An Open Localisation Interface to CMS using OASIS Content Management Interoperability Services

Aonghus Ó hAirt, Dominic Jones, Leroy Finn and David Lewis

Centre for Next Generation Localisation, Trinity College Dublin

NDP National Development Plan Transforming Ireland

sfi Science Foundation Ireland

DCU Dublin City University

UCD University College Dublin

University of Limerick

Trinity College Dublin

Challenges for Interoperability

- More iterative workflows
  - From push-based hand-offs
  - To change **notification & fine grained retrievals**
- Extending data management to support innovation
  - Statistical MT, named entity recognition, text analytics for QA and terminology management
  - All require **up-to-date relevant training corpora**
- Solutions must sit comfortably with technology of:
  - Content Management, Web Publishing & Localisation

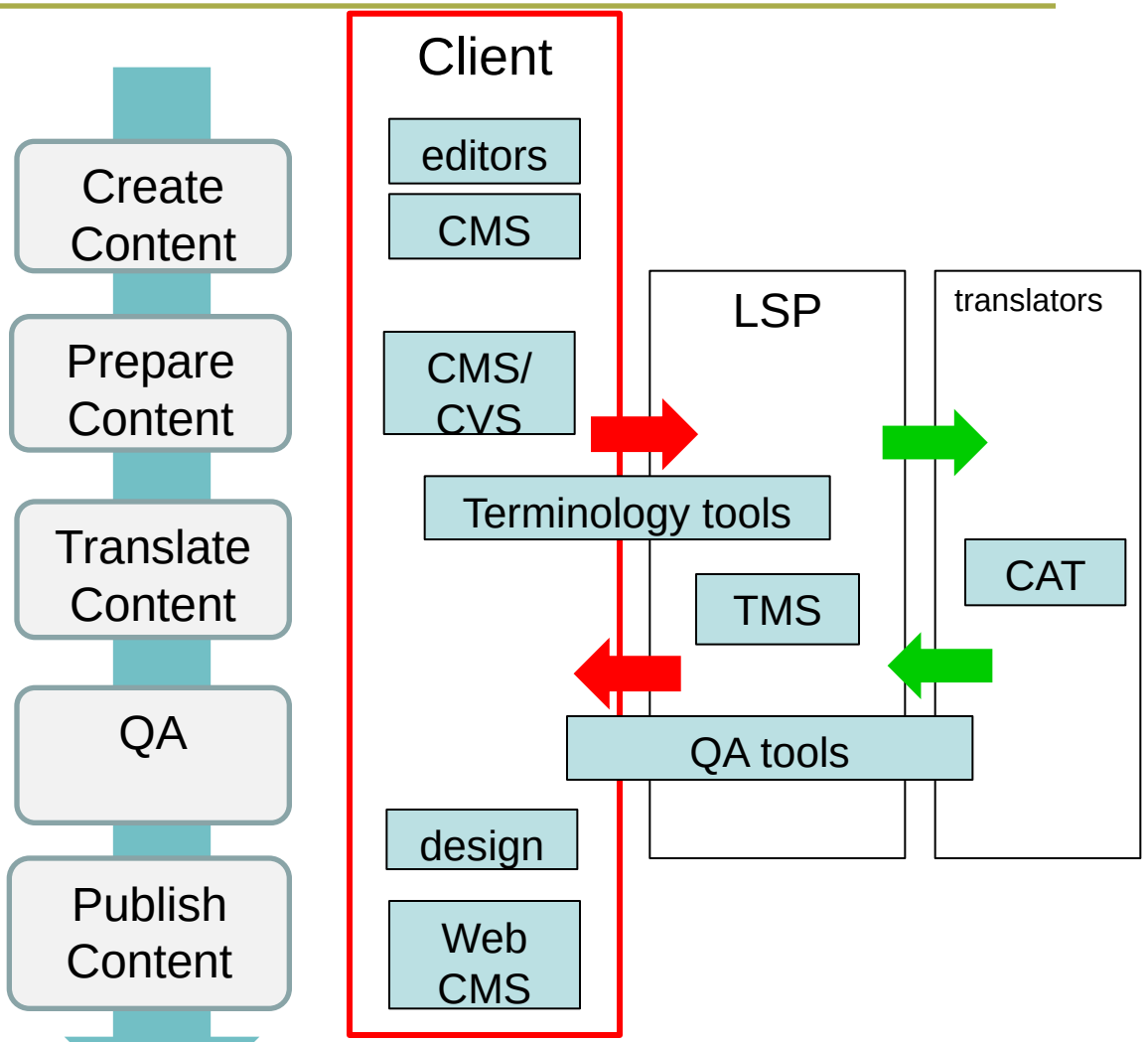
No one Standard can address it all

Integrating ITS and CMIS for L10n

& a little on XLIFF, RDF and Open Provenance

CMS-TMS Interoperability

- Interoperability roadblock
- High variety:
  - Content Management Systems
  - Content formats
- Increasingly dynamic
- Add language resource curation
  - Data driven MT, text analytics



Internationalisation Tag Set (ITS)

- Allows I18n and L10n tools to be instructed to treat specific text in specific ways
- Principles:
  - Minimise disturbance of original content
  - Don’t reinvent wheel
  - Link to existing meta-data before adding new
- Defined distinct, independent Data Categories
- Identify relevant text using:
  - Attributes to existing elements: LOCAL selection
  - Xpath selectors in a special element: GLOBAL selection

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

ITS 1.0 Data Categories

- Translate:
  - Mark whether the content of an element or attribute should be translated or not
- Localization Note:
  - Communicate notes to localizers about a particular item of content
- Terminology
  - Mark terms and optionally associate them with information, such as definitions
- Directionality
  - Specify the base writing direction of blocks, embeddings and overrides for the Unicode bidirectional algorithm
- Ruby
  - Provide a short annotation of an associated base text, particularly useful for East Asian languages
- Language Information
  - Express the language of a given piece of content
- Element within Text
  - Identify how an element behaves relative to its surrounding text, eg. for text segmentation purposes

ITS 2.0 Draft Data Categories

I18n

- Locale Filter
- External Resource
- Preserve Space
- Allowed Characters
- Storage Size
- ID Value

Language Technology

- Domain
- MT confidence
- Disambiguation
- Text Analysis Annotation

Provenance & QA

- Quality Issue
- Quality Précis
- Translation Provenance Agent
- Trans Revision Prov Agent
- Standoff Provenance

ITS and Content Management

- Global ITS rules can be defined in an external file
- Attribute applied to a node with following precedence:
  - LOCAL attributes
  - Embedded GLOBAL rules in reverse order
  - External GLOBAL rules in reverse order
- ITS allows tool-specific mechanisms for associating global rules with content – precedence not specified
- Common practice to apply a given set of rules to all documents in a project with the same schema
- Can this scale to multiple overlapping schema?

Can we use some CMS-level meta-data interoperability solution?

CMS Interoperability

- Integrating with CMS requires the use of an API. Until now, most CMS used proprietary APIs
- Proprietary interfaces to CMS lead to limited support, vendor lock-in and poor interoperability between CMS and with localisation tools
- Content Management Interoperability Service (CMIS) from OASIS offers a standardised API for interacting with CMS
- Localisation is out of scope for CMIS

How can CMIS facilitate the localisation of content across multiple CMS?

