D3.2: INITIAL L3DATA FEDERATION PLATFORM RELEASE

David Lewis, Leroy Finn, Kevin Koidl, Bruno Sanz del Campo, Alfredo Maldonado

Distribution: Public Report

Federated Active Linguistic data CuratiON (FALCON)

FP7-ICT-2013-SME-DCA

Project no: 610879

Document Information

Deliverable number:	D3.2 – revised version
Deliverable title:	Initial L3Data Federation Platform Release
Dissemination level:	PU
Contractual date of delivery:	31st May 2014
Actual date of delivery:	31st July 2014 – Updated 21st October 2014
Author(s):	David Lewis, Leroy Finn, Kevin Koidl, Bruno Sanz del Campo, Alfredo Maldonado
Participants:	TCD
Internal Reviewer:	DCU
Workpackage:	WP3
Task Responsible:	T3.2
Workpackage Leader:	XTM

Revision History

Revision	**Date**	**Author**	**Organization**	**Description**
1	6/5/2014	Leroy Finn	TCD	Initial platform documentation draft
2	31/7/2014	David Lewis	TCD	Review and add summary section
3	21/10/2014	David Lewis, Bruno Sanz del Campo, Alfredo Maldonado	TCD	2nd release, with updated introduction to section 2 and new section 3 on test results

Contents

1. Introduction

The L3Data Federation Platform implements the L3Data schema and system architecture and API defined by FALCON Deliverable D2.2: Initial L3Data Schema and Architecture. The L3Data store is implemented using existing linked data stores such as Sesame[1] and JENA[2] (with Jena being the configuration used in the original release). The implementation allows localisation project state to be recorded using the W3C Provenance RDF vocabulary, with specialisations taken from the CNGL Global Intelligent Content vocabulary as defined in D2.2. In this initial release the project state is captured from the XTM Cloud translation management tools in the form of a Translation Interoperability Package Protocol unit, and in particular the XLIFF1.2 file it contains. This is performed by a component called the Logger, which interacts with XTM Cloud using an open RESTful API. A further API is offered giving access to SQARQL query functionality over the project state. With this initial release the L3Data Federation Platform offers query, access and recording of L3Data to other WP3 components, primarily the translation management tool set centred on XTM Cloud. The APIs offered also enable the development of further web based tools to support monitoring of workflow and different LT component and human worker performance through visualisation of provenance queries conducted over federated L3Data stores, as well as to generate reuse audit reports. These tools will be developed as part of the FALCON showcase system.

This document provides an overview of the platform, as already specified in D2.2. This initial release will be made available via the FALCON Github repository[3]. While this document gives an overview of this release, those wishing to explore the platform in more detail are referred to the Github for the latest code and documentation.

2. L3Data Platform Release Overview

Figure 1 highlights where the L3Data Platform sits within the FALCON Showcase Systems architecture and its connections to other components. The L3Data platform consists of an open source linked data store, currently configured using the Jena-based store Fuseki[4], and a logger component that converts existing language resource and bi-text exchange formats into RDF according to the L3Data Schema. The logger is designed to flexibly support various bi-text formats, and is currently configured to support XLIFF1.2 files as generated by XTM Cloud, and such XTM files as included in TMS Interoperability Protocol Package format[5] as generated by XTM Cloud. However, it is also used to explore other mappings, and thereby allows the L3Data Schema to be expanded and validated for different use cases.

[1] http://www.openrdf.org/

[2] https://jena.apache.org/

[3] https://github.com/CNGL-repo/Falcon

[4] http://jena.apache.org/documentation/serving_data/

[5] https://code.google.com/p/interoperability-now/downloads/detail?name=The_TMS_Interoperability_Protocol_Package-1.4.1.pdf&can=2&q=

