

# Tweets Metadata

May 4, 2015

CS 4624 - Multimedia, Hypertext and Information Access

Department of Computer Science

Virginia Polytechnic Institute and State University

Blacksburg, VA 24061

# Project Personnel

## Principal Investigator:

- Dr. Edward Fox, Virginia Tech Department of Computer Science
  - email: [fox@vt.edu](mailto:fox@vt.edu)

## Clients:

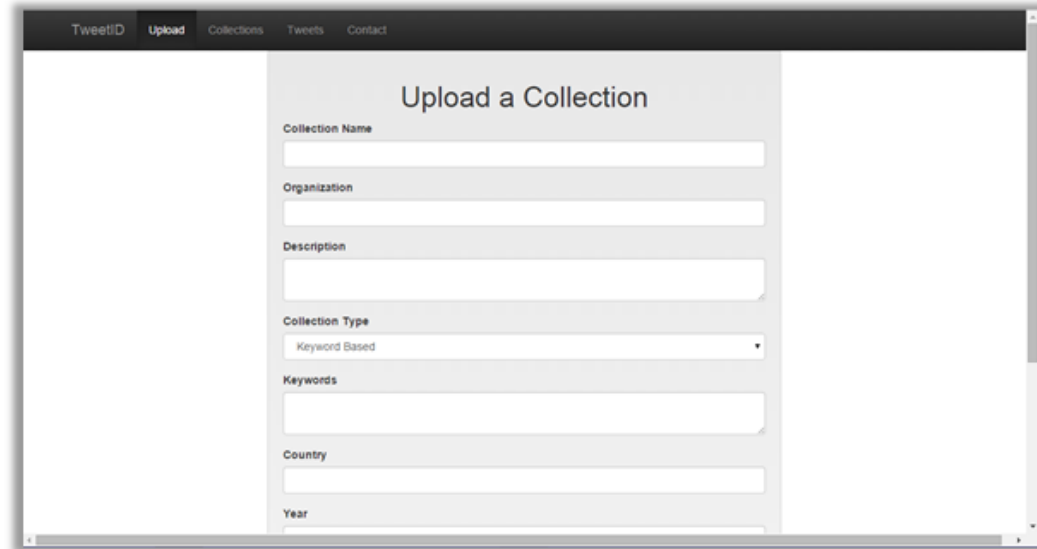
- Mohamed Magdy, Virginia Tech Department of Computer Science
  - email: [mmagdy@vt.edu](mailto:mmagdy@vt.edu)

## Student Team Members:

- Chris Conley
- Alex Druckenbrod
- Karl Meyer
- Samuel Muggleworth

# // Background

- CTRnet, QCRI, etc. collect and archive tweets surrounding events
- Desired some central database for tweet collections
- Michael Shuffett started project in 2014
- Implemented to support CTRnet and QCRI data formats



The image shows a screenshot of a web browser displaying the 'Upload a Collection' form. The browser's address bar shows 'TweetID' and the page title is 'Upload a Collection'. The form is titled 'Upload a Collection' and contains several input fields: 'Collection Name', 'Organization', 'Description', 'Collection Type' (a dropdown menu currently showing 'Keyword Based'), 'Keywords', 'Country', and 'Year'. The form is styled with a light gray background and white text.

