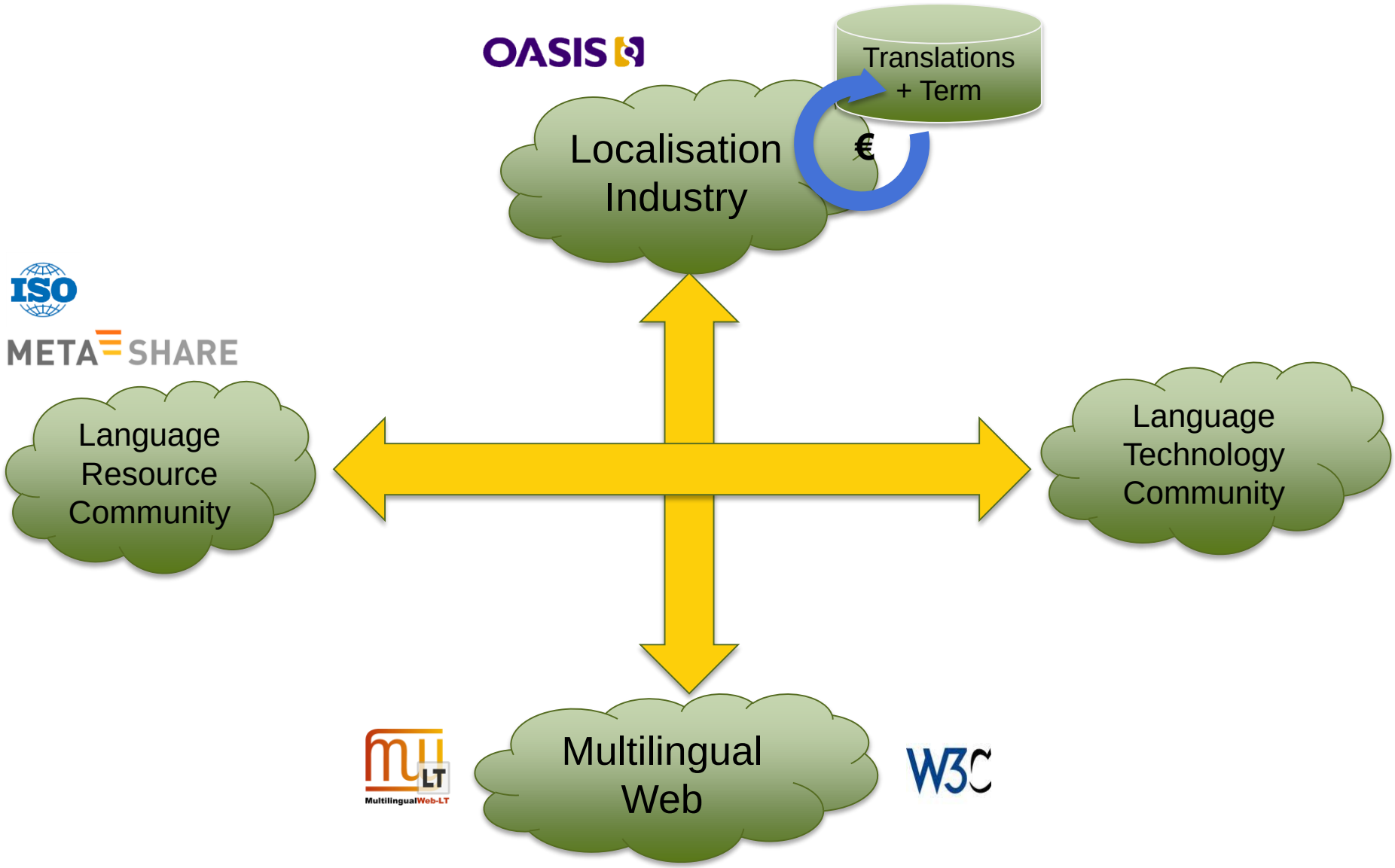
People • Content • Systems

Linked Data Reuse in the Language Services Industry

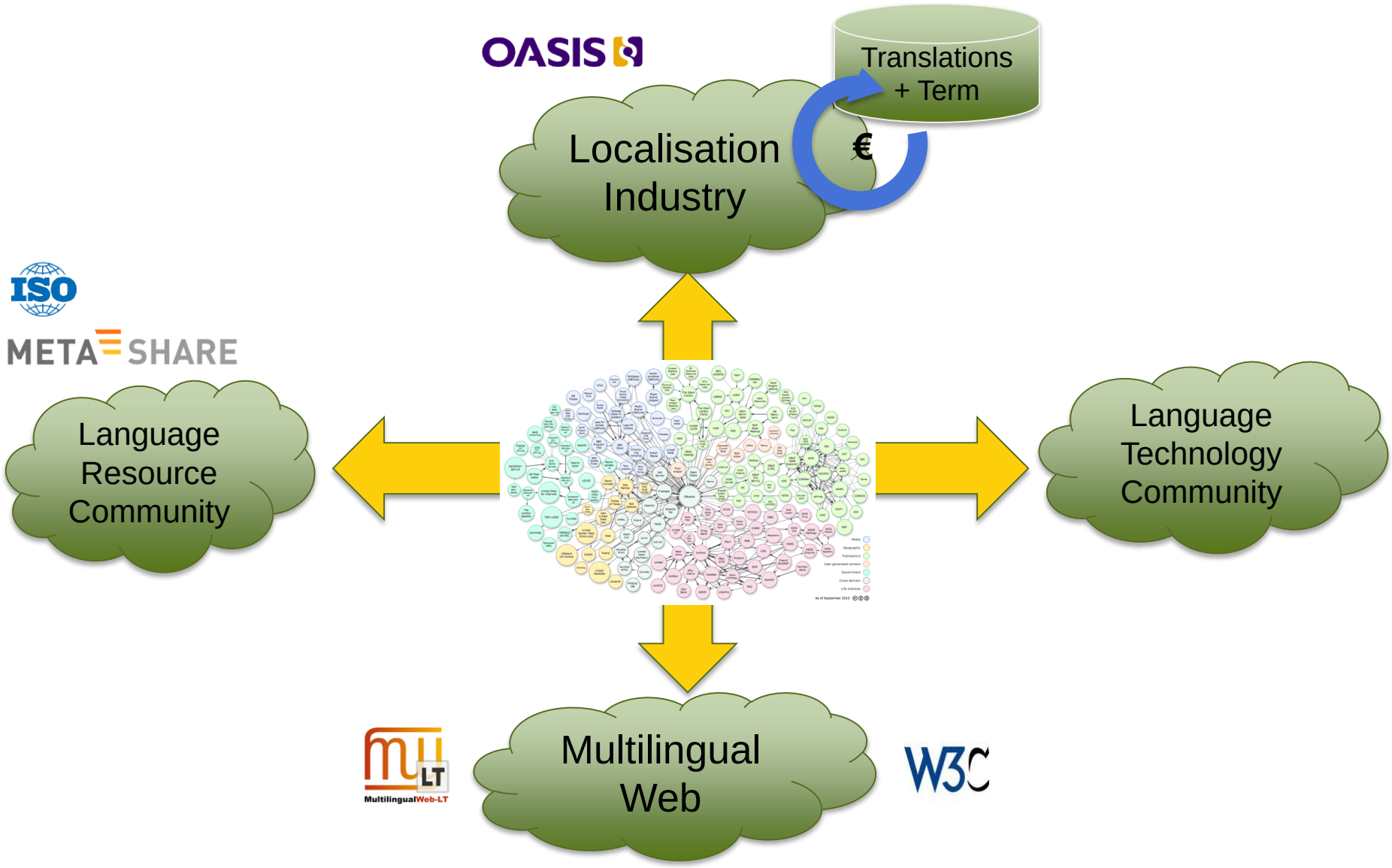
David Lewis, Leroy Finn, Dominic Jones

MultilingualWeb-LT

Language & Localisation: Meta-Data and Interoperability

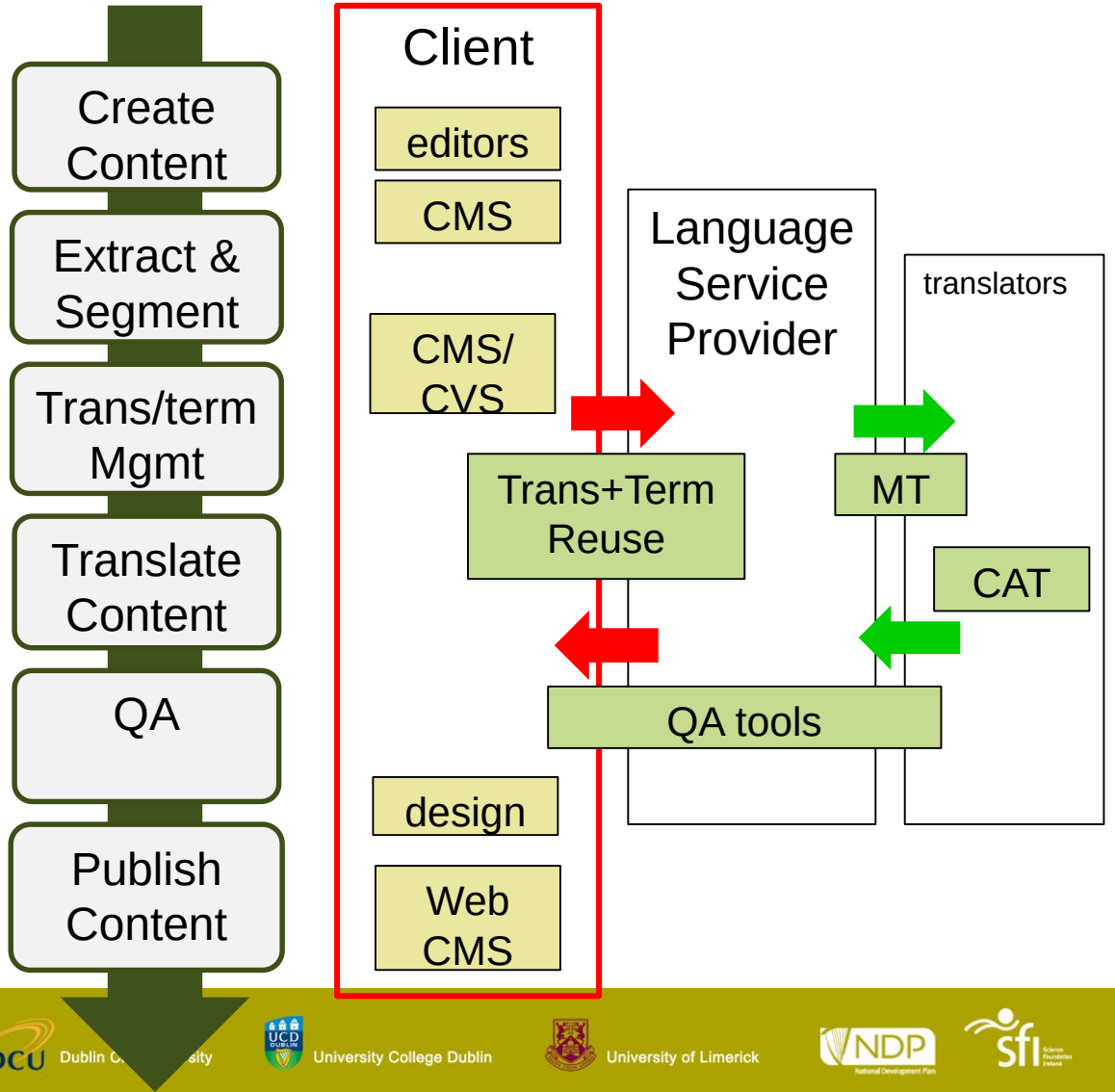


Language & Localisation: Meta-Data and Interoperability



Content Management-Localisation Workflow & Interoperability

- Interoperability roadblocks in high variety:
  - Content Management Systems
  - Content formats
- Increasingly dynamic
- Add language resource curation
  - Data driven MT & text analytics



Internationalisation Tag Set (ITS)

- Allows Localisation tools to be instructed to treat specific text in specific ways
- Annotates text in XML and HTML5 documents
- RDF Mapping – NLP Interchange Format
- Principles:
  - Minimise disturbance of original content
  - Don't reinvent wheel
  - Link to existing meta-data before adding new
- Defined distinct, independent Data Categories
- ITS2.0 is "Dublin Core" of the Multilingual Web

Local Approach

```
<para>
    Press the
    <uitext its:translate="no">START</uitext>
    button to sound the horn. The
    <uitext its:translate="no">MAKE-READY/ RUN</uitext> indicator flashes.
</para>
```

```
<para>
    Press the
    <uitext>START</uitext>
    button to sound the horn. The
    <uitext>MAKE-READY/ RUN</uitext>
    indicator flashes.
</para>
```

