

Geography of Slavery Conversion Process

Overview

This page documents work done to migrate XML data from the Geography of Slavery project into the Documents Compass Darma digital asset management system for use in the Person of the Founding Era Project. It consist of:

- data analysis and documentation of work done to address source data issues
- documentation of conversion and enrichment processes
- reports to support editorial work that will refine the data: disambiguation, link checking, etc.

Errors found in the source data during the course of the work were fixed in a working copy to improve the quality of our results. This may be referred to as the 'working copy' below.

The conversion process itself the following parts:

- **Owner/Subscriber Metadata Generation:** A process that generates an intermediate record for each owner mentioned in the source data and assembles or creates selected metadata that may be used to create DarmaXML records or support reporting processes.
- **Owner/Subscriber Metadata Disambiguation:** A process that programmatically disambiguates the records generated by looking at common slave links and name matches.
- **Owner Data Generation:** A process that uses both the source data and disambiguated owner metadata to assemble DarmaXML records for this class of persons.
- **Slave / Servant / Deserter / Free Person / Misc DarmaXML Record Generation:** A process that generates a DarmaXML record for each mention of a person in this class. This process also links these records to those of their owners, previous owners, and subscribers.
- **Slave / Servant / Deserter / Free Person / Misc Disambiguation:** A process that conflates the generated records by matching them on identifiers.

Specs for the classes of people to generate records for and the metadata to include were supplied in a sample xsl file, a Word document, and an Excel spreadsheet, all attached here. See also [2014-10-27 Telecon](#). A decision was made to include the records of free persons and miscellaneous types, the latter proving to be in most cases a person a slave/servant/deserter was believed to have run away with or associate with.

Also attached is a copy of the output for all classes of records. It may help to reference the source Masterfile and output when reviewing this information.

The programmatic processes for disambiguation and linking have known limitations but they are relatively accurate in most cases and offer huge gain over manually performing these processes. Reports have been generated from the source and converted data to support review of all records and complete the production of this data set. See **Remaining Work**.

Program Structure and Execution

The process authored is capable of creating data sets for all classes of persons or just selected ones.

1. The process to generate owner metadata and DarmaXML records is run.
2. The process to disambiguate owners is run. It generates to supplemental XML maps: 1) mapping owner records that link to the same slave; 2) mapping owner records with the exact same name. # 1 goes through a manual review process to finalize the map.
3. The reports in the previous step are utilized by programmatic process that conflates owner records. Note: There will still be some duplicates.
4. A process that generates DarmaXML records for the remaining classes runs. Each call for execution takes a class type as a parameter: male-slave, female-slave, unknown-slave, male-servant, female-servant, male-deserter, female-deserter, misc, amb. None of these has a dependency on the other (at this time), but all require that the owner metadata is generated first.
5. A final process is run to conflate records for the remaining classes into final data sets that will be loaded into DARMA.

General Notes

- Where there are multiple values for a text metadata field in Darma, the values are compiled as a semicolon or '/' separated list, the latter convention being used in text fields where semicolons might appear within values. The documentation that follows may refer to this as 'assembled as list' or 'as list'.
- Where there are multiple values for an xhtml field in Darma, the values are put into textual divisions with numbering that distinguish them.
- There is logic to process the names in the source data into the LastName, Title FirstName Middle Honorific syntax used by PFE, and to perform case handling (they appear as all caps or C&Ic in the source) but it is less than perfect. Some notes:
 - All single word references are changed to C&Ic
 - All name references (except owner/subscriber/powner) that contain (in any case form) the following have it moved to the LastName position: 'negro', 'negroe', 'mulatto', 'black'. The same is done for any record that begins with 'his' or 'her'. Owner/subscriber/powner records have logic to handle the appearance of 'jailer' or 'jailor'.
 - Two or three part names not matching the two prior conditions are change to the PFE syntax

- Any names matching the three prior criteria are formatted as titles, dealing with the articles 'a', 'an', 'the', 'another'
- Known issues of this part of the program:
 - May foul up honorifics (e.g. 'Jun.');
 - Irish names like O'connor (lc c) will need case fixing;
 - hyphenated names like 'Beveridge, Henry-valentine' need case fix;
 - single names such as 'Borden, Mrs.'.
 - See also related points in **Known Data Issues**
- Name logic is applied to `fullname`, `subjectname`, `(p)owner`, and `subscriber` name field values in the DarmaXML output but NOT `alias`. (OK?)
- All records are mapped to the appropriate collection by the process, i.e. alphabetical collections for the PFE record (based on lastname) and Biography collection for the bios.
- We can share the source code for the process via GitHub if anyone is interested. All that's required is that they have a GitHub account.

