# Identification of information sources and data processing techniques

## Executive summary

### Introduction

The current document " Research nr. 1.20. Internet users' behaviour analysis tool" is one of the reports produced within ERAF co-funded project "Information and communication technology competence centre". The research activity No 1.20 "Prototype Design" is led by a "IT Kompetences centrs SIA" and implemented by consortium of scientific and industrial partners.

Current document contains a study and evaluation of available data sources in global Internet. The activity starts with analysis of data availability, data quality, data validity. Current document serves as a fundamental element for further development stages and is a mandatory precondition for a practical subject linking algorithm development for data sources available. Moreover carrying out this research has brought acquisition of new practically applicable knowledge on data quality, data validity, data storage, data transformation, data structures, and prototype design.

### Aims of the Research and Methodology

Aims of the research are as follows:

(1) To identify available information sources/databases and data processing methods, providing various types of integrated information

(2) To acquire and present the new published and expert knowledge on Internet information sources available - data quality, and validity.

The research methodology has been chosen based on the aims of the research task, accessibility to information sources and practical business context. We selected the available sources of information in Facebook, Draugiem, Twitter, LinkedIn, CVK.

Research was carried out in three phases: (1) survey of available data sources (2) expert surveys and interviews (3) scientific literature review on data quality, data validity, data storage, data transformation, and data structures.

We used Enterprise Knowledge Development (EKD) approach to script and integrate the expert knowledge in form or reusable knowledge

## Research Process and Results

During the first phase researchers have surveyed five data sources: Facebook, Draugiem, Twitter, LinkedIn, and CVK.

The report is designed on bases on system specification documents, it is linked with following documents:

Examination of requirements;

LVS 72: 1996 Recommended Practice the Description of Software Design.

We used following research methods: Reesearch literature review, and Experts´ Views Collection.

Research literature surwey was used to evaluate the existing available big data research results on Big data quality, data validity, and design of data structures, data processing, and data storage.

We used Enterprise Knowledge Development (EKD) approach to script and integrate the expert knowledge in form or reusable knowledge is presented in Chapter 1, the detailed description is in Amendment 1. The concepts of data quality and validity is presented in Chapter 2. To identify the needed data structures and transformations we carried out the benchmarking of metodologies and selected Zachman Framework architecture.

Research results are presented in Chapter 3, data structures in chapter 4, data storage technical support in Chapter 5. Chapter 6 is prototype design and data quality characteristics.