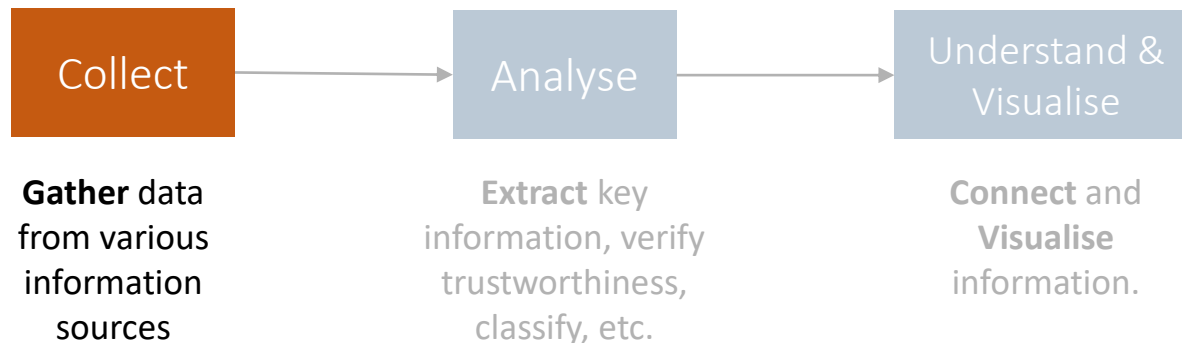


SMASAC - Data Collection and Filtering

GRÉGOIRE BUREL, MAYANK KEJRIWAL AND
PRASHANT KHARE



Social Media Usage during Crises

- **Social media usage during crises varies** depending on:

1. Type of information shared (e.g., affected individuals, caution and advice, donation or volunteering, message of support, etc.). (Olteanu et al, 2014)
2. Type of content shared (e.g., text, images, videos, links).
3. Content source (e.g., news organisation or journalist, eyewitness, government, NGO, company or for-profit organisation). (Olteanu et al, 2014)
4. Target audience (e.g., general public, other organisation, followers, friends/family).
5. Type of social media platform used (e.g., Facebook, Twitter, etc.)

