices, we will be using the Sun formatting format to format code within the Eclipse IDE for development.

2.1.2 Dependency management

Maven 2 will be used for dependency management, build and report generation.

Nature has the necessary infrastructure in place for Maven.

2.1.3 Continuous Integration

Hudson – Nature have a workflow that involves the use of Hudson to provide their continuous integration testing. Once the project is at a stage where this is necessary, this will be added.

2.1.4 IoC Layer

In accordance with Nature's Java Best Practice documentation, we will be using Google

Guice.

2.1.5 Unit testing

JUnit 4 – In-line with Nature's Java best-practices, which indicate a test coverage of 80%.

2.1.6 Logging

In accordance with Nature's Java Best Practices documentation, we will be using Log4J for logging and logging configuration.

2.1.7 Connecting to MarkLogic

Interaction with Marklogic will be performed using XCC 4.1.

2.1.8 Connecting to PubMed

The connection and querying of PubMed will be done using Eutils SOAP API.

2.2 Environment

2.2.1 Platform

2.2.1.1 Java Development

The majority of the development will be done using a Microsoft Windows machine connecting to the test instances of the MarkLogic databases within Nature.

2.2.1.2 MarkLogic

The version of MarkLogic that will be used for development, testing and production will be MarkLogic version 4.

To interact with MarkLogic from Java we will be using XCC library.

2.2.1.3 Semantico testing

Semantico predominantly use a virtual environment for testing, using Linux Ubuntu as a

standard operating system.

2.2.1.4 Nature testing

Testing on the Nature internal servers will be carried out on RedHat Linux.

2.2.1.5 Nature staging/production

The Nature staging and live environments will be running Unix.

2.3 Methodology

2.3.1 Interacting with MEDLINE

The interaction with PubMed will be carried out using the Java SOAP API <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/DOC/esoap_help.html>.

It has been recommended that the PubMed SOAP interface is used to query and retrieve records from the PubMed service.

This will help ensure that necessary upgrades to the PubMed service will be flagged early.

The authority XML "/thesauri/products.xml" in the Documents database, will be used as the definitive list of the ISSNs that will be used to query PubMed. The "/thesauri/products.xml" file will be retrieved from the MarkLogic database each time the application is executed to minimise the risk of data becoming out-of-date.

From initial investigation, we have determined that there are approximately 347,807 NPG articles in PubMed; of those, 190,110 have associated DOIs. The majority (325,455) have Mesh Terms. As discussed later in this document, there is a substantial shortfall of DOI associations.

2.3.2 Interacting with the Documents Database

The External Metadata schema, that will form the basis of the additional external metadata, has not yet been formally specified. The development of this is continuing within Nature to complete this work before this becomes necessary for work to continue. An example of this XML was supplied by Tony in the original MEDLINE Data Ingest Plan

document (please see Appendix 1).

The Documents database contains all NPG articles, it is the purpose of this project to populate this data with corresponding PMID and MeSH data. This data will sit in a "<external-metadata/>" element which is to be a new child element of "<record/>"

This data is to be updated daily to make sure that the data is up-to-date and there are no discrepancies.

Note: At the time of writing this, an initial meeting has determined that this is to be finalised shortly.

2.3.3 Logging and reporting

Attention to detail regarding the logging is of utmost importance. We will spend a proportion of the development time reviewing the logging and their levels.

Amongst others, we will need to log the following:

- Number of new articles.
- The actual query that was used against PubMed. This will also include the REST query for ease of debugging.
- Number of updates to existing articles.
- The count of articles in the Nature MEDLINE database with DOIs **before** the update is run.
- The count of articles in the Nature MEDLINE database with DOIs **after** the update is run.
- Number and PMID of articles with no DOI that consequently could not be used for an update.

