

Saffron: A Data Value Assessment Tool for Quantifying the Value of Data Assets [★] ^{★★}

Judie Attard¹[0000-0001-7507-1864], Jeremy Debattista²[0000-0002-5592-8936],
and Rob Brennan³[0000-0001-8236-362X]

¹ ADAPT Centre, Trinity College Dublin, Ireland
`judie.attard@adaptcentre.ie`

² School of Computer Science and Statistics, Trinity College Dublin, Ireland
`debattij@scss.tcd.ie`

³ ADAPT Centre, School of Computing, Dublin City University, Ireland
`rob.brennan@dcu.ie`

Abstract. Data has become an indispensable commodity and it is the basis for many products and services. It has become increasingly important to understand the value of this data in order to be able to exploit it and reap the full benefits. Yet, many businesses and entities are simply hoarding data without understanding its true potential. We here present *Saffron*; a Data Value Assessment Tool that enables the quantification of the value of data assets based on a number of different data value dimensions. Based on the Data Value Vocabulary (DaVe), Saffron enables the extensible representation of the calculated value of data assets, whilst also catering for the subjective and contextual nature of data value. The tool exploits semantic technologies in order to provide traceable explanations of the calculated data value. Saffron therefore provides the first step towards the efficient and effective exploitation of data assets.

Keywords: Data value · Data governance · Data value monitoring · Data value assessment · Linked Data · Explainability.

1 Introduction

“*Data is the new oil*” is a claim supported by many. Even though there are many things that differ between data and oil as a resource, such as their renewability and their effect on the environment, one cannot deny the similarities in their usage and utility potential, as well as in their nature of being indispensable commodities in today’s society. We are increasingly relying on data or data-based products and services, particularly in recent times, when the use of big

[★] This research has received funding from the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106), co-funded by the European Regional Development Fund and the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567.

^{★★} Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

data is ever so prevalent, and successful decision-making requires the effective contextual exploitation of information.

Whether one agrees with the above-mentioned claim or not, it is undeniable that data is, to different extents, valuable. But what is exactly meant by data value? Numerous publications in literature explore this term in various domains. Whilst the existing definitions of value might be somewhat similar, there is currently no consensus on the definition of “data value”, or on its representation. Moreover, it is inherently challenging to measure the value of data due to the subjective and contextual nature of value. In fact, to the extent of our knowledge, there currently exists no tool or framework that quantifies the value of data based on various data value dimensions (aspects that characterise data value, e.g. quality, cost, usage). In literature there are some approaches towards measuring one or two of these dimensions, such as [2–4], however these cannot be deemed as appropriate solutions to quantify data value since they do not cater for the highly heterogeneous nature of data value. While it is evident that the use of data has become a vital part of our everyday lives, only few are able to understand the usefulness of measuring of the value of data. In fact, many businesses are hoarding data without actually exploiting it or understanding its potential.

In order to target the niche in the topic of data value, our goal in this paper is to tackle the quantification of data value. This quantification is essential to the efficient and effective exploitation of data. We therefore propose our Data Value Assessment Tool *Saffron*; a customisable semantic-based tool that considers a number of data value dimensions to provide a comprehensive and context-aware data value quantification. *Saffron* connects to data governance centres to extract relevant metadata, uplifts it to a data value knowledge graph, and presents analysis and semantic driven traceable explanations of the calculated data value.

2 Saffron: The Data Value Assessment Tool

Our motivation for *Saffron* is to enable the optimisation of data value chains based on the quantification of data value. The tool therefore provides the capability of monitoring data assets as used within an enterprise, and uses the relevant metadata to calculate the value of the assets. Considering the lack of consensus on what characterises data value, we here designed *Saffron* to be extendible, and to calculate data value based on a number of different data value dimensions and the relevant metric groups and metrics as defined in [1]. We also take into consideration insight and feedback given by relevant stakeholders.

Figure 1 shows a diagram of the architecture of the *Saffron* tool. The latter enables its users to connect to one or more *data governance centres* through APIs. These centres include any methods used by an entity to manage their data, and the relevant metadata. *Saffron* is therefore able to extract the metadata on data assets as required.

