http://latc-project.eu

D4.2.1 Initial Sustainability Report

Project GA No.	FP7-256975
Project acronym	LATC
Start date of project	2010-09-01
Document due date	2011-08-31
Actual date of delivery	2011-08-31
Lead Partner	TALIS
Reply to	Keith Alexander, keith.alexander@talis.com
Document status	FINAL

Project GA No.	FP7-256975
Project acronym	LATC
Project full title	Linking Open Data Around The Clock
Dissemination level	PU
Number of pages	10
Task responsible	TALIS
Other contributors	DERI, VUA, FUB, INFAI
Author(s)	Keith Alexander
EC Project Officer	Stefano Bertolo
Keywords	sustainability

Table of Contents

1 Executive Summary

This document outlines the sustainability of the output of the LATC Support Action, as a whole, and of its component parts, after the end of the project in 2012.

No major obstacles are foreseen for sustaining the LATC 24/7 Platform or any of its components; Sindice Ltd have pledged to run any components not hosted and run elsewhere, until at least the end of the LOD2 project[1] in September 2014.

LATC partners have worked with the Open Knowledge Foundation through LOD2 to ensure sustainable hosting of those triplified EU datasets belonging to WP2 that are published at http://publicdata.eu. LATC consortium members will host other WP2 datasets sustainably directly.

[1] http://lod2.eu/

2 WP1 24/7 Platform

2.1 Platform

The EU FP7 project "LOD2 - Creating Knowledge out of Interlinked Data" will run until September 2014, and take responsibility for running the 24/7 Platform after the LATC project completes in 2012.

It may be possible to extend the usefulness of the 24/7 Platform by running it on a commercial basis. Many organisations spend time and money producing, sourcing, cleaning, and combining the data they need for their operations. Increasingly, organisations in both the private and public sector are becoming aware of the promise of easy data integration with numerous heterogeneous sources offered by Linked Data. And while Linked Data does make data-integration considerably easier, there are still sufficient challenges involved in linking between datasets, which the 24/7 Platform helps with, that organisations may be willing to pay for:

- Identifying candidate link targets and devising rules for linking. Although the Dataset Inventory and Silk Workbench will help people perform this task, organisations may benefit from consultancy services offering expert guidance.
- Quality Assurance; for serious use cases it is important for organisations to know how reliable a linkset is. It maybe also be useful for organisations to be able to determine how accurate a linkset is; depending on the use case a larger less accurate linkset may be preferred to a smaller less accurate one, and vice versa. A business offering based on the 24/7 Platform could provide a valuable QA service tailored to meet market needs as they emerge.
- Updates: The 24/7 Platform tracks updates to datasets and can regenerate linksets as the target datasets change, whilst also providing notification to linkset consumers; this is a useful and convenient service for which customers might pay.
- Access may be restricted to datasets of, for example, a commercially sensitive nature, and require authentication and/or payment. Dataset owners and consumers might pay for their data to be linked without being made public.
- Access to Linksets produced by a 24/7 Platform-derived business could be sold on a commercial basis through Talis's Data Marketplace, Kasabi[2].

2.2 Components

The 24/7 Platform produced by WP1 consists of either freely available open source software (such as Silk) or open interfaces run as an ongoing commercial concern (such as Sindice). This means that users may download and run it on their own hardware and its usefulness may exceed the lifespan of its deployment within the LATC project.

[2] http://kasabi.com/

2.2.1 Workbench

The Silk Workbench is a deployment of the Silk Link Discovery Framework software[3]; it will be hosted and run by Freie Universität Berlin until at least September 2014 as part of the LOD2 Project. The LATC customised Silk Workbench will be also kept up to date until at least September 2014.

The Silk Link Discovery Framework is developed by members of the Web-Based Systems Group at Freie Universität Berlin, Germany. In addition to being available to the LATC project, the Silk Link Discovery Framework will be under further development until September 2014 in the LOD2 Project. In LOD2 it will be enhanced with a Linking Assistant.
Furthermore the Silk Link Discovery Framework is part of the Linked Data Integration Framework (LDIF) and will be further developed in joint projects with ontoprise GmbH, like SMW+LDE.