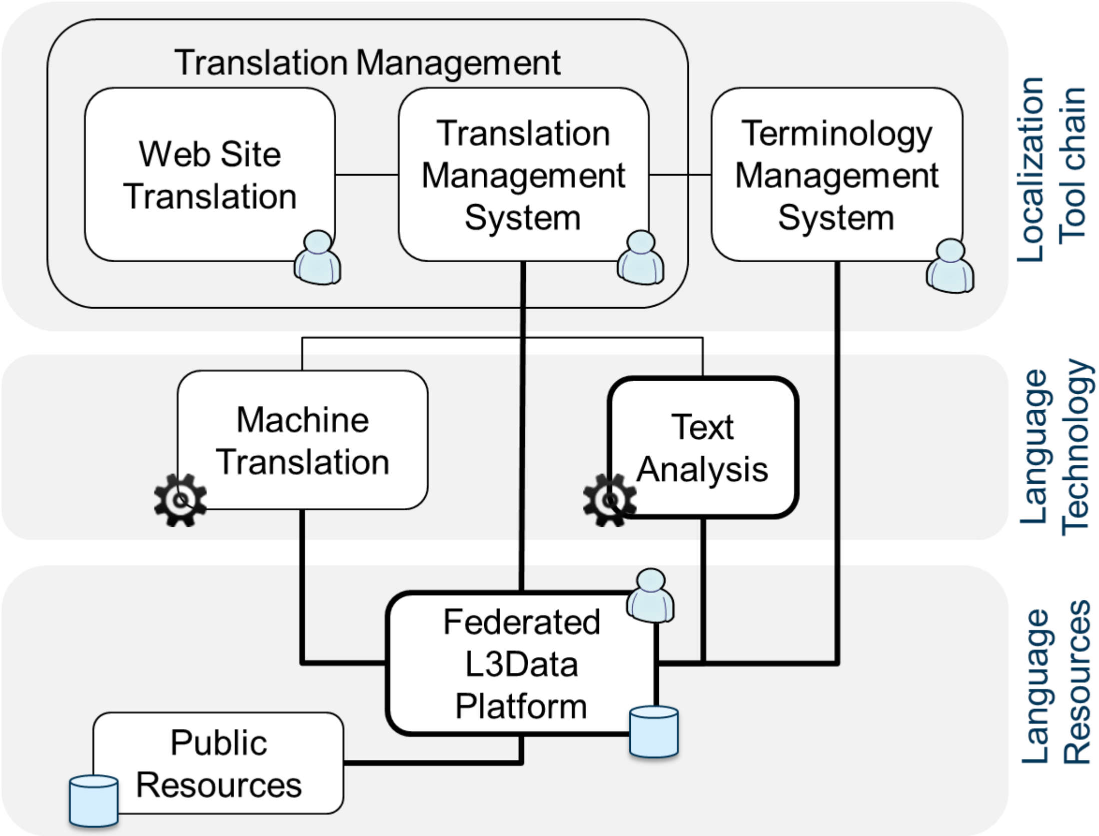
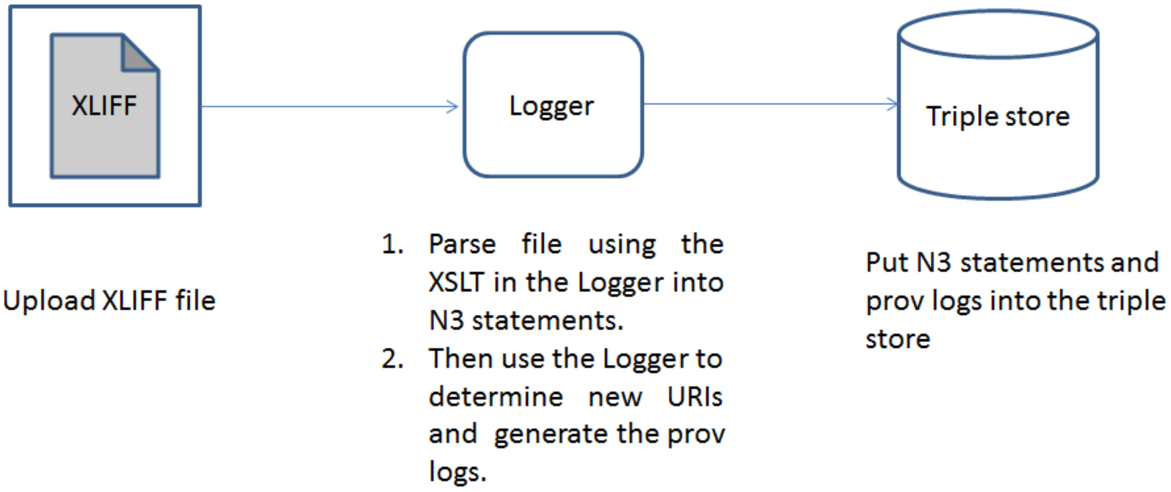


Figure 1: Elements of the Translation Tool Set reported in this document with components and elements highlighted

2.1. Logger Overview



Upload XLIFF file

1. Parse file using the XSLT in the Logger into N3 statements.
2. Then use the Logger to determine new URIs and generate the prov logs.

