Example “Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness”

<https://datorium.gesis.org/xmlui/handle/10.7802/1166>

Study description:



Internet source (e.g. web scraping) - requires workflow and protocol for information captured about the web scraping event. Transformation path from source to dataset

 Which type of source (HTML, specific social media platform, etc.)

 Selection of sources: Universe, sampling (optional), filter by criteria?

 Which data is being captured? Is it classified (procedure?)

 Coverage by topics (e.g. hashtags, keyword filtering etc)

 Geographic or IP-range coverage, date/time coverage

Specifics about the application (Twitter, Facebook etc) as a datasource, also specific APIs or tools/commercial services may be used

 Connecting information within the network if a “snowball” approach is used

 Legal and ethical considerations (informed consent, intellectual property, copyright)

**Which type of source (HTML, specific social media platform, etc.)**

ImplementedInstrument

* displayLabel “Python scripts for the data collection with the Twython API”
* uri “<https://datorium.gesis.org/xmlui/handle/10.7802/1504>”
* organizes Capture as RepresentedMeasurement
	+ source to ExternalControlledVocabularyEntry
		- content “web”
	+ displayLabel “Tweet ID”
* relatedMaterial to ExternalMaterial
	+ descriptiveText “Actively maintained, pure Python wrapper for the Twitter API. Supports both normal and streaming Twitter APIs”
	+ uri “<https://twython.readthedocs.io/en/latest/>”

**Selection of sources: Universe**

**Selection of sources: sampling (optional)**

**Selection of sources: filter by criteria?**