In addition to the reporting of the database status pre and post operation, runtime logging will be split into five levels:

1 **Debug** – This will output fine detail about the current process. This will include all information about which articles are currently being processed, and what the current

process is. For example:
DEBUG – Inserting MeSH Terms into External Metadata article – PMID 19295598 – DOI 10.1038/458293a

2 **Info** – The Info level will provide information at a higher level than debug and will output information about the main operations in the system. For example:
INFO – Starting to process 3000 articles for ISSN 1476-4687

3 **Warn** – The warn level will log any information that could result in non-fatal errors. This will be useful when tracking errors for missing data etc. An example of when the Warn level is used might be:
WARN – Could not find an article with DOI 10.1038/458293a in the External Metadata Database

4 **Error** – Error logging will be limited to only those problems that may result in the operation being aborted. This could be that a file cannot be accessed on the file system, or a network issue has occurred. The system will try and continue, but is likely to fail.
ERROR – Cannot access file /etc/nature/tmp/00221.xml – Access denied.

5 **Fatal** – These log entries are limited to only those operations that will result in the application aborting. If access to the databases or SOAP API is not possible, it will result in a fatal error.
FATAL – Network error – Cannot connect to http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/v2.0/soap_adapter_2_0.cgi *– response code 404.*

2.4 The process

2.4.1 Import and update of MEDLINE data

2.4.1.1 Initial import of MEDLINE data

This is a one-time process that requires us to retrieve all NPG articles from MEDLINE and store them verbatim into a MarkLogic database hosted at Nature.

Identifying the articles to ingest will be done using the ISSNs listed in the aforementioned products.xml file. Initially, this will not use any multi-threading, but the application will be designed in such a way that this can be added at a later date with minimum work.

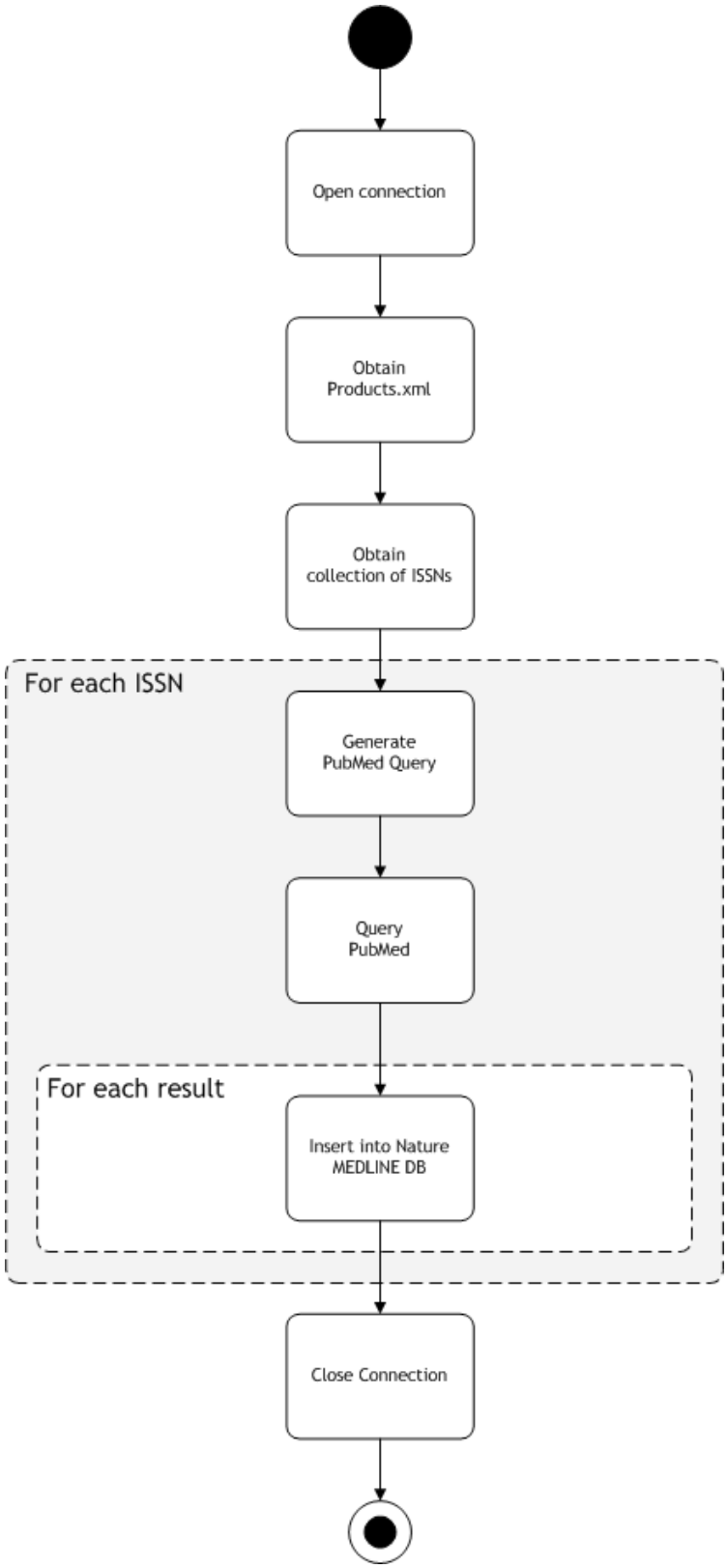
Once the internal MEDLINE database has been populated, we will carry out some tests to validate the data to ensure it contains all data we are interested in, and no more.

