PFE Workflow: Text to Database

Free Form Text from the DMDE

John Langdon: A politician and merchant from New Hampshire, he was originally a Federalist but became a Republican. He served for two terms in the U.S. Senate, retiring in 1801. He then served as governor of New Hampshire from 1805 to 1809 and again from 1810 to 1812, after which he retired from active politics. He married Elizabeth Sherburne. His niece, Caroline Langdon, married William Eustis.

Once we have the biographies from the papers projects they will be tagged in XML in PubMan according to our specifications.

On the left you can see some dummy tags representing most of the information found in the DMDE's biography of John Langdon.

I've also made 3 editorial decisions. First, the names of Langdon's wife and niece have been "standardized" to include their married names as inferred from the text. Second, I have removed Langdon's less than 1 year hiatus as governor. Third, I have [accurately] guessed that Langdon's began his first term as Senator in 1789. In real life, of course, all of these decisions would need to be fact checked, but they illustrate the importance of careful tagging (ex. New Hampshire governors only served 1 year terms, so I just erased Jeremiah Smith's term as governor from our database!).

PubMan CMS -- Tagged XML

```
<name>John Langdon</name>
<HomeState>New Hampshire</HomeState>
<occupation>merchant</occupation>
<occupation>politician</occupation>
<occupation>senator</occupation>
        <startyear>1789</startyear>
        <endyear>1801</endyear>
<occupation>governor</occupation>
        <startyear>1805</startyear>
        <endyear>1812</endyear>
<organization>Federalist Party</organization>
<organization>Republican Party</organization>
<organization>United States Senate</organization>
<wifename>Elizabeth Sherburne Langdon</wifename>
<neice> Caroline Langdon Eustis </neice>
<kinbymarriage> William Eustis </kinbymarriage>
```

Once the documents have been tagged in XML we will need to export that information to the database. We can presort the tagged information for each table in the database and then export that data in Comma Separated Value (CSV) files. CSV files are a simple and ubiquitous way to transfer tabled information between different computer systems. Once we have the CSV files we can load them directly into the database.

On the right you can see how the tagged data ends up in several (highly abbreviated) versions of our database tables.

Database Tables

Individuals
Name: John Langdon
HomeState: New Hampshire

MemberOf
Name: John Langdon
Organization: U.S. Senate
StartYear: 1789
EndYear: 1801

WorkedAs
Name: John Langdon
Status/Occupation: Governor of New Hampshire
StartYear: 1805
EndYear: 1812

WorkedAs
Name: John Langdon
Status/Occupation: merchant
StartYear: unknown
EndYear: unknown

RelatedTo
Name: John Langdon
Name: Elizabeth Sherburne Langdon
Relationship: Spouse

RelatedTo
Name: John Langdon
Name: Caroline Langdon Eustis
Relationship: Uncle/Niece

RelatedTo
Name: Caroline Langdon Eustis
Name: William Eustis
Relationship: Spouse

Jean Bauer
10/16/08