xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D8.1.4

Impact and Continuation Plans for xLiMe Services and Software

Editor:	Ronald Denaux, ISOCO
Author(s):	Ronald Denaux, ISOCO; Blaz Novak, JSI; Aditya Mogadala, KIT; Lei Zhan, KIT; Andreas Thalhammer, KIT; Dubravko Culibrk, UNITN
Deliverable Nature:	Report (R)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M18 – 30 April 2015
Actual Delivery Date:	M18 – 30 April 2015
Suggested Readers:	All project partners
Version:	1.0
Keywords:	Software; web services; data sets; dissemination; target audience; guidelines

Disclaimer

Full Project Title:	xLiMe – crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP8 Dissemination, Exploitation, and Community Building
Document Title:	D8.1.4 – Impact and Continuation Plans for xLiMe Services and Software
Editor:	Ronald Denaux, ISOCO
Work Package Leader:	Dubravko Culibrk, UNITN

Executive Summary

This document presents current and planned measures to maximise the lifetime and reusability of both software and data produced as part of the xLiMe project. We start by presenting various factors which affect the current impact and continuation of software and data, such as publication and quality, in the first year of the project. Then, we present our current plans for improving on the impact and continuation of the xLiMe software during and after the project, e.g. open sourcing (or commercially exploiting) software and publishing services or data. Finally, we analyse whether the current factors and plans adhere to the original dissemination plan and how the current delivery of software outputs can be improved to maximise the impact and continuation potential.

Table of Contents

Abbreviations

ASR	Automatic Speech Recognition
API	Application Programming Interface
EPG	Electronic Programme Guide
GPU	Graphics Processing Unit
GUI	Graphical User Interface
OCR	Optical Character Recognition
QoS	Quality of Service
SPARQL	Simple Protocol and RDF Query Language
TBD	To be decided

1 Introduction

During the xLiMe project, the xLiMe consortium has begun developing many software components, generating various datasets and setting up various data processing services.[1] In order to maximise the return of investment of the project, we look in this document at what is currently being done as part of the project to promote the impact of this software, data and services. The main goal is to assess which current activities can maximise the impact of the software being developed and to plan further activities with this focus. Besides the impact of the software output in terms of how much it is or can be reused by relevant academic and industrial communities, another way to maximise the return of the project is to try to extend the life of these software items beyond the life of the project. Therefore we also explore possible measures for making sure that valuable software, data and services remains usable after the xLiMe project has finished, consortium members have moved onto new projects and the xLiMe funding can no longer support the development of such technologies.

This document is delivered halfway into the xLiMe project (in month 18 of 36). At this point, most of the basic software has already been developed and integrated into various services. Also, initial data is being generated by the xLiMe toolkit. This means that we can perform a first assessment of current measures for impact and continuation and that we have time to tweak our dissemination strategy for the remainder of the project.

1.1 Scope and Relation to other Work Packages and Deliverables

This deliverable is part of task T8.1 “Dissemination and Community Building” in WP8 and the fourth of four deliverables in this task. While D8.1.1 and D8.1.2 gathered the initial dissemination materials and set up the project’s website for public dissemination of the outputs, D8.1.3 presented the main communication plan for the whole project, focussing mainly on plans for the dissemination of research results. This deliverable can be seen as a specialisation of D8.1.3 focused on the software outputs of the project as well as an initial assessment (and refinement) of the original plan with regards to these software outputs.

The other task in WP8 relates to the commercial exploitation of the outputs. Since D8.2.2 “Draft Business Plan” looks into how xLiMe outputs can be further commercially exploited, in this deliverable we focus on non-commercial impact on the three identified target audiences (academia, industry and general audience). Therefore we consider here only software components developed by the research partners JSI, KIT and UNITN, as software produced by use-case partners (VICO, ZATTOO and ECONDA) will be discussed as part of the draft business plan.

In terms of relations to other work packages, the regarded software and data outputs are being developed in WP 1 - 6. This document only provides short descriptions of the software and data outputs, in order to make it easy for readers to understand their functionality. For more detailed technical descriptions of each considered component we refer to the various technical deliverables in WP1 to WP5, as well as to deliverables in WP6 for an overview of the complete xLiMe architecture.

The scope of this document is describing (i) current activities in order to maximise the impact and quality of the produced software items and (ii) future measures ensuring their impact and continuation after the end of the project. Finally, although performance is an important factor that can affect the adoption of software, we do not go in too much detail here and only look at stability and performance in general terms. Detailed information about the performance of each component can we found in D7.1.1. “Early Benchmarking Report”.

[1] This is, of course, besides the generation and documentation of academic knowledge in the form of research systems, evaluation studies and research papers.

1.2 General Impact and Continuation Plans

As mentioned above, D8.1.3 “Communication Plan” already identified three main target audiences for the (research, software, data) outputs of xLiMe:

- Research community, in particular in the areas of computer vision, information extraction and semantics, etc.
- Industry and customers, in particular in the areas of new media, social web, content providers and business intelligence.
- General public, in particular those interested in advances in cross-media content analysis and related impacts on society.

D8.1.3 also identified a basic strategy for maximising dissemination and impact of the produced software comprising their source code, data sets and services:

- publishing open-source software components on publicly available repositories such as http://github.com/ with licensing and ownership information
- publishing datasets on the project’s website under permissive licenses such as Creative Commons
- making demonstrators and prototypes available to the target audiences.

