QLectives – Socially Intelligent Systems for Quality
Project no. 231200

Instrument: Large-scale integrating project (IP)

Programme: FP7-ICT

Deliverable D3.1.1

Techno-social "living-archive" with massive datasets from living labs and other sources

Submission date: 2010-08-01

Start date of project: 2009-03-01 Duration: 48 months

Organisation name of lead contractor for this deliverable: ETH Zurich

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	**x**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

DOCUMENT INFORMATION

1.1 Author(s)

Author	Organisation	E-mail
Sergi Lozaro	ETH Zurich	slozano@ethz.ch

1.2 Other contributors

Name	Organisation	E-mail
Nigel Gilbert	University of Surrey	n.gilbert@surrey.ac.uk
Nazareno Andrade	TUD	nazareno@gmail.com
Dario Taraborelli	University of Surrey	D.Taraborelli@surrey.ac.uk
Dave Hales	TUD	dave@davidhales.com

1.3 Document history

Version#	Date	Change
V0.1		Starting version, template
V0.2	7 July, 2010	Complete first draft
V0.3	26 July, 2010	Second draft
V1.0	28 July, 2010	Version for submission

1.4 Document data

Keywords	Techno-social system, collective dynamics, scientific communities, datasets, QScience, QMedia, Archive
Editor address data	slozano@ethz.ch
Delivery date	1 August, 2010

1.5 Distribution list

Date	Issue	E-mail
1 August 2010	**Consortium members**	qlectives@list.surrey.ac.uk
1 August 2010	**Project officer**	Jose.FERNANDEZ-VILLACANAS@ec.europa.eu
1 August 2010	**EC archive**	INFSO-ICT-231200@ec.europa.eu

QLectives Consortium

This document is part of a research project funded by the ICT Programme of the Commission of the European Communities as grant number ICT-2009-231200.

University of Surrey (Coordinator)
Department of Sociology/Centre for Research in Social Simulation
Guildford GU2 7XH
Surrey
United Kingdom
Contact person: Prof. Nigel Gilbert
E-mail: n.gilbert@surrey.ac.uk

Technical University of Delft
Department of Software Technology
Delft, 2628 CN
Netherlands
Contact Person: Dr Johan Pouwelse
E-mail: j.a.pouwelse@tudelft.nl

ETH Zurich
Chair of Sociology, in particular Modelling and Simulation,
Zurich, CH-8092
Switzerland
Contact person: Prof. Dirk Helbing
E-mail: dhelbing@ethz.ch

University of Szeged
MTA-SZTE Research Group on Artificial Intelligence
Szeged 6720, Hungary
Contact person: Dr Mark Jelasity
E-mail: jelasity@inf.u-szeged.hu

University of Fribourg
Department of Physics
Fribourg 1700
Switzerland
Contact person: Prof. Yi-Cheng Zhang
E-mail: yi-cheng.zhang@unifr.ch

University of Warsaw
Faculty of Psychology
Warsaw 00927, Poland
Contact Person: Prof. Andrzej Nowak
E-mail: nowak@fau.edu

Centre National de la Recherche Scientifique, CNRS
Paris 75006,
France
Contact person : Dr. Camille ROTH
E-mail:
camille.roth@polytechnique.edu

Institut für Rundfunktechnik GmbH
Munich 80939
Germany
Contact person: Dr. Christoph Dosch
E-mail: dosch@irt.de

QLectives introduction

QLectives is a project bringing together top social modelers, peer-to-peer engineers and physicists to design and deploy next generation self-organising socially intelligent information systems. The project aims to combine three recent trends within information systems:

- **Social networks** - in which people link to others over the Internet to gain value and facilitate collaboration
- **Peer production** - in which people collectively produce informational products and experiences without traditional hierarchies or market incentives
- **Peer-to-Peer systems** - in which software clients running on user machines distribute media and other information without a central server or administrative control

QLectives aims to bring these together to form Quality Collectives, i.e. functional decentralised communities that self-organise and self-maintain for the benefit of the people who comprise them. We aim to generate theory at the social level, design algorithms and deploy prototypes targeted towards two application domains:

- **QMedia** - an interactive peer-to-peer media distribution system (including live streaming), providing fully distributed social filtering and recommendation for quality
- **QScience** - a distributed platform for scientists allowing them to locate or form new communities and quality reviewing mechanisms, which are transparent and promote

The approach of the QLectives project is unique in that it brings together a highly inter-disciplinary team applied to specific real world problems. The project applies a scientific approach to research by formulating theories, applying them to real systems and then performing detailed measurements of system and user behaviour to validate or modify our theories if necessary. The two applications will be based on two existing user communities comprising several thousand people - so-called "Living labs", media sharing community tribler.org; and the scientific collaboration forum EconoPhysics.