Owner Metadata Generation

This process generates a metadata set that can be used to assist in generating reports and DarmaXML records. For each instance of a member of these groups:

(1)

- `name[@type = 'prevowner']`
- `name[@type = 'owner']`
- `name[@type = 'subscriberowner']`
- `name[@type = 'subscriber']`

(2)

- `owner`
- `subscriber`
- `powner`

a record is generated. While they represent the same thing, the groups are listed separately here since the source data was inconsistent in handling their markup and related metadata. Generated metadata looks like this:

Sample generated metadata for owners, subscribers, previous owners

```
<?xml version="1.0" encoding="utf-8"?>
<owners-subscribers>
  <person provenance="name-element" role="prevowner" source-record="v1798090063">
    <id>O_d2e187251</id>
    <name>David Fentress</name>
    <name-key>david fentress</name-key>
    <subject-name>Fentress, David</subject-name>
    <slaves>
      <slave-id link-provenance="source" link-type="prevowner"
source-record="v1798090063"
      target="pfe_GOS005209" />
    </slaves>
    <homes>
      <home source-record="v1798090063">Princess Anne</home>
    </homes>
  </person>
  <person provenance="subscriber-element" role="subscriber"
source-record="v1736091282">
    <id>d2e57</id>
    <name>Anne Smith</name>
    <name-key>anne smith</name-key>
    <subject-name>Smith, Anne</subject-name>
    <slaves>
      <slave-id link-provenance="generated" link-type="subscriber"
desc-type="single"
      source-record="v1736091282"
      target="pfe_GOS001347" />
    </slaves>
    <homes>
      <home source-record="v1736091282">Middlesex County</home>
    </homes>
  </person>
</owners-subscribers>
```

The provenances of the record is important owing the aforementioned differences in markup.

Group 1:

- Is generated from source data that has links to slave ids so the links to slaves have a high degree of certainty. No attempt to generate links is made and there is no worry if the person record is generated a group description.
- Has values for <home> generated from the <geography> element, which is not reliable as it is used for all types of locations.

Group 2:

- Have links to slaves that are generated from the data and thus will have a lower degree of certainty.
- Have more reliable values for <home>, although the logic varies between powner, subscriber, owner and for group records. See code for details.

For both groups names marked as owner-subscribers become owners.

A general rule about homes is that if it is a non-group record it is safe to use a value that precedes or follows the person's mention. In a group record only one that follows the mention.

The data in this file that is not used within the generated DarmaXML records can be drawn upon to create reports that assist data review and correction.

Owner DarmaXML Record Generation

Two records are generated for each instance of a person in the source data: a PFE name record and a Biography record. The generated PFE record (at this time) only has the owner name and a link to the biography record. Metadata fields are as shown in the sample here.

Sample owner/subscriber/prevowner metadata

```
<Object collectionID="c139" xml:id="pfe_O_d2e187251">
  <Name>Fentress, David; Fentress, Mr. David</Name>
  <Links>
    <Link type="Biography" targetID="bio_O_d2e187251"/>
  </Links>
</Object>
<Object collectionID="c365" xml:id="bio_O_d2e187251">
  <Name>Fentress, David; Fentress, Mr. David (1798-09-01; 1798-03-29;
1798-07-12)</Name>
  <Metadata>
    <Datum fieldName="biographies.SubjectName" fieldID="6" status="new">
      <Value>Fentress, David; Fentress, Mr. David</Value>
    </Datum>
    <Datum fieldName="Costa.docID" fieldID="204" status="new">
      <Value>v1798090063; v1798030017; v1798070049</Value>
    </Datum>
    <Datum fieldName="Costa.GoSLink" fieldID="217" status="new">
      <Value>http://www2.vcdh.virginia.edu/gos/</Value>
    </Datum>
    <Datum fieldName="workflow.BioType" fieldID="172" status="new">
      <Value>Advertisement</Value>
    </Datum>
    <Datum fieldName="workflow.AddDate" fieldID="137" status="new">
      <Value>2014-12-02-05:00</Value>
    </Datum>
    <Datum fieldName="Costa.Residence" fieldID="196" status="new">
      <Value>Princess Anne</Value>
    </Datum>
    <Datum fieldName="Costa.NewspaperCity" fieldID="195" status="new">
      <Value>Norfolk; Norfolk; Norfolk</Value>
    </Datum>
    <Datum fieldName="Costa.Dateline" fieldID="210" status="new">
      <Value>1798-09-01; 1798-03-29; 1798-07-12</Value>
    </Datum>
    <Datum fieldName="Costa.Publisher" fieldID="201" status="new">
      <Value>Willett and O'Connor; Willett and O'Connor; Willett and
O'Connor</Value>
    </Datum>
    <Datum fieldName="Costa.Title" fieldID="200" status="new">
      <Value>Norfolk Herald; Norfolk Herald; Norfolk Herald</Value>
    </Datum>
  </Metadata>
</Object>
```