Global Approach

```
<its:rules ... its:version="1.0">
    <its:translateRule selector="//uitext" translate="no"/>
</its:rules>
```

C. Lieske, F. Sasaki 2010

Local Approach

```
<para>
    Press the
    <uitext its:translate="no">START</uitext>
    button to sound the horn. The
    <uitext its:translate="no">MAKE-READY/ RUN</uitext> indicator flashes.
</para>
```

```
<para>
    Press the
    <uitext>START</uitext>
    button to sound the horn. The
    <uitext>MAKE-READY/ RUN</uitext>
    indicator flashes.
</para>
```

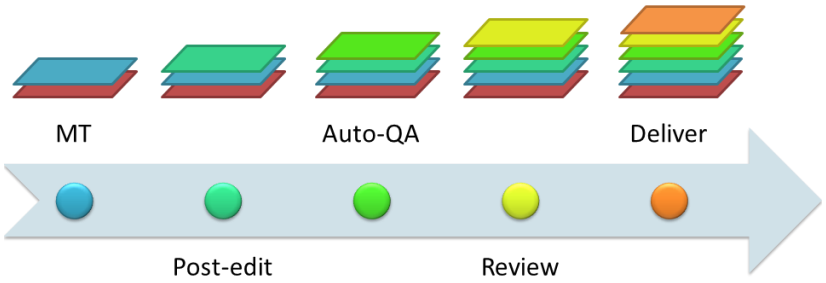
- Here Content is the Data
- So we must annotate textual content
  - RDFa and microdata better for annotating elements

Global Approach

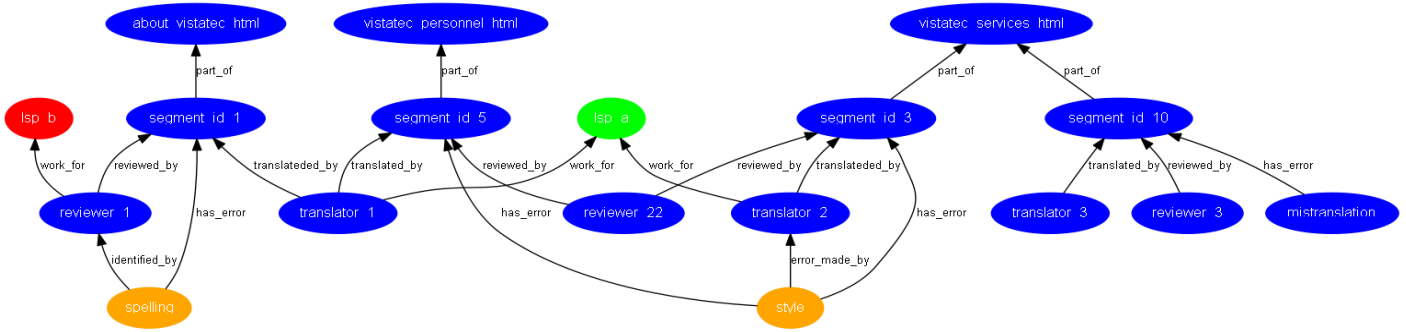
```
<its:rules ... its:version="1.0">
    <its:translateRule selector="//uitext" translate="no"/>
</its:rules>
```