Put N3 statements and prov logs into the triple store

Figure 2: Logger Overview

The Logger component (figure 2) is a servlet which receives successive versions of XLIFF files (provided by translation systems such as XTM Cloud upon completion of each process), transforms this XLIFF input into RDF statements and merges these statements into the triple store. It also generates extra RDF statements to track provenance information.

The Logger performs the following tasks for each input file it receives:

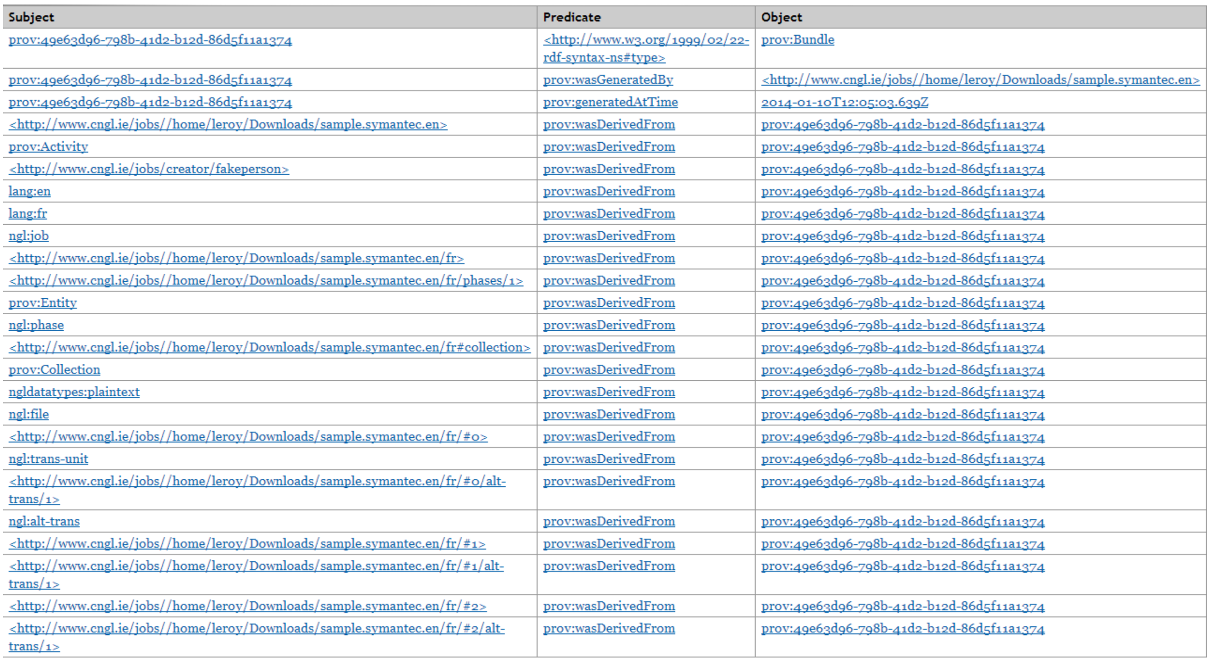
1. Transform the input XLIFF file into a set of RDF statements. This is done through an XSL transformation which processes the XLIFF document to produce N3 statements.
2. Merge these RDF statements into the triple store.
3. Add provenance information. A new "log" URI is created with associated statements, such as a timestamp and an input "job", and is linked using a provenance statement to each of the newly added URIs that have resulted from the merge.

2.2. Provenance logging Algorithm

1. Upload a XLIFF document.
2. Parse the XLIFF document using XSLT into N3 statements[6] (sample from triple store shown in Figure 3).
3. Then check the generated N3 statements URIs against the URIs in the triple store. This check is done at the merge stage where new URIs and data are added to the graph and old URIs are being merged with the graph along with old/new data
4. A prov id is generated by a Universally Unique Identifier (UUID) generator which is available **java.util.UUID** package. A UUID is required as the provenance URI has to be unique. Then do the following dependent on the results of step 3:

```
if(there are new URIs) then
        create a new prov record which indicates what file has generated this new
        prov record, the time the prov record is generated at and the new URIs that
        were added into the triple store
else then
        just make a prov record which record which indicates what file has generated
        this prov record and the time the prov record is generated at
```

