



# IMPLEMENTATION OF DYNAMIC DATA CITATION

James Duncan and Jennifer Pontius

Feb 13, 2017

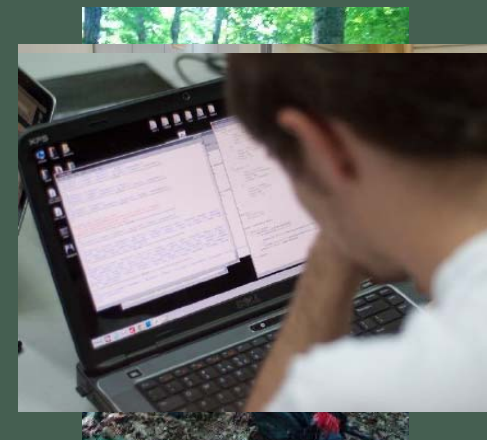
[james.duncan@uvm.edu](mailto:james.duncan@uvm.edu), [www.uvm.edu/vmc](http://www.uvm.edu/vmc)



# Vermont Monitoring Cooperative

Ecosystem Monitoring  
Collaborator Network

Data Archive, Access and  
Integration



**Soil**



**Water**



**Forest**



**Air**



**Wildlife**

# MANY DISCIPLINES, MANY CONTRIBUTORS

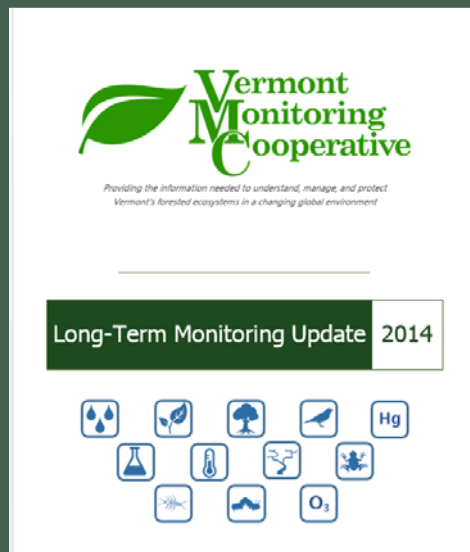
VMC houses any data related to forest ecosystem condition, regardless of affiliation or discipline





# 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

Management Home

Manage My Profile

MANAGE MY CONTENT

Project Management Home

Project Datasets

Dataset Management Home

Information

Dataset Attributes

Methods

Sampling

Site Description

Documents / Images

Notes

Versions

Data Management

PROJECT: TESTING PROJECT (ID=555)

DATASET: BURTON ISLAND MET (ID=2217)

## Manage Dataset Versions

Manage versions and dynamic subsets of the dataset. You can edit the data through your browser, upload a subset of the data for citation and documentation purposes. You can also download past versions of the dataset.

This versioning system is built upon the [Dynamic Data Citation best practices recommendations](#) of the International Geospatial Standards Committee.

Version 1 - **2217\_INIT** (created 2017-02-13 09:29:34, creator = Mike Finnegan) [VIEW](#)

Version 2 - **DATA\_2010** (created 2017-02-13 09:34:38, creator = Mike Finnegan) [VIEW](#)

Version 3 - **DATA\_2011** (created 2017-02-13 09:37:05, creator = Mike Finnegan) [VIEW](#)

Subset of Version 3 - **EARLY MORNING RAIN** (created 2017-02-13 09:40:05, creator = James Duncan) [VIEW](#)

Select An Option For Creating A New Version Of The Dataset

[Edit Data Online](#) [Define a Subset of the Data](#)

## Editing data for Burton Island Met

[<< Back to Dataset Versions](#)

Here you can make changes to your data. You can edit data, add new rows, or delete rows. A new version isn't created until you commit it. Once you have made all the changes, assign a version name and officially save the snapshot of the data.

