QLectives – Socially Intelligent Systems for Quality
Project no. 231200

Instrument: Large-scale integrating project (IP)
Programme: FP7-ICT

Deliverable D3.2.2
Datasets guide and manual – internal (living lab experiments) datasets

Submission date: 2013-02-28

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

Organisation name of lead contractor for this deliverable: UWAR

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
Michal Ziembowicz	University of Warsaw	ziembowicz@gmail.com

1.2 Other contributors

Name	Organisation	E-mail
Johan Pouwelese	TU Delft	j.a.pouwelse@tudelft.nl
Matus Medo	University of Fribourg	matus.medo@unifr.ch

1.3 Document history

Version#	Date	Change
V0.1	10 January, 2013	First draft
V0.9	18 February 2013	First final version
V1.0	28 February 2013	Approved version to be submitted to EU

1.4 Document data

Keywords	external datasets, living labs, data collection
Editor address data	ziembowicz@gmail.com
Delivery date	28 February 2013

1.5 Distribution list

Date	Issue	E-mail
	Consortium members	QLECTIVES@LIST.SURREY.AC.UK
	Project officer Roumen Borissov	Roumen.BORISSOV@ec.europa.eu
	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 quality.

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

The document contains descriptions of 10 databases collected within the QLectives project, as a part of living labs (QScience and QMedia). The data were used for validation of hypotheses developed in Streams 1 and 2. The report is divided into two main sections: QScience and QMedia, which in turn are divided into 3 batches. Each database description consists of 3 or 4 points:

- brief description – a short text describing the general characteristics of the data as well as the context in which they were collected;
- data format specification – describing available measures and rows' characteristics;
- information on data availability – describing the ways of obtaining the datasets, in most cases it is in the form of URLs;
- additional content – for some of the databases a set of scripts or preprocessed datasets are available and they are listed here.

Contents

1 Introduction

Qlectives was designed to rely on data collected each year by the "living labs". The living labs are two environments designed and ran within Qlectives that provide first-hand data for analysis. Each year a new portion of data (called "batch") was fed from Stream 3 and 4 into Streams 1 and 2. In the first year the data from living labs were not yet available and therefore the external sets were used. During subsequent years Batch1, Batch2 and Batch3 were produced. The aim of this report is to prepare a catalog of databases collected within Qlectives.

The document contains descriptions of 10 databases collected within the QLectives project, as a part of living labs (QScience and QMedia). The data were used for validation of hypotheses developed in Streams 1 and 2. The report is divided into two main sections: QScience and QMedia, which in turn are divided into 3 batches. Each database description consists of 3 or 4 points:

- brief description – a short text describing the general characteristics of the data as well as the context in which they were collected;
- data format specification – describing available measures and rows' characteristics;
- information on data availability – describing the ways of obtaining the datasets, in most cases it is in the form of URLs;
- additional content – for some of the databases a set of scripts or preprocessed datasets are available and they are listed here.

2 QMedia datasets

QMedia is an experimental media-sharing distributed software built around the concept of quality collectives. This database aims at describing how people link to others over the Internet to gain value and facilitate collaboration. The network of QMedia users serves as a living lab for QLectives, both as a test bed for developed algorithms and providing input for the design and evaluation of new algorithms.

The database is divided into 3 batches. All of them relate to different measures of BarterCast and Tribler data. More detailed description of the QMedia platform can be found in the D4.3.2 report. All datasets are publicly aviailable. Links are provided in the descriptions of each database.

Batch1

1) BarterCast graph

a) *Description*: This an instance of the graph build by BarterCast reputation mechanism crawled on 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). The BarterCast mechanism is based on building a weighted directed graph from the data transfers that have occurred among the peers, and on employing the Maxflow algorithm in this graph to evaluate reputations. Given the fact that the graph build 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.

b) *Data format*: Each row represents an exchange between 2 peers with the size of the transferred data. No information about the timing is available.