[6] http://www.w3.org/2000/10/swap/Primer



Subject	Predicate	Object
prov:49e63d96-798b-41d2-b12d-86d5f11a1374	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	prov:Bundle
prov:49e63d96-798b-41d2-b12d-86d5f11a1374	prov:wasGeneratedBy	<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en>
prov:49e63d96-798b-41d2-b12d-86d5f11a1374	prov:generatedAtTime	2014-01-10T12:05:03.639Z
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
prov:Activity	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs/creator/fakeperson>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
lang:en	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
lang:fr	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngl:job	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/phases/1>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
prov:Entity	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngl:phase	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr#collection>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
prov:Collection	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngldatatypes:plaintext	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngl:file	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#0>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngl:trans-unit	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#0/alt-trans/1>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
ngl:alt-trans	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#1>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#1/alt-trans/1>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#2>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374
<http://www.cngl.ie/jobs//home/leroy/Downloads/sample.symantec.en/fr/#2/alt-trans/1>	prov:wasDerivedFrom	prov:49e63d96-798b-41d2-b12d-86d5f11a1374

Figure 3: Sample provenance log

2.3. Updating the XLIFF to RDF transformation

The transformation of XLIFF files to RDF is done through the XSLT document in Logger/src/xliff2n3/xliff2n3.xsl. All statements other than the provenance statements (which are added programmatically in Logger/src/xliff2n3/Graph.java using URIs defined in Logger/src/xliff2n3/Constants.java) originate from this XSLT document, which can be tweaked to replace the grammar and the URIs used. Namespaces are included in this XSLT document too, and will be automatically added to the triple store server namespace list. After making any changes to the XSLT, you will probably want to clear the triple store and re-process the sample files to see the effect of your changes.

2.4. Deployment

Configuring for triple store server

Create a repository/dataset named for example CNGL. The type (DB or file-based) is up to you.

Configuring the Logger for various triple stores

- For Logger with Sesame, edit entries DFLT SESAME URL and DFLT REP ID in config.properties to match the URL of your Sesame Server and the name you have chosen in 3.3.1.
- For Logger with Jena, edit entries DFLT JENA URL SPARQL ENDPOINT and DFLT JENA URL UPDATE ENDPOINT in config.properties to match the URL of your Fueski Server and the name you have chosen in 3.3.1.

Using Java class

Check out Javadoc for more details of the classes and functions in the Logger Java project.

3. Tested Features and Interfaces

The following interfaces to the L3Data Platform have been implemented and tested:

- XLIFF and TIPP package from XTM Cloud: test files from translation jobs in XTM have been run through the Logger component to generated L3Data. This is how live translation project will be logged as L3Data.
- Public Terminology Resources: Samples of the EURVOC term base have been converted into L3Data using the LEMON vocabulary[7]. This shows how public terminology resources can be converted into L3Data.
- Public Translation Memory: Samples of the DG-T Translation Memory release[8] has been extracted and cross linked with a sample of EURVOC in order to evaluate the use of terminology frequency to prioritise machine translated segments for postediting. This test set was generated as Comma Separated Value files as is typical for LT evaluations, and these were mapped into the L3Data format, together with the cross-links to the above EURVOC entries.
- TBX Resources: FALCON has contributed to an open source java project to convert TBX into LEMON[9]. This has not yet been replicated in a configuration of the Logger component, but provides a consensus implementation of how the mapping should operate. This will enable L3Data updates of terminology to be received from TermWeb.

4. Summary and Next Steps

This initial release of the L3Data platform offers:

- Data federation capabilities through the exposure of translation project status as fine-grained linked data references via URLS;
- An open API conforming the TIPP and XLIFF standards for capturing translation project status, and thereby offering interoperability with platforms beyond XTM Cloud that implement these standards;
- An L3Data interface that enables relevant aspects of translation project workflows to be queried and thereby easily presented to project managers and LT component configuration personnel to assist decision making. This is based on a standard SPARQL query interface.
- The ability to extract meta-data from provided TIPP/XLIFF files, including the TIPP project meta-data and ITS2.0 meta-data such as source and target term or phrase annotations, machine translation confidence scores and translation provenance engine to track involvement of individual translators and MT engines. The XLIFF/ITS mapping resulted from collaboration with the W3C ITS Interest Group[10], including contribution from FALCON is refining the ITS2.0 RDF vocabulary[11]. The TIPP-RDF mapping is the result of a novel proposal

[7] http://lemon-model.net/

[8] https://open-data.europa.eu/en/data/dataset/dgt-translation-memory

[9] https://bitbucket.org/account/signin/?next=/vroddon/tbx2rdf

[10] http://www.w3.org/International/**its**/ig/

[11] http://lists.w3.org/Archives/Public/public-i18n-its-ig/2014Jun/0019.html

by FALCON. It is subject to on-going refinement in collaboration with the LinPort project[12], which is responsible for the meta-data structures in TIPP.