C. Lieske, F. Sasaki 2010

Process Monitoring in RDF

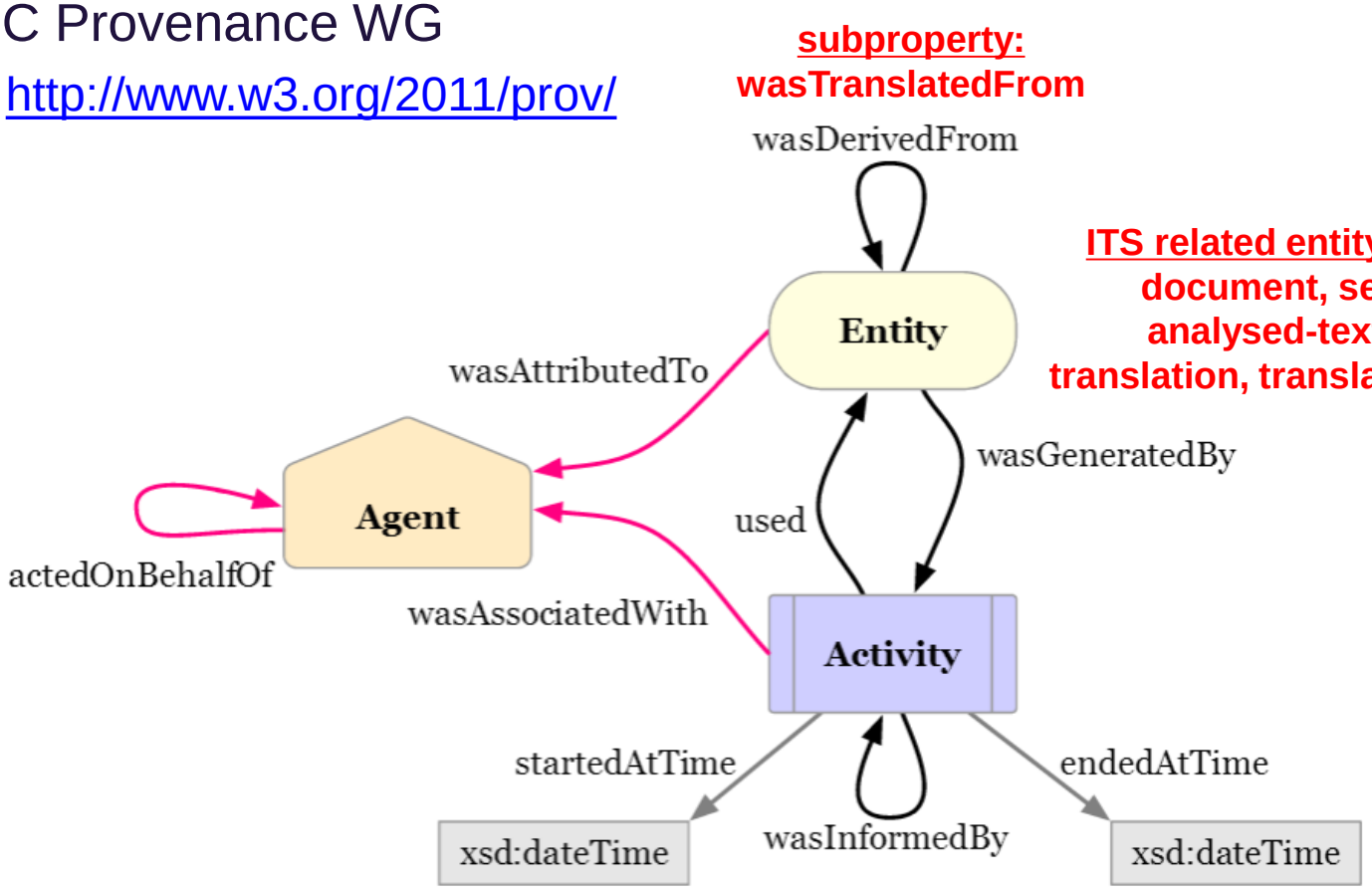


- Better up-front and real-time business decision support:
  - Use historic and aggregated metadata for inference
  - Content carries its own “audit trail”
- Metadata can be transformed to linked semantic data allowing:
  - Inferential queries across data siloes
  - New forms of data visualization
  - Extensible data relationships



