Terminology Metadata

W3C Multilingual Web LOD and MLW-LT Requirements Workshop

2012-06-12, Dublin, Ireland

Goal: standardize linking text with external knowledge

- Annotate fragments of content with human- and machine-readable references to external concepts that the fragment of text mentions.

Audiences, use cases

- Content authors: It should support the use case of terminology and controlled language
- Content localizers: accurate term and concept translation;
- Content management: indexing for information retrieval;
- MT providers: marking up input data for training MT systems

Relevant data categories

- Term
  - Mentions of terms linking to a terminology lexicon
- Named entity
  - Mentions of named entities, optionally disambiguating them to a ontology concept
- Disambiguation
  - Disambiguation of individual words, linking them to semantic networks
- Text analysis annotation
  - Which tool annotated this with what confidence?

Challenges

- No standard for integration with tools
- No support for web content (-> ITS-HTML5)
- Term tagging (-> text analytics)
- Contextual information
- Term matching in TM and CAT tools
- Different linking and de-referencing conventions for different data categories

De-referencing links?

- Term
  - ITS1.0: URI/XPath for the term entry
  - TBX: A term ID in a TBX terminology lexicon
  - TBX/RDF: URI
- Named entity
  - Entity concept: RDF URI
  - Entity type: RDF URI within a entity type repository (i.e. NERD) [ISSUE-3]
- Disambiguation
  - A synset ID within a semantic network (i.e. a wordnet) [ISSUE-94]

Consolidating the de-referencing protocol for all data categories

- Identify concepts/terms/synsets with URIs
- Identify term lexicons and semantic networks with URIs
- Current discussion on the mailing list: can we count on terminology resources and semantic networks being accessible in this matter

Text analysis annotation

- A special case of provenance:
  - its:annotationAgent = prov:Agent, prov:SoftwareAgent?
  - Its:annotationConfidence?

Examples

- Term
  - <span
    its-term=“http://www.isocat.org/rest/dc/132”
    its-terminology-lexicon=“http://www.isocat.org/...” >
    continuousAlphabeticalSequence</span>
- Named Entity:
  - <span
    its-concept=“http://dbpedia.org/resource/Dublin”
    its-entity-type=“http://nerd.eurecom.fr/ontology#Place”
    its-text-analysis-agent=“http://enrycher.ijs.si/”
    its-text-analysis-confidence=“0.96”>Dublin</span>
- Disambiguation
  - <span
    its-semantic-network=“http://purl.org/vocabularies/princeton/wn30”
    its-meaning=“http://purl.org/vocabularies/princeton/wn30/synset-chair-noun-1”
    its-text-analysis-agent=“http://enrycher.ijs.si/”
    >chair</span>

Serialization

- Inline annotations
  - Issues with complicating mark-up?
- Stand-off annotations
  - Pointing to arbitrary text fragments within HTML?