We opted not to store the advertisement in the biography field. It doesn't make sense to store this with owners as it can be traced through the records they are linked to (and virtually all are).

Slave / Servant / Deserter DarmaXML Record Generation

For each instance of a member of these groups two DarmaXML records are generated: male-slave, female-slave, unknown-slave, male-servant, female-servant, male-deserter, female-deserter, misc, amb. One is a PFE record and the other is a Biography record.

The following fields are only being generated when the person record is created from a non-group description (see also **Links and Remaining Work**): person.Forename2, person.Forename2Type, Costa.Subscriber, Costa.Age, Costa.Alcohol, Costa.Alias, Costa.Birth, Costa.Cloth, Costa.Disease, Costa.Escape, Costa.Ethnicity, Costa.Imp, Costa.Intent, Costa.Marks, Costa.Physical, Costa.Powner, Costa.Residence, Costa.Sale, Costa.Skill, Costa.Warning, Costa.Speech

Some metadata missing for records from group descriptions may be populated when the deduping process is completed, i.e. there may be conflation of single description and group description records for the same person.

Sample DarmaXML for slave record

› [Expand](#)

[source](#)

```
<?xml version="1.0" encoding="utf-8"?>
<Data xmlns="http://www.dataformat.com/darma"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.dataformat.com/darma
http://www.dataformat.com/local/darma_data_V2.xsd">
  <Header>
    <Collections>
      <Collection name="The Geography of Slavery" xml:id="c365"/>
      <Collection name="People of the Founding Era" xml:id="c1"/>
      <Collection name="A Surnames" xml:id="c134"/>
      <Collection name="B Surnames" xml:id="c135"/>
      <Collection name="C Surnames" xml:id="c136"/>
      <Collection name="D Surnames" xml:id="c137"/>
      <Collection name="E Surnames" xml:id="c138"/>
      <Collection name="F Surnames" xml:id="c139"/>
      <Collection name="G Surnames" xml:id="c140"/>
      <Collection name="H Surnames" xml:id="c141"/>
      <Collection name="I Surnames" xml:id="c142"/>
      <Collection name="J Surnames" xml:id="c143"/>
      <Collection name="K Surnames" xml:id="c144"/>
      <Collection name="L Surnames" xml:id="c145"/>
      <Collection name="M Surnames" xml:id="c146"/>
      <Collection name="N Surnames" xml:id="c147"/>
      <Collection name="O Surnames" xml:id="c148"/>
      <Collection name="P Surnames" xml:id="c149"/>
      <Collection name="Q Surnames" xml:id="c150"/>
      <Collection name="R Surnames" xml:id="c151"/>
      <Collection name="S Surnames" xml:id="c152"/>
      <Collection name="T Surnames" xml:id="c153"/>
      <Collection name="U Surnames" xml:id="c154"/>
      <Collection name="V Surnames" xml:id="c155"/>
      <Collection name="W Surnames" xml:id="c156"/>
      <Collection name="X Surnames" xml:id="c157"/>
      <Collection name="Y Surnames" xml:id="c158"/>
      <Collection name="Z Surnames" xml:id="c159"/>
    </Collections>
  </Header>
  <Object xml:id="pfe_GOS002448" collectionID="c146">
    <Name>Mary-anne (captured slave, 1776-03-01)</Name>
    <Metadata>
      <Datum fieldName="person.Fullname" status="new" fieldID="16">
        <Value>Mary-anne</Value>
      </Datum>
    </Metadata>
  </Object>
</Data>
```

```

<Datum fieldName="person.gender" status="new" fieldID="18">
  <Value>M</Value>
</Datum>
<Datum fieldName="person.DCCitation" fieldID="75" status="new">
  <html xmlns="http://www.w3.org/1999/xhtml">
    <head>
      <title>People of the Founding Era Citations</title>
      <link
href="http://dcdarma.dataformat.com/Resources/RadEditorConfigs/CSS/XHTML_PFE_II.css"
      type="text/css"
      rel="stylesheet"/>
    </head>
    <body>
      <div>
        <p>Runaway Advertisement, Virginia Gazette (1776-03-01); The