Executive Summary

Data availability is a key aspect in QLectives, since it drives the project's workflow, organized along the closed cycle formed by the four streams (Models and hypothesis; algorithms and simulations; empirical focus; implementation and deployment). This makes a requirement to have a common and standardized repository of datasets for all partners to share. The "Living- archive" is aimed to fulfil such a requirement.

This deliverable is the first one of two related to the "Living-Archive". It aims at: A) Describing the actual technical implementation of the archive; B) Listing the current available content and; C) Reporting about organizational activities and collaborations among QLectives' partners in relation to the Archive. Deliverable D3.1.3 (Final techno-social "living-archive") will provide a similar picture of the "Living-archive" at a stage closer to the end of the project (month 42).

From the implementation viewpoint, the "Living-Archive" is a part of QLectives' Wiki (http://www.qlectives.eu/wiki/index.php/Data_Archive), and is hosted by University of Surrey. It is organized along batches of data (from Batch 0 to, eventually, Batch 3). Batch 0 is an initial collection of data, contributed by partners in the Consortium from previous projects. The other batches (from Batch 1 to Batch 3) will include the data corresponding to each one of the loops in the ever-improving cycle on which QLectives' workflow is based. In terms of data format, it was agreed to use MySQL as a way of homogenising datasets from very different sources (like in the case of Batch 0). It was also agreed to define a common metadata specification. Taking into account the Archive particularities, DDI (Data Documentation Initiative) was the solution finally selected.

The "Living-Archive" contains, to date, 12 datasets (8 of them corresponding to Batch 0, and the other 4 to Batch 1). Section 3 in this document provides a short description of each one of them.

Regarding future data-collection from QLectives' living labs, in the short and mid term the

labs should be used directly to collect users’ behaviour data (e.g. before and after the introduction of a certain social functionality in a platform to evaluate their reaction). In a longer term, they could serve as the support for a wide variety of social scientific experiments (in a similar way that physical laboratories support ‘traditional’ behavioural experiments).

In addition, as the next steps in the development of QLectives’ living labs get clearer (i.e. as roadmaps are being refined), new platforms become more likely to be considered for data collection within the project. This includes, for instance, Living Science (http://www.livingscience.ethz.ch/) and QJournal (see QScience’s roadmap: http://www.qlectives.eu/wiki/index.php/QScienceRoadMap).

Contents

1. Introduction

Data availability is a key aspect in QLectives. As data drives the project’s workflow, organized along the closed cycle formed by the four streams (Models and hypothesis; algorithms and simulations; empirical focus; implementation and deployment), it is required to have a common and standardized repository where all datasets involved in the project can be stored and shared. The concrete implementation of such a common repository is called “Living-archive” (“the Archive” from now on). ETH Zurich (as leader of WP3.1) is responsible for its implementation and maintenance.

In agreement with the importance of the Archive, QLectives’ work plan includes up to 4 milestones on data collection (M1, M4, M7 and M10) and 2 deliverables about the Archive itself (D3.1.1. and D3.1.3). This document is the first one of these two deliverables. Its aim is threefold: Describing the actual technical implementation of the archive, listing the currently available content, and reporting about organizational activities and collaborations among QLectives’ partners in relation to the Archive.

The other related deliverable (D3.1.3 “Final techno-social ‘living archive’” (M42)), is expected to complement this one by providing information about technical changes introduced to the Archive during QLectives’ project development, as well as a summary of the information collected and a report on future plans for the Archive once the project is over.

The structure of the document is simple. In the next section, we provide details on the implementation and available content of the Archive at the moment. A description of planned data collection activities follows and closes the document.

2. Structure and data format in the "Living-archive"

Technical implementation and structure

The Archive is currently implemented as a part of QLectives' Wiki (http://www.qlectives.eu/wiki/index.php/Data_Archive), which is hosted by the Centre for Research in Social Simulation at the University of Surrey.