The query to retrieve the data form MEDLINE will be in two steps:

1. Generate a query against the eSearch service using a single ISSN to return a WebEnv.
2. The WebEnv will then be used against the eFetch service to retrieve the articles in their entirety.

Note: The MEDLINE database has an internal limit of 100,000 articles per transaction. Whilst is unlikely that this would be exceeded when querying using ISSN, the retrieval should allow for an incremental download approach.

2.4.1.1.1 Activity Diagram



2.4.1.2 Updating of External Metadata

After the initial population of MEDLINE data, a secondary process will run to locate corresponding records in the Documents database using the DOI (ArticleId IdType="doi") element listed in the MEDLINE data as their key. When a match is found the External Metadata will be updated with two new elements

1. Module element (listed as a <atom:entry> elements in Appendix 3.1 Example External Metadata XML), containing the MeSH Terms, and;
2. The PMID – the primary key of PubMed articles.

Note: An additional step may be required:

1. *As there is a substantial shortfall of DOIs in the NPG MEDLINE articles, a further step may be required to try and address this. Some additional investigative work will be carried out to try and reduce the shortfall.*

Note: We have discussed the possibility of adding the MeSH IDs into the MeSH term entry. For the time-being however, this has been deemed out-of-scope.

Please see diagram 2.4.3.1 for more details.