Geography of Slavery, University of Virginia (accessed 2014-09-01). (1)</p>
      </div>
      <div>
        <p>Runaway Advertisement, Virginia Gazette (1776-05-25); The
Geography of Slavery, University of Virginia (accessed 2014-09-01). (2)</p>
      </div>
    </body>
  </html>
</Datum>
<Datum fieldName="Costa.ID" status="new" fieldID="216">
  <Value>002448</Value>
</Datum>
<Datum fieldName="workflow.Status" status="new" fieldID="3">
  <Value>Advertisement</Value>
</Datum>
<Datum fieldName="workflow.Notes" status="new" fieldID="116">
  <Value>not in Windley</Value>
</Datum>
<Datum fieldName="workflow.AddDate" status="new" fieldID="137">
  <Value>2014-12-02-05:00</Value>
</Datum>
</Metadata>
<Links>
  <Link targetID="287667" type="Flourish"/>
  <Link targetID="284773" type="Occupation"/>
  <Link targetID="bio_GOS002448" type="Biography"/>
  <Link targetID="pfe_O_d2e55063" type="Servitude: Owned by"/>
  <?link-certainty high?><Link targetID="pfe_O_d2e57198" type="Subscriber"/>
  <?link-certainty high?></Links>
</Object>
<Object xml:id="bio_GOS002448" collectionID="c365">
  <Name>Mary-anne (captured slave, 1776-03-01)</Name>
  <Metadata>
    <Datum fieldName="biographies.SubjectName" status="new" fieldID="6">
      <Value>Mary-anne</Value>
    </Datum>
    <Datum fieldName="biographies.Bio" fieldID="4" status="new">
      <html xmlns="http://www.w3.org/1999/xhtml">
        <head>
          <title>Advertisements</title>
          <link
href="http://dcdarma.dataformat.com/Resources/RadEditorConfigs/CSS/XHTML_PFE_II.css"
          type="text/css"
          rel="stylesheet"/>

```

```

        </head>
        <body>
            <div>
                <h3>Virginia Gazette (Purdie), Williamsburg, March 1, 1776.
(1)</h3>
                <p>WILLIAMSBURG, Feb. 13, 1776. COMMITTED to James City jail, a
negro woman who calls herself MARY-ANNE, and says she belongs to Anthony Walker at
Petersburg. The owner is desired to take her away, and pay charges.</p>
                <p>
                    <em>not in Windley</em>
                </p>
            </div>
            <div>
                <h3>Virginia Gazette (Dixon & Hunter), Williamsburg, May 25,
1776. (2)</h3>
                <p>COMMITTED to Spotsylvania Gaol, a Negro Woman who says her
name is MARY-ANNE, and that she belongs to Mr. Walker of Petersburg. The Owner is
desired to prove his Property, pay Charges, and take her away. JACOB WHITLER,
Gaoler.</p>
                <p>
                    <em>not in Windley</em>
                </p>
            </div>
        </body>
    </html>
</Datum>
<Datum fieldName="biographies.Cite" fieldID="7" status="new">
    <html xmlns="http://www.w3.org/1999/xhtml">
        <head>
            <title>Advertisements</title>
            <link
href="http://dcdarma.dataformat.com/Resources/RadEditorConfigs/CSS/XHTML_PFE_II.css"
                type="text/css"
                rel="stylesheet"/>
        </head>
        <body>
            <div>
                <p>Runaway Advertisement, Virginia Gazette (1776-03-01); The
Geography of Slavery, University of Virginia (accessed 2014-09-01). (1)</p>
            </div>
            <div>
                <p>Runaway Advertisement, Virginia Gazette (1776-05-25); The
Geography of Slavery, University of Virginia (accessed 2014-09-01). (2)</p>
            </div>
        </body>
    </html>
</Datum>
<Datum fieldName="Costa.DescType" status="new" fieldID="211">
    <Value>single</Value>
</Datum>
<Datum fieldName="Costa.Nametype" status="new" fieldID="173">
    <Value>Captured Slave</Value>
</Datum>
<Datum fieldName="Costa.personID" status="new" fieldID="203">
    <Value>002448</Value>
</Datum>
<Datum fieldName="Costa.Image" status="new" fieldID="194">
    <Value>v2113.jpg // v2055.jpg</Value>
</Datum>