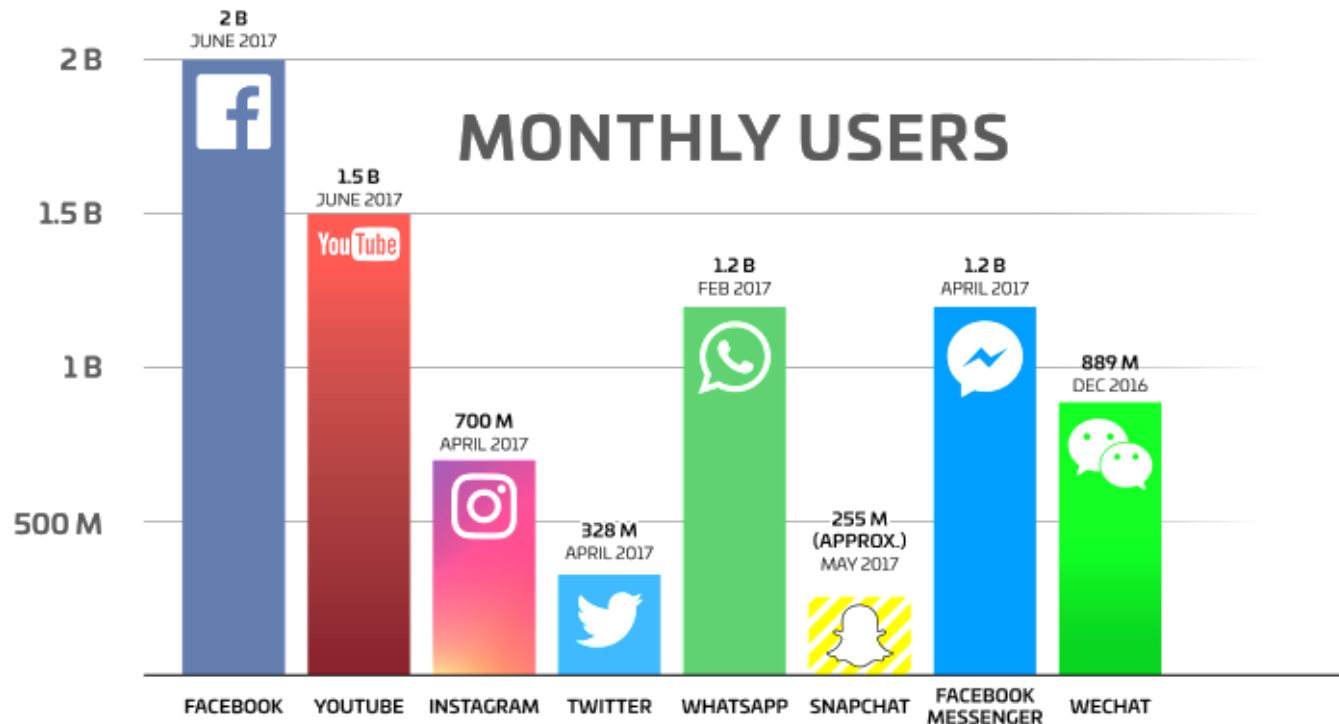


Access to Social Media

- Automatic access to social media data **can be restricted** in different ways:
 1. Public / Non-public data: Most social media websites do not allow access to the information posted unless reading access is given explicitly by the information creator.
 2. Query restrictions: Data access can be limited by **API restrictions** (e.g., rate limiting, query allowance).
 3. Data Sampling: High velocity data is sometimes **sampled** by social media companies. As result, it is only possible to retrieve a portion of the relevant information.
 4. Query Filtering: Often data is retrieved using **query parameters** (e.g., keywords, geolocation, etc.). Access to relevant data can be negatively impacted.











Most used Social Media Platforms







































Source: <https://techcrunch.com/2017/06/27/facebook-2-billion-users/>

Most used Social Media Platforms

	Platform	API Access	Usage	Audience	
	Facebook	Low	High	Mostly Personal	
	YouTube	High	Medium	General	
	Instagram	High	-	Personal/General	
	Twitter	High	High	Mostly General	
	WhatsApp	NA	High	Personal	
	Snapchat	Low	-	Personal	
	Facebook Messenger	Low	-	Personal	
	WeChat	Low	-	Personal	

Worldwide Social Media App Usage

Top App / Website	US	Canada	UK	France	Germany	Australia	Japan	China	India	Brazil	Mexico	South Africa	Saudi Arabia	Dubai	Jordan	Abu Dhabi
1																
2																
3																
4																
5																

Data Collection/Processing Methods

- Manual Methods

- Go to social media website or solicit contributions.
- Check for relevant content.
- **Copy/Paste** content into spreadsheet.
- Annotate spreadsheet.
- Visualise the data.
- Take action.

- Automatic Methods

- **Use platform APIs**
- Perform query using the API (e.g., keyword search)
- Automatically populate a database or spreadsheet.
- Annotate manually or automatically the data.
- Visualise the data
- Take action

Data Collection/Processing Methods

- Manual Methods

- Go to social media website or solicit contributions.
- Check for relevant content.
- **Copy/Paste** content into spreadsheet.
- Annotate spreadsheet.
- Visualise the data.
- Take action.

- Automatic Methods

- **Use platform APIs**
- Perform query using the API (e.g., keyword search)
- Automatically populate a database or spreadsheet.
- Annotate manually or automatically the data.
- Visualise the data
- Take action

Automatic Data Collection APIs (Twitter)