2.4.2 Ongoing updating of articles

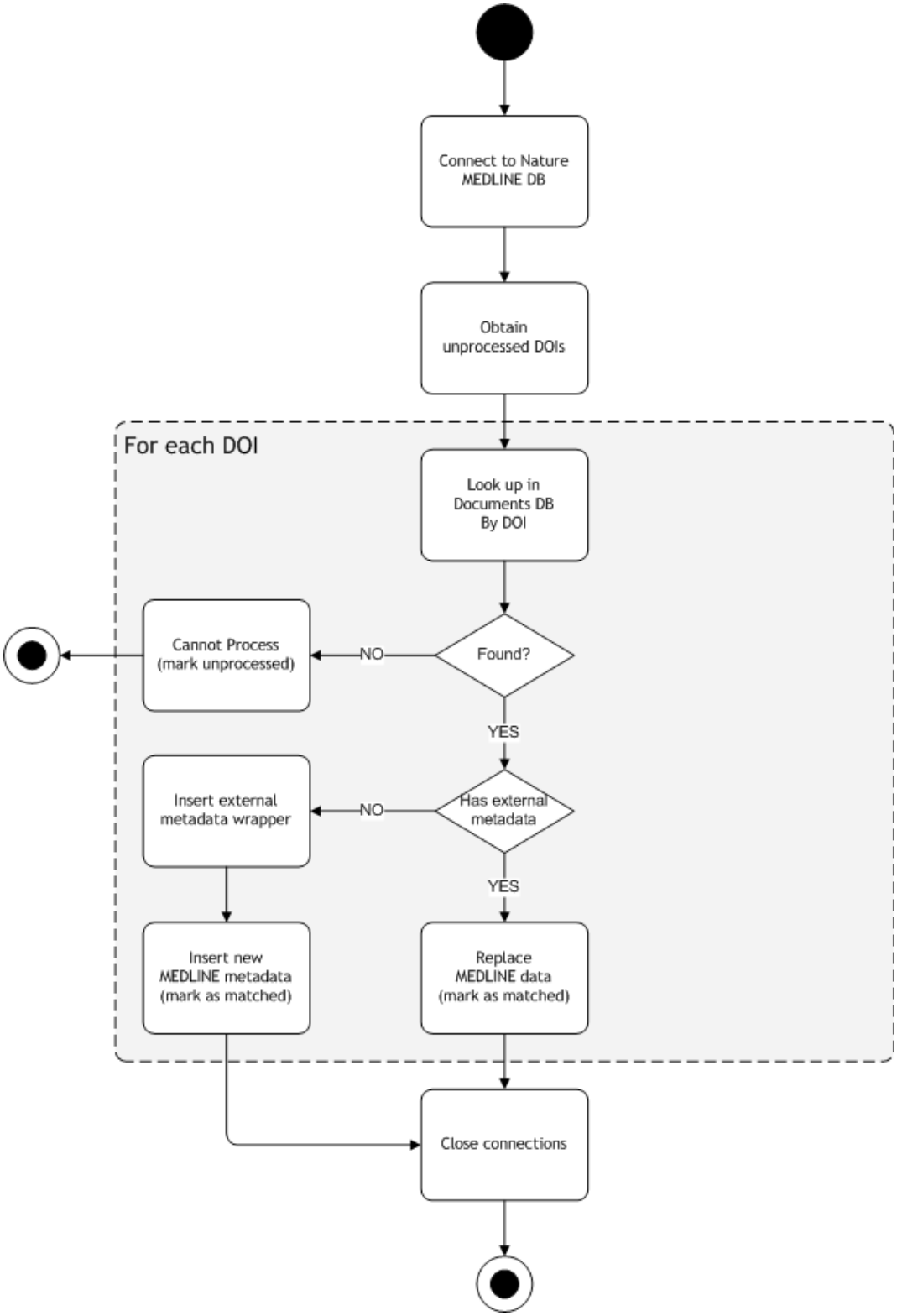
It is required that the process of querying PubMed and the subsequent updating of External Metadata records be an ongoing one.

The code that performs the initial import and updates should be designed in such a way to facilitate the periodical updates of articles. The regularity of these updates should be configurable per ISSN. Any ISSN that does not have an explicit value for how often it runs, should run at the default interval (also configurable). This will require a time-stamp to be stored against all ISSNs that have an explicit value. This can be done using a text file or internal (in-memory) database, such as HSQL.

When querying PubMed for updates a timestamp will be used to obtain only those that have been added or updated since the last time the update was run. This is also true if an ISSN has a specific value for it's regularity.

As mentioned in previous areas of this document, there is a significant shortfall of records that contain a DOIs. To address this, there will be a full week of investigative work to determine the effectiveness and best process to reduce the shortfall.

2.4.2.1 Activity Diagram



Note: Documents that cannot be processed will be addressed with the investigative work to be carried out to reduce the shortfall.

2.4.3 Reconciliation process

I addition to the processes above, it is also required that a further process is performed to reconcile between the MEDLINE records and those stored within the Nature's MarkLogic MEDLINE database. The primary goal of this step is as an insurance policy to ensure that no matching records have been missed.

This process will utilise the same code as elsewhere, but will take different arguments.

Using crontab to configure the regality of this process, it is likely that this will run on a weekly basis.

2.4.4 Reducing the shortfall between PMID/DOI associations

As mentioned elsewhere in this document, there will be some time allocated to investigate how to reduce the shortfall between PMID/DOI associations between the NPG Documents and MEDLINE databases.

From initial discussion there are some parameters that look promising (e.g. ISSN, volume and first page) to facilitate an 'exact match'. But, the investigation will determine this in more detail.