```

```
<Datum fieldName="Costa.Title" status="new" fieldID="200">
  <Value>Virginia Gazette</Value>
</Datum>
<Datum fieldName="Costa.docID" status="new" fieldID="204">
  <Value>v1776032113 // v1776052055</Value>
</Datum>
<Datum fieldName="Costa.NewspaperCity" status="new" fieldID="195">
  <Value>Williamsburg</Value>
</Datum>
<Datum fieldName="Costa.Publisher" status="new" fieldID="201">
  <Value>Purdie // Dixon & Hunter</Value>
</Datum>
<Datum fieldName="Costa.Source" status="new" fieldID="206">
  <Value>Virginia Gazette (Purdie), Williamsburg, March 1, 1776. // Virginia
Gazette (Dixon & Hunter), Williamsburg, May 25, 1776.</Value>
</Datum>
<Datum fieldName="Costa.Dateline" status="new" fieldID="210">
  <Value>1776-03-01 // 1776-05-25</Value>
</Datum>
<Datum fieldName="Costa.GoSLink" status="new" fieldID="217">
  <Value>http://www2.vcdh.virginia.edu/gos/</Value>
</Datum>
<Datum fieldName="workflow.BioType" status="new" fieldID="172">
  <Value>Advertisement</Value>
</Datum>
<Datum fieldName="workflow.AddDate" status="new" fieldID="137">
  <Value>2014-10-27</Value>
</Datum>
<Datum fieldName="Costa.Subscriber" status="new" fieldID="199">
  <Value> // Gaoler, Jacob Whitler,</Value>
</Datum>
<Datum fieldName="Costa.Residence" status="new" fieldID="196">
  <Value>Petersburg</Value>
</Datum>
```



```
</Metadata>
</Object>
</ObjectCollection>
```

Links

The following links are made:

- A link from the PFE record to a Biography record.
- For slave/servant/deserter records only: (candidate) links of the appropriate type are made to owner, previous owner, or subscriber records. Since we have done these for records generated from all (single/mixed/group) description types, the program denotes certainty levels. It is recorded in a processing instruction following each link that may be referenced by reporting processes. The criteria is currently:
 1. Links made from references already in the source data is 'high'
 2. Links generated for persons in a non group description are 'high'
 3. Links generated for persons in a group/mixed description that has one subscriber and no owners (695 out of the 1140 descriptions of this type) is 'high'
 4. All others are 'low'

Links between owners/subscribers and other classes are typed as follows: Servitude: Owned by, Servitude: Former Owner, Servitude: Indentured, Servitude: Served, Subscriber.

And for record candidates generated from non-group descriptions:

- Links to flourish records for persons meeting the following criteria: runaway, pregnant, captured, bearing punishment marks
- For slave records only: a link to an occupation record

Known Data Issues

General

- Data is not validated by any schema
- Markup practices vary by hand and over time
- Structural markup inconsistencies diagnosed during the course of work were fixed in the working copy.