Based on the experiences in developing this initial release, the platform will be advanced through the following collaboration, development, integration and test activities.

- The degree to which the project status meta-data can be captured in L3Data is restricted by the degree to which that meta-data is included in the XLIFF produced by other components that invoke the logging API. This will be evaluated and advanced through integration with the other FALCON components for translation management, machine translation, named entity and term recognition and terminology management that will be documented in deliverables D3.3, D3.4 and D3.5.
- Currently, L3Data support provenance meta-data associated with translation projects. While this is a source of LT component training data especially within a translation project, valuable L3Data can also be obtained by harvesting published language resources. FALCON is currently participating in activities to harmonise the format used for such language resources. It is collaborating with the Linked Data for Language Technology (LD4LT) W3C Community Group[13] to develop an RDF vocabulary for language resources based on the META-SHARE schema[14], the W3C DCAT Dataset meta-data vocabulary[15] and the W3C Provenance vocabulary[16]. The consensus from this work will inform the next iteration of the L3Data schema in deliverable D2.3.
- Federated access control will be included in the platform in upcoming iterations. FALCON is already involved in public consensus building activities related to access rights and rule for language resources[17] using the existing vocabularies such as Creative Commons[18] and the Open Digital Rights Language (ODRL) Ontology[19]. Results from this collaboration will inform the implementation of future iterations of the L3Data schema and platform implementation.
- The FALCON project initially aimed to use Named Entity Recognition for terminology extraction as outlined in D3.3. The project has also been evaluating the use of disambiguation service offered by the BabelNet[20] platform with promising results. Babelnet is a large multilingual lexical-encyclopaedic resource mined from lexical resources such as Wikitionay and Wordnet as well as encyclopaedic resources such as DBPedia and Wikipedia. As a third party resource, Babelnet is no under the direct control of potential L3Data users for the purposes of retraining. However, collaboration has begun with the developers of BabelNet to explore open update protocols between users of Babelnet who apply in it in professional linguistic roles, such as terminology management and translations. Such users are in a position to provide corrections or updates to the underlying resource that they

[12] http://www.linport.org/
[13] http://www.w3.org/community/ld4lt/
[14] https://www.w3.org/community/ld4lt/wiki/Meta-Share_OWL_metamodel
[15] http://www.w3.org/TR/vocab-dcat/
[16] http://www.w3.org/TR/prov-o/
[17] https://www.w3.org/community/ld4lt/wiki/Licensing_information
[18] http://creativecommons.org/ns
[19] http://www.w3.org/ns/odrl/2/
[20] http://babelnet.org/

encounter, however an open correction reporting protocol is required in order to motivate the integration of correction reporting features in existing tools. Specifically, FALCON is participating in the development of the Ontolex vocabulary by a W3C community group[21]. This may provide a suitable basis for such a reporting feature as the vocabulary will be an update of the LEMON vocabulary[22] currently used to structure data in BabelNet. Such a feature will therefore be considered in future updates to the Platform, and lexical features in the L3Data Schema will make use of LEMON and its Ontolex successor.

- The impact of L3Data on public sector language resources and public language technology services may most directly be made achieved through contribution to the development of guidelines for the procurement of future Connecting Europe Facility digital service for Automated Translation[23] envisaged by the European Commission. FALCON has made an initial contribution to the structuring of open data management requirements and guidelines[24] as an input to the MLi project[25], which is tasked with developing technical requirements for public automated translation services. This contribution has been published via the W3C ITS Interest Group with collaboration of the LIDER and QTLaunchpad projects, and has already received input and comment from TAUS, GALA and the EU publication office. It is hoped that this initiative will provide a platform for aligning L3Data with wider stakeholder requirements and the showcase the L3Data platform as a proof of concept of open data management for public MT services. It will provide a focus both for the interaction with lexical-conceptual services described above and possible best practice guidelines for published bi-text as linked open data[26], as being considered by the W3C Best Practice in Multilingual Linked Open Data (BPMLOD) Community Group.

[21] http://www.w3.org/community/ontolex/

[22] http://lemon-model.net/

[23] https://ec.europa.eu/digital-agenda/en/connecting-europe-facility

[24] http://www.w3.org/International/its/wiki/Open_Data_Management_for_Public_Automated_Translation_Services

[25] http://mli-project.eu/

[26] http://www.w3.org/community/bpmlod/wiki/Draft_Guidelines_on_Bitext_as_Linked_Data