OASIS Content Management Interoperability Services (CMIS)

- “defines a domain model and Web Services and Restful AtomPub bindings that can be used by applications to work with one or more Content Management repositories/systems.” (CMIS standard)
- Published in 2010
- Participation from Adobe, Alfresco, EMC, IBM, Microsoft, Oracle, SAP, and others.

CMIS Implementations

Alfresco 3.3+
Apache Chemistry InMemory Server
Athento
COI
Day Software CRX
EMC Documentum
eXo Platform with xCMIS
Fabasoft
HP Autonomy Interwoven Worksite
IBM Content Manager
IBM FileNet Content Manager
IBM Content Manager On Demand
IBM Connections Files
IBM LotusLive Files
IBM Lotus Quickr Lists
ISIS Papyrus Objects
KnowledgeTree 3.7+
Maarch 1.3
Magnolia (CMS) 4.5
Microsoft SharePoint Server 2010
NCMIS
NemakiWare
Nuxeo Platform 5.5
O3spaces 3.2+
OpenIMS
OpenWGA 5.2+
PTC Windchill
SAP NetWeaver Cloud Document
Seapine Surround SCM 2011.1
Sense/Net 6.0+
TYPO3
VB.CMIS

CMIS Objects

- A repository is a container of objects.
- Objects have four base types:
  - **Document object** – “elementary information entities managed by the repository”
  - **Folder object** – “serves as the anchor for a collection of *file-able* objects”
  - **Relationship object** – “instantiates an explicit, binary, directional, non-invasive, and typed relationship between a *Source Object* and a *Target Object”*
  - **Policy object** – “represents an administrative policy that can be enforced by a repository, such as a retention management policy.”

(CMIS Specification)

CMS-L10n Interoperability: Two Requirements

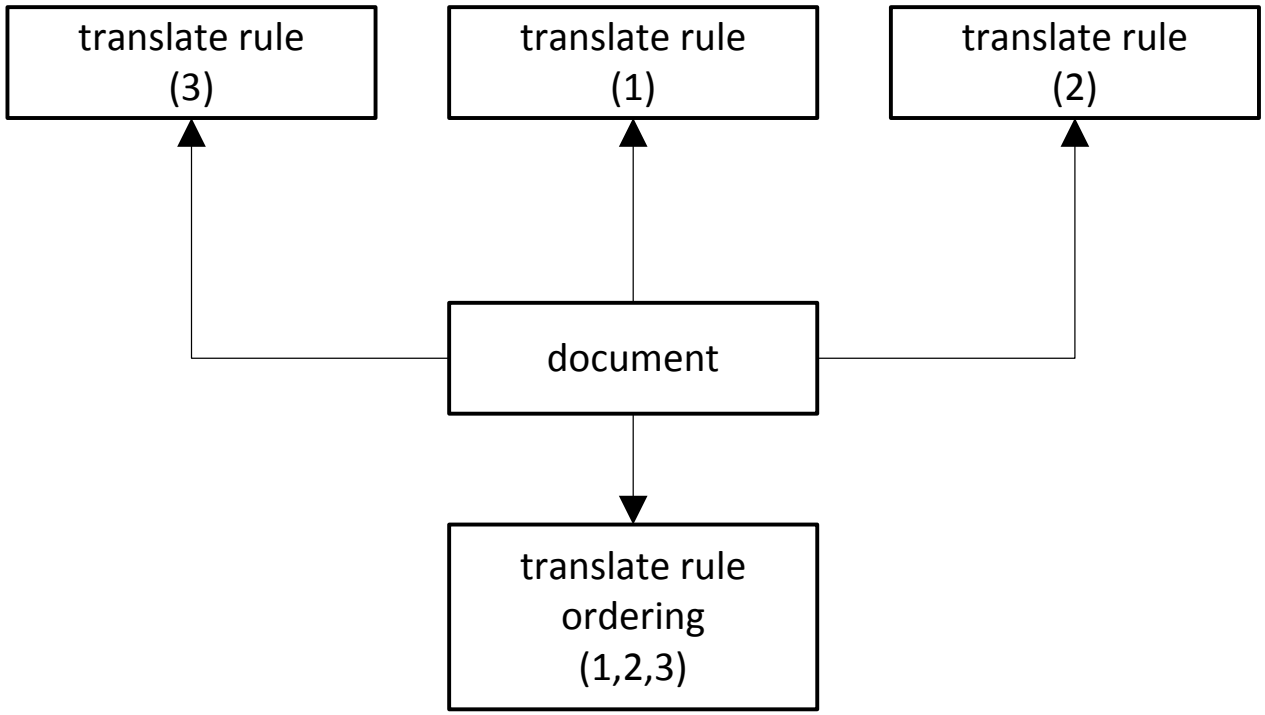
- Flexible ITS rule to document bindings
  - The same rule to be applied to multiple documents
  - Multiple rules to be applied to individual documents
  - Specify the precedence order in which rules are processed for a document
- Aim to support external ITS rules via CMIS

- Need to signal L10n-relevant updates to documents
- MLW-LT (ITS2.0) workgroup identified a requirement for such ‘readiness’ signalling
- Aim to support open asynchronous change notification for CMIS

Design: Extending CMIS Implementations

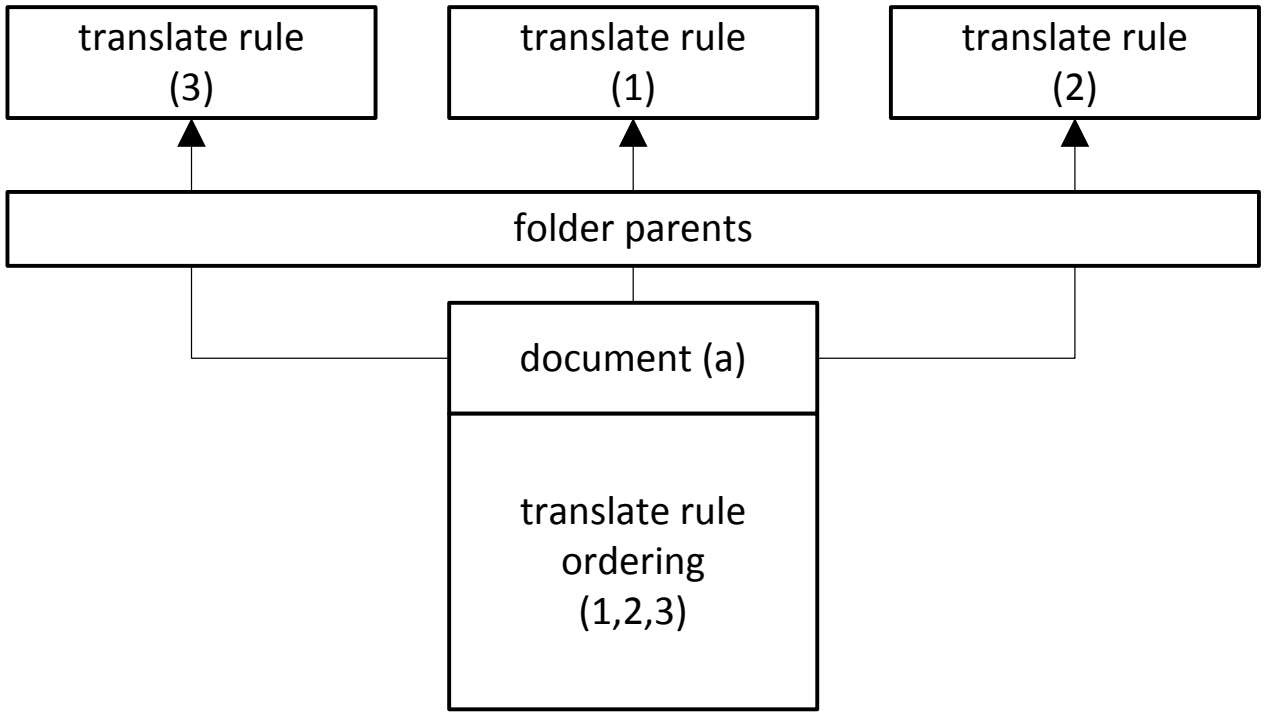
- Two approaches to modelling the localisation information:
  - Custom content modelling
  - Alfresco aspects
- Implementation in repository
  - Alfresco (primary)
  - Nuxeo (basic testing)