- Automatic data collection generally relies on JSON APIs and OAuth credentials. For example, for Twitter, you need to:
 - Create a Twitter account (<https://twitter.com>).
 - Obtain an OAuth access credentials (i.e., access token, access secret, consumer key and consumer secret) (<https://apps.twitter.com/app/new>).
 - Use Search API for collecting tweets (<https://developer.twitter.com>).
 - Save Tweets in JSON or other format for later analysis.



```
Example
$ gem install twurl
$ twurl authorize --consumer-key key \
                  --consumer-secret secret
$ twurl authorize --consumer-key key \
                  --consumer-secret secret
$ twurl "/1.1/search/tweets.json?q=earthquake"
{
  "statuses": [
    {
      "created_at": "Mon Feb 12 10:58:42 +0000 2018",
      "id": 963004430915289100,
      "id_str": "963004430915289088",
      "text": "RT @Independent: Magnitude 4.4 earthquake strikes near Beijing https://t.co/usyGy0kyA",
      "truncated": false,
      "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [
          {
            "screen_name": "Independent",
            "name": "The Independent",
            "id": 16973333,
            "id_str": "16973333",
            "indices": [
              3,
              15
            ]
          }
        ]
      },
      "urls": [
        {
          "url": "https://t.co/usyGy0kyA",
          "expanded_url": "http://www.independent.co.uk/news/world/asia/beijing-earthquake-latest-update-china-magnitude-hebei-tremors-capital-a8206336.html",
          "display_url": "independent.co.uk/news/world/asi...",
          "indices": [
            63,
            96
          ]
        }
      ]
    }
  ]
}
```

Twitter Data and Crises



Wildfire



People of NSW, be careful because there's fires spreading! Stay safe everyone!

CRISIS



Hundreds of volunteers in Mexico tried to unearth children they hoped were still alive beneath a school's ruins



Earthquake

Floods



Two trucks and one car in the water after a road collapse at Hwy 287 and Dillon. #cowx #boulderflood

Volume

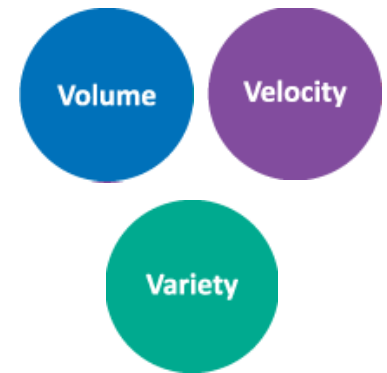
Variety

Velocity

Veracity

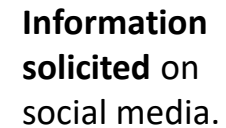
Accessing Relevant Information

- Information overload:
 - During crises, a flood of data gets generated. For example:
 - Over a million tweets generated during Hurricane Harvey 2017.
 - 500% increase in the tweets bandwidth during 2011 Japan earthquake.
 - The characteristics of social media posts such as short length, colloquialism, syntactic issues pose additional challenges of processing the data.
 - Almost impossible to manually absorb and process the sheer volume.



How do we **separate crisis-related content** from irrelevant information?

How to **ensure** that a wide range of **crisis related topics across diverse crisis situations** are **filtered in**?

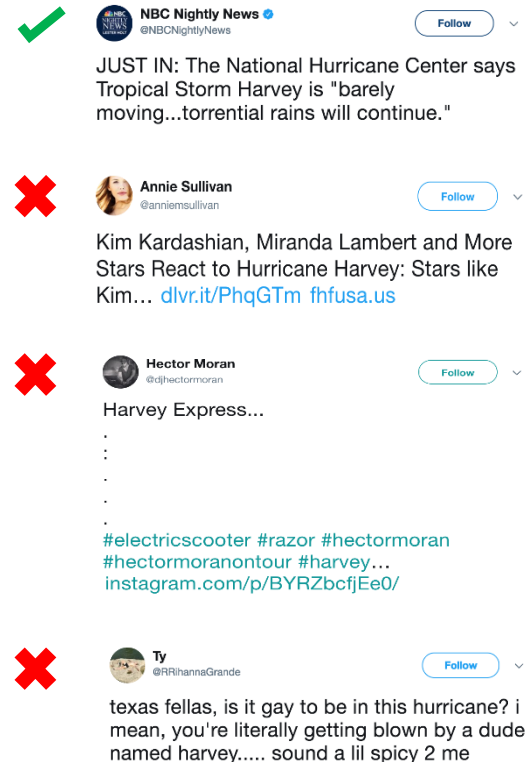


Google Sheets

12

Data Curation (Filtering)

- Content Curation:
 - Methods and processes used for **gathering relevant information** relevant to a particular topic of interest.
- Information Filtering Systems:
 - Approaches for removing irrelevant and unwanted information from an information stream using partially-automated or computerized methods **prior to presentation to a human user.**



The screenshot displays four tweets from a social media feed, each preceded by a green checkmark or a red X to indicate its status in a curation process.

