Provenance

MLW-LT Requirements workshop

Dublin

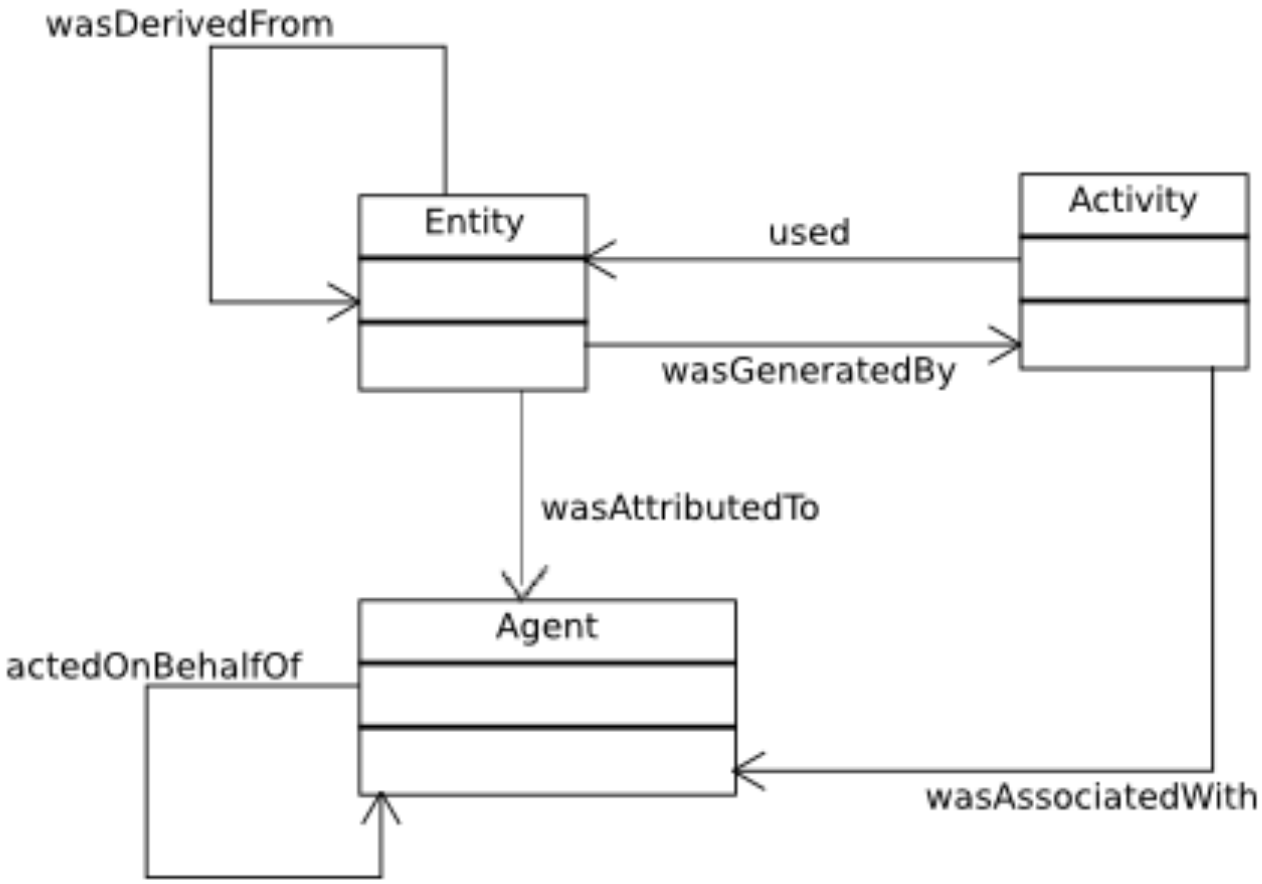
13 June 2012

Provenance use cases

- Localisation job progress monitoring
  - E.g. crowdsourcing
- Synchronising parallel source revision and translation
- Localisation process improvement
  - Translator/MT/QA efficacy
- Low cost assembly of quality parallel text
- Distributed quality auditing
  - Client/Provider Quality data synchronisation

W3C Provenance WG

- Data Model - PROV-DM
  - RDF/OWL
  - Text record
  - XML
  - query
- Time line – Jan 2013



ITS Linking options

- Document-level outbound `its-prov-ref`
  - Global tag ref to PROV element representing document
  - Segement level or other entity in PROV can be indicate via wasDerivedFrom (NIF also has properties for linking segemetn in order)
- Segment-level outbound
  - Requires segment level spans for completeness
  - Could also be used for term level markup
- Document-level inbound
  - Document URL in PROV record
- Segment-level inbound
  - Can do segment level provenance using NIF patterns
    - http://nlp2rdf.org/nif-1-0#toc-nif-1-0-uri-recipes
  - May still need document level outbound to indicate existance or PROV record
- Combination would be useful

ITS usage example

```
<span its-prov-ref=http://www.eg.org/prov-ex1.txt
   its-prov-ent="e1">My hovercraft is full of eels.</span>
```

Translated to

```
<span its-prov-ref="http://www.eg.org/prov-ex1.txt"
   its-prov-ent="e2">Mon aéroglisseur est plein
   d'anguilles.</span>
```

http://www.eg.org/prov-ex1.txt contains

```
entity(e1)
entity(e2)
wasGeneratedBy(e1, a1, 2011-11-16T16:05:30)
activity(a1, 2011-11-16T16:05:00, 2011-11-16T16:06:00,
   [its-prov-process-type="authorContent", its-source-
   lang="en"] )
wasGeneratedBy(e2, a2, 2011-11-16T16:07:30)
activity(a1, 2011-11-16T16:07:00, 2011-11-16T16:08:00,
   [its-prov-process-type="mTranslate"] )
```

http://lists.w3.org/Archives/Public/public-multilingualweb-lt/2012May/0065.html

Defining Agents

```
agent(Trevor, [ prov:type="Person",
  its-prov-agent-type="author" ] )

wasAssociatedWith(a1, Trevor)

agent(matrex-eng1234,
  [ prov:type="SoftwareAgent",
  its-prov-agent-type="smt",
  its-prov-src-lang="en",
  its-prov-tgt-lang="fr" ] )

wasAssociatedWith(a2, matrex-eng1234)
```

Alternative error reporting

```
entity(e3, [its-ent-type="qa-error-report",
   its-qa-err-severity="0.5",
   its-qa-err-note="suspect terminology")

wasGeneratedBy(g1, e3, a3, 2011-11-16T16:08:30)

wasDerivedFrom(e2, a3, g1)

activity(a3, 2011-11-16T16:08:00, 2011-11-16T16:09:00,
[its-prov-process-type="translateQA", its-prov-qa-
   ruleset="LISAQA"] )

wasAssociatedWith(a3, Pierre)

agent(Pierre, [ prov:type="Person", its-prov-agent-type="trans-
   QA-checker" ] )
```

Issues

- Stand off mark up for provenance?
- Use of PROV WG output
  - Probably requires non-normative usage profiles
- In line mark-up for provenance?
  - Verbosity
  - Can't have multiple record per element
- Agents with implicit activity
  - its-author="Trevor"
  - its-revision-agent
  - Its-translation-agent
  - agent in Quality profile