1.3 Document Overview

In order to survey these general guidelines for the target non-commercial components and derive suitable measures, we will first look in the remainder of this document at each of main software items (source code, data and services) produced by the research partners, focusing on the current state and factors affecting the impact and continuation potential of each of these software items. Then, in section 3 we will describe our current plans to further improve the impact and continuation opportunities for the software items. Finally in Section 4 we will discuss general lessons that we can draw from our current and planned impact and continuation measures, before concluding the deliverable in Section 5.

2 Current Impact and Continuation Factors

In this section, we present the current state of the main software items. In particular we look at software and data publication, as well as service availability. For each software item we provide:

- Its name and a short description to provide context within the xLiMe project
- A short description of the history of the software item, in order to indicate whether it was developed specifically as part of the xLiMe project or whether the item was originally developed prior to the project.

For each software item we also look at aspects which can affect the uptake of the software item in research and industry:

- Availability: who can access the source code, data and services?
- Licensing issues: can interested parties reuse the software item (and are there issues that potential users need to take into account?).
- Stability: the assumption here is that the more stable a software item is, the more likely it is to be adopted, since there are less costs associated with handling changes or unstable software.
- Reusability: likelihood that people outside of the xLiMe project will want to reuse the software item and how easy will it be for them to reuse it?
- Documentation: has the software item been documented? The assumption here is that better documented software items are easier to adopt and can have a higher impact.

Note that in this section we only look at the current state of the software items around month 18 of the xLiMe project. In Section 3 we will look our plans for all of these software items until the end of the project and in Section 4 we will discuss in more detail how current and planned measures can be improved.

2.1 JSI Newsfeed

Description: Feed of multilingual news and blogpost.

History: Developed as part of various EU projects and extended as part of xLiMe

2.1.1 Software

Availability: Proprietary, only available to JSI members

Repository: internal JSI repository

License: Proprietary

Stability: Production Quality

Reusability: Very high, as demonstrated by its use in various EU projects.

Documentation: Provided at the project's website http://newsfeed.ijs.si/

2.1.2 Service and Data

History: The web-service API has been the main way to access this component.

Availability: Public web service accessible for research use

URL: http://newsfeed.ijs.si/stream/

Stability: Data format has been stable for the last few years, but it will be updated this year due to consolidation. The distribution side of the service has sufficient resources and is not expected to be a problem, however the processing side of the service (scraping, linguistic analysis and indexing) runs at 100% resource utilization; hence a code rewrite and server infrastructure replacement are planned.

Availability: >99.7%

Response time: <50ms

2.2 Social Media Annotation

Description: Linguistic components tailored to handle social media as part of D2.3.2

History: Developed as part of the xLiMe project

2.2.1 Software

Availability: Code is not available yet.

Repository: private repository at JSI

License: TBD

Stability: Alpha, i.e. in active research and development

Reusability: Medium. The main functionality is relevant to researchers and industry, but installation and execution of the software is currently cumbersome due to various dependencies.

Documentation: None yet as software is still under heavy development.

2.2.2 Service

Description: Web service API for accessing the social media annotation functionality

History: Set up as part of the xLiMe project

Availability: API available for project use (not public)

URL: Private

Stability: API and data format still subject to change as the software is still being developed.

Availability: ~30%

Response time: <10ms

2.3 Speech to Text API Wrapper

Description: Custom wrapper around commercial ASR services (VecSys and Pervoice) in order to adapt their input and output to the xLiMe context.

History: Developed during the xLiMe project.

2.3.1 Software

Availability: Closed source

Repository: private repository at JSI

License: TBD

Stability: Near production quality.

Reusability: Little, each ASR service has their own API and customizations have to be performed for each new ASR service to wrap.

Documentation: None.

2.3.2 Service

Description: Receives a list of Zattoo TV channels for which the speech stream needs to be converted to text. Schedules and distributes the speech to text tasks to the available ASR services and translates the output into the xLiMe datamodel.

History: Set up as part of the xLiMe project

Availability: API available for project use (not public)

URL: Private

Stability: stable in terms of data format.

Availability: ~60% since relying on VecSys cloud service, which is only available part-time.

Response time: ~2 minutes

2.4 Event Registry

Description: Aggregates annotated media-items into high-level clusters (e.g. events) and provides services such as search, monitoring and analytics.

History: Developed as part of previous EU-projects such as xLIKE and to be extended as part of the xLiMe project.

2.4.1 Software

Availability: Closed source.

Repository: Private repository at JSI.

License: Closed source.

Stability: Production quality.

Reusability: Medium to High. While the original intent of the EventRegistry was to find clusters of news articles to aggregate them into events, in xLiMe this will have to be extended into aggregating general media-items (not necessarily news related) into “topics” (not necessarily events). We expect many of the underlying implemented algorithms for clustering, indexing, searching and analytics to be useful, but some changes will have to be made.

Documentation: The software itself has internal documentation at JSI. The web-based API is documented alongside the python client library on github: https://github.com/gregorleban/event-registry-python.git

2.4.2 Service

Description: Web-based API for accessing the Event Registry

History: Originally set up as part of the xLIKE project, but planned to be extended as part of xLiMe.