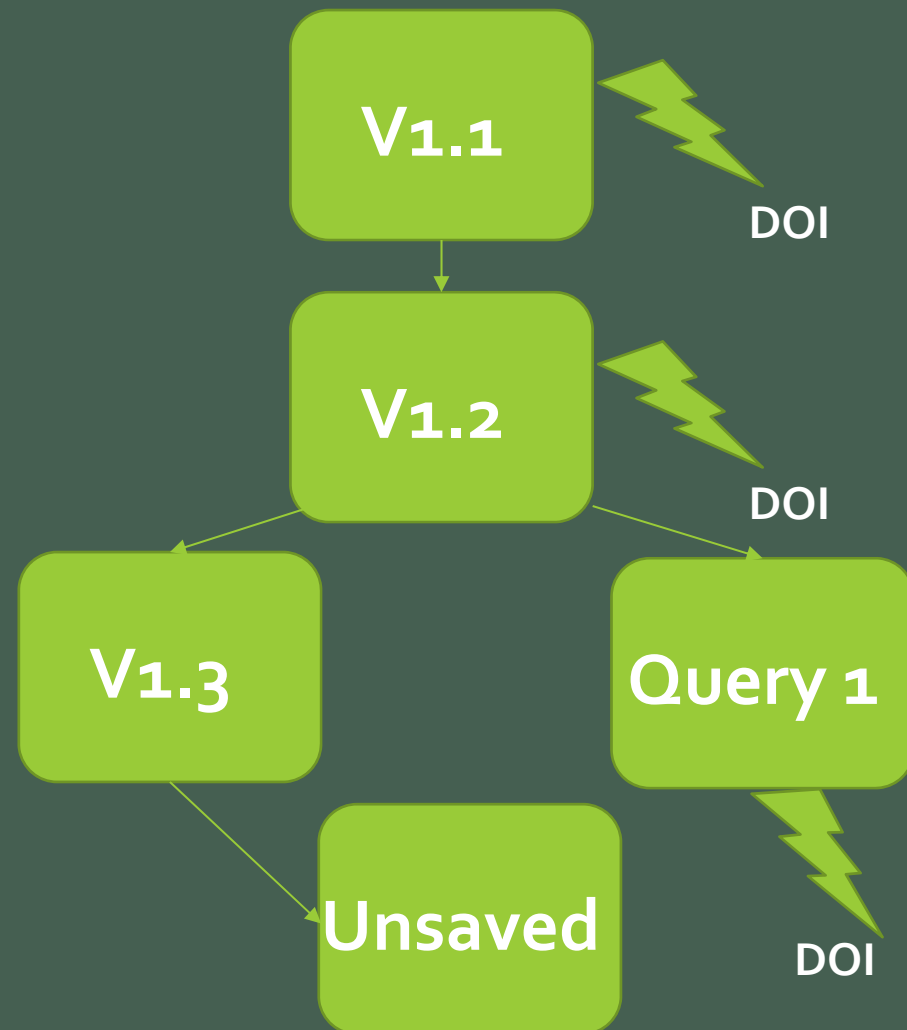
[Show changes](#) [Save to a new version](#)

	2010-09-30 21:30:00	2017-09-13	273	2130	2010-01-01	16.42
	2010-09-30 21:45:00	2017-09-13	273	2145	2010-01-01	16.69
	2010-09-30 22:00:00	2017-09-13	273	2200	2010-01-01	16.77
	2010-09-30 22:15:00	2017-09-13	273	2215	2010-01-01	16.82



# 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

Version ID	Dataset ID	Version Name	Version ID	Person ID	Query Hash	Result Hash	Time stamp	Version Type	Parent Version	DOI
23456	3525	Version 1.5	3		....	....	....			
23574	3525	Unsaved	-1		....	NULL	....			

## Step Tracking Table (Child of Version Info)

Step ID	Version PID	Step Type	Forward	Backward	Order
983245	23574	delete	DELETE FROM...	INSERT INTO...	1
983245	23574	update	UPDATE SET site="Winhal l"...	UPDATE SET site="Lye Brook"...	2

# 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

db_table_pk *	RECORDTIME	AIRTEMP	RELHUMID	BAROPRESS	SAMPLEDAY	SAMPLETIME	PYRAN
1	2010-09-30 07:30:00	15.650	86.6	1009	273	730	
2	2010-09-30 07:45:00	15.700	86.1	1009	273	745	
3	2010-09-30 08:00:00	15.630	86.6	1009	273	800	
4	2010-09-30 08:15:00	15.460	89.0	1009	273	815	
5	2010-09-30 08:30:00	15.530	89.2	1009	273	830	
6	2010-09-30 08:45:00	15.430	90.2	1008	273	845	46.17
7	2010-09-30 09:00:00	15.480	90.5	1008	273	900	58.48
8	2010-09-30 09:15:00	15.600	90.3	1008	273	915	92.30
9	2010-09-30 09:30:00	15.750	90.5	1007	273	930	92.30
10	2010-09-30 09:45:00	15.870	90.9	1007	273	945	113.90
11	2010-09-30 10:00:00	16.060	90.4	1007	273	1000	92.30
12	2010-09-30 10:15:00	16.240	90.1	1007	273	1015	126.20
13	2010-09-30 10:30:00	16.410	89.7	1007	273	1030	123.10
14	2010-09-30 10:45:00	16.500	89.9	1006	273	1045	144.70
15	2010-09-30 11:00:00	16.390	90.4	1006	273	1100	109.30
16	2010-09-30 11:15:00	16.290	92.3	1006	273	1115	103.10
17	2010-09-30 11:30:00	16.230	91.7	1006	273	1130	92.30



# ACKNOWLEDGMENTS

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



MacArthur Foundation

- 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!