- Green Checkmark:** A tweet from NBC Nightly News (@NBCNightlyNews) with the text: "JUST IN: The National Hurricane Center says Tropical Storm Harvey is 'barely moving...torrential rains will continue.'" This tweet is marked as relevant.
- Red X:** A tweet from Annie Sullivan (@anniesullivan) with the text: "Kim Kardashian, Miranda Lambert and More Stars React to Hurricane Harvey: Stars like Kim... [dlvr.it/PhqGTm](#) [fhfusa.us](#)". This tweet is marked as irrelevant or unwanted.
- Red X:** A tweet from Hector Moran (@jhectormoran) with the text: "Harvey Express...
.
.
.
.
.
#electricscooter #razor #hectormoran
#hectormoranontour #harvey...
[instagram.com/p/BYRZbcfJEe0/](#)". This tweet is marked as irrelevant or unwanted.
- Red X:** A tweet from Ty (@RHhannaGrande) with the text: "texas fellas, is it gay to be in this hurricane? i mean, you're literally getting blown by a dude named harvey..... sound a lil spicy 2 me". This tweet is marked as irrelevant or unwanted.

Filtering Methods



CrisisLex

- Query filtering using social media APIs:
 - Use hashtags, keywords, crisis specific phrases or lexicon (impacted location name, canonical form of disaster name - e.g. *Hurricane Harvey*).
- Post collection filtering (before or after storage):
 - Text search (similar to above).
 - Semantic search (requires entity extraction)*.
 - Automatic categorisation / Tagging (clustering approaches, topic modelling, Machine Learning models)*.

* Entity extraction methods and automatic categorization is discussed later on in this tutorial

flood crisis, victims, flood
victims, flood powerful,
powerful storms, hoisted
flood, storms amazing,
explosion, amazing rescue,
rescue women, flood cost,
counts flood, toll rises,
braces river, river peaks,
crisis deepens, prayers,
thoughts prayers, affected
tornado, affected, death
toll, tornado relief, photos
flood, water rises, toll,
flood waters, flood appeal,
victims explosion, bombing
suspect, massive explosion,
affected areas, praying
victims, injured, please
join, join praying, prayers
people, redcross, text
redcross, visiting flood,
lurches fire, video
explosion, deepens death,
aid, help flood,
died explosions, marathon
explosions, flood relief

Filtering using Social Media APIs

- Twitter API:
 - Use hashtags, keywords, crisis specific phrases or lexicon (e.g., CrisisLex lexicon).
 - Use Geolocation constrains (but many documents do not have geolocation information).



Filtering using APIs: Information Vs Noise

- Hashtag Hijacking:

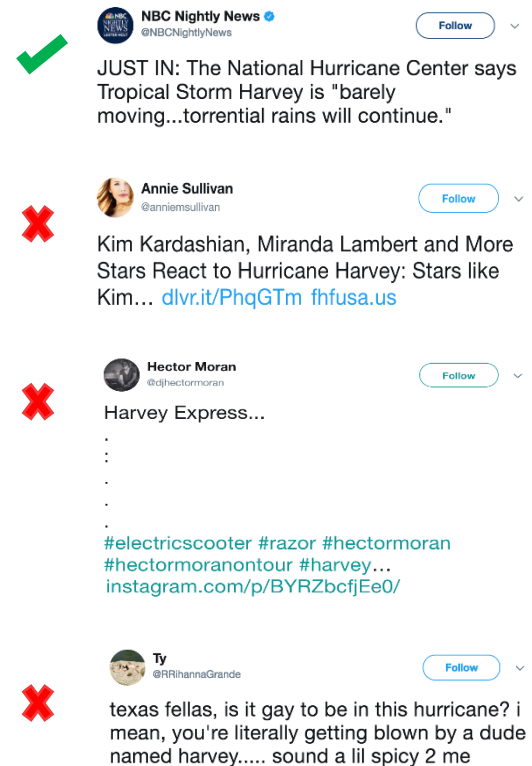
- Hashtags, Keywords can be a convenient way to create a broad set of data, but they **do not always reflect the crisis related information**.

