

CS6604

Twitter Metadata

Michael Shuffett
Virginia Tech
Blacksburg, VA
shuffett@cs.vt.edu

Primary Client: Mohamed Magdy, (mmagdy@vt.edu)

Background

- ◇ Large number of tweet collections about events
 - ◇ CTRNet
 - ◇ IDEAL
 - ◇ QCRI

No collection level metadata standard

+

No easy merging solution

=



















Poor collaboration support

Project Goals

- ◇ Develop metadata standards for tweet collections
 - ◇ start, end timestamps
 - ◇ geographic coverage
 - ◇ details of how collection was prepared
 - ◇ Filtering
 - ◇ Cleaning
 - ◇ Enriching
- ◇ Create software package that merges and describes multiple collections

IDEAL Project@VT - Tweet Archive 1

Total number of archived tweets: 205,600,960

Archive ID	Keyword / Hashtag	Description	Tags	Screen Name	Count	Create Time	
1	#egypt	Tweets for Egyptian revolution		ctrnet	10531836	Sun, 07 Oct 2012 18:28:33-0400	
2	#libya			ctrnet	1281309	Sun, 07 Oct 2012 18:28:52-0400	
3	#blacksburg			ctrnet	28397	Sun, 07 Oct 2012 18:29:01-0400	
4	#jan25			ctrnet	823655	Sun, 07 Oct 2012 18:29:08-0400	
5	#bahrain			ctrnet	18639693	Sun, 07 Oct 2012 18:29:18-0400	
6	#yemen			ctrnet	1664897	Sun, 07 Oct 2012 18:29:26-0400	
7	japan earthquake			ctrnet	506321	Sun, 07 Oct 2012 18:29:37-0400	
8	#syria			ctrnet	14380321	Sun, 07 Oct 2012 18:30:00-0400	
9	OccupyWallStreet			ctrnet	304980	Sun, 07 Oct 2012 18:31:26-0400	
10	#nrv	new river valley (blacksburg) related tweets		ctrnet	54119	Sun, 07 Oct 2012 18:32:04-0400	
11	virginia tech			ctrnet	677067	Sun, 07 Oct 2012 18:32:26-0400	
12	iran earthquake			ctrnet	108327	Sun, 07 Oct 2012 18:33:15-0400	
13	diabetes	health category		ctrnet	7146040	Sun, 07 Oct 2012 18:33:37-0400	
14	heart attack	health category		ctrnet	14147056	Sun, 07 Oct 2012 18:33:56-0400	
15	foursquare			ctrnet	40659079	Sun, 07 Oct 2012 18:34:18-0400	
16	#Isaac	hurricane Isaac in Aug. 2012		ctrnet	79881	Sun, 07 Oct 2012 18:34:48-0400	
17	turkey syria	violence between Turkey and Syria in Oct. 2012		ctrnet	621956	Sun, 07 Oct 2012 18:35:27-0400	
18	emergency preparedness			ctrnet	148337	Sun, 07 Oct 2012 21:49:13-0400	

#egypt - Tweets for Egyptian revolution

Created on Sun, 07 Oct 2012 18:28:33 -0400 and total number of tweets = 10531836

tags:

START DATE	END DATE	ORDER	VIEW LIMIT	FROM USER	TWEET TEXT	LANGUAGE	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>
			10				remove RTs

query

HTML Permalink = <http://spare05.dlib.vt.edu/archive.php?id=1>
RSS Permalink = <http://spare05.dlib.vt.edu/rss.php?id=1>
Excel Permalink = <http://spare05.dlib.vt.edu/excel.php?id=1>
Simple Table Permalink = <http://spare05.dlib.vt.edu/table.php?id=1>
JSON API = <http://spare05.dlib.vt.edu/apiGetTweets.php?id=1>