As it is described in section B.1.3 of QLectives' description of work, data collected within the framework of the project is grouped into *batches*. Batch 0 is an initial collection of data, contributed by partners in the Consortium from previous projects. It is intended to work as a trigger for the initial modelling activities, and as a back up along the project life in case the data collected from QLectives living labs is not enough to performed data analyses in WP 3.3. The other batches (from Batch 1 to Batch 3) should include the data corresponding to each one of the loops in the ever-improving cycle on which QLectives is based on. The Archive's implementation reflects such an organization by presenting each batch in a separate wikipage.

Common standard formats for data

In order to have a common format for all the data contained in the repository, independently of its original source, it was agreed to store it as MySQL relational databases. This option seems to be adequate for the current available datasets, which do not include much metadata but basically simple tables. However, as data collection progresses within the project, and richer datasets are collected by different partners using diverse means (from direct sampling of users' behaviour to survey and, eventually, experiments), the Archive will probably need a common metadata specification for the whole Consortium.

Such a common metadata specification should be flexible enough to cope with the abovementioned diversity of data sources and types but, at the same time, its implementation should be simple enough for a rather small repository as the Archive. The concrete solution chosen for this project is DDI (Data Documentation Initiative) (DDI, 2010).

DDI expects to become an international standard for describing social science data. Currently, there are two operative versions of DDI’s specification, namely v2.1 and v3.1. Since DDI plans to maintain both versions in the future, and tools for translation between versions are already available, deciding for one particular version simply depends on the particular needs of the project to adopt the metadata specification.

DDI 3.0 supports the entire life cycle of social science datasets (from the study concept and data collection, to its archiving and analysis). It has a modular design, thought to support grouping and comparing of large amounts of studies, and is especially suitable for highly complex data files (obtained, for instance, from massive surveys). These features have made DDI 3.0 a perfect choice for supporting repositories of several agencies performing large scale demographic studies and providing developing WEB tools for comparative analysis of the resulting massive datatsets.

DDI 2.1, on the contrary, is focused on data collection and storage, so study conceptualization and data analysis are not implicitly covered by it. Its design is hierarchical, based on a single file including the following information: DDI document header (with its description and information about it); Study description (information about the context of the data production and distribution (creators, methodology, abstract, keywords, etc.); Data files description (information about the data file or files (format, size, number of cases, etc.); Variable description (information about the data items or rows and columns in a tabular data file/s) and Other study materials (such as inline reference materials or references to external reference materials (coding schemes, thesauri, citations to publications).

Taking into account that the main functionality of the Archive is to store sets of data, and that the Consortium partners asked for a solution requiring few extra formatting efforts, DDI 2.1 has been chosen.

The project is currently preparing metadata descriptions of the Batch 1 datasets using DDI 2.1. The project is also working towards a generic metadata specification capable of representing terms of quality for any kind of digital object (see Deliverable 4.4.1). While the

objectives of DDI and our generic metadata specification are entirely different, the relationship between them is a matter under current consideration.

3. Datasets currently available

In the following we provide a brief description of each one of the available datasets, organized by batches. Most of them already comply with the common formats described in the previous section, ETH Zurich (with the assistance of other partners) is working to format the rest and inform all the other partners regarding future uploads of data to the Archive.

Batch 0

Data contributed by different partners from previous projects.

Filelist.org sample*:* A collection of real traces of unique peers from the Filelist.org BitTorrent tracker obtained in January 2006. It is possible to trace the behaviour of unique peers over multiple sessions and multiple swarms with this tracker. This is not possible with public trackers such as mininova.org. The traces record the size of the files that are shared in each swarm and the connectability of peers (i.e. if they are behind a firewall or freely connectable). The traces capture the realistic high churn rates found in deployed P2P systems. Each trace records approximately 23,000 unique events making a total of 23000 events. More information about the collection of these traces can be found in (Rahman *et al.*, 2009).

PGP contacts graph: List of edges of a snapshot of Pretty-Good-Privacy algorithm's web of trust as it was on July 2001. Only bidirectional signatures (i.e., peers who have mutually signed their keys) have been considered. This filtering process guarantees mutual knowledge between connected peers and makes the PGP network a reliable proxy of the underlying social network. The complete undirected network is composed by 57243 vertices with an average degree $<k>=2.16$. The giant component (GC) of this network, i.e., the largest connected subnetwork, (which is the data actually provided) comprises 10680 vertices and its average degree is $<k>_{GC}=4.55$.

URV email traffic graph: List of edges of the network formed by e-mail interchanges between members of the University Rovira i Virgili (Tarragona, Spain). In this network, each email address is a node and a link between two of them implies an email communication.