- Overly specific query terms:

- The real crisis information is contained in a **very diverse** form.
- Document **geolocation** may be **inaccurate or missing**.

- Unknown situation / insufficient context:

- The subjects in the information range from health/well-being of affected individuals, infrastructure, donations etc.
- Obviously, various concepts together form a context which reflects relevance with the crisis situations.



✓ **NBC Nightly News** @NBCNightlyNews
JUST IN: The National Hurricane Center says Tropical Storm Harvey is "barely moving...torrential rains will continue."

✗ **Annie Sullivan** @anniesullivan
Kim Kardashian, Miranda Lambert and More Stars React to Hurricane Harvey: Stars like Kim... dvr.it/PhqGTm fhfusa.us

✗ **Hector Moran** @djhectormoran
Harvey Express...
.
.
.
.
.
#electricscooter #razor #hectormoran
#hectormoranontour #harvey...
instagram.com/p/BYRZbcfjEe0/

✗ **Ty** @RRihannaGrande
texas fellas, is it gay to be in this hurricane? i mean, you're literally getting blown by a dude named harvey..... sound a lil spicy 2 me

Post Collection Filtering

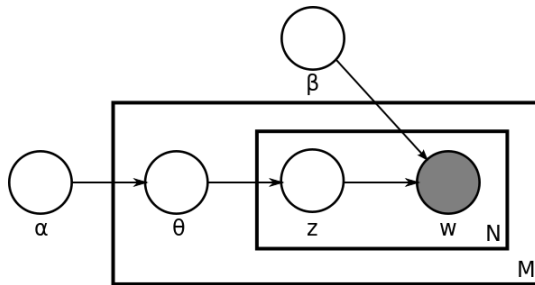
- Text search:
 - More complex queries can be used after indexing the data that may be not doable using the APIs (e.g. tokenisation, query expansion, etc.)
- Semantic search (requires entity extraction)*:
 - Semantic search can be used after extracting entities of interest from documents (e.g., place names, actors, etc.)
- Automatic categorisation / Tagging (clustering approaches, topic modelling, Machine Learning models)*:
 - Topic modelling and clustering can discover hidden topics in the collected documents.
 - Machine learning can be used for filtering data (e.g., relatedness, event type, information type).



* Entity extraction methods and automatic categorization is discussed later on in this tutorial

Text Clustering

- A simple approach for filtering textual data is to **group related documents automatically (clustering)**.
 - E.g., K-Nearest Neighbour (KNN), Latent Dirichlet Allocation (LDA) (Blei et al., 2003).



- Text clustering approaches are unsupervised: **manual annotations are unnecessary.**
- **Filtering** can then be done using the clusters generated by those models.
- Unsupervised approaches are limited and may not always produce relevant clusters.



Semantic Search

Query based systems, where the **relevancy of a document is established based on the context of a query** (Mangold et al., 2007):

- ***Describe events as semantic queries and use knowledge graphs and ontologies to map the data and the query.***
- ***Use of Natural Language Processing techniques and external Knowledge Graphs.***
- ***Can be tailored to a certain type of events or broad category of events.***



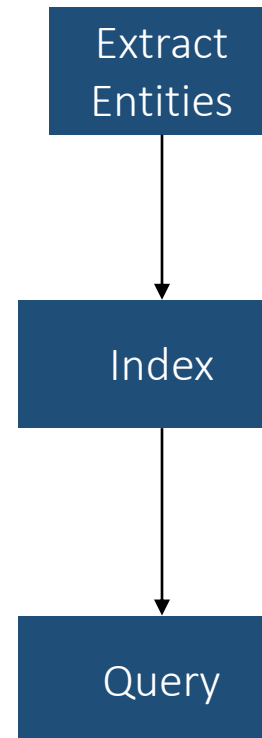
WGS841

Semantic Search - Motivation

- Complex event/information types are non-trivial to be described by single entity or keywords.
- N-grams or bag of words do not directly contain semantic information.
- Search for content based on a theme/context instead of precise information.