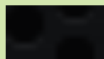
@shosholovemido RT @Shorouk_News: شيك وشبعات .. واتق من نفسه.. شيك وشبعات في مجلس إدارتها #CBCeXtra <http://t.co/TxhzP2m5tk>.. «فيديو.. بعد لقاءهم #المشير.. إعلاميون في وصف «#السياسي»: «متواضع وهادئ.. واتق من نفسه.. شيك وشبعات في مجلس إدارتها»

Sun May 04 07:20:39 +0000 2014 - tweet id 462854302869241856 - #1



@CBCeXtra هذا الصباح:الهيئة العامة للتأمينات تختار ممثلي أصحاب المعاشات في مجلس إدارتها #CBCeXtra #Eqypt

Sun May 04 07:20:37 +0000 2014 - tweet id 462854291200688128 - #2



@egy_updates RT @leloveluck: Sabahi announces 'emergency' economic plan: <http://t.co/MEcSibXUv2> involves 21 projects, including a solar energy plan. #Eq...

Sun May 04 07:20:29 +0000 2014 - tweet id 462854259282042880 - #3



@Ghorabz «فهمي»: واشنطن طلبت مني عدم الحديث عن زيارتي لـ #روسيا «#Eqypt هي واشنطن دي هي اللي بت.....كك» <http://t.co/DLA3Xcsqk1>

Sun May 04 07:20:20 +0000 2014 - tweet id 462854219725549568 - #4

```
1 {
2   "archive_info": {
3     "id": "84",
4     "keyword": "oklahoma tornado",
5     "description": "",
6     "tags": "",
7     "screen_name": "ctrnet",
8     "user_id": "247438424",
9     "count": "454458",
10    "create_time": "1369084318"
11  },
12  "tweets": [
13    {
14      "archivesource": "twitter-search",
15      "text": "Two Killed by Tornado in Small Oklahoma Town - A tornado killed two people in... http://t.co/
16      "to_user_id": "",
17      "from_user": "GiterDoneNews",
18      "id": "462844779617734656",
19      "from_user_id": "1403574307",
20      "iso_language_code": "en",
21      "source": "<a href=\"http://www.ajaymatharu.com/\" rel=\"nofollow\">Tweet Old Post</a>",
22      "profile_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
23      "geo_type": "",
24      "geo_coordinates_0": "0",
25      "geo_coordinates_1": "0",
26      "created_at": "Sun May 04 06:42:49 +0000 2014",
27      "time": "1399185769"
28    },
29    ...
30  ]
31 }
```

