MDQual – (Meta)-Data Quality Workshop

Dimitris Gavrilis[1](✉) and Christos Papatheodorou[2]

[1] Department of Electrical and Computer Engineering,
University of Patras, Patras, Greece
gavrilis@gmail.com

[2] Department of Archives, Library Science and Museology,
Ionian University, Corfu, Greece
papatheodor@ionio.gr

1 Introduction

It is well known that we are rapidly moving towards a data driven world, where all aspects in our everyday lives are data driven. In all domains, from healthcare to retail and finance, data is collected, analyzed and used to make decisions, usually utilizing machine learning techniques. Data Science involves collecting, cleansing and integrating data prior of analysis. The quality of this data is critical and directly affects the outcome of all data science related tasks. Moreover, metadata is used to annotate data and facilitate data organization and retrieval. Metadata quality also directly affects retrieval and other operations (such as data integration) and workflows that are metadata driven.

Although various metrics have been proposed to measure metadata and data quality, in most cases they are highly subjective and/or domain specific. Moreover, they are directly related to the intended use of the data, meaning that a dataset could be of high quality for one use and of low quality for another. In all cases, (meta)data quality has a tremendous impact on data science related tasks and ultimately in everyday life. The proposed workshop aims at exploring the various quality issues found in people working with both data and metadata across domains. An inter-disciplinary workshop where data scientists across different domains will meet and:

- share and exchange experiences regarding (meta)data quality;
- identify patterns in (meta)data quality;
- share methodologies and metrics that will help to measure (meta)-data quality;
- share/propose tools that can be used effectively in improving (automatically) (meta)-data quality.

This initiative aimed at bringing together a community of data scientists that have expertise in a diverse set of domains, such as archives and libraries, healthcare, biology, humanities, computer science and engineering, environment, agriculture, economics, etc. Apart from sharing metrics and methods to identify and resolve quality issues and evaluate datasets, the workshop aimed at promoting the use of tools and services for the automatic measurement and improvement of (meta)data quality. Although few such

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 671–672, 2017.
DOI: 10.1007/978-3-319-67008-9

tools are available in the market, a good number of standalone micro-services are available and can be used to automatically improve (meta)data quality.

We welcomed position papers expressing the data and metadata quality needs from content providers (libraries, archives, museums, public and private sector organizations that manage multimedia content). Moreover, we welcomed research papers that described methods, metrics, services and tools for measuring and ensuring quality. The workshop provided a session for demonstrating implemented systems and services in order to trigger discussions on real world needs and running systems.

2 Topics

Indicative topics of the Workshop were:

- Data and metadata quality measurement methods
- Data and metadata quality requirements for e-research, health, education and digital humanities, etc.
- Metrics for data quality measurement in for e-research, health, education and digital humanities, etc.
- Metrics for metadata quality measurement in for e-research, health, education and digital humanities, etc.
- Tools and services for measuring quality
- Tools and services for improving quality
- Services for automatic data and metadata enrichment

3 Website

More information can be found at http://qualitics.org/mdqual2017.