InterSocialDB: An Infrastructure for Managing Social Data
Dimitris Souravlias, Georgia Koloniari and Evaggelia Pitoura
Computer Science Department,
University of Ioannina, Greece
{dsouravl, kgeorgia, pitoura}@cs.uoi.gr

Abstract
Today, online social networks generate content of unprecedented scale. In this short paper, we describe the basic components of an infrastructure for storing, managing and analyzing social data.

1. Introduction
The rapid proliferation of online social networks (OSNs), such as Facebook, Twitter and LinkedIn, has resulted in creating an influx of timely data on any topic of interest. Most popular OSNs offer APIs that allow applications to access their underlying social graph and user-generated content. The analysis of social data attained from such OSNs has applications in many areas including mobile services, retail, advertisement, manufacturing, and social and life sciences.

Unlike traditional data, content generated by social networks is inherently dynamic. After the initial posting of an item, users continue to generate new related content by for example posting comments, ratings or replies. As time passes, the perceived popularity, interpretation and quality of published data may change significantly. Another major difference with traditional data is quality. In traditional database systems, although there may be some inconsistencies (such as those caused by typing or other errors), the typical range of quality is substantially narrower than that in social media where the quality of content varies from very high to very low, including postings that constitute spam or advertisements.

Social data have rich, highly interrelated content. In addition to document content and link structure, a wide variety of metadata is available regarding authorship, time, preferences, and interactions with other content and users. Natively, social data is not in any unified structured format; for example, tweets are pieces of text augmented with links, hashtags and other metadata, Facebook pages contain a variety of different types of content, including images and videos, while in all cases, along with data there is social information (such as friends, followers, interest groups) and popularity related information (such as likes or mentions).

Collecting, modeling, storing and analyzing social data poses many challenges [1]. In this paper, we outline an infrastructure for a social data management service for providing storage and processing functionality to applications that would like to analyze social data.

2. The InterSocialDB Framework
Our goal is to develop an infrastructure for managing social data. The components of this infrastructure can be distinguished into (Fig. 1):

(a) a data acquisition component that collects (either continuously or on demand), pre-processes, models, aggregates and stores data generated from online social networks, and

(b) a data processing component that performs various analytical tasks on social data and presents the results of this analysis to the users.

The data acquisition component can be partitioned into three phases, namely collection, pre-processing and integration.
During data collection, data are gathered from social networking sites. Such collection may be through ad-hoc, one-time queries or through continuous monitoring queries. Besides specifying which data to collect, it is central to determine appropriate metadata and provenance information for annotating the collected data. Provenance is a general term referring to recording and tracing the origins and alterations of data stored in a database [2]. Take for example a simple query about the appearance of term “Patras”. Annotations may include metadata about the appearance of the term, such as information about the related social media post and its creator as well as provenance information including who, when and where posed the query about “Patras” and when and how the collection of data was performed. After the collection of raw data and its annotations, extra processing is required for extracting related information and data cleaning. Since the original data do not follow any specific format, we need to translate data to a common data model. There are many alternatives for storing social data [3]. Candidates include (a) traditional relational database systems, such as MySQL, (b) key-value stores (for example, HBase), (c) document databases (such as CouchDB) and (d) native graph databases (such as Neo4j). As a final step, before storing the collected data, some integration-related activities are applicable. Since we will support data generated from more than one social networking site, for example, we need to identify references to the same object.

The data processing part refers to the analysis of the collected data so that important information is extracted. The analysis may use both stored data and live streaming data. Data presentation is also central for identifying potential interesting events.

3. Conclusions
In this short report, we argued that besides the analysis of social data, a variety of data management tasks, including modeling, integration and storage are central. We are currently in the process of implementing the related infrastructure.

References