Semantic Search - Approach

- Entity extraction systems are used such as TextRazor, Dbpedia Spotlight, etc.
- An indexing and search system such as Lucene/Solr is used.
- A graph database store uniquely defined properties for each document

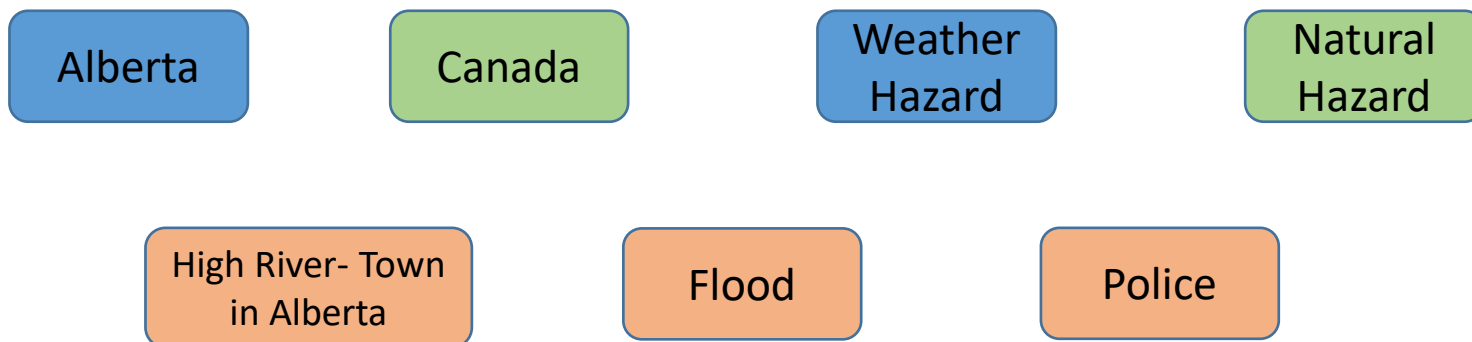


Entity extraction is used for obtaining fine grained information within documents.

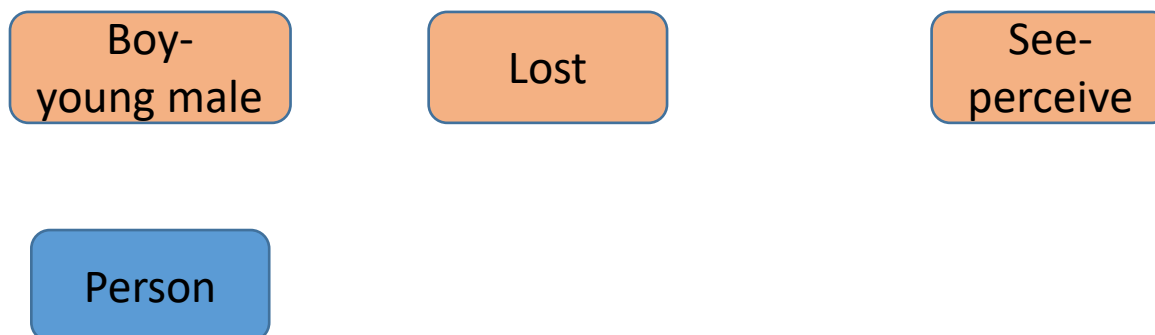
The data is **indexed** using regular information retrieval tools or stored in a graph database.

Relevant information is retrieved using entities and relations between related content.