Data headers:

- "Upload" – user1 id (an integer number),
- "Link" – user2 id (an integer number),
- "Download" – volume of information exchanged by user1 with user2 (an integer number).

c) *Data availability:* The data is available in the form of an .sql database accompanied by .xml file with the metadata specification under these URLs:

- *http://www.qlectives.eu/wiki/images/5/5b/Bartercast-data.sql,*
- *http://www.qlectives.eu/wiki/images/6/65/B1_QMedia_001.xml.*

d) *Additional content:* **BartercastPeerStats**: A processed dataset with the statistics of uploads and downloads of peers. Each row represents a peer.
Data headers:
- "numUpload" – number of uploads
- "numDownload" – number of downloads
- "totalUpload" – total size of uploads
- "totalDownload" – total size of downloads
- "meanUpload" – average size of an upload
- "meanDownload" – average size of a download
- "sdUpload" – standard deviation of the upload size distribution
- "sdDownload" – standard deviation of the download size distribution

The database in .csv form is available here:
- *http://www.qlectives.eu/wiki/images/c/ca/BartercastPeerStats.csv.*

2) Anonymized sample of download history from a set of Tribler peers

a) *Description*: History of downloaded files as reported by a set of Tribler clients crawled in the Tribler network.

b) *Data format*: The variables define who downloaded content from a particular swarm (indicated by a torrent id) at a given time.
Data headers:
- "peer id" – peer ID,
- "torrent id" – file ID,
- "timestamp" – when the file was downloaded by the peer.

c) *Data availability:* The data are available in the form of a compressed text file accompanied by an .xml file with the metadata specification under these URLs:
- *http://qlectives.eu/wiki/datastore/anonymized_download_activity.bz2,*
- *http://www.qlectives.eu/wiki/images/d/d8/B1_QMedia_002.xml.*

d) *Additional content*: Time-sliced statistics tracking peers' activities over time - available on request (Chih-Chun Chen - *c.chen@abmcet.net*).

3) Anonymized sample of Bartercast records crawled from Tribler peers

a) *Description*: We crawled the Tribler network from 20 June 2009 until 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.

e) *Data format:* In the case that "peer_id_from" and "peer_id_to" be 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. Data headers:

- "peerid" – the id of contacted peer by the crawler,
- "peer_id_from" – the id of the peer in the BarterCast record,
- "peer_id_to" – the id of the peer in the BarterCast record,
- "downloaded" – the amount of data that "peer_id_from" has downloaded from "peer_id_to",
- "uploaded" – the amount of data that "peer_id_from" has uploaded to "peer_id_to",
- "last_seen" – the time that peerid has got informed about the activity of "peer_id_from" and "peer_id_to download",
- "remote_peer_time – the local time of the "peerid" before replying to crawler,
- "crawler_time" – the local time of the crawler when it receives a reply from remote peer ("peerid").

b) *Data availability:* The data are available in the form of a compressed text file accompanied by an .xml file with the metadata specification under these URLs:

- *http://www.qlectives.eu/wiki/images/f/f1/Anonymized_records.bz2,*
- *http://www.qlectives.eu/wiki/images/d/d8/B1_QMedia_003.xml.*

Batch2

4) Anonymized sample of download history from a set of Tribler peers

a) *Description*: Anonymised data about user behavior crawled in Tribler's network from January 1 to July 1, 2010. The dataset was created by parsing the logs of the superpeers. These servers are used in the Tribler network, for bootstrapping newcomers). Furthermore, the superpeers try to contact a peer once every 4 hours. The responses of the peers are written to a file, which were used to create the dataset.

b) *Data format:* In total 4343 peer's download behavior were logged. Below a short description of the meaning of the fields in the database is given, where permid stands for an anonymized identifier of a user, tid stands for an anonymized file identifier, they are both integer numbers. Data headers:

- "permid_tid" (permid, tid, timestamp) – first known timestamp for which this permid has downloaded this tid. An older version was published in Batch 1,
- "permid_first_last" (permid, first_seen, last_seen) – a file containing the first and lastseen dates of each permid,
- "permid_nr" (permid, num_downloads) – a file containing the number of total downloads for each permid,
- "above_average_sorted" (rank, permid, num_downloads) – a file containing all peers with above average number of downloads, sorted by num_downloads desc and added a rank,
- "top_downloader_when" (timestamp, peer_...) – the number of files downloaded at this timestamp by the top-50 downloaders. Each peer has its own column,
- "permid_active" (timestamp, version_..., event) – file containing the number of ative peers that specific day (a peer is active if a superpeer has received a message from it),
- "tid_nr_sorted" (rank, tid, num_downloads) – file containing the number of downloads foreach tid, sorted by num_downloads desc.

c) Data availability: The data are available in the form of a compressed text file accompanied by an .xml file with the metadata specification under these URLs:

- *http://www.qlectives.eu/wiki/images/3/36/Qmedia_batch2.sql,*
- *http://www.qlectives.eu/wiki/images/8/8f/B2_QMedia_001.xml.*

Batch3

5) Qmedia BarterCast reputation dataset

a) *Description:* The dataset collected for Qlectives as one of the living labs was extracted from real-world usage of Qmedia by over 73,000 participants. It focuses on the reputation mechanism. Reputation mechanisms are widely used in online networks to rank users or products, but despite their importance, very few studies have been done or published on their real behavior. This dataset studies an Internet-deployed distributed reputation mechanism called BarterCast that is specifically designed for peer-to-peer file-sharing systems. In the paper referred below, we study this mechanism from the network perspective and we provide a detailed analysis, which includes such network topology measures as the degree distribution, node interconnectivity, the clustering coefficient, community structure, and distance measures. Besides, we study the geographical spread and content sharing behavior of the system participants and correlate the results with their connectivity in the network. We interpret each evaluated measure in the scope of reputation and file-sharing mechanisms and propose relevant implications and prospective applications for future

b) All the measurements are based on data that we have collected during two years of crawling the QMedia file-sharing network, which employs BarterCast as its reputation mechanism. In the BarterCast work-graph, an edge indicates the amount of data transferred from one peer to another. The directed graph contains of 73,201 nodes and 352,042 edges. Furthermore, since most of the graph measures, like the clustering coefficient, only make sense when the underlying graph is connected, we consider the Largest Connected Component (LCC) of the work-graph. In total there are 939 connected components, out of which 780 contain only two nodes. References:

> *Rahim Delaviz, Niels Zeilemaker, Johan A. Pouwelse, and Dick H.J. Epema, "A Network Science Perspective of a Distributed Reputation Mechanismâ", submitted for publication at IFIP networking 2013, Brooklyn, New York.*

c) *Data format:* The data are stored in the Pajek NET format. First part of the file is a list of nodes, in the second part in each row there are 3 numbers:

- first peer ID,
- second peer ID,
- amount of transferred data.

d) *Data availability:* The complete BarterCast graph dataset (anonymized) in the PAjek NET format can be found at
   - *http://www.qlectives.eu/wiki/images/1/1f/Latest_records_part_16.txt.Looped.Pajek.net.gz*.

3 QScience datasets

Batch1

6) Econophysics Forum data1

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

b) *Data format*: Each row represents a pair:
- "user_id" – an ID of a user (an integer number)
- "paper_id" – an ID of a paper (an integer number)

a) *Data availability:* The data is available in the form of an .sql database accompanied by .xml file with the metadata specification under these URLs:
- *http://www.qlectives.eu/wiki/images/a/a2/EF08_04-09.sql,*
- *http://www.qlectives.eu/wiki/images/7/76/B1_QScience_001.xml.*

c) *Additional content*: In addition, there are two processed datasets of
- number of papers per user – available in .csv file under this URL: *http://www.qlectives.eu/wiki/images/3/38/EcoPhysT1PapersPerUser.csv*
- number of users per paper – available in .csv file under this URL: *http://www.qlectives.eu/wiki/images/7/70/EcoPhysT1UsersPerPaper.csv.*

7) American Physical Society's bibliometric data (additional dataset)

a) *Description*: This dataset is external to Qlectives, however it was included in Batch1 as part of QScience research material. This corpus comprises over 450,000 articles and dates back to 1893; its size and extent make it attractive for use in research about networks and the social aspects of science. The first data set consists of all pairings of articles in which one article cites another within the collection. The second set contains basic metadata about each article in the collection. Researchers may learn more about the data sets and request access to them by visiting *http://publish.aps.org/datasets/*. Requests will be quickly reviewed and, if approved, the data will be made available for download. Questions may be sent to*: data-requests@ridge.aps.org.*

b) *Additional content*: A java library is available for retrieving publication details from the metadata. Contact: Chih-Chun Chen *(c.chen@abmcet.net)* for details.

