A Semantic Model for End-to-End Multilingual Web Content Processing

David Lewis, Alex O’Connor, Dominic Jones
Centre for Next Generation Localisation,
Trinity College London
Challenges for the Localisation Industry

- Demand reflects shifts in global trade
- Web & mobile, Apps & SaaS, perpetual beta, long tail
- User Generated Content

Localisation Industry
Centre for Next Generation Localisation

DCU
- Machine Translation
- Translation Technology
- Multilingual Information Retrieval
- Semantic Model Evolution

UCD
- Speech Synthesis
- Speech Recognition
- Semantic Model Generation

UL
- Localisation Research Centre
- Localisation Standards
- Localisation Workflow

TCD
- Personalisation
- Adaptive Hypermedia
- Text Analytics
- Interaction Design
- Service Integration & Management
- Semantic Interoperability

© 2010 TCD
Supporting the Global Customer

- Multi-lingual IR
- Real time social media translation
- Enterprise MT for user forum
- Assisted spoken interaction with users
- Community rating/post-editing
- LSP offer quality on-demand
Conventional Localisation Interoperability

- Hand-off standards in XML
- Import & Export functions in monolithic tools
Web Services – Potential Benefits for Localisation

- Views sought from developers, LSP and tool vendors
- Scalability through automated interfaces
  - Coping with unpredictable workloads in an agile world
  - Building blocks for automated workflows
- 24x7 availability for a geographically distributed workforce
- 'Pay as you use' models
- More granular cost management
- Easy to “rip and replace” components
- Easy deployment and version consistency
- Guarantees from automated workflows
- Support for specialist companies delivering 'niche' services
  - Language Technology and Language Resource Curation
Web Services: Needed Now - Need to Innovate

1. Machine Translating
2. QA checking
3. Reporting
4. File Parsing
5. Leveraging
6. Segmentation
7. Pseudo translating
8. Updating
9. Generate localised file
10. Archiving
11. Artwork/Multimedia processing
12. Translating
13. Language reviewing
14. Testing
Industry Survey – Barriers to Adoption

- Need for open standards
  - Reuse tools solutions
  - Building industry consensus
- Performance
- Reliability and robustness
- Security: need buy-in from IT departments at content developers and LSPs
- Maturity of workflow engines
Web Service Interoperability

- Tools vendors and service providers already offering web services
- Repurposing handoff standard requires careful **profiling** and definition of **processing expectations**

---

**Language Web Services**

- **Content Developer**
  - Content Management System

- **LSP**
  - Translation Management System

- **Translator**
  - Computer Assisted Translation

**Tools vendors and service providers already offering web services**

**Translation Management System**

"XLIFF"
Semantic Web to the Rescue?

- Scalable in extension, good for incremental adoption
- Link XML content and RDF meta-data (RDFa)
- Leverage existing vocabularies, e.g. Dublin Core, Open Provenance Model, Creative Commons
- Maturing Tools & Stores
- RDF(S)
  - Subject-Property-Object
  - Name Spaces
  - URI
  - Classes and Subclasses
  - Range and Domain
## Relational Data

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Title</th>
<th>Author</th>
<th>Publisher ID</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>596002637</td>
<td>Practical RDF</td>
<td>S. Powers</td>
<td>7642</td>
<td>350</td>
</tr>
<tr>
<td>596000480</td>
<td>Javascript</td>
<td>D. Flanagan</td>
<td>3556</td>
<td>936</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Row = Subject

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Title</th>
<th>Author</th>
<th>Publisher ID</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>596002637</td>
<td>Practical RDF</td>
<td>S. Powers</td>
<td>7642</td>
<td>350</td>
</tr>
<tr>
<td>596000480</td>
<td>Javascript</td>
<td>D. Flanagan</td>
<td>3556</td>
<td>936</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

This example from http://research.talis.com/2005/rdf-intro/
Column = Property

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Title</th>
<th>Author</th>
<th>Publisher ID</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>596002637</td>
<td>Practical RDF</td>
<td>S. Powers</td>
<td>7642</td>
<td>350</td>
</tr>
<tr>
<td>596000480</td>
<td>Javascript</td>
<td>D. Flanagan</td>
<td>3556</td>
<td>936</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Cell = Value

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Title</th>
<th>Author</th>
<th>Publisher ID</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>596002637</td>
<td>Practical RDF</td>
<td>S. Powers</td>
<td>7642</td>
<td>350</td>
</tr>
<tr>
<td>596000480</td>
<td>Javascript</td>
<td>D. Flanagan</td>
<td>3556</td>
<td>936</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Graph View

