XSLT Ingest Example: Appendix C

Appendix C

The next topic concerns the problems in the source data.

```
<ROW id='1011001'>
        <NETID>dag065</NETID>
        <FNAME>David</FNAME>
        <MNAME>augustus</MNAME>                               0
        <LNAME>Green</LNAME>
        <DEGREE>Masters</DEGREE>
        <YEAR>1985</YEAR>
        <INSTITUTION>University of Kansas</INSTITUTION>
        <MAJOR>Oriental Art History</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-10-24 13:10:59.0</LAST_UPDATED>
</ROW>
<ROW id='1011002'>
        <NETID>dah3507</NETID>
        <FNAME>Don</FNAME>
        <MNAME>A</MNAME>
        <LNAME>Horsham</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>2005</YEAR>
        <INSTITUTION>Cornell University</INSTITUTION>
        <MAJOR></MAJOR>                                       1
        <MINOR></MINOR>
        <LAST_UPDATED>2012-08-29 17:09:42.0</LAST_UPDATED>
</ROW>
<ROW id='1011003'>
        <NETID>dah3507</NETID>
        <FNAME>Don</FNAME>
        <MNAME>A</MNAME>
        <LNAME>Horsham</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>1996</YEAR>
        <INSTITUTION>Smith College</INSTITUTION>
        <MAJOR>Architecture</MAJOR>
        <MINOR>Italian Language</MINOR>
        <LAST_UPDATED>2012-08-29 17:10:40.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 1 - Figure 18

- [F18H0] The middle name is all in lower case. If name parts are all shifted to upper case or all shifted to lower case before comparison then this becomes unimportant.
- [F18H1] The **MAJOR** data field is empty. This is needed to properly label the **EducationalTraining** instance and so this row will be rejected.

```
<ROW id='1011004'>
        <NETID>dhv010</NETID>
        <FNAME>Denise</FNAME>
        <MNAME>H</MNAME>
        <LNAME>Valencia</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>1992</YEAR>
        <INSTITUTION>SUNY OSwego</INSTITUTION>                    0
        <MAJOR>Vocational Technical Education</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-07-03 13:56:40.0</LAST_UPDATED>
</ROW>
<ROW id='1011005'>
        <NETID></NETID>                                          1
        <FNAME>Denise</FNAME>
        <MNAME>Hortense</MNAME>
        <LNAME>Valencia</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>1988</YEAR>
        <INSTITUTION>SUNY Cortland</INSTITUTION>
        <MAJOR>Education</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-07-03 13:56:40.0</LAST_UPDATED>
</ROW>
<ROW id='1011006'>
        <NETID>dag065</NETID>
        <FNAME>David</FNAME>
        <MNAME>Augustus</MNAME>
        <LNAME>Green</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>1980</YEAR>
        <INSTITUTION>State University at Albany</INSTITUTION>
        <MAJOR>Studio Art</MAJOR>
        <MINOR>English</MINOR>
        <LAST_UPDATED>2012-10-24 13:10:59.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 2 - Figure 19

- [F19H0] The __INSTITUTION__ field has mixed case characters which will cause a mismatch if character for character equality is required for organizational matching. However, if all characters are shifted to the same case before comparison this will not be a problem.
- [F19H1] The __NETID__ field is missing. This causes uncertainty when comparing names. Since it is not uncommon for there to be several distinct people with exactly the same name parts, a token like netid is perfect for disambiguation of people. This is because netid is guaranteed to be unique even when a person's name has changed due to choice, marriage or divorce. A missing netid weakens the association of a degree record and a person.

```
<ROW id='1011007'>
        <NETID>cad2616</NETID>
        <FNAME>Cathrine</FNAME>
        <MNAME>a</MNAME>                                               0
        <LNAME>Dale</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>1999</YEAR>
        <INSTITUTION>SUNY Oswego</INSTITUTION>
        <MAJOR>Environmental and Organismal Biology</MAJOR>
        <MINOR>Business Administration</MINOR>
        <LAST_UPDATED>2012-12-07 12:18:49.0</LAST_UPDATED>
</ROW>
<ROW id='1011008'>
        <NETID>cad2616</NETID>
        <FNAME>Cathrine</FNAME>
        <MNAME>A</MNAME>
        <LNAME>Dale</LNAME>
        <DEGREE>Associate</DEGREE>
        <YEAR>1995</YEAR>
        <INSTITUTION>Tompkins Cortland   Community College</INSTITUTION> 1
        <MAJOR>Liberal Arts and Sciences</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-07 12:19:17.0</LAST_UPDATED>
</ROW>
<ROW id='1011009'>
        <NETID>ask317</NETID>
        <FNAME>Andrea</FNAME>
        <MNAME>S</MNAME>
        <LNAME>Killian</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>2004</YEAR>
        <INSTITUTION>Cornell University</INSTITUTION>
        <MAJOR>Animal Science</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-09-10 08:17:53.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 3 - Figure 20

- [F20H0] See Figure 18, highlight 0.
- [F20H1] Extra embedded whitespace. The XPATH function normalize-space deals effectively with this sort of issue.