Availability: Public web site and web service

URL: http://eventregistry.org/

Stability: Stable in terms of data format and API for news events. Also, the current infrastructure is sufficient to handle the current and expected use of the service. The API is expected to be extended as part of the xLiMe project in order to accommodate multimedia items and topic (non-news-event) clusters.

Availability: >99.7%

Response time: <20ms

2.5 Usage Analytics Module

Description: Will provide services for aggregating and analysing the usage of the xLiMe data.

History: To be developed as part of T5.2 of the xLiMe project.

2.5.1 Software

Availability: Not developed yet.

Repository: TBD

License: TBD

Stability: Planning.

Reusability: Little. This software module will focus on the xLiMe data and it usage at the various use-cases. Hence, we do not expect it to be reusable for other data streams in other settings.

Documentation: TBD.

2.5.2 Service

Description: This software component is expected to be exposed as both a web-based interface and a web-based service.

History: To be set up as part of the xLiMe project

Availability: TBD (probably public web service)

URL: TBD

2.6 Cross-lingual Semantic Annotation

Description: Adds cross-lingual tags to texts in various languages using a DBpedia-based cross-lingual lexicon, this is the main result of D3.3.1.

History: Originally developed as part of the xLIKE project and further refined as part of the xLiMe project. It originated as an extension to Wikipedia Miner which has been extended regarding different approaches to mention detection, graph-based disambiguation and cross-linguality.

2.6.1 Software

Availability: Semi-open source (available to project partners)

Repository: https://svn.aifb.uni-karlsruhe.de/external/xlike/gwifi/

License: To be decided (planned to be GPLv2)

Stability: Production quality, although new features (e.g. emergent entities) will need to be tested and improved.

Reusability: Highly reusable. The code has already been used to set up various deployments at JSI and VICO, albeit with support from original developers.

Documentation: The main design and architecture of this software component has been documented in research papers, but detailed user and developer documentation is missing; hence new developers need to spend time understanding the source code.

2.6.2 Data: Indices

Description: The data required to provide the cross-lingual semantic annotation includes indices generated from Wikipedia and DBpedia dumps. The indices are currently not available but can be generated using the software and the original Wikipedia and DBpedia dumps.

History: Generated as part of the xLiMe project

2.6.3 Service

History: Set up at VICO as part of the xLiMe project

Availability: Private web service at VICO's premises.

URL: n/a

Stability: Stable API, additional features may be added during the rest of the project (e.g. emergent entities).

Availability: >90%

Response time: ~5ms per query

2.7 Video OCR

Description: Performs optical character recognition on video in order to extract text from video streams.

History: Developed for the xLiMe project based on pre-existing software by Carnegie Mellon University (CMU). Besides various refinements to the original software, this component includes character recognition handled by Tesseract[2].

2.7.1 Software

Availability: Only available as a service as part of the xLiMe toolkit

Repository: Internal repository at UNITN

License: Closed source

Stability: alpha (may contain bugs and will need further development in order to increase performance and stability).

Reusability: Medium. While this software component has potentially high reusability due to its task, in practice, the software is not mature enough to be used in production settings. For research purposes, the reusability is better, since it can be used as a baseline to implement new approaches to improve text detection in video.

Documentation: No public documentation yet other than the general description of the approach in xLiMe deliverable D2.2.1. Privately, current documentation consists of source code comments.

[2] http://code.google.com/p/tesseract-ocr available under the Apache License 2.0

2.7.2 Service

Description: xLiMe service which uses the video OCR software to analyse Zattoo video streams and convert output to the xLiMe data format.

History: Set up as part of the xLiMe project

Availability: Web service available to project partners via the produced output pushed into Kafka.

URL: n/a

Stability: alpha, software is still in development, so it is updated frequently. However, the data format and API are expected to be stable.

Availability: >60%

Response time: ~1s (6 frames per second)

2.8 Video Annotation

Description: Software component for recognising various object types in video streams and images.

History: Developed for the xLiMe project, as an extension of the Caffe deep learning framework[3]. The extension consists of trained models for brand recognition.

2.8.1 Software

Availability: Private, available to project partners via services linked to the xLiMe platform.

Repository: Private repository at UNITN

License: TBD

Stability: Beta, the core deep learning framework is stable, but the trained models may need to be improved in order to reach production quality.

Reusability: Low to medium. The trained models will focus on specific brands as required by the xLiMe use-cases, models for new brands will need to be trained from scratch.

Documentation: General approach described in D3.2.1. Privately, source code contains comments.

2.8.2 Service

Description: Service wrapper around the video object detection software to apply the detection to Zattoo video streams and push outputs to the Kafka broker.

History: Set up as part of the xLiMe project

Availability: Private web service, available to xLiMe partners via the Kafka output.

URL: n/a

Stability: Stable in terms of data format, API and expected accuracy.

Availability: >60%

Response time: <100ms

[3] caffe.berkeleyvision.org/ available under a BSD 2-Clause License.

2.9 Media-item Cross-linking

Description: Multi-Modal Correlated Centroid Space for Multi-lingual Cross-Modal Retrieval

History: Developed from scratch for the xLiMe project, but using various open source libraries.