Silk is published as open source under the terms of the Apache Software License. Future releases will be announced on the Silk mailing list[4], the source code as well as installation archives are available at the major open source repository Assembla[5].

2.2.2 Metadata Store (MDS)

The LATC Metadata Store (MDS) is hosted on the Talis Platform. In addition to being available to the LATC project components which read from, and write to, it, (including the Dataset Inventory which makes the data publicly available and browsable), the Metadata Store will also be available through Kasabi Data Marketplace (a Talis Group business), where application developers will be able to use, monitor, and merge it with other datasets. The interfaces by which other LATC components communicate with the MDS, is run as a Managed Service by Talis Consulting, and Talis intend to run this ongoing for at least 3 years after the end of the LATC project, beyond which, the decision to continue maintaining it will be based on its usage.

2.2.3 Dataset Inventory (DSI)

The Dataset Inventory (DSI) is developed using an open source software project called Puelia[6], which implements the Linked Data API[7], and has several deployments at data.gov.uk subdomains to serve Linked Data. Puelia is used by Talis Consulting in providing bespoke Linked Data hosting for their customers, and by Kasabi, so the software project is very likely to continue to be maintained. The Dataset Inventory as a deployment will continue to be maintained for at least 3 years after the end of the LATC project, with assessment thereafter based on its usage.

[3] http://www4.wiwiss.fu-berlin.de/bizer/silk/
[4] https://lists.sourceforge.net/lists/listinfo/silk2-discussion
[5] http://www.assembla.com/spaces/silk/
[6] http://code.google.com/p/puelia-php/
[7] http://code.google.com/p/linked-data-api/

2.2.4 Console

When the LATC project ends in 2012, DERI will host the Console until at least September 2014. The source code is openly licensed and available on Github[8]. VUA plan to maintain the Console as necessary (under the limit of one developer-day a week) until September 2013.

2.2.5 Crawler & Indexer

The Crawler & Indexer exists as part of Sindice, available through Sindice's APIs, and is expected to continue being available for the foreseeable future (certainly until the end of the LOD2 project), run by DERI.

2.2.6 Runtime

The Runtime consists of Silk Link Discovery Framework (Map Reduce version) and Hadoop[9], both of which can be expected to developed and maintained into at least the medium term (Silk, as mentioned previously, will continue to be developed within LOD2). DERI will run and maintain the deployment until at least the end of the LOD2 project in September 2014.

2.2.7 QA Module

The architecture of the QA Module has changed since the Description of Work was written, and now consists of "Internal QA", which is done between both the Runtime and the Workbench, and "External QA"[10].

Internal QA. The deployment of the Internal QA does not exist separately from the Workbench and the Runtime; sections 2.2.1 and 2.2.6 describe the sustainability of those components.

External QA. The external QA will be run at http://qa.linkeddata.org/ by InfAI for at least 2 years after the end of the LATC project (until 2014). The code is open source and available on Github[11].

2.3 External Dependencies

CKAN[12] is directory of datasets available on the web run by the Open Knowledge Foundation[13]. The Metadata Store (MDS), and consequently the Dataset Inventory (DSI) and Workbench, rely upon CKAN to provide descriptions of datasets. This requires that:

- CKAN remains live.

[8] https://github.com/LATC/24-7-platform/tree/master/latc-platform/console
[9] http://hadoop.apache.org/
[10] See D.1.4.1 First Deployment of QA Module
[11] https://github.com/AKSW
[12] http://ckan.net
[13] http://okfn.org

- Linked Data publishers (and/or that others in the Linked Data community) register their datasets with CKAN and keep the metadata updated as necessary.
- The MDS can continue to synchronise metadata about the curated datasets with CKAN using the CKAN API.

