TagOntology

Tom Gruber
Co-Founder and CTO, RealTravel
tomgruber.org
Good Morning
“Let’s share tags.”

- What would we actually share?
  - stuff that only people can read, one by one
  - data that makes for pretty graphs and clouds
  - information that has value when shared

- How do we create value by sharing?
  - increase the population of contributors
  - create cross-links and multiple perspectives
  - enable new computational riffs
Levels of agreement on technical infrastructure -> value enabled

- **Formats**: Data can be accessed
- **Schema**: Data can be queried and stored
- **Ontology**: Data can be interpreted, aggregated, composed

- Application/service: Data and functionality can be shared (requires a monopoly)
Semantic agreement enables constructive composition

composition  noise?
Example: formal match, semantic mismatch

- System A says a tag is a property of a document.
- System B says a tag is an assertion by an individual with an identity.
- Does it mean anything to combine the tag data from these two systems?
  - “Precision without accuracy”
  - “Statistical fantasy”
Semantic agreement enables useful composition

- Systems A, B, C, & D agree that a tagging is an assertion tagged(term,item,agent)
  - *they also must agree on details such as how to determine equality of terms, items, and agents*
- System B, C, & D agree, in addition, that the assertions include polarity (+ or -)
- All systems can count up tags on an item
- Systems B & C can merge voting data
- System D (anti-spam) knows more about agents. It can riff on B’s and C’s data to give some agents more weight, and it can make inferences about agent validity.
Ontology is a mechanism for making semantic agreement

- Independent of data model, format, application
- Can be stated in many equivalent forms
  - Languages like OWL
  - Semantic Web has tools for translation, validation, and serialization into XML formats
- Allows for partial, minimal commitment
  - only hard requirement is logical consistency
- Enables data translation, and lets you know which inferences can be made on the data
TagOntology – core terms

- **Term** – a word or phrase that is recognizable by people and computers
- **Document** – a thing to be tagged, identifiable by a URI or a similar naming service
- **Tagger** – someone or thing doing the tagging, such as the user of an application
- **Tagged** – the assertion by Tagger that Document should be tagged with Term
Tag Terms

- **Term.name** – a function from Terms to text strings.
- **TermEquals(name1, name2)** – true when a string matches a term with equality.
  - if TermEquals(term1.name,term2.name) then term1 is identical to term2.
- **choice**: is TermEquals invariant over case, whitespace, punctuation?
Better: “tagged object”? 

Document.id – function from documents to universally scoped identifiers (URI or URL)

- If URIEquals(doc1.id, doc2.id)
  then doc1 == doc2

choice: is document one-to-one with URI identity? (Are alias URLs possible?)
Taggers

- Taggers are users of systems, writers of blogs, etc. The intent is that they reflect individual human judgment.
- Taggers need id’s too.
- **choice:** can tagging be done without taggers?
  - if $\text{Tagged}(\text{document}, \text{term})$ then there is some $\text{tagger}=f(\text{document})$ such that $\text{Tagged}(\text{document}, \text{term}, \text{tagger})$
  - This implies that tagging the same document with the same terms more than once adds no information.
The Tagged relation

- Tagging is represented as a relation
  \textbf{Tagged}(\texttt{document}, \texttt{term}, \texttt{tagger})
- There is no way to refer to the tuple itself
- Negation is like untagging:
  - \textit{it is impossible for the same document to be tagged and not tagged by the same tagger with the same term}
- Disagreement is relative to a tagger
Polarity – “voting” for an assertion

- Tagged\((document, \text{term}, \text{tagger}, +or-)\)
- + is the “default”
- Can’t have both + and – for same \((d, t, t)\)
- *Polarity is logically different than negation*
Scope and sources

- **Source** is a site, community, or organization that anchors a namespace. Source.id is a URI.
- Scope can be individual, source, or universal

**Choices:**
- Scope of `document.id`: universal? URI or URL?
- Scope of `tagger.id`: universal (URI) or rel to source?
- Scope of `term.name`: universal or rel to source?
- Scope of `tagged` assertion: universal or rel to source?
Defaults on the tagged relation

- Tagged(doc, term, tagger)
- Tagged(doc, term, tagger, +or-)
- Tagged(doc, term, tagger, +or-, source)
- **Choice:** What do these mean?
  - Tagged(*, term, tagger, +or-, source)
  - Tagged(doc, term, tagger, +or-, source)
Metatagging

- Is there a difference between $\text{Tagged}(\text{document, term, tagger})$ and $\text{Tagged}(\text{term, term, tagger})$?
  - Syntactically no, semantically YES!

- What about: $\text{Tagged} (\text{tagger, term, tagger})$?
  - Unless we can agree on what these mean at some level, we can’t compute on other people’s data.
Applications that could use this ontology

- Collaboratively filtered search: Wink, ...
  - Find things matching Q that my tagging buddies think matches Q

- Semistructured query
  - “all hotels in Barcelona tagged with “real pool”

- Micro reviews
  - “all hotels rated 5 on “real pool”
Possible results

- A TagOntology that defines a coherent conceptual model of all this
- 2 Subsets that are ready for buy-in
  - “core tagging” and “collaborative tagging”
- Outer levels that need work
  - “metatagging”
- Tight coordination with proposals at the data and formats levels