Batch2

7) Econophysics Forum data 2

a) *Description*: User behaviour in EconophysicsForum (*www.unifr.ch/econophysics*) from July 5 to November 11, 2010. The dataset is fully anonymised

b) *Data format:* Each row represents a timestamped interaction of a user and a paper.
Data headers:
- "user id" – an ID of a user (an integer number),
- "paper id" – an ID of a paper (an integer number),
- "user action" – an action of the user (a character 'D' – download or 'V' – view),
- "timestamp" – the date and time of the interaction (datetime sql format)

c) *Data availability:* The data is available in the form of an .sql database accompanied by .xml file with the metadata specification under these URLs:
- *http://www.qlectives.eu/wiki/images/9/93/EFanonymized_data.sql,*
- *http://www.qlectives.eu/wiki/images/4/4a/B2_QScience_001.xml.*

8) Interviews with Scientists (additional dataset)

a) *Description*: This dataset was collected within Qlectives, yet it is not part of the living labs. It was included in Batch2. This dataset contains transcripts (N=18) from interviews (N=19) conducted with scientists in the natural sciences talking about quality in science. The data was collected between October and December 2010 and were anonymised.

b) *Data availability:* Available upon request from: Maria Xenitidou (*M.Xenitidou@surrey.ac.uk*).

9) GoodReads rating data (additional dataset)

a) *Description*: This dataset is external to Qlectives, however it was included in Batch2 as part of QScience research material. This dataset contains (i) the ratings for 10000 books from the Goodreads community book-sharing service: *http://www.goodreads.com/* and some user metrics (e.g. number of books, number of friends, demographics) and (ii) a subnetwork of the Goodreads social network.

b) *Data availability:* Available upon request from: Chih-Chun Chen (*c.chen@abmcet.net*).

Batch3

10) Econophysics Forum data 3

a) *Description*: User behaviour in EconophysicsForum from July 6 2010 to December 31, 2012. The dataset is fully anonymised. to improve the anonymity, paper uploading actions are omitted from the data file EF_anonymized.dat - only abstract views (V) and downloads (D) are stored. Data from 4-10 November 2012 are missing because the corresponding weblogs haven't been stored properly.

b) *Data format:* Each row represents a timestamped interaction of a user and a paper.
Data headers:
- "user" – an ID of a user (an integer number),
- "paper" – an ID of a paper (an integer number),
- "action" – an action of the user (a character 'D' – download article or 'V' – view abstract),
- "days from 06/07/2010" – number of days from the beginning of data set (an integer number)
- "date and time" – the date and time of the interaction (datetime sql format)

c) *Data availability:* The data is in the form of an .dat file ('*EF_anonymized.dat*') containing tab-delimited columns of data. It is available on request from: Matus Medo (*matus.medo@unifr.ch*)
File '*paper_metadata.dat*' contains metadata of relevant papers - papers which do not have any metadata on the Econophysics Forum are listed in '*paper_metadata-errors.dat*'. Both files are not publicly available due to privacy concerns.

11) Focus Groups with Scientists (additional dataset)

a) *Description*: This dataset was collected within Qlectives, yet it is not part of the living labs. It was included in Batch3. This dataset contains audio and transcripts from focus groups (N=4) conducted with academics and researchers in the social and natural sciences asking them to discuss activities related to collaboration and assessing publications in focus group sessions. The data was collected between June and July 2012.

b) *Data availability:* Available upon request from: Maria Xenitidou (*M.Xenitidou@surrey.ac.uk*).