Open Knowledge Foundation plan to continue running and maintaining CKAN and its API(s) on an ongoing basis.

Since CKAN has been used as the canonical source of data about datasets by the producers of the LOD cloud diagram, and since the diagram is such a widely used tool in Linked Data advocacy and education, Linked Data publishers have been motivated to register their datasets with CKAN and keep the record up to date.

In addition, it may also be possible to make it easier for publishers to create and update their CKAN record by providing a tool to synchronise the CKAN description with the metadata provided in their VoID document[14]. Talis and/or DERI will seek to provide such a tool by the end of the LATC project unless this need is met by others, or in some other way.

CKAN recognise the importance of the stability of their API and have adopted the practice of 'versioning' their API when they make changes that would otherwise break consuming applications. DERI collaborate with the OKF directly and via the LOD2 project[15] to ensure the dataset metadata is provided sustainably.

Talis will keep the data in the Metadata Store up to date as part of the maintenance. Ideally this will continue to be done through synchronising with CKAN, but if in the future CKAN were to become unavailable or unsuitable, Talis would seek to update the Metadata Store from other sources (for example, VoID documents published by dataset publishers).

[14] Metadata about a Linked Data datset written using VoID: Vocabulary of Interlinked Datasets (http://rdfs.org/ns/void)
[15] http://lod2.eu

3 WP2 Test-bed for Data Intensive Applications: EU Data

3.1 Business Related Datasets

Linked Data versions of EURES[16] and CORDIS[17] have been published by FUB (EURAXESS has still to be published). The conversions run on an automated schedule that keeps the datasets up to date and FUB aim to continue to support them after the LATC project ends.

INFAI have published FTS, hosted at http://fintrans.publicdata.eu, which will continue to be maintained by Sindice.

3.2 Legal & European Institutions Datasets

When published (due M18), these datasets will be hosted sustainably (for example at http://dataincubator.org or http://publicdata.eu), the conversion code will be released under an open source license, and the primary data provider will be contacted with a view to encouraging them to publish Linked Data themselves.

3.2.1 Eurostat

Eurostat will be published at http://eurostat.linked-statistics.org; the domain is owned by Richard Cyganiak at DERI, and the server is run by DERI. The code that converts Eurostat to Linked Data is available on github[18]. DERI plans to continue to publish Eurostat as Linked Data indefinitely; keeping the data in synch with the original source[19]. DERI plan to show http://eurostat.linked-statistics.org to contacts at Eurostat as a demonstration of how they can publish their data on the web.

3.2.2 Eventseer

VUA have created a Linked Data version of Eventseer[20], an Academic events and networking site. The Linked Data is currently published at http://linkeddata.few.vu.nl/eventseer; VUA are also in contact with Eventseer's developers to help them publish Linked Data themselves.

[16] http://www4.wiwiss.fu-berlin.de/eures/
[17] http://www4.wiwiss.fu-berlin.de/cordis/
[18] https://github.com/LATC/EU-data-cloud/tree/master/institutions/Eurostat
[19] http://epp.eurostat.ec.europa.eu/
[20] http://eventseer.net

4 Conclusions

Sustainability has been an important aim in the design of LATC; the investment has been in enhancing and co-ordinating existing software projects and resources, rather than on creating new ones that depend upon LATC funding to run.

The most significant part of LATC's sustainability is on aligning the project with LOD2, which runs for two years longer than LATC, and gives project partners scope to continue supporting LATC's outcomes. LATC project partners have committed individually to sustaining those 24/7 Platform components not supported by LOD2 for at least as long, or even longer. Beyond that, it may be possible to make the Platform generate revenue to sustain itself by building a business model (or business models) around it (and/or its parts).

The datasets in WP2 will be similarly well-supported for the medium term, both by LOD2 and individual project partner commitments. In the longer term, the most sustainable (and desirable) solution is that the original data providers begin to publish Linked Data themselves; LATC will engage with the original data providers to try to help them do so if possible, using the WP2 triplified datasets as demonstration tools.