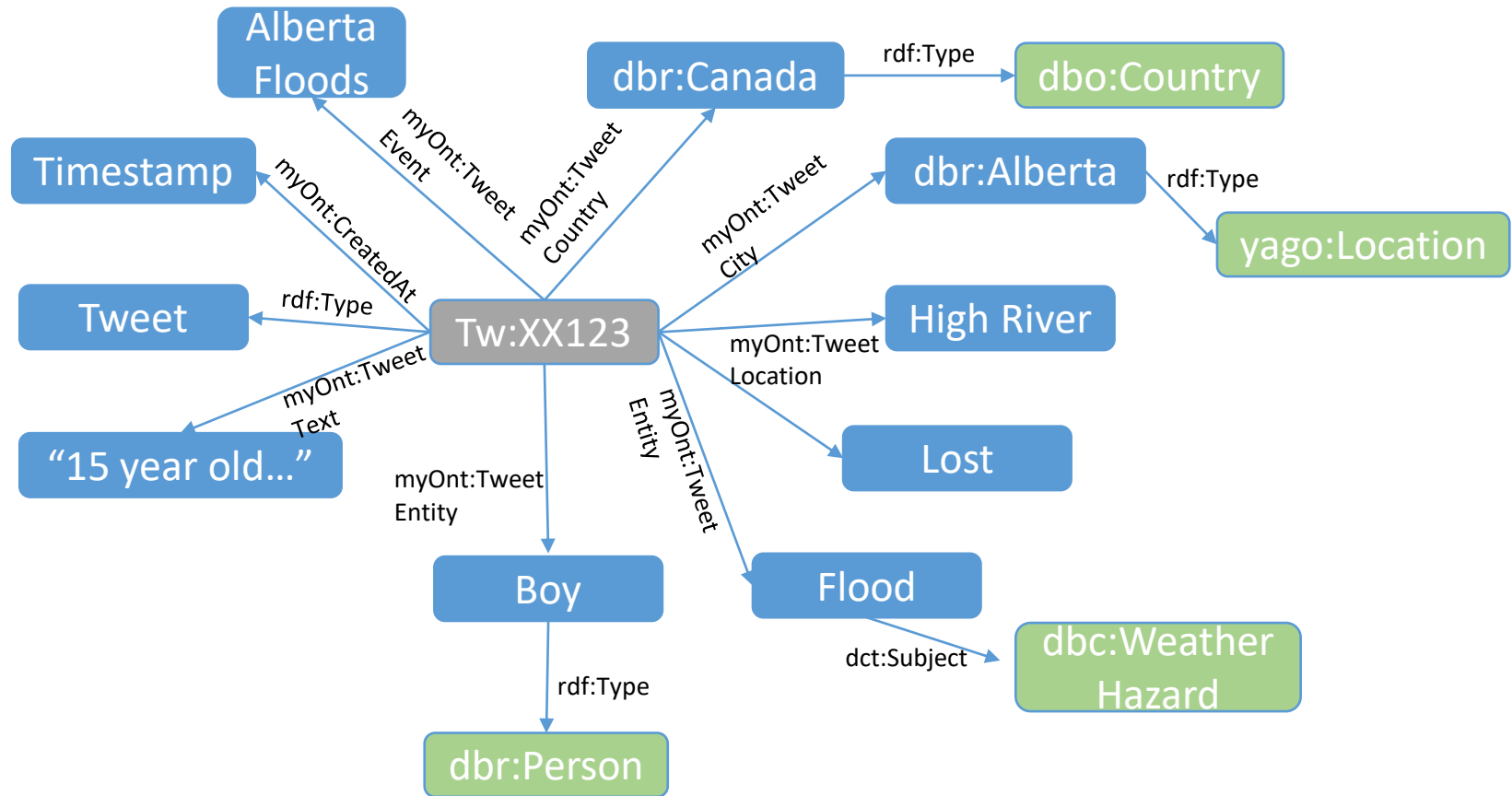
Document Entities - Example



"A 15 year old High River boy is missing due to flood. Call police if you see Eric St. Denis"

