The Dacura Data Curation System

Kevin Feeney(✉)

Knowledge and Data Engineering Group,
School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
kevin.feeney@cs.tcd.ie

Abstract. This paper describes the Dacura system which uses linked-data technologies to monitor and constrain the quality of datasets assembled from heter-ogeneous sources and managed by distributed teams of domain experts.

Keywords: Linked data quality metrics · Data curation · Visualization · Semantic web

1 Introduction

Consistency and accuracy of data is a very real concern in scaling datasets which leverage the web of data [1]. Harvesting and analyzing linked data can prove to be challenging due to schematic differences, incomplete and inaccurate information, entropy and the rapid pace of change [2]. If the data collected is to be used to support real-world analysis and decision making, the quality of the data is a critical factor in determining confidence in the analytic results. Furthermore, if we are to build applications which are based upon such linked datasets, data-quality assurance guarantees (for example about property completeness and schematic conformance) facilitate software development and the construction of robust applications.

Collecting linked datasets which leverage the web of data and other big-data sources is, in general, an open ended challenge in that such datasets are rarely complete. The continuous emergence of new information sources and the publication of new datasets mean that there is a steady supply of new data that can be incorporated into harvested datasets to improve accuracy and coverage and to reflect changes over time. In this context, we are interested in scalability, not in terms of the size of the dataset that we are collecting, but scalability in terms of the volume of data that we leverage in constructing our dataset. This concept of scalability is particularly relevant in the linked data domain, where interlinking facilitates the construction of datasets based on multiple externally managed datasets [3]. Improving productivity and efficiency of data quality control is critical to achieving this scalability, both by enabling the incorporation of automated data-harvesting approaches into data-collection efforts and by minimizing the time and effort required by human dataset managers and by domain experts.

B. Bozic et al. (Eds.): CHDDH 2016, IFIP AICT 482, pp. 15–20, 2016.
DOI: 10.1007/978-3-319-46224-0_2

The general problem that Dacura addresses is that we want to collect and publish linked datasets and build applications that depend on that data. However, we also want to continually improve, refine and extend our dataset by incorporating more sources of data over time.

2 Use Case and Requirements

This section describes our use case, and how our quality-measurement requirements were derived. We wish to support an internationally distributed community of humanities and social science scholars who are collecting, improving, sharing, exploring and analyzing time series data describing historical human societies. This community is collaborating on the Seshat Global History Databank project [4] (Fig. 1), which aims to encode time-series datasets describing the most important features of all human societies since Neolithic times. The scope of the project – over 100 researchers and approximately $10 million in total funding, divided across multiple autonomous collaborating projects, with a 10 year time-frame – is such that the data-collection process is necessarily incremental. All data will have the potential to be progressively improved and extended, to take advantage of the potentially large community of volunteers, and the large number of potential sources of data available on the web 00 from structured datasets in DBpedia, to details of archaeological finds encoded in databases, or described in academic publications stored in electronic archives.

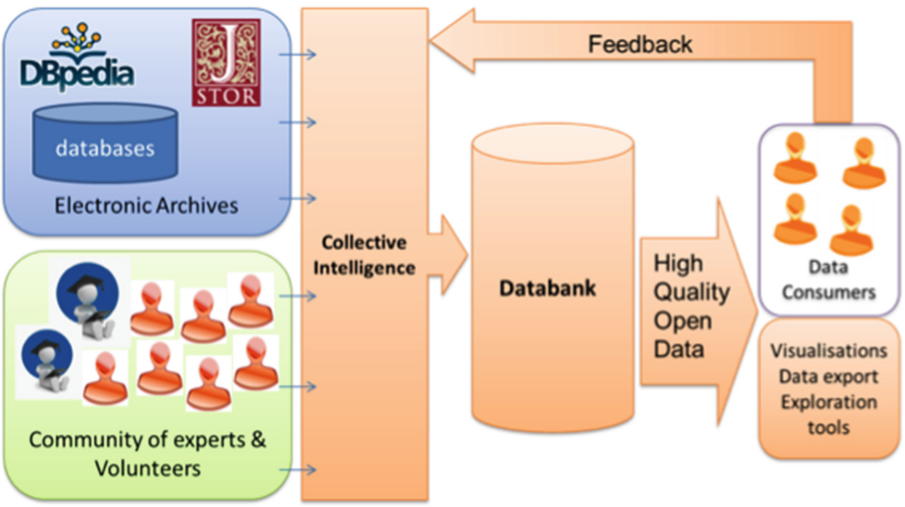

Fig. 1. Use Case – scholars creating, sharing and improving HSS data

The overarching goal of our work is to harness the input of the community of human experts to most efficiently and productively transform the wealth of

knowledge and data available into high-quality datasets and to provide visualisations, analysis and modelling, data-export and a variety of other tools based upon that data. The system must be dynamic because a requirement of the research program is to iteratively publish datasets which cover specific regions and time-slices and subsets of the Seshat variables and to evolve the datasets so that they improve progressively over time as their coverage is extended. From this goal, we derived the following general system objectives of Dacura:

O1. To maximize our ability to expand and improve the dataset over time, including evolving the schema, without compromising tools built on top of the dataset.
O2. To minimise the effort required to incorporate new human expertise and new sources of data into the curated dataset.
O3. To detect inconsistencies in the dataset, areas where it is incomplete, areas where it does not adhere to best practice in linked data publication.
O4. To provide mechanisms to improve the dataset over time in a controlled way, including generation of suggestions for amendments.
O5. To track the progress on the dataset in terms of the rate of change and time to completion.

The specific requirements of the domain dictated further requirements:

Req. 1: RDF-based storage and publication of shared Linked Data datasets.
Req. 2: Workflows and tools for distributed authoring and enrichment of consistent, high-quality datasets.
Req. 3: Domain-specific data visualisations of historical dynamics through a range of temporal and geographic frames.
Req. 4: Data quality analysis and enforcement functions that assure high quality data is produced.
Req. 5: Processes and tools that support the lifecycle of data, allowing for controlled change of datasets and schemata.
Req. 6: Application-aware processes and tools that constrain data changes such that data consuming applications continue to function despite an evolving dataset.

The system under study consists of data acquisition, data enrichment, data lifecycle management, data publication and data visualisation, exploration and analysis tools built on top of the data. Data acquisition initially focused on semantic uplift of previously collected data in structured formats such as spreadsheets or databases to a common Linked Data vocabulary. The next stage of data acquisition will support the semi-automated generation of new datasets from raw historical sources such as online newspaper archives. This second acquisition stage must minimise the effort required from domain experts to create the dataset by automating retrieval/conversion tasks and allowing less knowledgeable contributors to validate and incrementally improve candidate time series data before experts finally authorise it.

In order to develop tools which could meet these requirements, data-quality analysis was identified as a critically important capability. Monitoring of data-quality as it passes through the various lifecycle stages is required if the final published data is to remain of high enough quality to support meaningful statistical analysis. Furthermore, to facilitate the development of robust visualisation and exploration tools on top of the data, certain thresholds of quality must be met (for example, completeness of a location property is often a pre-requisite for datasets to be meaningfully mapped) and data-quality monitoring is required in order to ensure that the curated dataset continues to meet these thresholds as it changes over time. The rest of this paper describes the data-quality measurement features that we developed through the Dacura platform in order to allow us to create work-flows to meet these requirements.

3 Dacura System and Quality Control

The Dacura system [5] developed at TCD is a data curation system designed to support the harvesting, assessment, management and publication of high-quality Linked Open Data. The Dacura workflow is outlined in Fig. 2.

Data architects define schemas which describe the information contained within external sources. These sources can range from other Linked Open Data sources to unstructured text such as books, journal articles, and newspapers. These sources are entered into the candidate queue. Architects can define what processing steps, such as human review and automated data quality checking, are required for candidates to be accepted into the Dacura datastore as reports. These reports can then be analysed by domain experts, in order to transform potentially incomplete, incorrect, or incompatible reports into coherent summaries of the information which they contain. These expert-generated facts can then be published in various formats, such as data dumps, browseable catalogues, or interactive visualisations. The published representations are then viewed by data consumers, who can offer suggestions for modifications to the underlying schemas and corrections to the generated data. Dacura provides tools at all steps of the data lifecycle to support users in these various roles.

Automated data quality analysis is deployed to ensure that the data harvested by this process is high quality and to improve the usefulness of generated datasets [5]. Updates to data are inspected to ensure that they contain valid content and formatting. Messages are characterised by operation type and graph, and then tested against semantic rules which describe data quality issues. The system finally checks user permissions to determine if users can directly update the datastore or add items to the report processing queue.

Experiments using the Dacura tool to measure the speed of production and quality of political violence datasets indicated large variability between the quality of data produced by different human volunteers [5] which can significantly affect the overall quality of the dataset. As a consequence, it is evident that there is a need to implement live data quality control, something that provides data curators with tools to assess the quality of the data being produced during the