1 | 336620771126415361 Devastating Before-And-After Pictures Of One Of The Elementary Schools Wrecked In The Oklahoma
Tornado: One of... http://t.co/F056I2wBYl 2013-05-20 23:13:40 Mktg4theFuture 10000 10000 http://bit.
ly/16GvMrA \N \N
2 | 336620770472112128 Praying for Oklahoma! 2013-05-20 23:13:40 meg_hurrell 38.8381 -104.821 \N \N \N
3 | 336620771042553857 pray for Oklahoma 2013-05-20 23:13:40 angelafraser87 10000 10000 \N \N \N
4 | 336620771243872256 Ray Manzarek dead of bile duct cancer. Gigantic tornado in OK. Asinine legal case in FL. I'm
going to call for a do-over on today. 2013-05-20 23:13:40 galduric 42.2774 -83.7333 \N \N \N
5 | 336620771017359360 My thoughts & prayers with everyone in moore oklahoma :(Sad day): 2013-05-20 23:13:40
vthompson2010 38.8816 -77.091 \N \N \N
6 | 336620771449380864 RT @fatbellybella: Prayers and lots of light for Oklahoma. 2013-05-20 23:13:40 SassySmartSgRHO
10000 10000 \N \N \N
7 | 336620771629727744 All I want is a pain pill right now! This weather is stressing me out everywhere we go to avoid
the tornado it goes that way! #nothanks ? 2013-05-20 23:13:40 SarahMarae95 10000 10000 \N \N \N
8 | 336620771843661824 I live in Tulsa Ok, Please Pray for Oklahoma 2013-05-20 23:13:40 kellicolleen 36.1333
-95.9756 \N \N \N
9 | 336620772422463488 RT @PPRedCross: Concerned abt loved ones in #tornado damage area? Go to http://t.co/yIN2jt7gK0
then click on SEARCH. 2013-05-20 23:13:40 mlsimmons 10000 10000 http://Redcross.org/safeandwell \N \N
10 | 336620772145627137 ??75????????????????????????????????(???)???\My heart goes out to everyone
in Oklahoma.Pray for Oklahoma! 2013-05-20 23:13:40 rose_hisayo 10000 10000 \N \N \N
11 | 336620772539920384 For those asking about helping those hit in, #Oklahoma City" to donate. Text "redcross" to
90999 or visit http://t.co/DzQOLGzInl 2013-05-20 23:13:40 robynd323 42.4485 -73.254 http://redcross.org \N \N
12 | 336620772720250880 RT @nicholenordeman: Oklahoma. Heartsick.\\Jesus, please be unimaginably near. Mercy, Lord. 2013-
05-20 23:13:40 kristindjustice 10000 10000 \N \N \N
13 | 336620772598628352 Great way to start the week! On my way to Oklahoma! 2013-05-20 23:13:40 AJ_Nennig6 10000 10000
\n \N \N
14 | 336620772615405568 "@MichaelSkolnik: 75 to 100 horses killed at Orr Farms in Moore, Oklahoma. (via @kfor) #tornado"
so sad :(2013-05-20 23:13:40 Chillimouth 10000 10000 \N \N \N
15 | 336620771738796032 RT @JHarden13: Praying for everyone in Oklahoma City. This is crazy! 2013-05-20 23:13:40
CoooooLin11 35.4866 -96.6859 \N \N \N
16 | 336620773001269248 RT @GeriCanedo: Pray for Oklahoma. 2013-05-20 23:13:41 neetusandher 10000 10000 \N \N
\n

Tweet-Level Metadata

- ◇ All tweets originate from API
- ◇ Leverage standard present in API
 - ◇ <https://dev.twitter.com/docs/platform-objects/tweets>
 - ◇ <https://dev.twitter.com/docs/api/1/get/statuses/show/:id>
- ◇ Namespace JSON