2.9.1 Software

Availability: Open source

Repository: https://github.com/adityamogadala/CSquareSUR

License: GNU GPL V3

Stability: Beta (fairly stable, but bugs may be discovered and possible changes may be implemented to improve performance during the project).

Reusability: Requires access to Matlab, which is a proprietary software with prices ranging from 35 euros for a basic student version to 2000 euros for a standard license (although a free trial version is available for 30 days).

Documentation: The software repository includes a readme file describing how to execute the code, it does also documents what the input or output files are and what the software component does in technical terms by linking to a research paper describing the component.

2.9.2 Data

Description: Contains topics extracted from English, German and Spanish files and SIFT histogram Image features from the data-set described in the paper. Each of the folders contains 10, 100 or 200 topics. Also contains text files present in Spanish and German from which these features were extracted.

History: Generated as part of the xLiMe project

Availability: Open Source

URL:

- http://people.aifb.kit.edu/amo/data/raw_features_text_images.zip
- http://people.aifb.kit.edu/amo/data/Text-Ger-Spa.zip

License: GNU GPL V3

Reusability: Data is strongly linked to the software, but can be used according to the specified licence; in particular, attribution is requested in order for use to be able to track impact.

Documentation: All relevant details and description of the data are present in README files of the folders.

2.9.3 Service

Description: Provides web-based access to the media-item cross-linking capabilities.

History: Set up as part of the xLiMe project

Availability: Public web service

URL: http://km.aifb.kit.edu/services/xlimesearch/kitsearch?q=

Stability: Stable API that will not change at the current URL. Any changes required during the rest of the project will be applied to a different version of the Service.

Availability: >90%

Response time: 7ms per API transaction

2.10 Y1 Zattoo Demo Infrastructure

Description: Integration code to provide the zattoo.xlime.eu website

History: Developed as part of the xLiMe project.

2.10.1 Software

Availability: Closed source.

Repository: Internal at Zattoo

License: Closed source.

Stability: Early Beta. New versions of the Zattoo demo will have to be developed based on updated business cases and capabilities of the xLiMe toolkit.

Reusability: Little. This is tailored for the Y1 Zattoo use-case and the Y1 capabilities of the xLiMe toolkit, we expect both to change and hence the software will need to be redesigned.

Documentation: None.

2.10.2 Service

Description: Suggests related media contents for a given zattoo stream at zattoo.xlime.eu

History: Set up as part of Y1 of the xLiMe project.

Availability: Private website, available for project partners and some Zattoo users.

URL: http://zattoo.xlime.eu

Stability: Stable but incomplete set of features.

Availability: ~60%

Response time: <50ms (for requests, but service has 5 minute latency in relation to real-time)

2.11 Semantic Search

Description: Provides search services on top of the xLiMe data. It is comprised of a small number of pre-existing services such as SPARQL querying as well as microservices such as autocompletion and summarisation of entities.

History: Custom component developed as part of the xLiMe project.

2.11.1 Software: DBpedia Autocomplete

Availability: Open Source

Repository: https://github.com/steffenthoma/dbpedia-autocomplete

License: GPLv2

Stability: Beta

Reusability: High, the service is often required by applications which need to provide users with a way to select entities from DBpedia. The github page provides links to the datasets required for setting up a MySQL database which can be used to provide this service. Only data for the current release of DBpedia (version 3.9) is available and there are no instructions to help users generate the pagerank data from other versions of DBpedia (or for different datasets).

Documentation: the github repository provides basic but sufficient instructions on how to set up the service.

2.11.2 Software: Entity Summarisation

Availability: Closed Source

Repository: n/a

License: TBD

Stability: Alpha (still in development)

Reusability: High. The need to provide a short subset of information about an entity in a semantic dataset is very common in semantic user interfaces.

Documentation: documented in research paper[4].

2.11.3 Service: xLiMe SPARQL Endpoint

History: Set up as part of the xLiMe project based on Virtuoso Open Source Edition

Availability: Private web services, available for project partners

URL: http://km.aifb.kit.edu/services/xlime-sparql

Stability: Stable, few or no API changes expected.

Availability: >80%

Response time: depends on query

2.11.4 Service: xLiMe Autocomplete

History: Set up as part of the xLiMe at KIT

Availability: Public web services

URL: http://km.aifb.kit.edu/services/xlime-autocomplete/search.html

Stability: Stable, few or no API changes expected.

Availability: >80%

Response time: <100ms per request

[4] http://www.aifb.kit.edu/images/e/e4/Paper_92_final.pdf

2.11.5 Service: Entity Summarisation

History: Set up as part of the xLiMe by KIT

Availability: Public web services

URL: http://km.aifb.kit.edu/services/summa/

Stability: Stable, few or no API changes expected.

Availability: >80%

Response time: <500ms

3 Impact and Continuation Plans

In this section we go through the various software and services that are being (or will be) developed as part of the xLiMe project and we discuss our plans for maximising awareness about the software and services as well as for maintaining the availability of these services and promoting the further development of software after the xLiMe project. For each software item we discuss:

- Target audience and dissemination: this identifies the specific audience which may be interested in the software item, which helps in planning the way in which the software item can be disseminated. We also discuss how the software item will be promoted.
- Planned stability: this indicates a target software quality for the software item. This can be different than the current software quality and also includes issues such as a target stability, how well the software will be tested and how diverse and mature the community will be around the software project.
- Planned data availability: This indicates how the required and produced data will be available by the end of the project and beyond, including any plans for publication, replication, promotion, etc.
- Planned service availability: This indicates our plans for making software services available until the end of the project and beyond. When possible, we discuss how easy it will be for third parties to deploy a similar service on their own.