ITS rules using Policy Objects



Translate rules as policy objects

ITS Rules as Folders



Translate rules as folder objects

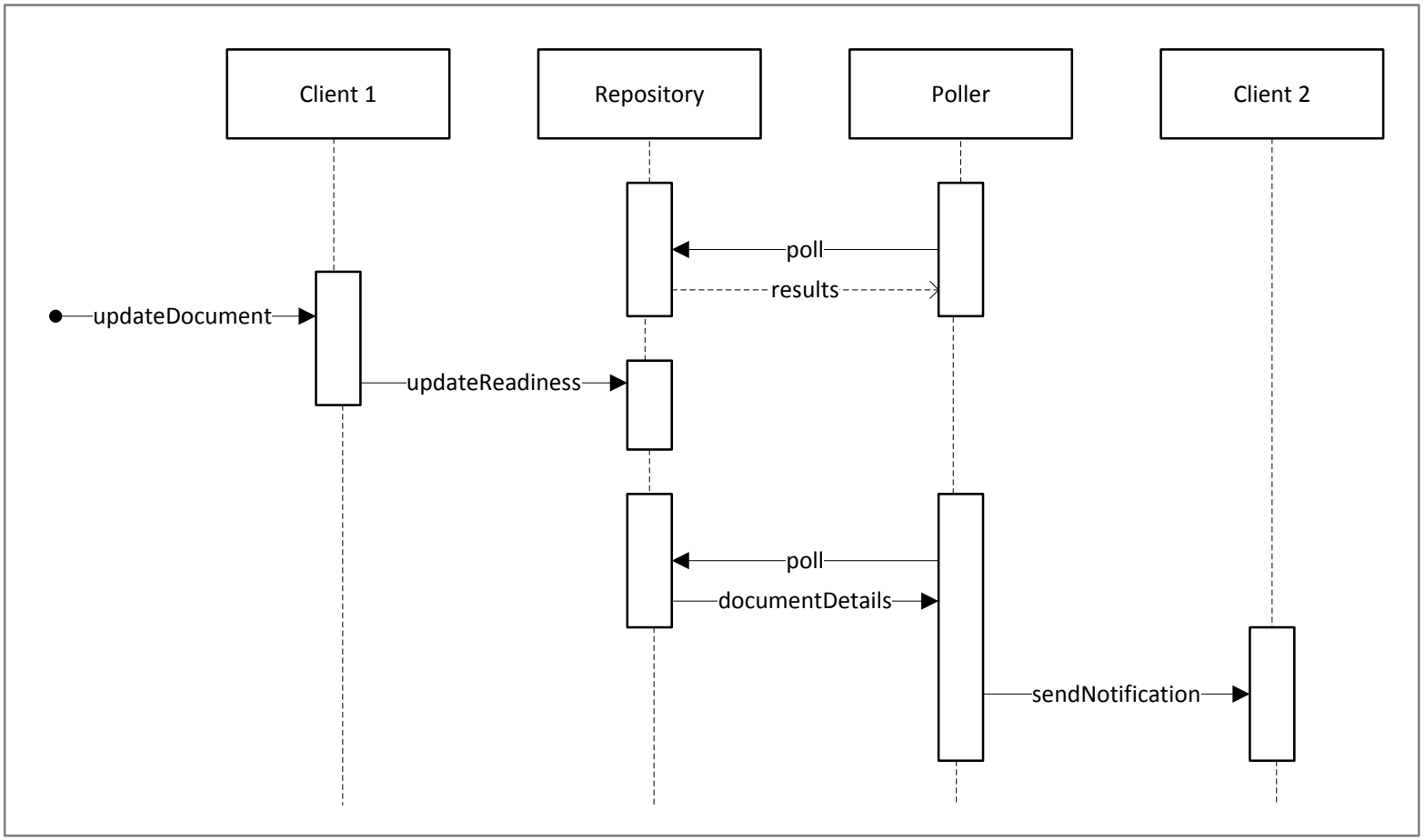
Signalling Readiness from CMS

- Readiness meta-data
  - Indicates the readiness of a document for submission to L10n processes or provide an estimate of when it will be ready for a particular process
- Data model
  - **ready-to-process** – type of process to be perfomred next
  - **process-ref** – a pointer to an external set of process type definitions used for ready-to-process
  - **ready-at** – defines the time the content is ready for the process, it could be some time in the past, or some time in the future
  - **revised** – indicates is this is a different version of content that was previously marked as ready for the declared process
  - **priority –** high or low
  - **complete-by –** indicates target date-time for completing the process

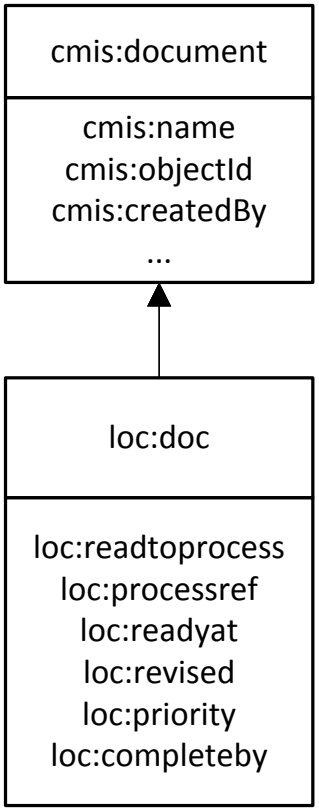
Polling extension to CMIS

- Polling schemes describe the way in which documents are polled for updated readiness properties
  - scheme name / ID
  - polling interval
  - notification method
  - notification target / host
  - port (for network connection)
  - readiness property
  - readiness value