Collection-Level Metadata

Descriptive
Metadata

- What

Provenance

- Who
- When
- How

Collection-Level Metadata

Descriptive
Metadata

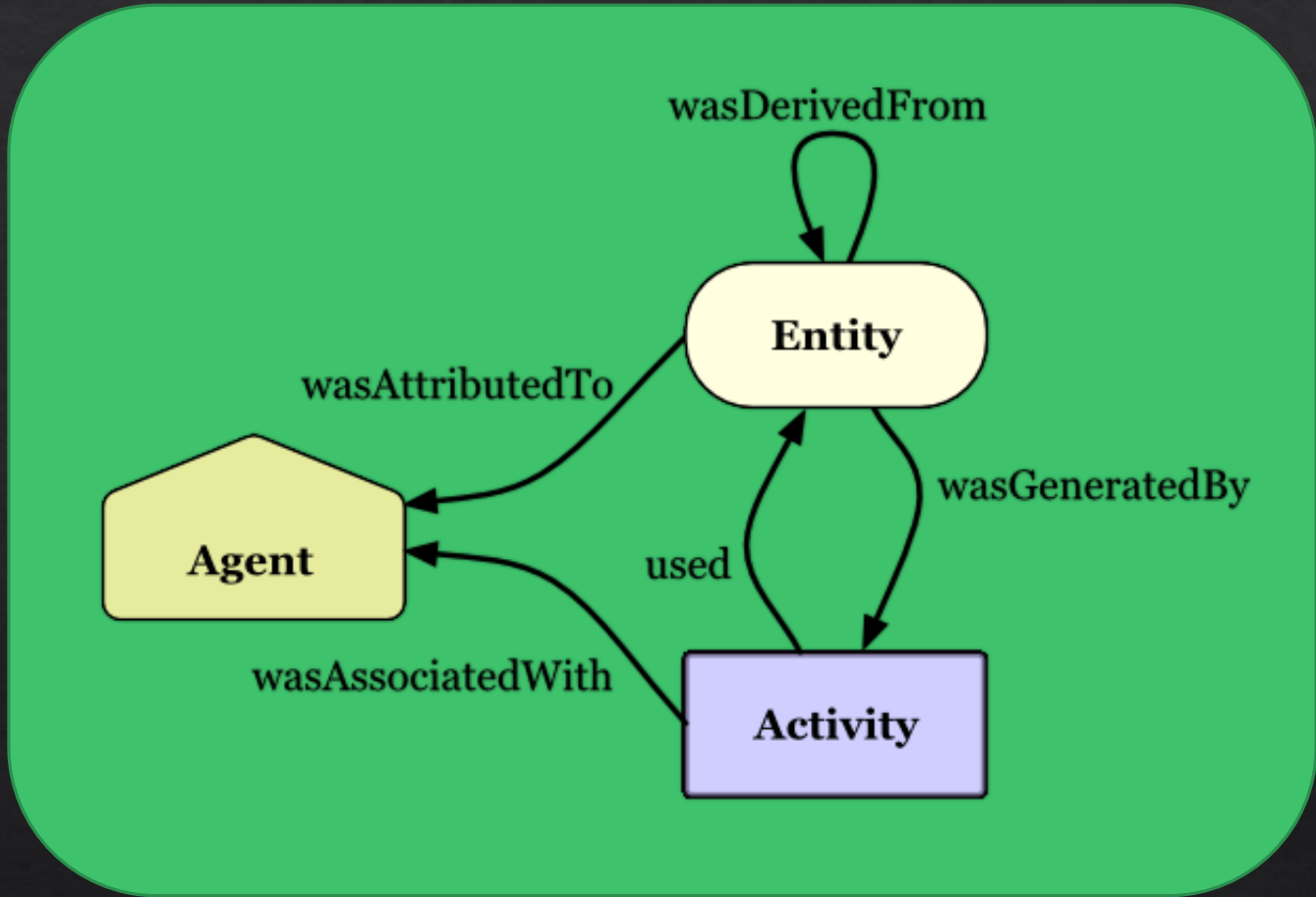
- What

◇ Dublin Core Terms

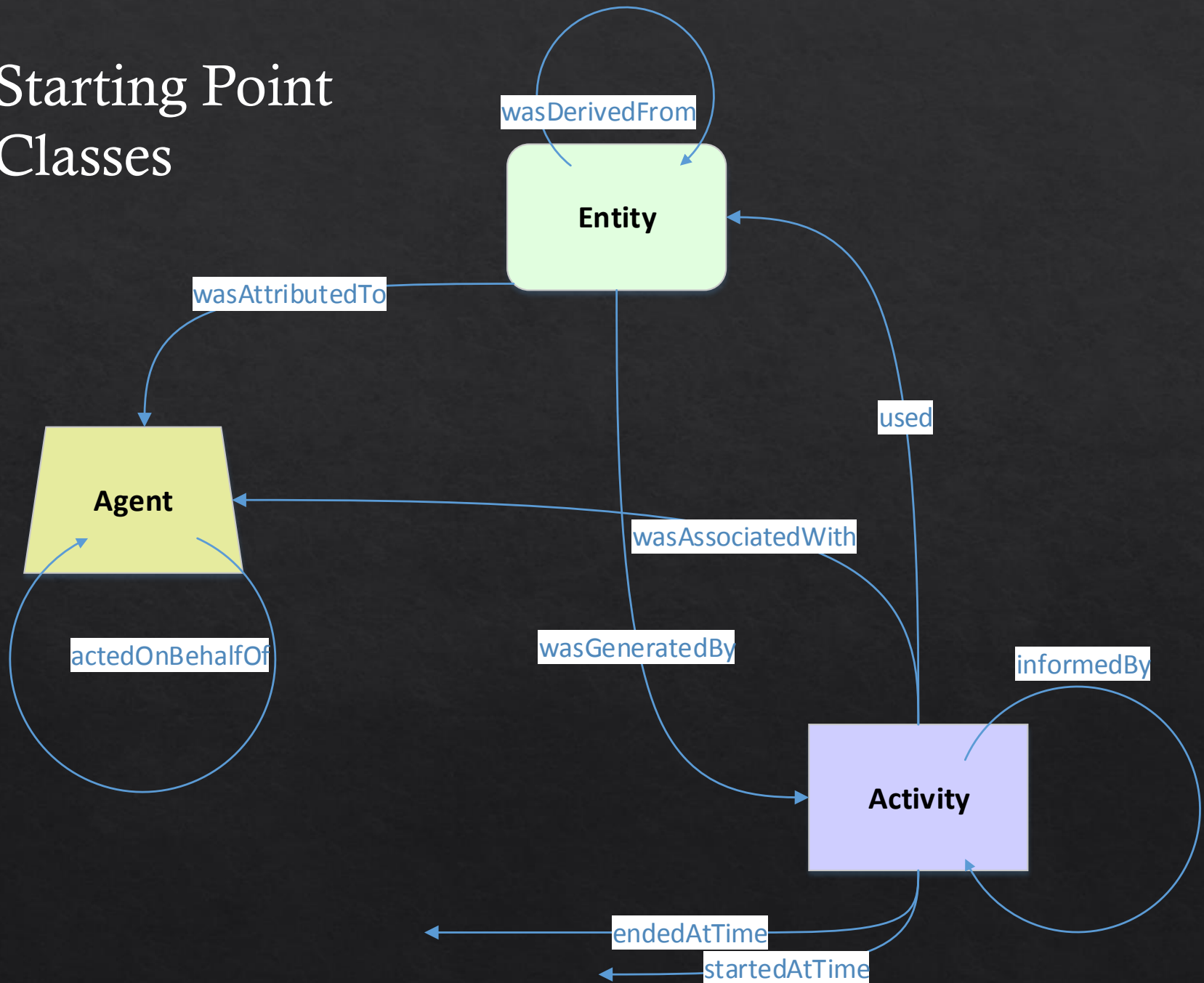
◇ title, description

Provenance

- Who
- When
- How



Starting Point Classes



```
1 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
2 @prefix dcterms: <http://purl.org/dc/terms/> .
3 @prefix prov: <http://www.w3.org/ns/prov#> .
4 @prefix tid: <http://www.tweetid.org/tweet/> .
5 @prefix : <http://example.com/> .
6
7 :vt-oklahoma-tornado
8   a prov:Collection;
9   dcterms:title "Oklahoma Tornado"
10  dcterms:description "Tweets about Oklahoma Tornado";
11
12  prov:hadMember
13    tid:295450465339064321, tid:3954504653346069327, tid:5829504653346069829, ...;
14
15  prov:wasAttributedTo :mark;
16  prov:wasAttributedTo :vt;
17
18  prov:atLocation "US OK" # ISO 3166-2
19
20  prov:wasGeneratedBy :keyword-query;
21 .
22
23 :mark a prov:Person, prov:Agent, prov:Entity .
24
25 :vt a prov:Organization, prov:Agent, prov:Entity .
26
27 :keyword-query
28   a prov:Activity;
29   prov:startedAtTime "2013-06-07T16:28:17Z"^^xsd:dateTime;
30   prov:endedAtTime "2013-06-07T16:28:17Z"^^xsd:dateTime;
31   prov:used :keyword-list;
32 .
33
34 :keyword-list
35   a prov:Collection;
36   prov:hadMember
37     "oklahoma tornado", "oklahoma storm", "#okc flood" ...;
38 .
```

Summary of Collection-Level Metadata

- ◇ Dublin Core
 - ◇ Title, Description
- ◇ PROV-O
 - ◇ Starting Point Classes
 - ◇ collection, organization, hadMember, atLocation
- ◇ ISO 3166-2 for locations
- ◇ W3/XMLSchema#dateTime