In the Semantic Data Management component, *Saffron* uses the Data Value Vocabulary⁴ (DaVe) to construct a knowledge graph containing information such

⁴ <http://theme-e.adaptcentre.ie/dave/>

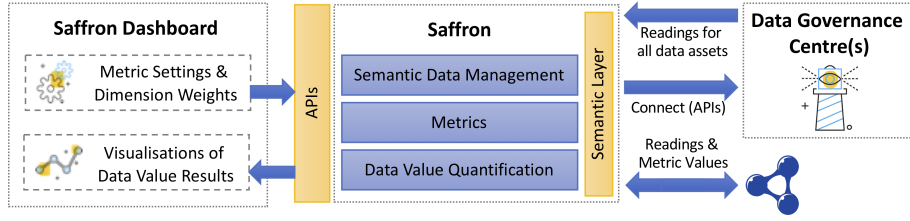


Fig. 1. Architecture of Saffron: The Data Value Assessment Tool

as the name of the data asset, its description, and other metadata required to calculate the implemented metrics. We refer to the latter as data asset *readings*.

As a proof of concept, we here implemented four different dimensions to characterise data value, namely *Infrastructure*, *Usage*, *Data*, and *Quality*. For each of these dimensions we implemented a number of metrics, totalling to eight metrics over the four dimensions. Table 1 provides an overview based on the hierarchy used in the DaVe vocabulary. Each of these metrics require one or more data asset readings. For example for the *Created By* metric we require the ID of the person who created the data asset. These readings are then used within the respective formulas of each metric to calculate the metric value. These results are added to the data asset knowledge graph and persisted to a triple store.

Table 1. Implemented Dimensions, Metric Groups, and Metrics

| Dimension | Metric Group | Metric |
|----------------|--------------|------------------------|
| Usage | User | Created By |
| | | Class of User |
| | Data | Last Modification Date |
| | | Created On |
| Quality | Intrinsic | Completeness |
| | | Accuracy |
| Data | Extrinsic | Trust |
| Infrastructure | Data | Data Management |

For the quantification of the data value of data asset, we take into consideration the metric values calculated as described above, as well as any *Metric Settings* and *Dimension Weights* specified by the user through the *Saffron Dashboard*. The metric settings are ‘assumptions’ required to cater for the subjective nature of data value. For example, one might consider an older data asset to be more valuable, but the opposite might also stand true. Therefore these settings are used in order to tailor the overall data value calculation according to the specific use context. Similarly, the dimension weights are used to cater for the contextual nature of data value, where one dimension might be considered to be relevant in one context, but less in another. For example, the usage dimension

would be considered less important than the quality dimension (particularly a timeliness metric) for weather forecast data. It is important to note that the metric calculations are not affected with the dimensions weights, and are therefore objective.

Through the Saffron Dashboard the user is able to access a number of interactive visualisations, including: (1) The overall data value of a project (consisting of a number of data assets); (2) The data value for specific assets, including a breakdown of the dimension values; (3) The metric values for specific assets; (4) The historic metric values for specific assets as they changed over time; and (5) The project dimensions weights' current settings.

In the Saffron Dashboard the user is also able to view an explanation of how the data value was calculated. This explanation is generated within the Semantic Data Management component, where asserted knowledge about the data asset (from the knowledge graph) and the user set weights are coupled with the terminology concepts about data value as defined in the DaVe vocabulary. This enables us to present the user with a concise explanation of *why* and *how* Saffron provided the given result as the data value of a data asset.

3 Conclusion

In this paper we presented the *Saffron*: Data Value Assessment Tool; the first tool that enables users to quantify the value of their data assets based on a number of dimensions. The tool is extendible and caters for the subjectivity and context dependence of data valuation through the use of weights and settings. Whilst still a proof of concept with a limited amount of implemented dimensions and metrics, the Saffron tool is already being validated and evaluated with stakeholders. Saffron is a concrete step towards quantifying the value of data assets and enabling their effective and efficient exploitation.

References

1. Attard., J., Brennan., R.: A semantic data value vocabulary supporting data value assessment and measurement integration. In: Proceedings of the 20th International Conference on Enterprise Information Systems - Volume 2: ICEIS., pp. 133–144. INSTICC, SciTePress (2018)
2. Klann, J.G., Schadow, G.: Modeling the Information-value Decay of Medical Problems for Problem List Maintenance. In: Proceedings of the 1st ACM International Health Informatics Symposium. pp. 371–375. IHI '10, ACM, New York, NY, USA (2010)
3. al Saffar, S., Heileman, G.L.: Semantic Impact Graphs for Information Valuation. In: Proceedings of the Eighth ACM Symposium on Document Engineering. pp. 209–212. DocEng '08, ACM, New York, NY, USA (2008), event-place: Sao Paulo, Brazil
4. Ying, Chen: Information Valuation for Information Lifecycle Management. In: Second International Conference on Autonomic Computing (ICAC'05). pp. 135–146 (Jun 2005)