If – as is expected – a solution for matching records is found, the external metadata will hold a value to identify how it had been matched, to indicate the confidence of accuracy. This will be especially valuable when more than one method (excluding DOI) has been used to match records.

Whilst primarily the work undertaken will be to investigate the options, it should be made clear that there are deliverables. This will of course depend on how successful the findings are. The best scenario – If initial investigations determine that an 'exact match' can be relied upon early on – then this should be worked-in to the overarching application. But, in the worst-case-scenario – where no definitive method of finding a reliable 'exact match' could be found – documentation should be supplied outlining what had been tried and the reasons that it was unsuccessful.

Another area of discussion was the possibility of using the Nature SQL database, but has currently been deemed outside of project scope.

2.5 Data Storage

2.5.1 Nature MEDLINE Database

The naming structure of the medline database will be:

- Test – “Medline”
- Staging – “Medline”
- Live – “Medline”

The URI scheme for the MEDLINE articles will be “/medline/<pmid>” and indexed on PMID and DOI.

Note: Further changes to indexes will evolve as the project progresses.

2.5.2 Nature Documents Database

At the time of writing this document, the schema for the External Metadata in the MarkLogic Documents database has not been finalised.

3 Summary of work and durations

Below is a very high-level list of the planned activities, and is not intended to be a list of deliverables.

The majority of the work will be carried out by Declan Newman, with Steve Mallen assisting with XML/MarkLogic specific tasks. It is worth bearing in mind that the work will not necessarily be undertaken sequentially as some can be carried out concurrently.

Resources have currently been assigned for 4 days per week. However, this may change from week-to-week. E.g. 3 days one week and five the next.

1.Application configuration module – 3 days

2.PubMed Query generator and ingestion into MarkLogic MEDLINE database – 5 days

3.Logging and reporting module – 4 days

4.Interaction and query with Nature MarkLogic Documents database – 6 days

5.Testing – 2 days

6.Investigative work to reduce shortfall of PMID/DOI associations between Nature Documents and MEDLINE databases – 5 days

Total: 25 days

Note: The 2 days of testing will enable us to address any performance and/or environment issues, and is in addition to ongoing unit testing that is part of the development process.

4 Appendix

4.1 PubMed XML

Taken from Tony Hammond's original document

```
<PubmedArticle>
  <MedlineCitation Owner="NLM" Status="MEDLINE">
    <PMID>19295598</PMID>
    <DateCreated>
      <Year>2009</Year>
      <Month>03</Month>
      <Day>19</Day>
    </DateCreated>
    <DateCompleted>
      <Year>2009</Year>
      <Month>04</Month>
      <Day>28</Day>
    </DateCompleted>
    <Article PubModel="Print">
      <Journal>
        <ISSN IssnType="Electronic">1476-4687</ISSN>
        <JournalIssue CitedMedium="Internet">
          <Volume>458</Volume>
          <Issue>7236</Issue>
          <PubDate>
            <Year>2009</Year>
            <Month>Mar</Month>
            <Day>19</Day>
          </PubDate>
        </JournalIssue>
        <Title>Nature</Title>
        <ISOAbbreviation>Nature</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Dinosaurs: Fuzzy origins for feathers.</ArticleTitle>
      <Pagination>
        <MedlinePgn>293-5</MedlinePgn>
      </Pagination>
      <AuthorList CompleteYN="Y">
        <Author ValidYN="Y">
```

```
            <LastName>Witmer</LastName>
            <ForeName>Lawrence M</ForeName>
            <Initials>LM</Initials>
          </Author>
        </AuthorList>
        <Language>eng</Language>
        <PublicationTypeList>
          <PublicationType>Comment</PublicationType>
          <PublicationType>Historical Article</PublicationType>
          <PublicationType>News</PublicationType>
        </PublicationTypeList>
      </Article>
      <MedlineJournalInfo>
        <Country>England</Country>
        <MedlineTA>Nature</MedlineTA>
        <NlmUniqueID>0410462</NlmUniqueID>
      </MedlineJournalInfo>
      <CitationSubset>IM</CitationSubset>
      <CommentsCorrections>
        <CommentOn>
          <RefSource>Nature. 2009 Mar 19;458(7236):333-6</RefSource>
          <PMID>19295609</PMID>
        </CommentOn>
      </CommentsCorrections>
      <MeshHeadingList>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Animals</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">China</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Dinosaurs</DescriptorName>
          <QualifierName MajorTopicYN="Y">anatomy &amp;amp; histology</QualifierName>
          <QualifierName MajorTopicYN="N">classification</QualifierName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="Y">Evolution</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Feathers</DescriptorName>
```

```
            <QualifierName MajorTopicYN="Y">anatomy &amp;amp; histology</QualifierName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Fossils</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">History, Ancient</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Phylogeny</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Skin</DescriptorName>
            <QualifierName MajorTopicYN="Y">anatomy &amp;amp; histology</QualifierName>
        </MeshHeading>
    </MeshHeadingList>
</MedlineCitation>
<PubmedData>
    <History>
        <PubMedPubDate PubStatus="entrez">
            <Year>2009</Year>
            <Month>3</Month>
            <Day>20</Day>
            <Hour>9</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="pubmed">
            <Year>2009</Year>
            <Month>3</Month>
            <Day>20</Day>
            <Hour>9</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="medline">
            <Year>2009</Year>
            <Month>4</Month>
            <Day>29</Day>
            <Hour>9</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
    </History>
```

```
      <PublicationStatus>ppublish</PublicationStatus>
      <ArticleIdList>
        <ArticleId IdType="pii">458293a</ArticleId>
        <ArticleId IdType="doi">10.1038/458293a</ArticleId>
        <ArticleId IdType="pubmed">19295598</ArticleId>
      </ArticleIdList>
    </PubmedData>
</PubmedArticle>
```