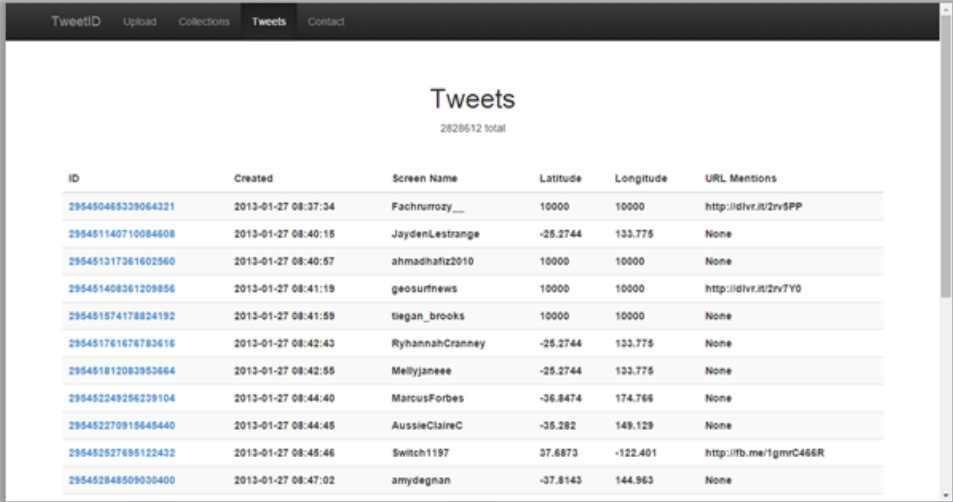
Figure 1: Michael Shuffett's Upload information page

# // Goals

- Tweet and tweet-collection metadata standard
  - Enable collection sharing and consistency
  - Standard exists at collection and tweet levels
- To implement methods for merging such collections
- To create a web-app tool that will allow a user to upload new collections and execute merging of collections.

# // Code Base, Technologies

- Shuffetts TweetID tool as starting point
  - Implemented upload and merging specifically geared towards IDEAL and QCRI collections
- Technologies used:
  - Python 2.7.9 - scripting
  - SQLite - database
  - HTML5/CSS/Bootstrap - web development
  - jQuery/Flask-script - client side scripting
  - jinja2 - dynamic HTML templating/rendering
  - Flask/WTFORM - forms and upload



Tweets  
2828512 total

ID	Created	Screen Name	Latitude	Longitude	URL Mentions
<a href="#">295450465339064321</a>	2013-01-27 08:37:34	Fachrumrozy_	10000	10000	<a href="#">http://divr.it/2rv5PP</a>
<a href="#">295451140710084608</a>	2013-01-27 08:40:15	JaydenLestrangle	-25.2744	133.775	None
<a href="#">295451317361602560</a>	2013-01-27 08:40:57	ahmadhafiz2010	10000	10000	None
<a href="#">295451408361209856</a>	2013-01-27 08:41:19	geosurfnews	10000	10000	<a href="#">http://divr.it/2rv7Y0</a>
<a href="#">295451574178824192</a>	2013-01-27 08:41:59	tiegan_brooks	10000	10000	None
<a href="#">295451761676782616</a>	2013-01-27 08:42:43	RyhannahCranney	-25.2744	133.775	None
<a href="#">295451812083953664</a>	2013-01-27 08:42:55	Mellyjaneee	-25.2744	133.775	None
<a href="#">295452349296239104</a>	2013-01-27 08:44:40	MarcusForbes	-36.8474	174.766	None
<a href="#">295452270915645440</a>	2013-01-27 08:44:45	AussieClaireC	-35.282	149.129	None
<a href="#">295452527695122432</a>	2013-01-27 08:45:46	Switch1197	37.6873	-122.401	<a href="#">http://fb.me/1gmrC466R</a>
<a href="#">295452848509030400</a>	2013-01-27 08:47:02	amydegnan	-37.8143	144.963	None

Figure 2: Shuffett's tweet listing page

# // Discoveries

- All files converted to .tsv
- expected a 'text' field, but never writes it to database
- Using collection.name as principal key in database
- Supports only:
  - One schema type
    - 9 fields: 7 data fields + 2 \N fields\*
  - Two file types
    - .csv or .tsv

# // What we did

- Determined standard
  - What data is necessary?
- Updated Interface
- Created collection.id field as principal key
- Allowed multiple schemas/formats
  - Request schema from uploader
  - Map and record only standardized data to our schema.
  - Original merge process unchanged