3.1 JSI Newsfeed

Target audience and dissemination: This software is intended both for the research and industrial communities, since it provides core functionality for media monitoring services. JSI is actively promoting this service as a stand-alone service and as part of various research and industrial demonstrators. The xLiMe use-case demos and showcase interface will further help to create awareness of this resource.

Planned software stability: By the end of the project we expect the software to have improved production quality. Although the current version can already be used in production, performance can be improved. By the end of the project we expect to have rewritten part of the software in order to achieve performance improvements, as a result of this work we expect to decrease response times from <50ms to <20ms.

Planned data availability: We plan to continue with the current data availability approach (available for research purposes only). There is currently no commercial availability for this service, but this can be looked into if there is enough demand from industry.

Planned service availability: After the end of the xLiMe project, we expect this service to be further maintained for many years as part of future research and commercial projects. Furthermore, since the current infrastructure is near its limits, we expect to be able to update this infrastructure to be able to handle increasing numbers of requests at faster speeds. We expect to increase availability from >99.7% to >99.9%. This service will not be available for redeployment by third parties, since to source code is proprietary and it depends on various proprietary libraries and services.

3.2 Social Media Annotation

Target audience and dissemination: This software is intended primarily for the research community, although industrial communities can also be interested, since annotation of social media is increasingly requested in industry. The main dissemination for this component will be in the form of research publications at relevant conferences and journals in order to reach the intended research community. In order to reach industrial communities and maximise research impact, we also plan to make this software component available as open-source to maximise its reusability and to try to create a software community around this project.

Planned software stability: By the end of the project we expect the software to be in beta: most bugs will have been fixed and the software will be fit for purpose, but different users may want to tweak the

software in order to make it applicable to their particular needs. Documentation about the software will be made available as part of the software distribution.

Planned data availability: Any required data for execution of the software will be distributed and documented alongside the source code as part of the software distribution.

Planned service availability: After the end of the xLiMe project, the service wrapper around this software will only be further maintained as long as the xLiMe toolkit needs to be active. However, if this service can be reused in future projects by JSI, the service will be maintained and updated accordingly. Third parties with basic technical skills in python and C++ will be able to deploy their own service based on the open-sourced software distribution.

3.3 Speech to Text API Wrapper

Target audience and dissemination: This software is intended primarily at the research community. Industrial organizations that may be interested in this type of software can interact directly with ASR component providers such as VecSys and Pervoice to acquire their services. Dissemination plans for this component are limited to research papers discussing the integration of text to speech services in the xLiMe platform.

Planned software stability: By the end of the project we expect the software to have production quality.

Planned data availability: The data required to run the services are proprietary and the data produced by this text to speech services will be made available as part of the xLiMe data dumps.

Planned service availability: This service will only be available for a limited time after the xLiMe project finishes as the ASR service providers will not extend their licenses after the project. Third parties will not be able to deploy their own service without contacting and purchasing ASR licenses.

3.4 Event Registry

Target audience and dissemination: This software is intended for both the research and industrial communities. The research community can be interested in the big-data challenges and approaches showcased by this component, while the industry community may be interested in the monitoring and business intelligence applications. The multi-media and aggregation additions that are planned as part of the xLiMe project will be disseminated via research publications as well as updated demonstrators to be presented at research and industrial conferences and social media.

Planned software stability: By the end of the project we expect the software to have production quality. This is ensured by having automatic tests and having a continuous delivery approach where the core Event Registry functionality is continually updated and available at the main website. New functionality is available via a staging version of the website and will be integrated and evaluated as part of the xLiMe toolkit.

Planned data availability: The Event Registry data will be available via its web-service API.

Planned service availability: After the end of the xLiMe project, JSI plans to keep the Event Registry website and services alive as part of future research and commercial projects. Any new features developed as part of xLiMe are planned to be incorporated into Event Registry and made available via the web-API. There will be no option for third parties to deploy their own version of Event Registry, since the source code is closed source and the effort and cost of deployment and maintenance is relatively high; however, the python client-api is open source and is encouraged.

3.5 Usage Analytics Module

Target audience and dissemination: This software will be intended primarily at the research community, although industrial communities can also be interested, since it will provide common functionality in gathering and analysing usage analytics of the xLiMe data; this functionality may be adapted for analysis of

other related data. We plan to disseminate this software via research papers and using demonstrations as part of the xLiMe showcase interface and xLiMe use-case prototypes.

Planned software stability: By the end of the project we expect the software to have production quality. This will be guaranteed by the use of automatic unit and integration tests as well manual integration tests as part of the xLiMe showcase. We plan to make this software component open source under a permissive license (e.g. Apache) to encourage contribution from third parties. However, since the project will be strongly linked to the xLiMe platform, we expect contributors to either request access to the xLiMe toolkit or to be able to plug into their own data providing platform.