Only email communications that are bidirectional (A sent an email to B and vice-versa) and sent to less than 50 recipients have been considered. The dataset corresponds to the giant component of such a network, which presents a clustering coefficient C=0.254 and an average shortest path length d=3.606. (Guimerà et al, 2003) provides further information about the collection process and structural features of the resulting network.

SLACER graphs: Different outcome samples of SLACER algorithm's (Hales and Arteconi, 2006). The algorithm follows a link-based incentive approach, that is, nodes make and break links in the network to minimize the effects of others nodes’ selfish behavior. SLACER implements a simple local adaptation rule: Nodes can selfishly increase their own performance (or utility) in a greedy and adaptive way by changing their links and strategy. They do this by copying nodes that appear to be performing better and by making randomized changes with low probability.

WIKI data: Dump including daily snapshots of the Special:Statistics page for 11,500+ of the largest known MediaWiki-based wikis, collected between August 2007–April 2008 by polling a publicly-available web service at http://s23.org/wikistats/largest_html.php. The dump contains the following data for each entry: *unique ID, URL, title, number of content pages, total number of pages, edits, number of admins, number of users, number of images, stub ratio, timestamp*. In (Roth et al, 2008), the authors used these data to assess the content and population dynamics of a large sample of wikis.

Epistemic hypergraphs (scientific collaboration, socio-semantic networks): Datasets on co-evolving socio-semantic networks corresponding to the following data.

a. MedLine data, request on "zebrafish" for a 1950-2004 range on PubMed.
b. MedLine data, sample PubMed request on "bronchodilators for treating asthma in children", March 2007
c. Data featuring abstracts of accepted papers at ECCS 2005 and 2006

Flickr groups (group demographics, governance features): (3 subsets) Several dumps containing data about the structure, demographics and evolution of a sample of Flickr groups obtained via http://dev.nitens.org/flickr/group_trackr.php. The dump includes daily snapshots of groups since their registration to the service, including the following data: *number of users,*

number of admins, number of photos, throttling type, privacy type, moderation type (first and second subsets). For a subset of 9360 groups, supplementary data is available collected between June and July 2009, including: two snapshots of the *unique IDs of the entire population* of each group; two snapshots of the *unique user IDs of the contacts of each group member*; two snapshots of the *unique group ID of the affiliations of each group member*. A table with a summary of social network metrics for this selected dataset is also available, including the following measurements calculated on the network of group members defined via contact links: average of (directed) *user indegree*, average of *clustering coefficient*, *proportion of reciprocated links*, average *membership spread* (number of other group affiliations per member). The table also includes data on group membership turnover between the two snapshots: *number of new members* and *number of lost members* (third subset). Fresh datasets can be obtained on demand from the Web service.

Seeders' resource allocation in two BitTorrent communities: Measurements of current and possible resource allocation in two BitTorrent communities. Allocations are measured at instants; there are 100 such measurements for Bitsoup and 88 for Filelist. For each timestamp, there are three xml files: one which expresses the numbers of seeders and leechers in each torrent, and has no seeder unallocated, and two files where torrents are listed with no seeder in them and seeders are listed with their capacity not allocated and with the list of torrents they can seed at the measurement instant. The two files represent different estimation strategies for files currently owned by each seeder. Such strategies are described in (Capotă et al, submitted).

Citation graphs in arxiv.org (hep-th and hep-ph): Lists of citing - cited papers in categories *hep-th* and *hep-ph* of arxiv.org. The data includes all internal citations until 1 May 2003, and paper ids can be also used as timestamps to study citation dynamics within such a period. More information about the collection of these data can be found in http://www.cs.cornell.edu/projects/kddcup/datasets.html

Batch 1

Data obtained from the living-labs at some point during the first year of QLectives.

BarterCast data: This is an instance of the graph build by the BarterCast reputation mechanism crawled in September 2009. BarterCast is an epidemic protocol that allows peers to exchange information about their contribution in the network (namely, their upload and download rates). Given the fact that the graph built by BarterCast is sparse and disconnected, this dataset contains only the largest connected component consisting of 1163 nodes and 3268 edges. An edge between two nodes is directed and represents the information exchange (uploaded or downloaded information) between these two nodes.

Anonymized sample of download history from a set of Tribler peers: History of downloaded files as reported by a set of Tribler clients crawled in the Tribler network.

