XSLT Ingest Example: Source Data

The Source Data

Start Next

Suppose that a query is run against a RDBMS to get our source data. Example rows from the XML result set source file are shown below in Figure 1. The entire file can be found in the result set file **EduRS.xml** found in the**example/source** directory of the online example. Also see Appendix C. There is frequently more than one row for a given person and the data may or may not be complete as indicated in the Figure 1 below.

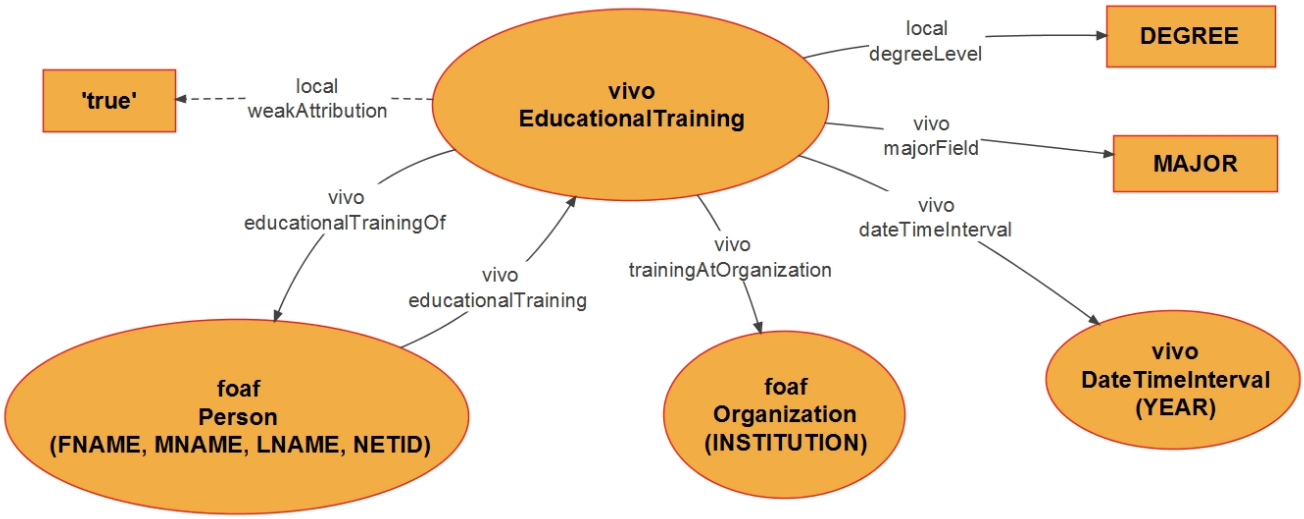
Notice also, that there is a 'row id' number (guaranteed to be unique) associated with each row and a unique **NETID**associated with most rows (see the result set file description for a full list of source abnormalities in the Appendix C). We are interested only in rows that have a first name (**FNAME**), a last name (**LNAME**), a degree level (**DEGREE**), a major (**MAJOR**), a year (**YEAR**) and an institution (**INSTITUTION**). We will try to deal appropriately with missing netids (**NETID**), character case discrepancies, and extra white space. We will ignore the **MINOR** and**LAST_UPDATED** data. Part of the process will involve result set row rejection when there is no sense in its inclusion. For example, notice that the second record has no listed major. Clearly this row is a good candidate for rejection just as a missing institution would also be.

Row rejection could be performed as part of our SQL query. This is also true of correcting rows with case and whitespace issues. However, in an institutional environment with complex data stewardship, we may not be in control of the query and just have to deal with whatever result set that we are given. In our work we have used this as an opportunity to pass the rows that we reject back to the managers of the original data source for repair.

```
<ROWS>
<ROW id='1011001'>
        <NETID>dag065</NETID>
         <FNAME>David</FNAME>
         <MNAME>augustus</MNAME>
         <LNAME>Crccn</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>
        <MINOR></MINOR>
        <LAST_UPDATED>2012-08-29 17:09:42.0</LAST_UPDATED>
</ROW>
```

Result Set Fragment Figure 1

The following figure illustrates the RDF data model we intend to populate using the source data from the tags in the result set rows. Note that the predicate **local:weakAttribution** is conditional.



RDF Data Model Figure 2

In Figure 2 the rectangles represent data; the oval shapes represent object classes and the labeled arrows predicates. A namespace abbreviation is also given e.g. **vivo** or **foaf**. The string **local:degreeLevel**, like **local:weakAttribution**, is the name of a yet to be defined predicate - more later. The central object will be of type **vivo:EducationalTraining** as mentioned above.

Start Next