Planned data availability: The data that will be produced by this module includes data usage statistics, detected trends and trend predictions. Current plans are to make these data available only via the xLiMe toolkit and interfaces (both as web service APIs and web based user interfaces). Once this component has been implemented we will assess whether this data can be exported and included as extensions to the xLiMe media annotation data dumps.

Planned service availability: We plan to activate this service around the end of the second year of the project. Initially the service will be available only to project partners via a web service API and a web-based user interface. After the end of the xLiMe project, this service will be kept alive as long as the xLiMe toolkit needs to be kept alive for demonstration purposes. Although the software will be made open source and third parties will be able to set up their own service, this will depend on them having access to the xLiMe toolkit (many parts of which will not be publicly accessible due to security issues).

3.6 Cross-lingual Semantic Annotation

Target audience and dissemination: This software is intended primarily at the research community, although industrial communities can also be interested, since it provides core functionality for cross-lingual textual annotation. As such, any dissemination of the xLiMe platform will showcase some of the functionality of this component. Furthermore, this software component is included in other demonstrators separate from the xLiMe project to maximise its exposure. Finally, an important dissemination strategy is the open-sourcing of the software, when we will move the source code to a public repository (e.g. on github).

Planned software stability: By the end of the project we expect the software to have production quality (including new features added during the xLiMe project). This is ensured by means of integration tests, code reviews and automatic builds with dependency management (using apache maven). While the architecture and approach of the software has been published in the scientific literature, some technical aspects need to be clarified as part of the open-sourcing of the software (when we will migrate the software repository from KIT's SVN server to github or similar public repository). Although the software is currently developed by a single developer at KIT, we hope that a community will form around the software once it has been open-sourced.

Planned data availability: We plan to make the index data required to run the service available in order to facilitate execution and deployment without needing to generate the indices from scratch; these indices will be distributed along with the source code repository.

Planned service availability: After the end of the xLiMe project, KIT will keep this service alive on the current URL under the same or improved QoS as long as the xLiMe platform is kept up to date. Furthermore, this service will outlive the xLiMe platform as long as new projects require this service; this is likely to occur, since cross-lingual annotation is likely to be a core service in many future projects. Furthermore, even if the service is changed to a private address, all the source code and data necessary to deploy this service will be available to enable third parties to set-up their own service.

3.7 Video OCR

Target audience and dissemination: This software is intended primarily at the research community, as such, the planned dissemination includes publication of the main approach and evaluation results at top conferences and computer vision journals. Furthermore, we will open source the code for this software component on github under a BSD-style license. We will also submit this work to the ACM Multimedia Open Source track, which should provide visibility to this software component and its results, as well as promoting its reuse.

Planned software stability: By the end of the project we expect the software to be in beta quality. It will still be a research prototype, but we expect to have resolved the main bugs. Also, the developer community will consist of around 5 developers from both UNITN and CMU (as this is an extension to their original framework). The produced documentation (research papers, readme files and source code comments) should be sufficient for skilled users, who have access to the right hardware, to execute the software on their machines. We expect therefore that this software will continue to be developed well after the end of the xLiMe project, either as a direct continuation of the software project or serving as the basis for further improvements.

Planned data availability: The data produced by this software component will be available as part of the xLiMe annotation data dumps. All the required data for running the software will be available as part of the github open source.

Planned service availability: After the end of the xLiMe project, we do not expect to keep this service running for a long time. This is due to the fact that it requires the use of expensive resources (GPUs), which will probably be required for other research projects. Since the source code will be available, third parties will be able to set up their own video OCR service, although they may need to provide bridges to their own video streams if they do not have access to Zattoo's video streams. Also, third parties will need to have access to suitable GPUs in order to achieve comparable performance to the xLiMe service.

3.8 Video Annotation

Target audience and dissemination: This software is intended primarily at the research community and early adopters from industry. Planned dissemination activities include presenting the research based on this software component at research conferences and journals. The software will be made available as open source on github under the BSD 2-Clause license to promote the reuse of the software; especially among the large Caffee development community, which includes researchers and early adopters from industry.

Planned software stability: By the end of the project we expect the software to have near-production quality, but only for the specific xLiMe use-cases (i.e. for detecting specific brands and object types). This will be ensured by including unit tests and research evaluations. Also, the main approach will be documented in research papers, which will enable developers (along with readme files and source code comments) to execute the software on their machines. This means that the Caffee development community (up to 1,000 developers) will be able to reuse and contribute to the further development of this software.

Planned data availability: The annotation data produced by this component during the xLiMe project will be made available via annotation data dumps. The data required to execute this code: trained models, training dataset, etc. will be included as part of the open source project on github.

Planned service availability: After the end of the xLiMe project, we do not expect to keep this service running for a long time, for the same reason as for the Video OCR service: this service requires the use of expensive resources (GPUs), which will probably be required for other research projects. Since the source code and trained models will be available, third parties will be able to set up their own video annotation service, although they may need to provide bridges to their own video streams if they do not have access to Zattoo's video streams. Also, third parties will need to have access to suitable GPUs in order to achieve comparable performance to the xLiMe service.

3.9 Media-item Cross-linking

Target audience and dissemination: This software is intended for the research community, hence the software and data required to run this service will only be promoted to the research community by means of academic papers and demonstrations at conferences. Further dissemination relies on making the source code available with its documentation on Github.com (this makes the software easy to find by researches and developers). Although the software is currently developed by a single developer, the project on github is open for new contributors.