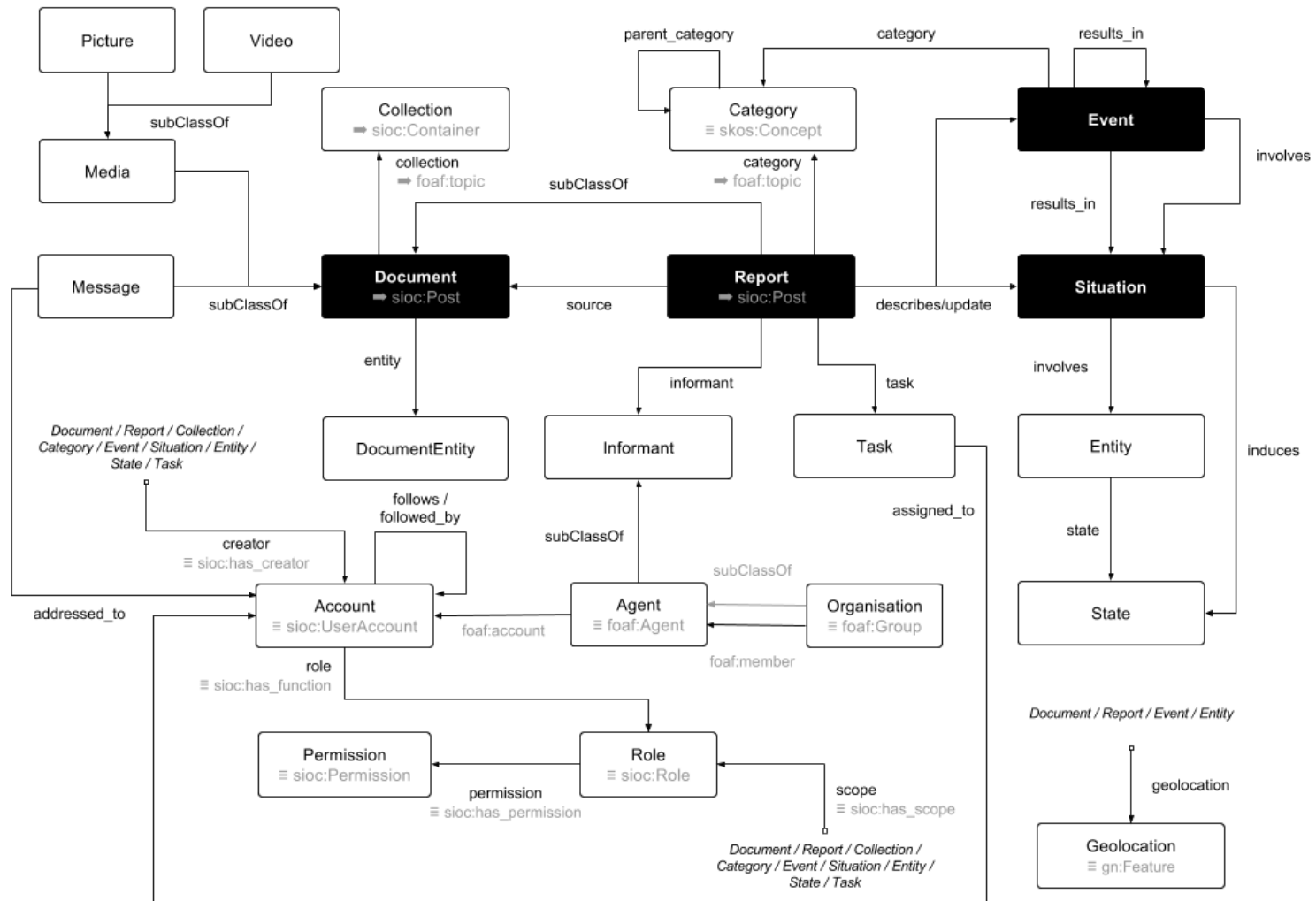


Graph Representation of Tweet



Concept Source: Tonon, A., Cudré-Mauroux, P., Blarer, A., Lenders, V. and Motik, B., 2017, May. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In *European Semantic Web Conference* (pp. 138-153). Springer, Cham.

Ontological Representation



Automatic Data Collection on Twitter

Hands-on

Summary

- Data **collection can be done automatically** using the APIs of different social media platforms.
- **Social media platform usage varies** greatly across countries and demographics (e.g., target, information, etc.).
- Access to social media data **can be limited** using APIs restrictions.
- Multiple methods can be used for **filtering quickly irrelevant information** or focusing on specific content of interest.