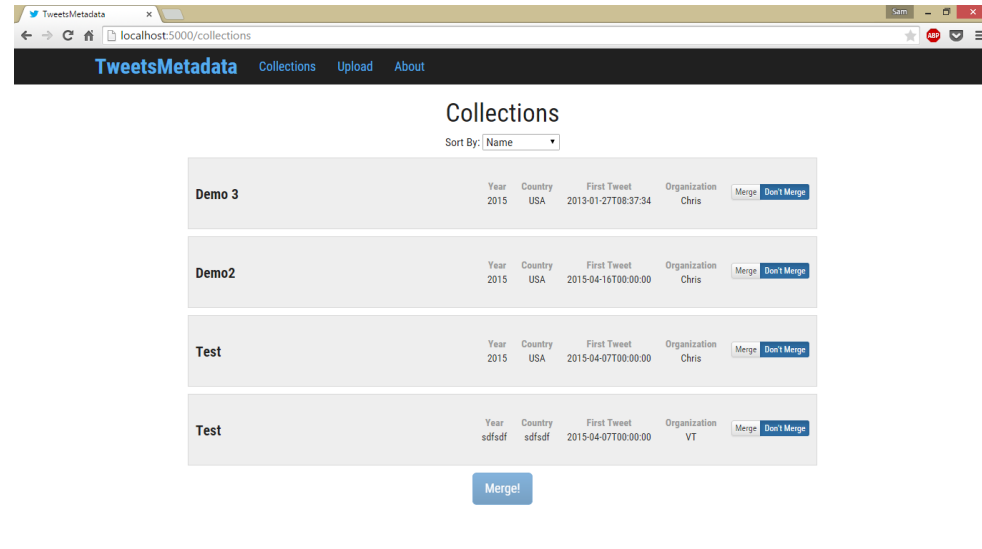


Figure 3: Our Updated Collection Page

# // Standards

	id	created_at	screen_name	latitude	longitude	url_mentions
	Filter	Filter	Filter	Filter	Filter	Filter
1	1	April 7, 2015	chrisc93	51.5033630	-0.1276250	www.google.com
2	2	April 24, 2015	chrisc93	51.5033630	-0.1276250	www.google.com
3	3	April 16, 2015	chrisc93	51.5033630	-0.1276250	www.google.com

Figure 4: Tweet metadata as captured from our test database.

	id	name	organization	description	collection_type	keywords	country	year	tags	first_tweet_date	last_tweet_date
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	0	Test	TweetsMetadata	This is a test	keyword	none	USA	2015	tags	2015-04-07T00:...	2015-04-07T00:...
2	1	Test	VT	sdad	keyword	sdfasdf	sdfsdf	sdfsdf	sdfs	2015-04-07T00:...	2015-04-07T00:...

Figure 5: Collection metadata as captured from our test database.



# // Lessons Learned

- Attempted to switch to Python3
  - Learned about incompatible libraries
- Commenting code is very important
  - Continuing a project is much easier with documentation
- Be aware of database size
  - Better to implement for large database from the start
  - I.e. avoid iterating through the entire database

# // Future Plans

- Implement autonomous capabilities
  - More complex schemas
  - Less manual mapping/schema info
- Become largely more flexible
  - Less reliant on specific formatting
  - Accept more file types
- Persist Merges
- Scale the technologies appropriately depending on size of project.

# // References

- "Developer Policy." *Developer Policy*. Twitter, Inc., 22 Oct. 2014. Web. <<https://dev.twitter.com/overview/terms/policy>>.
- Shuffett, Michael. "Twitter Metadata." *Twitter Metadata*. VTechWorks, 10 May 2014. Web. 09 Feb. 2015. <<https://vtechworks.lib.vt.edu/handle/10919/47949>>.
- "Twitter Terms of Service." Twitter, Inc., 8 Sept. 2014. Web. <<https://twitter.com/tos?lang=en>>.
- "QCRI." Qatar Computing Research Institute, 2015. Web. <<http://www.qcri.com>>.
- "CTRnet - Events Archive." Virginia Polytechnic Institute and State University. Web. <<http://www.ctrnet.net/>>



??? Questions ???