Provenance Data: RDF-based logging

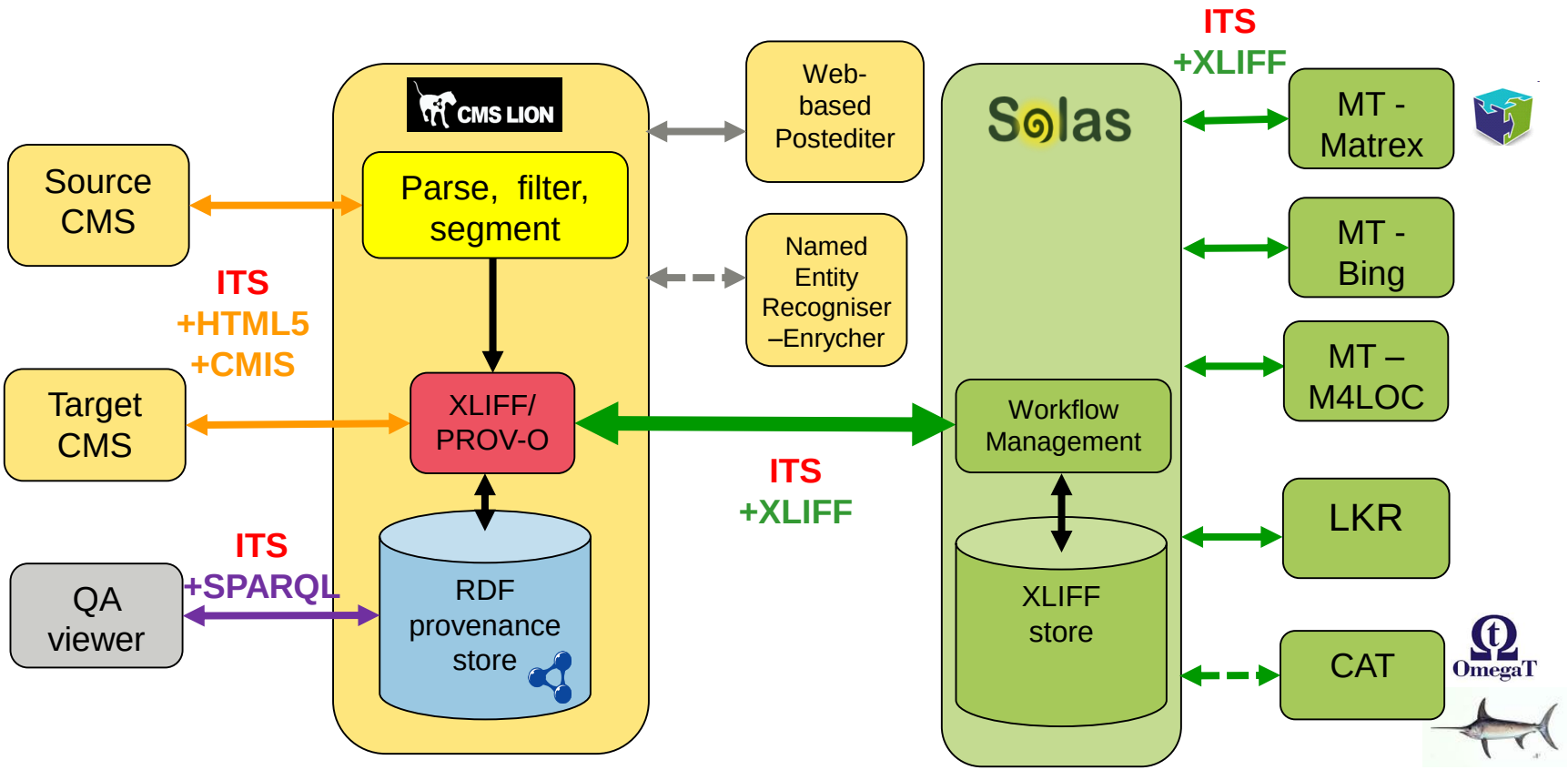
- W3C Provenance WG
  - http://www.w3.org/2011/prov/

