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
  - CTRNet
  - IDEAL
  - QCRI

No collection level metadata + 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
Open Provenance Model

- A wasTriggeredBy P1
- P1 wasDerivedFrom P2
- A1 wasGeneratedBy P1
- A wasControlledBy Ag
- P wasGeneratedBy A
- P used R
- P1 used R
- A1 used R
- A2 used R

Adapted from http://openprovenance.org/tutorial/slides/2-opm-
Challenges

- Complete provenance is not typically provided in standard collection data
- How to merge collections that come from any number of formats