Anonymized sample of Bartercast records crawled from Tribler peers: Crawl of the Tribler network from 20 June 2009 to 9 September 2009. The crawler asks each discovered peer to send its BarterCast records with timestamp later than 20 June. The discovered peers are asked every hour for at least 50 records that they have not sent to the crawler yet.

The data is structured in a relational database with two tables. The main table is *anonymized_bartercast_records*. Its columns are *peerid, peer_id_from, peer_id_to, downloaded, uploaded, last_seen,,remote_peer_time* and*, crawler_time*. The other table is called *amended_bcast_records*. Its structure is the same as *anonymized_bartercast_records*, except that the columns "last_seen" , "remote_peer_time" and "crawler_time" have been replaced by the column "relative_time", and that crawler_time = remote_peer_time - last_seen + crawler_time. In the case "peer_id_from" and "peer_id_to" being the same, then download and uploaded values show the total amount of data that "peer_id_from" has download and uploaded (from/to both Tribler and non-Tribler peers) until that point of time.

Econophysics Forum data: This dataset contains 45k user-paper pairs which capture who (user) was reading what (paper) on the Econophysics Forum. The data was collected in 2008 from April till September and the present dataset is fully anonymised.

4. Data-collection for further development of the “Living-archive”

General strategy for data collection

The future lines of work regarding data collection from QLectives’ living labs were discussed in a working group during the project meeting in Zurich last June.

For the short and mid term, it was agreed that living labs should be used directly as providers of users’ behaviour data. This would include, for example collecting data before and after the introduction of a certain social functionality in a platform to evaluate users’ reaction, or combining user’s behaviour data with assumptions (e.g. collect users’ requests to a service providing scientific production indices in order to make visible ‘friendship’ or ‘rivalry’ relationships among scholars).

In a longer term, the possibility of developing a wide variety of sociological experiments using the living labs as a support (in a similar way to physical laboratories supporting ‘traditional’ behavioural experiments) was discussed.

QMedia

QLectives’ TUD team hosts QMedia. As they have direct access to the system, they can easily organize data collection actions (alone or in collaboration with other partners in the project) and deposit data in the Living Archive. As an example, TUD plans to verify the efficiency of certain mechanisms for indirect reciprocity among users as a way to enhance their cooperation, by collecting data about their behaviour before and after introducing such mechanisms.

QScience

QScience’s web instance and database are currently hosted by University of Fribourg, and data collection has been performed in the past via full access to the corresponding weblogs (i.e. files storing different aspects of users activity in the site). The IT service at University of Fribourg is working on a new, flexible and powerful tool for weblog access and analysis,

which is expected to be deployed very soon (even possibly before the final publication of this report).

The project is also considering the option of using TUD’s webservers to host QScience, mainly in order to avoid the dependence on technical staff external to the project. Such an alternative would possibly benefit data collection actions in terms of flexibility and format integration across living labs.

Other sources

Finally, as next steps in the implementation of QLectives’ living labs get clearer, new platforms become more likely to be included in further data collection actions within the project. Two relevant examples of such platforms would be the already existing Living Science (http://www.livingscience.ethz.ch/) and the planned QJournal (see QScience’s roadmap: http://www.qlectives.eu/wiki/index.php/QScienceRoadMap).

5. References

Capotă, M., Santos, F., Andrade, N. Vinkó, T. Brasileiro, F. and Pouwelse, J. (Submitted) “QSeeding”.

DDI (Data Documentation Initiative). http://www.ddialliance.org/

Guimera, R, Danon, L., Díaz-Guilera, A., Giralt, F. and Arenas A. (2003) “Self-similar community structure in a network of human interactions”, Physical Review E 68, 065103.

Hales, D. and Arteconi, S. (2006) "SLACER: A Self-Organizing Protocol for Coordination in Peer-to-Peer Networks," IEEE Intelligent Systems, vol. 21, (2), pp. 29-35,

Rahman, R. Hales, D., Meulpolder, M., Clements, M., Heinink, V., Pouwelse, J. and Sips, H (2009) “Robust vote sampling in a P2P media distribution system” ipdps, pp.1-8, 2009 IEEE International Symposium on Parallel&Distributed Processing, 2009

Roth, C., Taraborelli, D. and Gilbert, N. (2008) “Measuring wiki viability: An empirical assessment of the social dynamics of a large sample of wikis” Proceedings of the 4th International Symposium on Wikis and Open Collaboration WikiSym.