Observations on the Challenge of Identity

Bitcurator Natural Language Processing
Chapel Hill
February 2017
A Broad View of Identity

• Identity is a ubiquitous challenge, an essential epistemological challenge

• From the perspective of cultural heritage it is a challenge for the primary objects of interest: the artifacts that we curate
  • How do we identify them? How do we document or describe them?

• But also for the agents or actors that create the artifacts
  • How do we identify them? How do we document or describe them?

• Both are very much related in that the two challenges are similar

• Given limited time, today I will focus on the agents, and of the different kinds (persons and groups, and arguably also delegate agents), I will focus on persons
Observation One: Social Phenomena

• The objects of interest for the cultural heritage community are social objects

• Artifacts by definition, are human created objects, and thus social objects: they occupy a shared space of human experience and of human understanding and discourse

• While the artifacts have a physical existence, it is the use of the physical to carry human expression, to communicate from one person to another that makes of them social objects
Observation One: Agents and Social Agents

- The human creators (individuals or groups) of the artifacts are social agents, that is, social constructs.
- It is not difficult for us to recognized that groups are social agents:
  - Groups are associated with physical spaces, they occupy them.
  - But buildings and offices are not the groups themselves.
  - Nor are groups the same as the sum of the members at any given time.
  - Members come and go, and yet we recognize a persistent entity, and enduring "thing."
- But are persons social agents?
Observation Two: Persons are Physical Agents

• A person is a physical thing, a biological physical thing to be precise

• Each biological person has certain properties
  • Birth date and eventually death date
  • Born and die in particular places
  • Sex: male, female, or multiple other possibilities (determined by biomedical experts)
  • Features of various textures and color: skin, eye, hair ...
  • And while always in a constant state of change, is identifiable from moment to moment, and often throughout most of the stages of life

• It is the physical person (in part, and with some social adornment) that we identify when we “people watch”
Observation Three: Persons are also Social Agents

- While the physical body must serve as a “host,” must effect the activities of life and work, the person has other properties that are socially acquired
  - Name or names
  - Engages in activities that are socially recognized, that is, that take place in and are recognized in a space shared with other persons, other physical-social persons
    - Language or languages (spoken or written)
    - Education
    - Has an occupation, position, vocation ...
    - Produces artifacts, alone or with others: archival records, published books, paintings, movies, music, buildings and other structures, social accomplishments, ... and on and on
- This is the social person, this is the social identity, distinct from the physical identity
- It is a fragment of the social person (name) that we encounter when reading a document and with Name Entity Recognition techniques
Observation Four: One Physical Person, Multiple Social Persons

• Each person has one and only one physical identity
• Most of the time, each person has one predominate social identity
  • The social person overlaps from end to end with the physical person
  • The two are coextensive
• Because it is common, we tend to conflate the two, see them as one and the same
• But ... quite often one physical person may be the “host” for more than one social person
• The social person must be contained within the life span boundaries of the physical person, but need not be coextensive with it

• Multiple social persons generally come in three types
  • Predominate social person - coextensive with physical person
  • Alternative social person - contained within temporal span of physical person, not coextensive, and may overlap temporarily with other social persons (coexist)
  • Predominate social person replaces at a moment in time a previously existing predominate social person

• Examples of each:
  • George Washington
  • Samuel Clemons, Mark Twain (though the latter eclipsed the former)
  • A change in gender and frequently name: different social persons, though same (fundamentally) same physical person
Some Concluding Observations

• CIDOC-CRM and FRBR-LRM both conflate the physical person and the social person (in everyday language, they have a bias for the “real person”)
• Library community traditional has emphasize “bibliographic reality” and thus the social person; ISNI also
• Evidence is more often than not fragmented
• To form a ”reliable” identity, we must triangulate across multiple sources providing mutually corroborating facts and contexts, assembling fragments into a constellation of facts that together “identify” the person
• NER is the first step, and then ... the challenge of Identity Resolution (IR)
• IR is exceptionally challenging, never perfect (for people or computers), and never ending, as new facts emerge, new judgement (sound or not) are made ...
• I fear there is a great deal more to be explored and observed ...