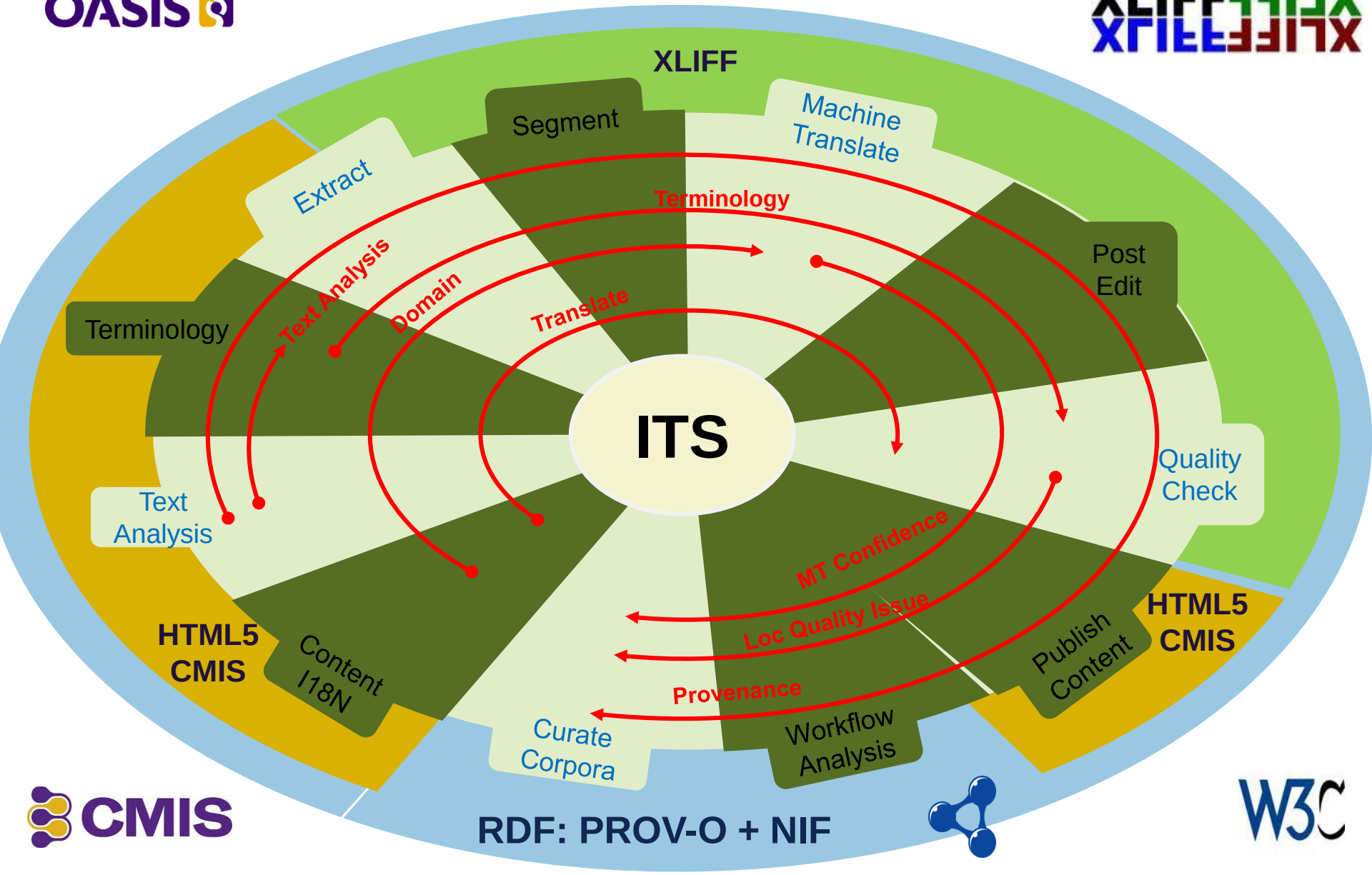


ITS related entity subclass:
document, segment, analysed-text, term, translation, translation-revision

From: http://www.w3.org/TR/prov-primer/

CMS to Localisation Workflow Roundtrip




OASIS
XLIFF
XLIFF
Segment
Machine Translate
Extract
Terminology
Post Edit
Text Analysis
Domain
Translate
Terminology
ITS
Quality Check
Text Analysis
MT Confidence
Loc Quality Issue
HTML5 CMIS
Content I18N
Publish Content
HTML5 CMIS
Provenance
Curate Corpora
Workflow Analysis
CMIS
RDF: PROV-O + NIF
W3C

Active Curation

- Localisation Industry:
  - €33bn global industry 12% growth
  - Established commercial models for translation and term reuse
  - High interoperability costs
  - Long tail industry – 99% SMEs
  - SMEs struggle to benefit from leveraging reuse
  - Challenged to adapt machine translation and text analytics technology
- **Active Curation:** *On-demand, quality-driven* assembly of training corpora
  - Collect human linguistic judgements from across the workflow
  - Professional translator/reviewer, crowds, end user
  - Already shown 25% MT improvement via in-job retraining
- **Linked Language and Localisation Data**
  - Extending existing tools to generate process provenance data
  - Pooling/Bokering of data across value chains

Where Next?

- Help EC and other public bodies to bootstrap data
  - per year EC produces 11 million pages in 22 language => 14 billion triples
- Attribution/license annotation and access control – Controlled Linked Data
- Increase industrial engagement
  - Media localisation – audio, video, images
  - Non-localisation MT and TA applications
  - Multilingual content personalisation
  - YOUR USE CASE HERE
- Join new W3C Community Group:

Multilingual Linked Open Data

http://www.w3.org/community/bpmlod/

CNGL
People • Content • Systems

Thank You