4.2 Example External Metadata XML

Taken from Tony Hammond 's original document

```
<record version="1.3">
  <metadata>
    <meta:id>458293a</meta:id>
    <meta:doi>10.1038/458293a</meta:doi>
    <meta:authors>
      <meta:author>
        <meta:first>Lawrence M.</meta:first>
        <meta:last>Witmer</meta:last>
        <meta:full>Lawrence M. Witmer</meta:full>
      </meta:author>
    </meta:authors>
    ...
  </metadata>
  <external-metadata>
   <atom:entry
   xmlns:atom="http://www.w3.org/2005/Atom"
>
  <atom:title>Medline</atom:title>
  <atom:author><atom:name>NLM</atom:name></atom:author>
  <atom:summary>Medline ...</atom:summary>
  <atom:id>urn:uuid:e0c1ddea-3464-41b7-ac5b-3e5145cddff3</atom:id>
  <atom:link href="http://dx.doi.org/10.1038/458293a" rel="related" />
  <atom:updated>2006-12-14T00:00:00Z</atom:updated>
  <atom:content type="application/xml">
          <entry:metadata xmnlns:entry=...>
      <PMID>19295609</PMID>
      <MeshHeadingList>
        <MeshHeading>
```

```
          <DescriptorName MajorTopicYN="N">Animals</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">China</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Dinosaurs</DescriptorName>
          <QualifierName MajorTopicYN="Y">anatomy &amp; histology</QualifierName>
          <QualifierName MajorTopicYN="N">classification</QualifierName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="Y">Evolution</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Feathers</DescriptorName>
          <QualifierName MajorTopicYN="Y">anatomy &amp; histology</QualifierName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Fossils</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">History, Ancient</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Phylogeny</DescriptorName>
        </MeshHeading>
        <MeshHeading>
          <DescriptorName MajorTopicYN="N">Skin</DescriptorName>
          <QualifierName MajorTopicYN="Y">anatomy &amp; histology</QualifierName>
        </MeshHeading>
      </MeshHeadingList>
    </entry:metadata>
  </atom:content>
</atom:entry>
   </external-metadata>
   <content type="npg_dtd">
    <article id="458293a" language="eng" publish="issue" relation="no">
    ...
    </article>
   </content>
</record>
```