Planned software stability: By the end of the project we expect to have fixed most of the bugs in the software as well as optimised the algorithms, resulting in production quality code. In order to ensure the quality of the final version, we are using automated unit testing and manual integration tests as part of the xLiMe platform. Furthermore, the code will be reviewed before the end of the xLiMe project.

Planned data availability: The data produced by this service is included as part of the xLiMe platform and thus will be available via dumps and via the XLiMe query services. Data dumps will include readme files in order to aid users in understanding the data.

Planned service availability: This service will be kept available on the current or updated URL under the same or improved QoS for a limited period of time (to be decided) by KIT. Reported bugs will only be fixed during the life of the xLiMe project.

By the end of the project, documentation will be made available for this service in terms of a User and Deployment manual. This documentation will make the final service easy to deploy by technical users who will need to clone the source code from the public repository, install publicly available dependencies such as MongoDB and follow the instructions in the deployment manual.

3.10 Zattoo Demo Infrastructure

Target audience and dissemination: This software is intended primarily as validation of the Zattoo use-case within the project. It can also have a role in dissemination of the project's result in industry. The research community is not a target group for this software component. Dissemination of this software component is not planned other than as an illustrating example of the xLiMe toolkit. Internally, Zattoo can use this software component to developers and managers in order to focus their development of new tools.

Planned software stability: By the end of the project we expect the software to be in Beta quality. Most obvious bugs will have been resolved and the main features will work enough in order to validate the business-cases identified by Zattoo. However, it is not a goal of the project to develop a tool that can be deployed to all Zattoo users without first going through a product development cycle.

Planned data availability: Data related to this software component is either proprietary or is part of the generated xLiMe data. The xLiMe data will be made available via dumps and via the standard search interfaces provided by the xLiMe toolkit.

Planned service availability: From Y2 of xLiMe, this service will be the responsibility of Zattoo and they will be responsible of maintaining this until they have validated their business case.

3.11 Semantic Search

Target audience and dissemination: This software is intended mainly for the research community, although some industry developers may also be interested. The software and data required to run the semantic search services will be promoted to the research community by means of academic papers and demonstrations at conferences. Since the search functionality will also be available via the xLiMe showcase and use-case demonstrators, these services will also be promoted in industry. Further dissemination relies on making all the data and source code available with its documentation on Github.com (this makes the

software easy to find by researches and developers). Although the software is currently developed by a single developer, the project on github is open for new contributors. To further make this software visible to the research community there are currently efforts to involve a more stable research community around the entity summarization component, e.g. by organizing the SumPre2015 workshop and to build the portal http://entitysummarization.org

Planned software stability: By the end of the project we expect to have fixed most of the bugs in the software as well as optimised the algorithms, resulting in production quality code. In order to ensure the quality of the final version, we are using manual integration tests as part of the xLiMe platform. Furthermore, the code will be reviewed before the end of the xLiMe project and we are using automated dependency management in order to ensure that the software is easy to build and deploy.

Planned data availability: The data required by this dbpedia autocompletion service is available from its github page. Any updates on this data will also be published.

Planned service availability: These services will be kept available on the current or updated URL under the same or improved QoS for a limited period of time (to be decided) by KIT. Reported security bugs will be fixed during and after the life of the xLiMe project.

By the end of the project, documentation will be made available for these services in terms of readme files in the open source code repositories; which will include usage and deployment instructions as well as links to any required data. This documentation will make the final service easy to deploy by technical users.

4 Discussion

Sections 2 and 3 have discussed both the current and planned impact and continuation measures for the various software items developed as part of the xLiMe project. In this section we will summarise how well we are promoting the software items to the target audiences and we will discuss whether we are adhering to the original communication plan. This discussion will be used as a starting point in order to define additional guidelines in order to maximise the impact and continuation potential of the software produced during the project.

4.1 Dissemination

From sections 2 and 3 we can conclude that most of the software items is available to at least project partners in one form or another (as a software service attached to the xLiMe platform, as some web-based API, as data available from a URL or on request, or as source code which can be accessed on request or via a software repository). In many cases, the software is also available to the general public for free or to industry with few restrictions imposed by the open source licensing terms. There is some room for improvement in this regard: some software is initially developed by the research partners and only released as open source after publication of research papers and cleaning up of the source code.

In terms of documentation, most of the software is well documented in terms of research papers which describe the main design idea and evaluate the approach. However, available source code, data and services are often hard to use when only given the URL of the software item. That is: research papers are good for disseminating the research, but not very suitable for reproducing or executing the software items. This has especially an impact on software which is intended for industrial audiences, which may not be interested in only reading the research papers, but which may be interested in trying out the software with their own datasets. The lack of documentation in these cases can make it very hard or impossible for industrial users to try the software items, thus reducing the impact.

Finally, in terms of software quality, we can see from Sections 2 and 3 that xLiMe is producing fairly high quality software in terms of stability, quality of source code, APIs etc. Most components are at a stable beta or production-quality level, which means both researchers and industrial stakeholders should be able to reuse the software items without worrying too much about whether they would need to invest much time further developing the software. An important factor here is that this quality is not always readily apparent to stakeholders just visiting the source code repositories (in part due to the lacking documentation).