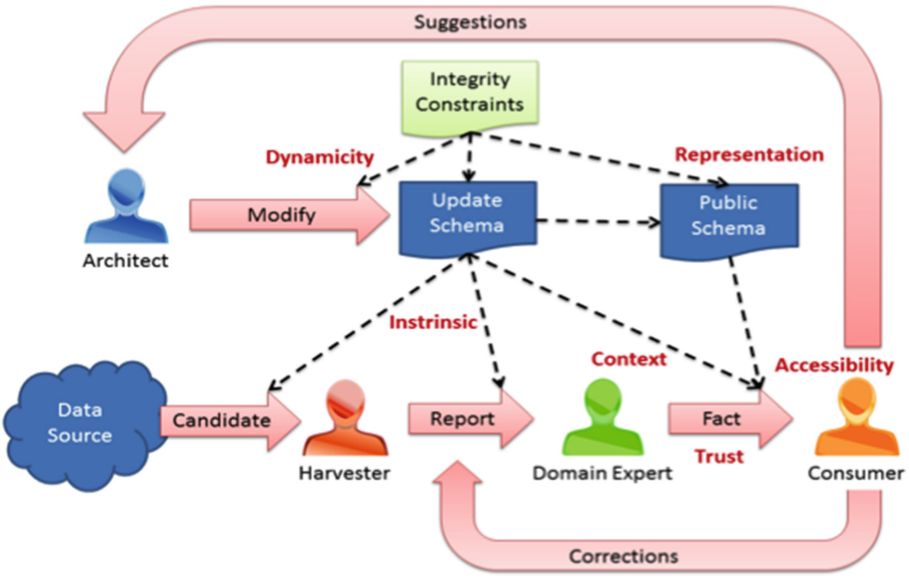

Fig. 2. Basic dacura workflow and roles.

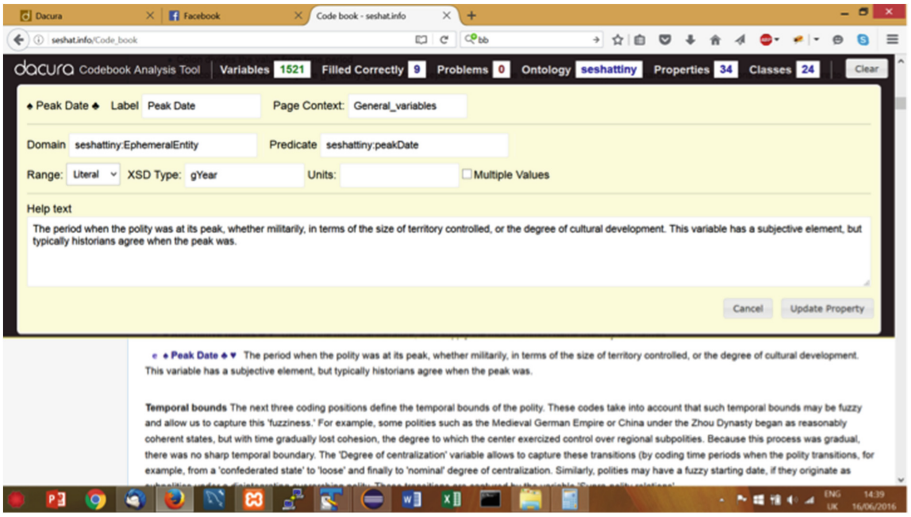

Fig. 3. Screenshot of the Dacura codebook analysis tool.

harvesting process itself, and not only after the final results are present. This could ensure that good and reliable Linked Data will be available during all the data lifecycle processes (Fig. 2).

Currently, development of the Dacura system is focused on providing tool support for the Seshat Global History Databank. This is a large-scale effort

to provide historical time-series data for the entirety of human history from the development of agriculture to the present day, in order to test and explore models of social structures and change. Seshat currently uses a wiki to collect data, but this approach makes it difficult to collate the collected information in a format amenable to statistical analysis. Dacura provides a set of web-based tools to allow Seshat researchers to more quickly and easily enter their data, as well as storing it in a format which allows easy manipulation in order to test historical hypotheses.

4 Conclusion

In this paper we provide an overview of the Dacura system. This paper outlines the requirements that led to its construction, a brief overview of how it functions, and some details as to how it integrates with the Seshat wiki. The Dacura system is under continuous development, and will be undergoing trials with users in the autumn of 2016. These trials will be used to further develop and improve the system.

References

1. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., Hitzler, P.: Quality assessment methodologies for linked open data. Semant. Web J. **7**, 63–93 (2013)
2. Millard, I., Glaser, H., Salvadores, M., Shadbolt, N.: Consuming multiple linked data sources: challenges and experiences. In: First International Workshop on Consuming Linked Data (COLD2010), Shanghai (2010)
3. Schultz, A., Matteini, A., Isele, R., Mendes, P., Bizer, C., Becker, C.: LDIF - a framework for large-scale linked data integration. In: 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France (2012)
4. Turchin, P., Brennan, R., Currie, T., Feeney, K., Francois, P., Hoyer, D., Manning, J., Marciniak, A., Mullins, D., Palmisano, A., Peregrine, P., Turner., E.A.L., Whitehouse., H.: Seshat: the global history databank in Cliodynamics: J. Quant. Hist. Cult. Evol. **6**(1) (2015)
5. Feeney, K., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. Int. J. Semant. Web Inf. Syst. **10**(2), 35–62 (2015)