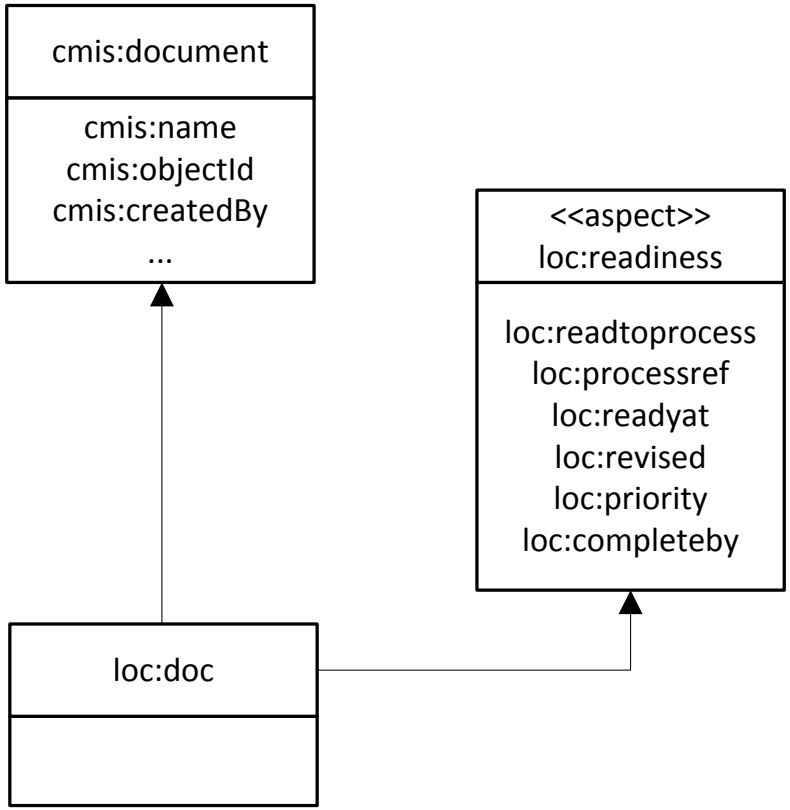
Polling sequence



Readiness

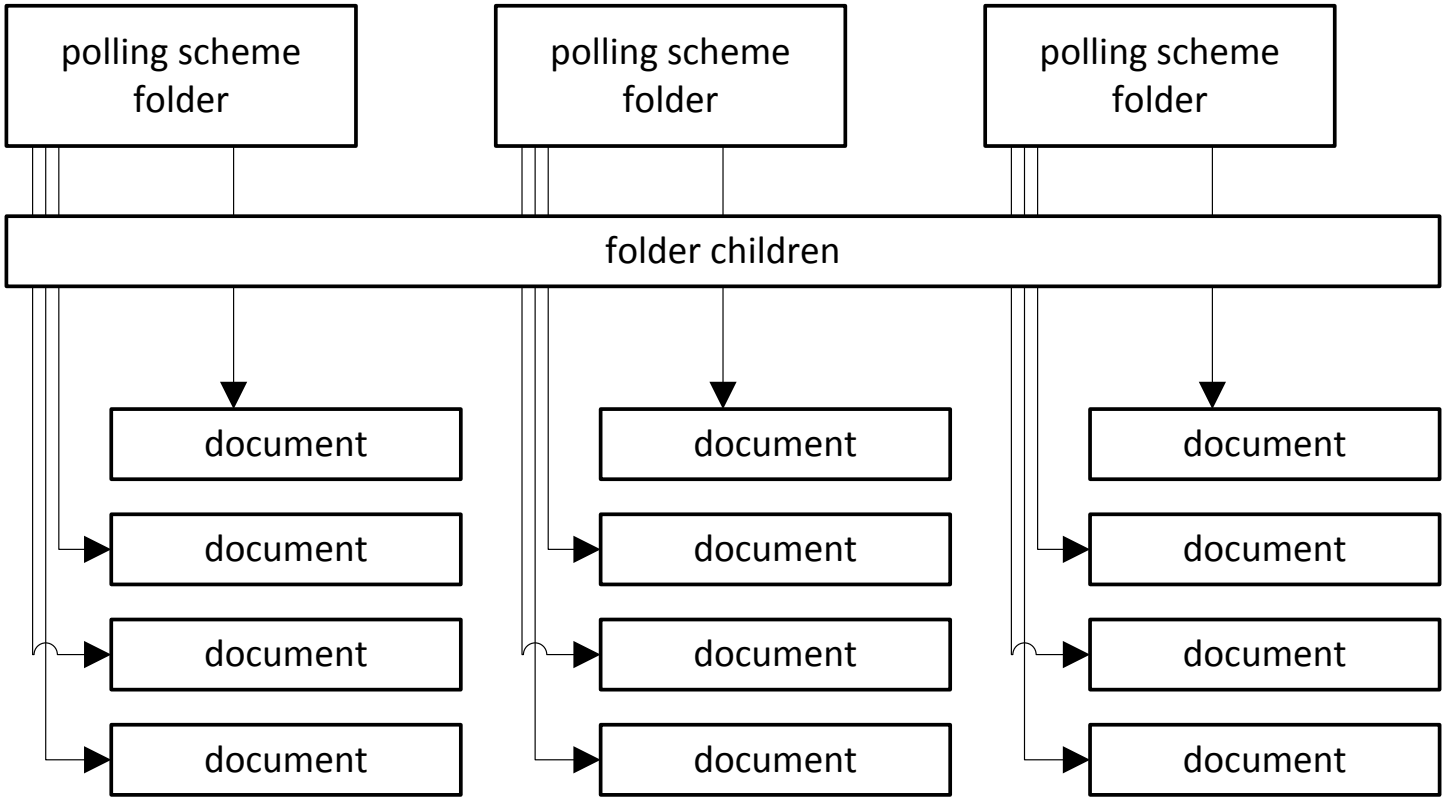


Readiness modelled as custom object

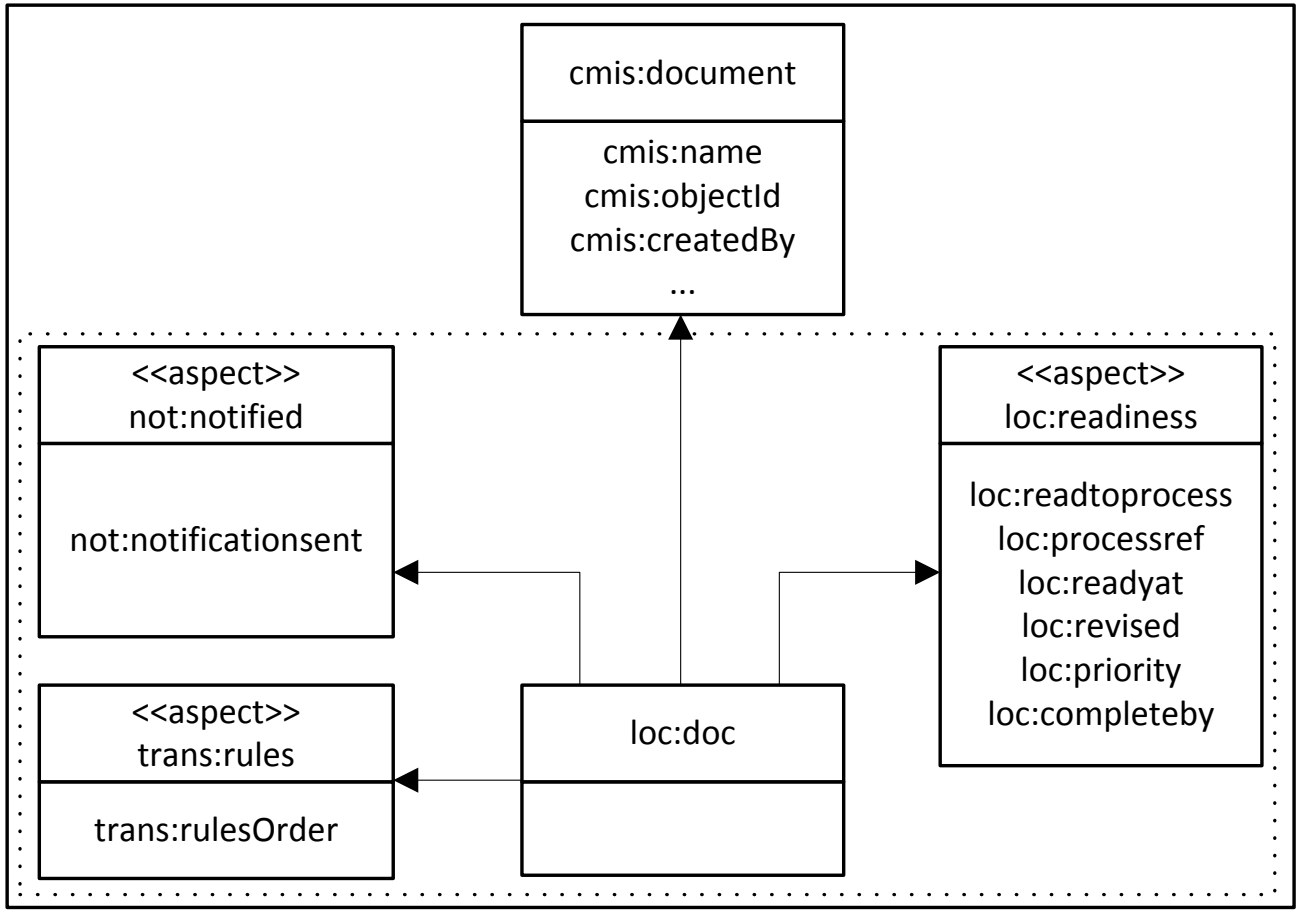


Readiness modelled with an aspect

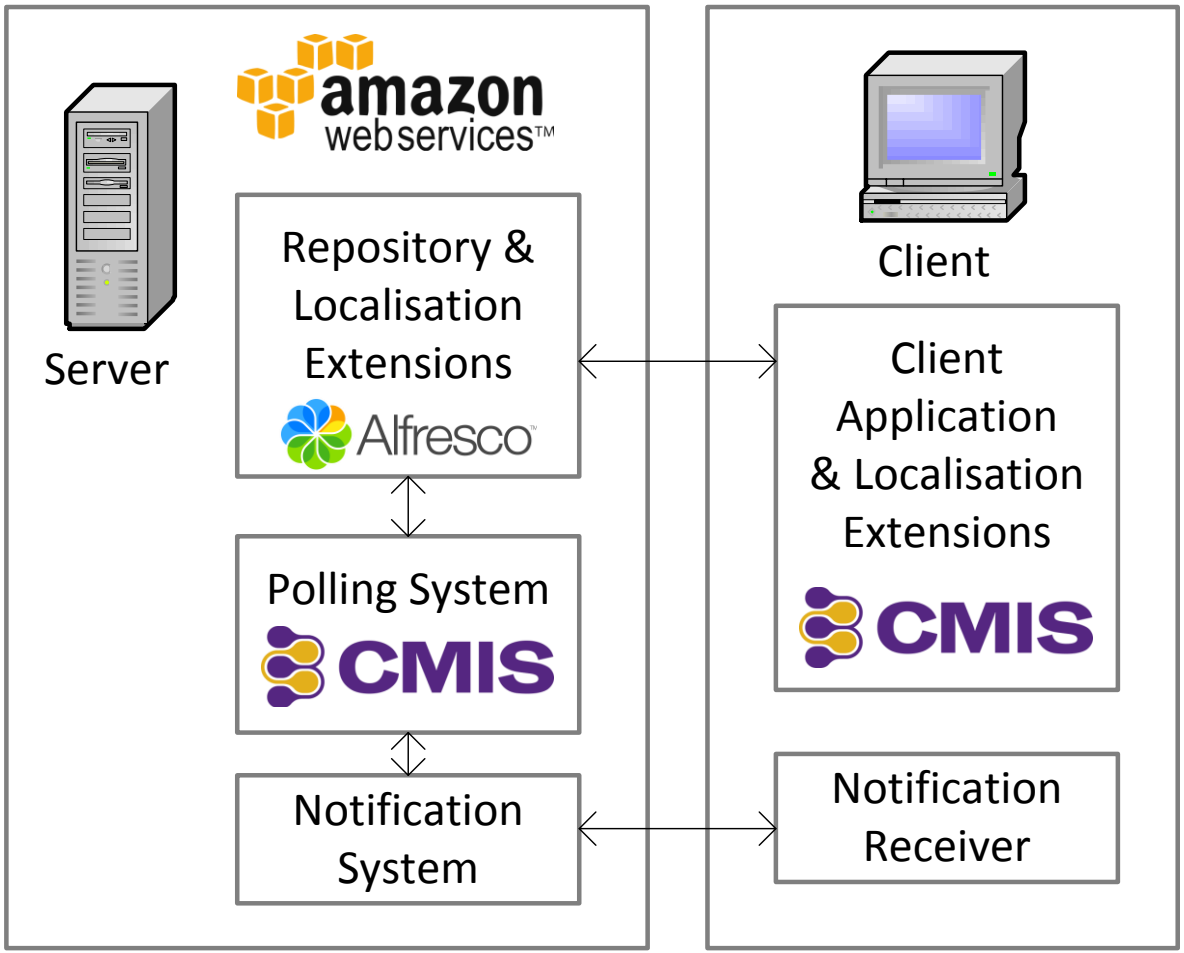
Polling Schemes



Document model with localisation



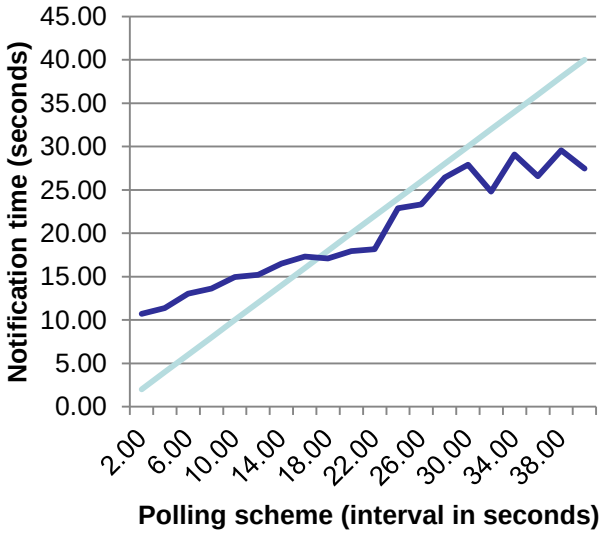
Technical setup

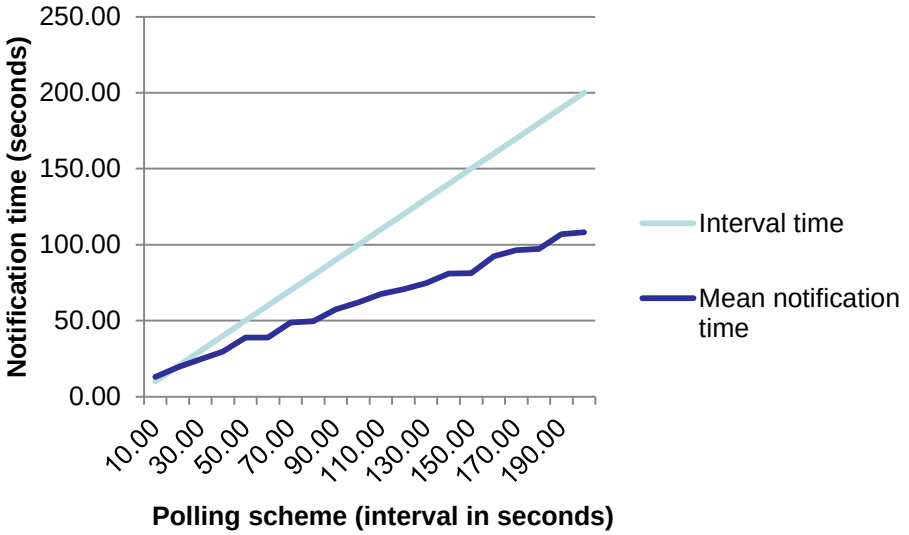


- Repository browser tool
- Polling system
- Notification system