4.2 Continuation

In terms of software continuation, we can see a distinction between core software (JSI newsfeed, Event Registry, video annotation) and non-core software (speech to text API wrapper, Zattoo-demo infrastructure). Core software has already been identified as such and continuation plans are set in place in order to reuse and further develop these components as part of future research and commercial projects.

The future of non-core software items is less certain. When possible, these software items are made available as open source in order for third parties to be able to take the software projects further after the original authors have moved on to new projects. A danger in this regard is that open sourced software is often lacking in documentation tailored for software development and adaptation to new contexts (e.g. non xLiMe data). Also, in many cases, software and services cannot be deployed independently by third party stakeholders due to missing (or xLiMe-specific) dependencies which are not documented or abstracted properly in the source code. As a result it will be difficult for developer communities to form around such software items. A right balance has to be found between the amount of time that we need to spend documenting and promoting such components (time which is not directly contributing to xLiMe scientific objectives) while making sure that they can be further developed.

Regarding the continuation plans for the xLiMe services, we think that the full xLiMe toolkit will not be available (i.e. processing new information) for much longer after the end of the project. In particular the more resource intensive annotation components such as Video OCR, Video Annotation and Speech-to-Text

components are likely to be the first components to be shut down. In the case of Speech-to-Text, licensing agreements will need to be arranged with the ASR component providers and computing resources will need to be secured. These are currently being set up at a computing centre in KIT, but since these resources are primarily meant for research purposes, it is unclear whether they will be available after the end of the xLiMe project (when research results have been shown). A similar situation occurs with the Video OCR and Annotation services, which rely on expensive GPU hardware at UNITN, which will likely be claimed by new research projects. Hence, unless some of these services are moved to one of the commercial partners (most likely Zattoo) before the end of the project, it is likely that the xLiMe toolkit will only continue to execute as a limited toolkit analysing mostly text-based media.

Another option for the continuation of the xLiMe services is for third parties to set up their own version of the xLiMe toolkit. We believe that this will be possible based on the various open-sourced components as well as the detailed description of the integration architecture in xLiMe deliverables. This option will require substantial investment of resources by third parties, since it will require various web servers and, depending on the which media needs to be analysed, new data sources and components may need to be developed.

In terms of data continuation, the historical annotation (and derived aggregated) data gathered during the xLiMe project will be published as dumps, which will be made available from the project website starting in Y2 of the project, with regular updates. An automatic pipeline will be set up in order to publish this historical data without requiring much human effort. This should ensure the availability of this data for many years after the end of the project. Eventually, it would also be good to publish and promote this data on external dataset repositories such as datahub[5]. We expect this data will be of value primarily to the research community as we are not aware of any similar cross-media, cross-lingual dataset. The main drawback of this dataset will be that the source multimedia streams will not be available, since they are only available for a limited time from Zattoo and also licencing issues may apply. However, it is possible that in the future, researchers will be able to access historical media streams from other resources based on the stored EPG metadata.

[5] http://datahub.io

5 Conclusions

This deliverable presented and discussed current and planned impact and continuation measures for software, data and services (that will be) produced as part of the xLiMe project. In general, the current approach for disseminating software items is appropriate given the limitations in terms of research and data licensing issues. The various software products are being communicated to mainly the research community and plans are in place for reaching also industrial stakeholders in the second half of the project. In terms of continuation of the software items, we have identified core components for which clear continuation plans exist in both research and industrial projects. We have also identified non-core components for which we have discussed options for maximising their continuation potential. The discussions resulted in additional guidelines which are presented below. These guidelines will be distributed to the project partners and will be taken into account during the rest of the project.

5.1 Guidelines for Software and Data Impact & Continuation

Software that can be open-sourced, should take into account the following checklist in order to maximise impact and continuation potential:

- Is license suitable for intended audience? E.g. if intended audience includes industry, a more permissive license such as Apache or BSD should be chosen, rather than GPL.
- Is license compatible with dependencies?
- Does the project include a clear README file that can be used by the intended audience to execute the code? E.g. this can be verified within the xLiMe consortium by asking another partner to check-out and execute the code.
- Does the project name, project description and README file provide enough information to make the project easy to find by people searching for similar projects? E.g. check that searching on github or google returns the project page.
- Are links to published papers, the xLiMe project and relevant demos included in the README file?
- Is the software as easy to execute as possible? E.g. can resources be provided to simplify installation, such as by providing pre-compiled packages and required data?
- Has the software release been advertised on the xLiMe website, relevant mailing lists and social media?
- Are there links from the xLiMe showcase and use-case prototypes back to the software component page?

xLiMe datasets that can published, should take into account the following checklist:

- Is the license suitable for the intended audience?
- Does the license allow third parties to copy and redistribute the data?
- Is data bundled together with metadata in order to make data format easier to understand and reuse?
- Can the data be linked or uploaded at central repositories which are popular among the target audience?
- Has the dataset release been advertised on the xLiMe website, relevant mailing lists and social media?
- Are there links from the xLiMe showcase and use-case prototypes back to the various datasets?

Finally, for xLiMe services which have been open sourced, the following additional checklist should be taken into account in order to facilitate third-party deployment and further development of similar services:

- Does the documentation indicate specific hardware requirements?
- Does the documentation explain how to adapt the software to non-xLiMe data?