Specific

- Some <desc> elements are not typed properly. See v1782120068 (should be group not single). Query `//desc[@type = 'single'][count(descendant::age) > 1]` finds some of these. Have been fixed in the working copy of the source data.
- There are some empty name elements. See ids 005611, 005898, 005768. Records for these have the name 'No Name in Data'
- It's possible that the original project did not entirely disambiguate slaves, at least in a few cases. Check name reports of non-owner persons to see if all records with multiple names add up.
- Age ranges handled poorly, both min and max age marked separately instead of just marking the range. See v1770081673. See also errors in age markup as at v1796080004. Signs of programmatic errors from the past. Fixed/addressed in working copy.
- Pregnancy sometimes not marked. Saw <physical>she is with Child</physical> in data instead of <characteristics type='pregnant'>. Can be located with `//physical[contains(lower-case(.), 'with child')]`. Modified conversion query criteria to pick up on these.
- Data for subscribers and owners is inconsistent: some marked <subscriber> or <owner> with no link(s) to slaves. Others marked <name type='subscriber' slaveIDRef=''> or <name type='subscriber' slaveIDRefs=''>. The working copy has been changed to consistently use @slaveIDRefs since that is all that is nec. ID values are whitespace separated so they can be parsed. Another problem with owners/subscribers is that none of them have IDs, therefore the possibility exists there may be multiple parties with the same names; or, the same party with name variants. Also, since they did not use authority records to store the links to the slaves they own, the link data could have errors or need to be compiled from multiple instances of the same owner/subscriber (i.e. the information could be different each time it is entered).
- <home> does not always mark a home, see v1737041655 (where <home> is used to mark the place of a robbery, not the home of the owner—or is it?). See also v1745121399 where the 'court-house' (where he ran away from) is marked as home.
- <geography> has a broader use than <home> (e.g. is used to mark jails), so is less reliable, but is all we have in records from a certain time frame/source, i.e. it was used in lieu of <home> for that an other purposes. This is the case in records where <name> is used to mark subscribers and owners.
- Subscribers not always individually marked, see 'John Lewis, and John Heylyn' at v1737041298. Note also in that record that John Lewis is mentioned as an owner but is NOT marked as owner.

- Sometimes non-name data included in name elements, as in <powner> content like 'Jamers Porter in Maryland', 'Mrs. Bunting, the widow of....'
- There was a duplicate id in the source data records, see v1775021956. Retained, appended '_A' to id value.
- <image> and <notes> sometimes wrongly embedded in <desc> instead of as siblings (8 instances). Corrected in working copy of the source.
- Fixed instances (3) of subscriber being outside <desc>.
- Found completely unwrapped description at v1803070099.
- Instances of notes found inside desc (3) fixed. (Should follow as sibling.)
- Inconsistent use of <dateLine> vs. <dateline> (corrected to <dateline> in working copy).
- Links to slaves formalized in some data, but not in all

Statistics

Persons with programmatically created records (1 PFE, 1 Bio for each):

Class	Count
Owner/Subscriber	4,853
Female Slaves	596
Male Slave	3,242
Female Servants	83
Male Servants	1,125
Female Deserters	1
Male Deserters	520
Unknown	2
Free Persons	14
Misc	70
Total	10,486*

* Actual total is slightly less owing to membership in multiple classes by a small number of records (which were conflated).

Records programmatically conflated:

- Owner candidates: 7,052. Post conflation: 4,853 (as above).
- See report FinalReport_Owners_Spreadsheet.xlsx for details on method.

Remaining Work

After the [data attached](#) is loaded to Darma:

- Review owner records:
 - check those that were conflated from a name match only (filter attached report).
 - check those missing links (just three)
 - review names for further conflation
- Review records for other classes of persons (slaves, servants, etc.):
 - review low certainty links
 - Check name reports of non-owner persons to see if all records with multiple names add up (i.e. conflated records).

Misc. Notes, Questions, and Ideas

- We did find non-owner persons under different names (with the same ID)
- We might be able to assist the name review of owners with a report:
 - where one whole name matches another partially (e.g. john mayo is a part of john mayo esq.)
 - where last name matches
- There's no metadata distinguishing the role(s) of owner records. Should there be?

- The discussion of retrieving more metadata from group records can be taken up if needed.
- We could add `Ran Away With` links between slaves and misc persons. There may be few enough of these to consider doing manually during review and correction.
- We could add an occupation of `free person` for `amb` record types.
- The slave records are the place to disambiguate owners, as the articles will be there, along with the links, which will show all the possible owner names
- ALL we have to disambiguate owners on (beyond the link and name processes already run): residence (not usu.), article metadata: paper, date, publisher; text in article (e.g. if says 'brother of...' or some such)