- Test tools

Evaluation

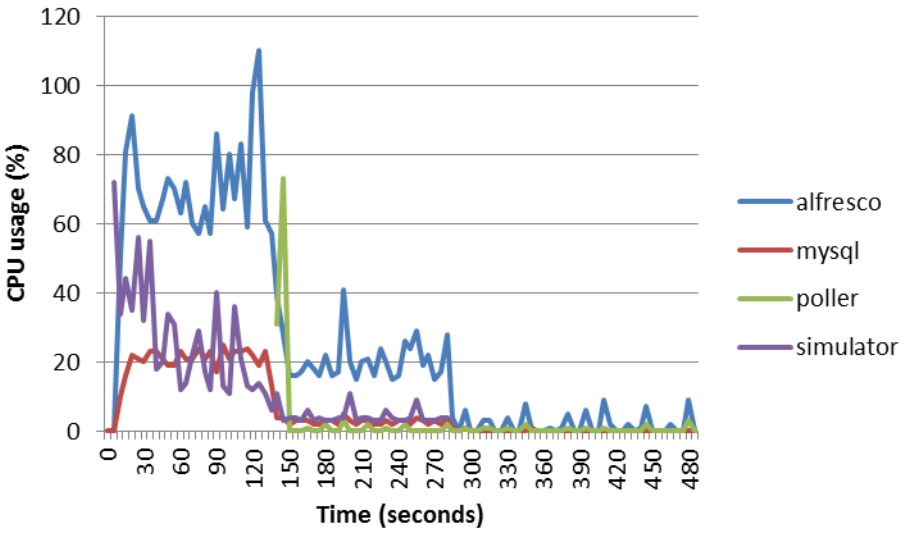
- Notification response time

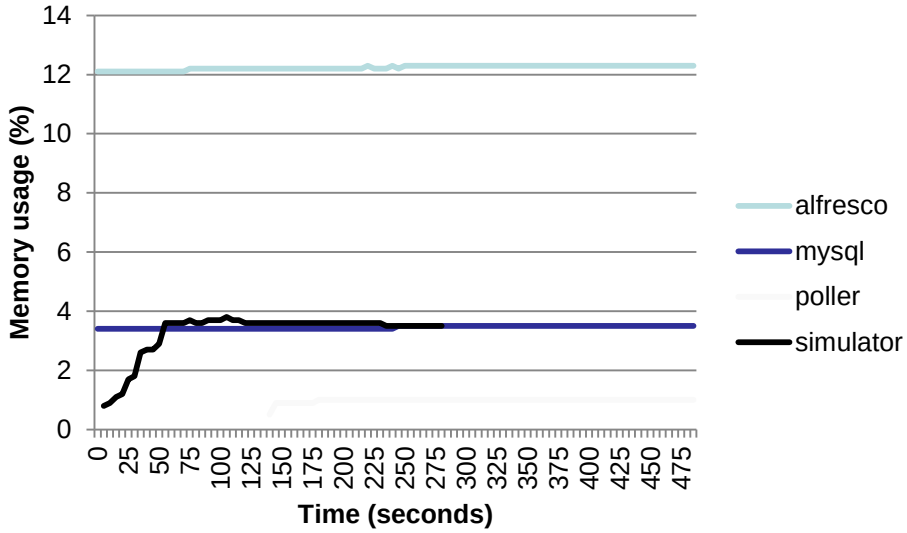




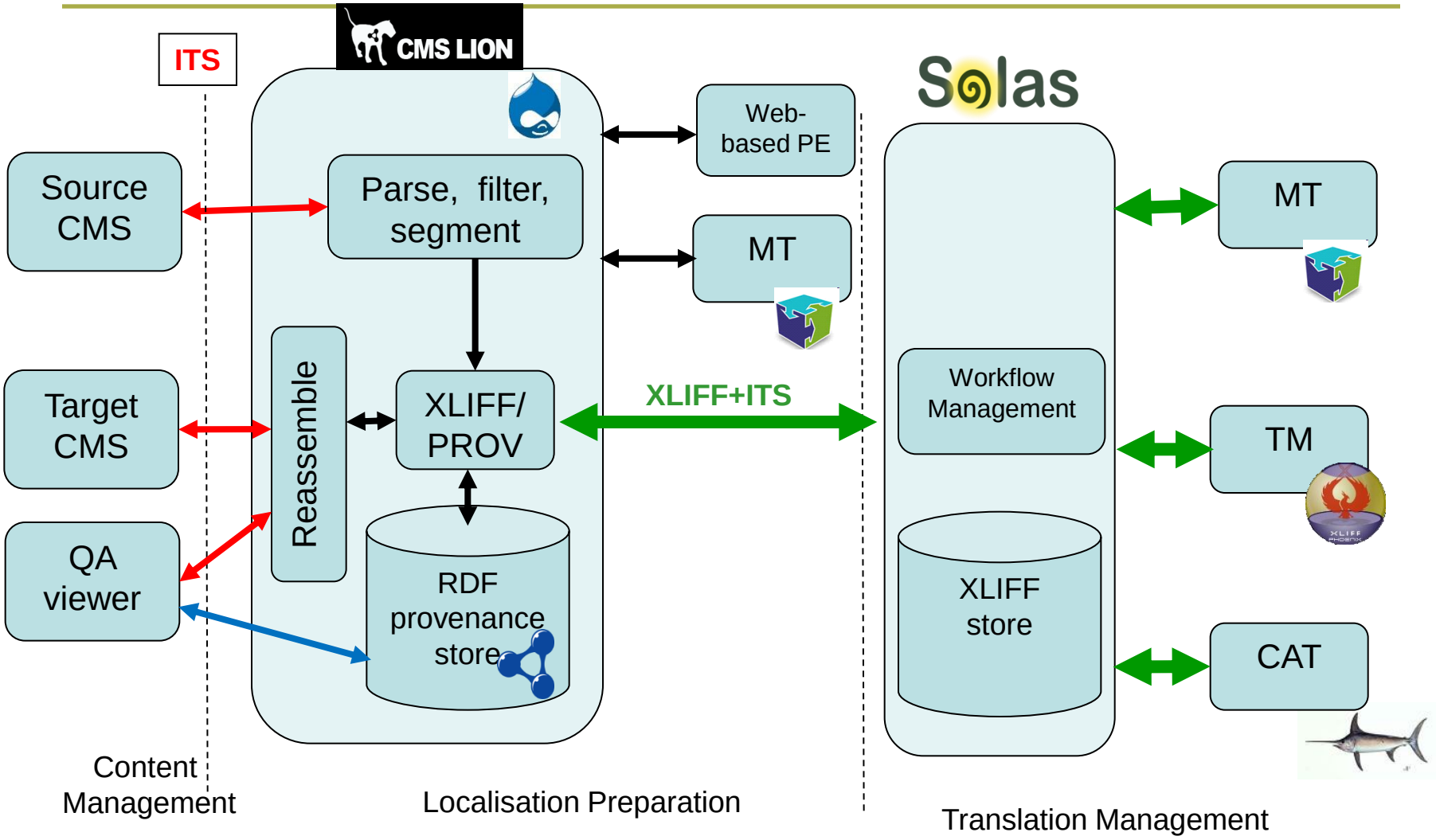
Evaluation

- Performance evaluation





Content Management - L10n Workflow Integration

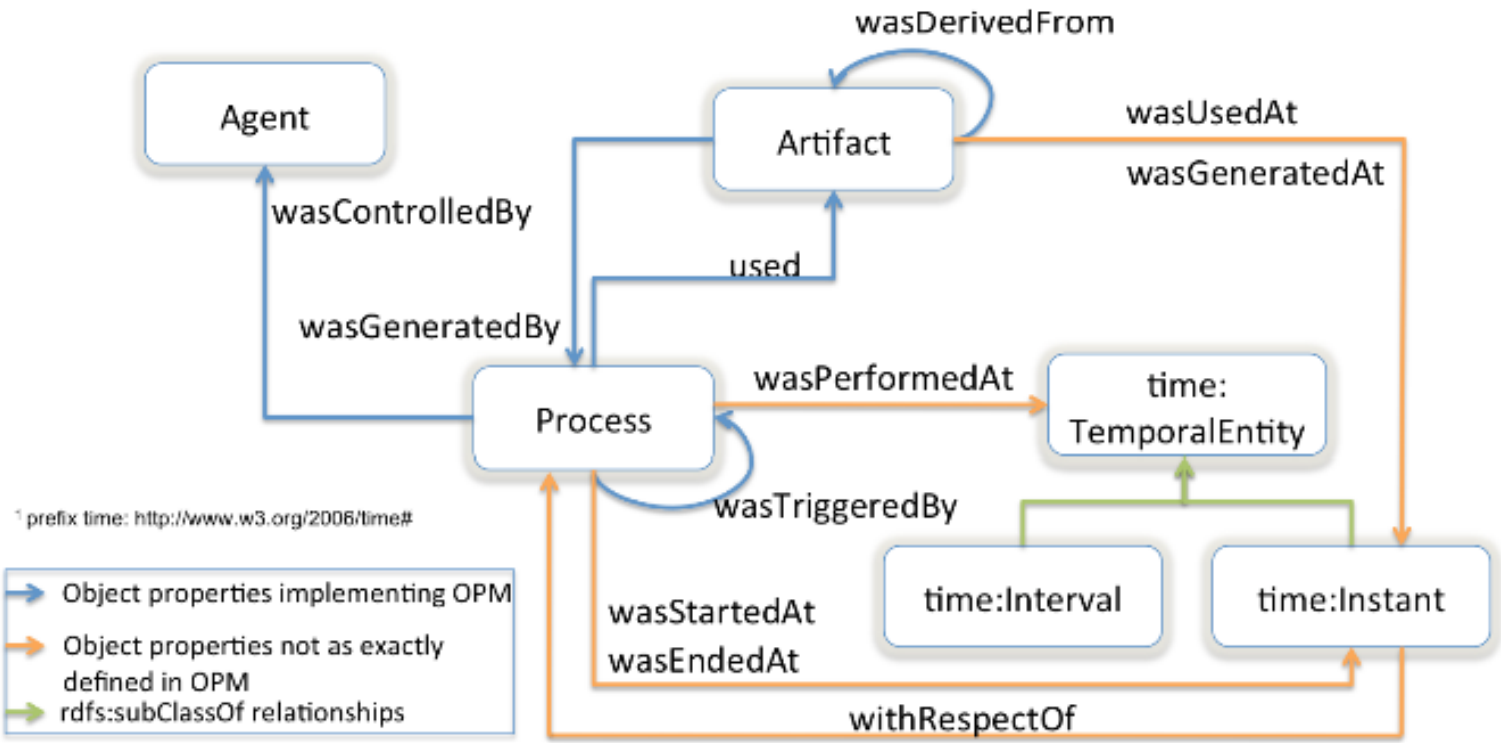


XLIFF and Open Provenance

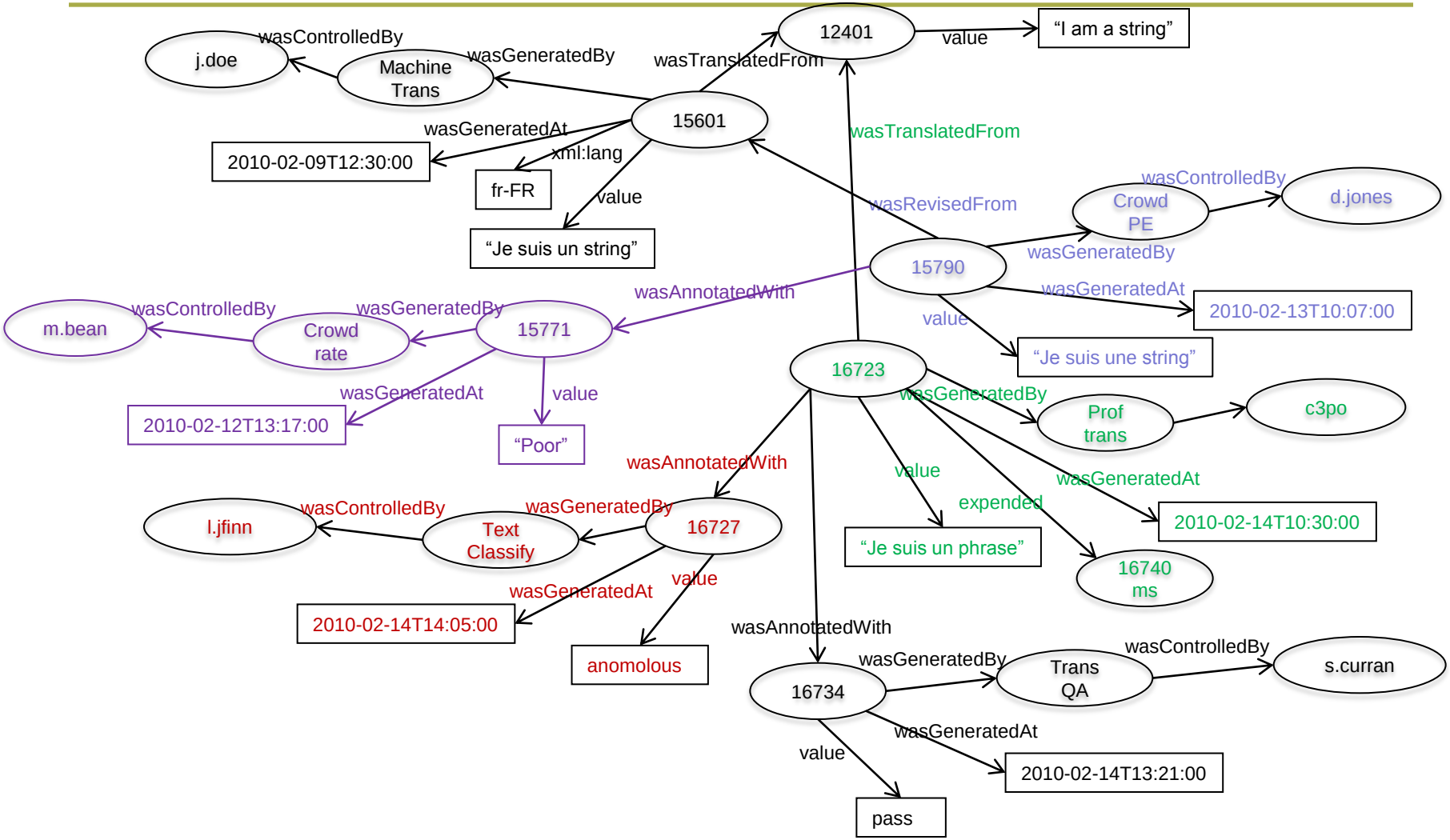
- Capture XLIFF **transformations** that operate on content and its meta-data as the result of **content processing** by different localisation workflow services
