IMPLEMENTATION OF DYNAMIC DATA CITATION

James Duncan and Jennifer Pontius

Feb 13, 2017

james.duncan@uvm.edu, www.uvm.edu/vmc
MANY DISCIPLINES, MANY CONTRIBUTORS

VMC houses any data related to forest ecosystem condition, regardless of affiliation or discipline
WHY WE NEED IT

- Continually evolving datasets
- Some errors not caught till next field season
- Frequent reporting and publishing
DYNAMIC DATA CITATION – FEATURES NEEDED

- Light footprint on database resources
- Works on top of existing catalog and metadata
- Works in an institutionally managed PHP/MySQL environment
- User-driven control of what quantity of change constitutes a version
- Integration with management portal
- Track granular changes in data
TWO FORMS OF VERSIONING

- Dynamic Subsetting
  - Storing unique and repeatable query linked to correct state of the dataset

- Provenance tracking
  - Storing successive states of the dataset
TECHNOLOGY STACK AND CATALOG

- Linux, Apache, MySQL database, PHP, JQuery
- Implements a Project Catalog, where projects contain datasets, and datasets are a single data object with metadata
- Data Objects can be database tables, binary files, images, etc.
  - Only tables can be subsetted
  - All have provenance tracked
USER WORKFLOW– DATA EDITING

- Modify a dataset (append, replace, edit)
  - Changes tracked
  - Original data table unchanged

- Commit to version, assign name
  - Computes result hash (table pkid, col names, first col data) and query hash
  - Updates data table to new state
  - Assign DOI and URL
  - Commits version
USER WORKFLOW - SUBSETTING

- Define query using a builder or typing in SQL
- Commit to version, assign name
  - Computes result hash (table pkid, col names, first col data) and query hash
  - Updates data table to new state
  - Assign DOI and URL
  - Commits version
USER WORKFLOW - RECOVERING

- Restore previous version
  - Creates new version table from current data table state
  - Compiles query steps from VersionStep
  - Walks table back to prior state using stored SQL
DEALING WITH THE UNVERSIONED

- Allowing users to turn off versioning
  - Securing those steps already versioned
  - Allowing more changes to dataset without tracking them

- Allowing users to turn on versioning
  - Basically, not allowed
## TABLE STRUCTURE

### Version Info Table

<table>
<thead>
<tr>
<th>Version ID</th>
<th>Dataset ID</th>
<th>Version Name</th>
<th>Version ID</th>
<th>Person ID</th>
<th>Query Hash</th>
<th>Result Hash</th>
<th>Time stamp</th>
<th>Version Type</th>
<th>Parent Version</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>23456</td>
<td>3525</td>
<td>Version 1.5</td>
<td>3</td>
<td></td>
<td>....</td>
<td>....</td>
<td>....</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23574</td>
<td>3525</td>
<td>Unsaved</td>
<td>-1</td>
<td></td>
<td>....</td>
<td>NULL</td>
<td>....</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Step Tracking Table (Child of Version Info)

<table>
<thead>
<tr>
<th>Step ID</th>
<th>Version PID</th>
<th>Step Type</th>
<th>Forward</th>
<th>Backward</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>983245</td>
<td>23574</td>
<td>delete</td>
<td>DELETE FROM...</td>
<td>INSERT INTO...</td>
<td>1</td>
</tr>
<tr>
<td>983245</td>
<td>23574</td>
<td>update</td>
<td>UPDATE SET site=&quot;Winhal...&quot;</td>
<td>UPDATE SET site=&quot;Lye Brook&quot;...</td>
<td>2</td>
</tr>
</tbody>
</table>
IMPLEMENTATION CHALLENGES AND QUESTIONS

Challenges

✓ Large updates
✓ Re-creation of past versions, in terms of garbage collection and storage
  - Binary files

Questions

Query uniqueness checking and query normalization

Efficient but effective results hashing strategies
✓ Linear progression of data, versus branching network
STILL TO COME

- Garbage collection
- Better handling of result hashes
- Web-based data editing validation
ACKNOWLEDGMENTS

- Adoption seed funding - MacArthur Foundation and the Research Data Alliance

- The US Forest Service State and Private Forestry program for core operational funding of the VMC

- Fran Berman, Yolanda Meleco and the other adopters who have been sharing their experiences.

- All the VMC cooperators that contribute
THANK YOU!