This example from http://research.talis.com/2005/rdf-intro/
Graph View

This example from http://research.talis.com/2005/rdf-intro/
Classes, Properties, Specialisation and Name Spaces yield Scalable Extensibility

@prefix ulo: <http://www.eg.org/ontologies/upperLayer#>

Upper Layer Ontology: Reusable concepts and associations

@prefix dsm: <http://www.domain.org/ontologies/domainSpecific#>

Domain Specific Model: Concepts multiply inherit from upper layer and extend

@prefix sdsm: <http://www.metoo.org/ontologies/yaDomainSpecific#>

Subsequent Domain Specific Model: Refer to upper layer for consistency, easing association with other domain specific models

© 2010 TCD
Semantics for Web Service Interoperability

- Web service input and output passed as structured data: XML, JSON
- Semantic Annotations
  - Operation/interface to service category
  - Input/Output to semantic structures
- W3C Semantic Annotation for WSDL (SAWSDL)
- WSMOLite for RESTful

Semantic

Syntactic

Client

input

lifting

lowering

http

output

Service category

© 2010 TCD
Propose RDFS Semantic Models

Content

- The **core processed content taxonomy** that we use to record the content transformations of various types (including training) delivered by services

Service Model

- Taxonomy service category classifications of different services, dovetails with the content model
NGL Content: Seed Taxonomy

- :Managed Content
  - :Generated Content
    - :Grouped Content
      - :Counted Content
    - :Analysed Content
      - :Serialised Content
    - :Prepared Content
      - :Prepared For Localisation
    - :Localised Content
      - :Automatically Processed Content
      - :Manually Processed Content
      - :Asset Content
  - :Personalised Content
  - :Presented Content
  - :Retrieved Content
NGL Content: Seed taxonomy for localised content

- Localised Content
  - Segmented Content
  - Translated Content
  - Reviewed Content
  - Reengineered Content
  - Reassembled Content
    - Leveraged from TM
      - Postedited
      - Machine Translated
      - Human Translated
NGL Service Model: Seed Taxonomy

- :NglService
  - :Generate Content
    - :Author Text
      - :Recognise Speech
    - :Machine Translate
      - :Leverage TM
        - :Synthesise Speech
  - :Transform Content
    - :Classify Text
  - :Annotate Content
    - :Rate Content
    - :Tag Content
  - :Process Group Content
    - :Order Group
    - :Filter Group
    - :Select From Group
  - :Create Service
    - :Create MTSvc
    - :Create TM LeverageSvc
    - :Create Composite Svc
    - :Create Personalised Svc
Model Refinement

- Seed Semantic Model
- Apply to Data Flow of specific process models
- Extend semantic model
- Capture semantic annotation for service contracts
- Revise core model
Fine-grained Roundtrips

Customer

Content Developer

Content Management System

LSP

Translation Management System

Translator

Computer Assisted Translation

MT TA TM MT TA MT TA TB

Language Technology Web Services

© 2010 TCD
25 RDF triples

- Around 3M RDF links

http://en.wikipedia.org/wiki/Linked_Data


As of September 2010 ©
Interoperability through Linked Data

Content Developer

CMS
xml:tm, termlink, RDFa

LSP
Translation Management System

Translator
Computer Assisted Translation

Language Technology Web Services

MT TA
TM MT TA
MT TA TB
Content State Transformation

- Content semantics aim to express the state transformation that operate on content and its meta-data as the result of content processing by different services.

- A provenance-oriented model based on the Open Provenance Vocabulary is used to capture both process transformation and details of the agents and content of those operations.

- This allows processes to be defined in terms specific transforms (additions, changes) on content.
Linked Localisation Data

- Open Provenance Vocabulary
  - Lightweight version of Open Provenance Model
  - http://openprovenance.org/
1: Register source Document (typically from CMS)  
2: Register doc for localisation and segment  
3: annotate each segment with source QA and then revise specific source segments as appropriate
Next Steps

- Applying and Revising semantic model to different use cases
  - CNGL Web+LT+Loc demonstrators
  - Annotating available APIs
- Semantic sandpit
- Content mark-up for links - RDFa
- Avoid standardising semantics
  - Linked Data vocabularies established through adoption
- Stress testing semantic technology
- Access control to federated triple stores
- Service pre-conditions and effects with rules
Conclusions

**Extensible** semantic models offers interoperability without stifling innovation

**Semantic Annotation** offers improved service-oriented interoperability

**Provenance-based** linked data for fine-grained process roundtrips

**Quality-driven** language resource management
Visit:
www.cngl.ie

Watch:
http://www.youtube.com/user/nextgenlocalisation

Contact:
Dave.Lewis@cs.tcd.ie