- A **provenance** model used to capture process operations
  - agents and properties of those processes
- Support **managing & auditing quality** of processes
  - correlating output of individual steps with professional, crowd and consumer judgement
  - support end-to-end process management
  - terminology management
- **On-demand** language resource assembly
  - e.g. for parallel text for MT training

Linked Localisation Data: RDF-based logging

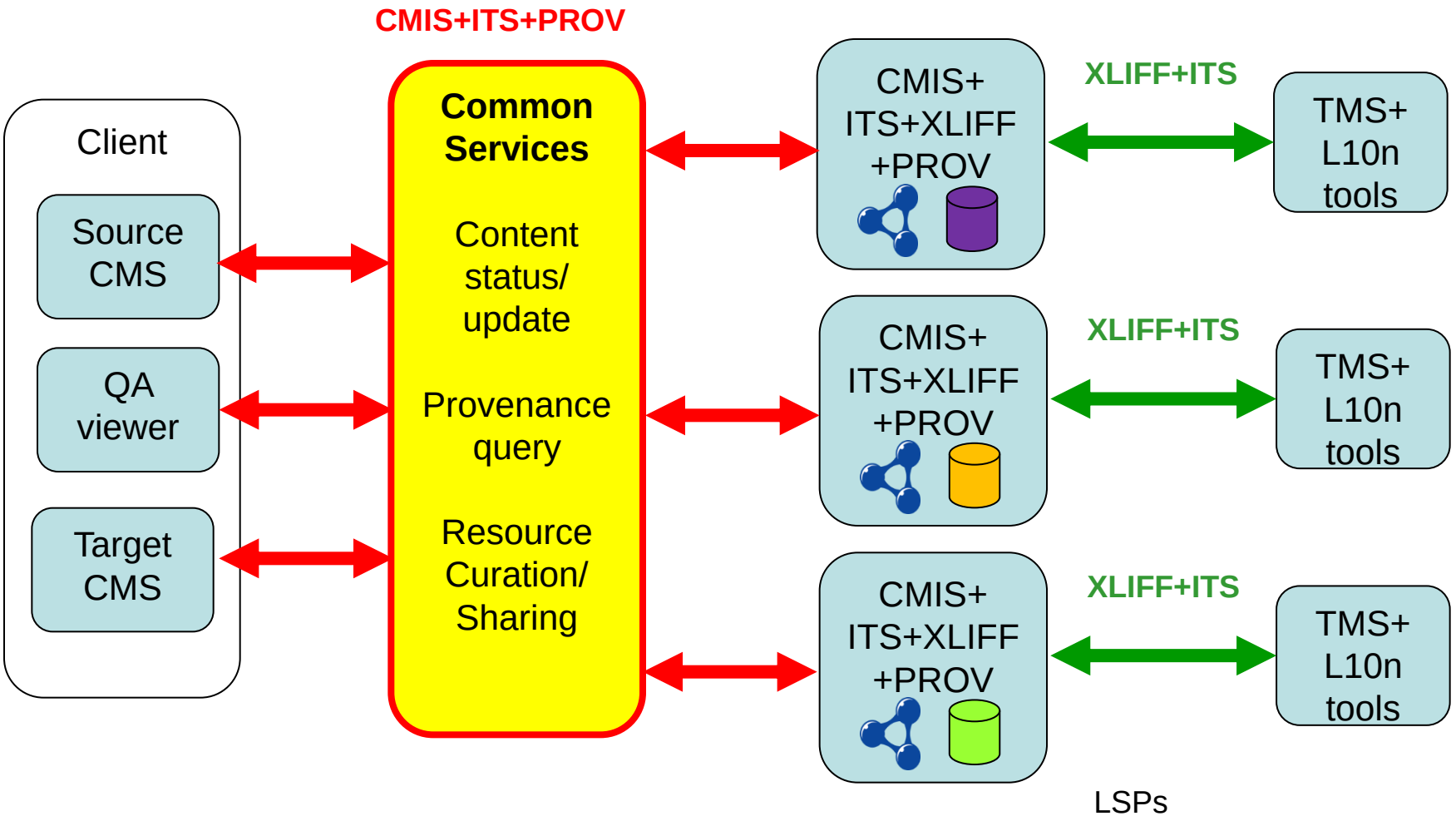
- Open Provenance Vocabulary
  - http://openprovenance.org/
  - Active W3C Provenance working group



LT Assisted Localisation Process Provenance



Future LSP-Neutral Open Service



Conclusion

- Have extended CMIS to support:
  - Document level ITS rules
  - Open document change notification mechanism
- Strong potential to streamline CMS-L10n integration in combination with XLIFF and PROV
- Achieved with current CMIS specification
  - Custom extension to folder object
  - Custom extension to policy object may be better
- Next Steps
  - Combining standards for vendor-neutral CMS integration
  - Aligh with ITS2.0 and XLIFF2.0
  - Discuss extensions with CMIS-compliant vendors

Questions.

THANK YOU.

Follow ITS Use Case at:

http://www.w3.org/International/multilingualweb/lt/wiki/CMS_Neutral_External_ITS_Rules_and_Readiness

Follow XLIFF+ITS mapping at:

http://www.w3.org/International/multilingualweb/lt/wiki/XLIFF_Mapping