```
<ROW id='1011010'>
        <NETID>ask317</NETID>
        <FNAME>Andrea   </FNAME>                                        0
        <MNAME>S</MNAME>
        <LNAME>Killian</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>2012</YEAR>
        <INSTITUTION>Pennsylvania State University</INSTITUTION>
        <MAJOR>Business Administration</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-09-10 08:17:27.0</LAST_UPDATED>
</ROW>
<ROW id='1011011'>
        <NETID></NETID>                                                 1
        <FNAME>Andrea</FNAME>
        <MNAME>S</MNAME>
        <LNAME>Killian</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>2006</YEAR>
        <INSTITUTION>Michigan State</INSTITUTION>
        <MAJOR>Youth Development</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-09-10 08:15:56.0</LAST_UPDATED>
</ROW>
<ROW id='1011012'>
        <NETID>arf72</NETID>
        <FNAME>Arthur</FNAME>
        <MNAME>R</MNAME>
        <LNAME>Fuller</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>2009</YEAR>
        <INSTITUTION>Kansas State University</INSTITUTION>
        <MAJOR>Public Administration</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-04 11:05:12.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 4 - Figure 21

- [F21H0] Trailing whitespace.
- [F21H1] Missing netid. See Figure 19 Highlight 1.

```
<ROW id='1011013'>
        <NETID>sb016</NETID>
        <FNAME>Suzy</FNAME>
        <MNAME></MNAME>
        <LNAME>Beltaine</LNAME>
        <DEGREE>Associate</DEGREE>
        <YEAR>1992</YEAR>
        <INSTITUTION>Tompkins Cortland Community College</INSTITUTION>
        <MAJOR>Liberal Arts</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-07-03 13:56:40.0</LAST_UPDATED>
</ROW>
<ROW id='1011014'>
        <NETID>arf72</NETID>
        <FNAME>Arthur</FNAME>
        <MNAME>R</MNAME>
        <LNAME>Fuller</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>2004</YEAR>
        <INSTITUTION>Kansas State University</INSTITUTION>
        <MAJOR></MAJOR>                                              0
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-04 11:05:12.0</LAST_UPDATED>
</ROW>
<ROW id='1011015'>
        <NETID></NETID>                                             1
        <FNAME>Arthur</FNAME>
        <MNAME>R</MNAME>
        <LNAME>Fuller</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>2004</YEAR>
        <INSTITUTION>Kansas State University</INSTITUTION>
        <MAJOR>Animal Husbandry</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-04 11:05:12.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 5 - Figure 22

- [F22H0] Same as [F18H1]. This row will also be rejected.
- [F22H1] Same as [F19H1].

```xml
<ROW id='1011016'>
        <NETID>arf27</NETID>
        <FNAME>Arthur</FNAME>
        <MNAME>R</MNAME>
        <LNAME>Fuller</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>2004</YEAR>
        <INSTITUTION>Kansas State University</INSTITUTION>
        <MAJOR>Animal Husbandry</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-04 11:05:12.0</LAST_UPDATED>
</ROW>
<ROW id='1011017'>
        <NETID>arf33</NETID>
        <FNAME>Arthur</FNAME>
        <MNAME>R</MNAME>
        <LNAME>Fuller</LNAME>
        <DEGREE>Master</DEGREE>
        <YEAR>2006</YEAR>
        <INSTITUTION>Kansas State University</INSTITUTION>
        <MAJOR>Animal Husbandry</MAJOR>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-12-04 11:05:12.0</LAST_UPDATED>
</ROW>
<ROW id='1011018'>
        <NETID>dah256</NETID>
        <FNAME>Don</FNAME>
        <MNAME>A</MNAME>
        <LNAME>Horsham</LNAME>
        <DEGREE>Bachelor</DEGREE>
        <YEAR>1990</YEAR>
        <INSTITUTION>Smyth College</INSTITUTION>
        <MAJOR>Oriental Art History</MAJOR>
        <MINOR>Farsi</MINOR>
        <LAST_UPDATED>2012-08-29 17:10:40.0</LAST_UPDATED>
</ROW>
```

EduRS.xml Fragment 6 - Figure 23