A Semantic Data Value Vocabulary Supporting Data Value Assessment and Measurement Integration

Judie Attard[1] and Rob Brennan[1]
[1]*KDEG, ADAPT Centre, School of Computer Science and Statistics, OReilly Institute, Trinity College Dublin, Dublin 2, Ireland*
{*rob.brennan, attardj*}*@cs.tcd.ie*

Keywords: Data Value, Data Value Chains, Ontology, Linked Data, Data Governance

Abstract: In this paper we define DaVe: a data value vocabulary that allows for the comprehensive representation of data value. This vocabulary enables users to extend it using data value dimensions as required in the context at hand. DaVe caters for the lack of consensus on what characterises data value, and also how to model it. This vocabulary will allow users to monitor and asses data value throughout any value creating or data exploitation efforts, therefore laying the basis for effective management of value and efficient value exploitation. It also allows for the integration of diverse metrics that span many data value dimensions and which most likely pertain to a range of different tools in different formats. This data value vocabulary is based on requirements extracted from a number of value assessment use cases extracted from literature, and is evaluated using Gruber's ontology design criteria, and by instantiating it in a deployment case study.

1 INTRODUCTION

Data has become an essential part of products and services throughout all sectors of society. All data has social and commercial value (Attard et al., 2017), based on the impact of its use in different dimensions, including commercial, technical, societal, financial, and political. Despite the growing literature on data as an asset and data exploitation, there is little work on how to directly assess or quantify the value of specific datasets held or used by an organisation within an information system. For example, existing literature on data value chains simply describe processes that create value on a data product, however they do not actually discuss how to measure or quantify the value of data. Without assessment, effective management of value and hence efficient exploitation is highly unlikely (Brennan et al., 2018). Data value assessment involves the monitoring of the dimensions that characterise data value within a data value chain, such as data quality, usage of data, and cost. In real-world information systems this involves integration of metrics and measures from many sources, for example; log analysis, data quality management systems, and business functions such as accounting.

This value assessment and integration task is further exacerbated by the lack of consensus on the definition of data value itself. Part of this is due to the complex, multi-dimensional nature of value, as well as the importance of the context of use when estimating value. This indicates the need for terminological unification and building a common understanding of the domain, both for practitioners and for integrating the results of value assessment tools. Some variety of term definitions are due to the interdisciplinary nature of this field. However, current data value models, dynamics, and methods of categorisation or comparison, are also highly heterogeneous. These differences stem not only from the different domains of study, but also the diverse motivations for measuring the value of data (i.e. information valuation). Examples of these purposes include; ranking of results for question answering systems (Al-Saffar and Heileman, 2008), information life cycle management (Chen, 2005; Jin et al., 2008), security risk assessment (Sajko et al., 2006), and problem-list maintenance (Klann and Schadow, 2010).

The aim of this paper is to answer the following research question:

"To what extent can Data Value be modelled to act as basis for data value assessment and measurement integration?"

By studying this question we aim to gain insight into data value and data value metrics, provide a common models for exchange of data value metadata and enable the creation of data value assessment frameworks

or toolchains built on many individual tools that assess specific value dimensions. In this paper we therefore define the Data Value Vocabulary (DaVe); a vocabulary that enables the comprehensive representation of data value in an information system, and the measurement techniques used to derive it. The Data Value Vocabulary is expressed as Linked Data so that tools or dataset owners can easily publish and exchange data value metadata describing their dataset assets. In order to ensure interoperability of the vocabulary, we reuse concepts from existing W3C standard vocabularies (DCAT (Maali et al., 2014) and DataCube (Cyganiak et al., 2014)). Moreover, in order to cater for this rapidly evolving research area, and also for the extensive variety of possible contexts for information valuation, we designed DaVe to allow users to extend the vocabulary as required. This will allow users to include metrics and data value dimensions as needed, whilst also keeping the defined structure. In this paper we also gather together a set of data value assessment use cases derived from literature, and provide evaluation of the model through a structured evaluation of the ontology under Gruber's ontology design criteria, as well as through an example instantiation of the data value model in a deployment case study.

The rest of this paper is structured as follows: Section 2 describes a set of use cases for data value assessment metadata and derives common requirements, Section 3 discusses related work with respect to the requirements, Section 4 presents the data value vocabulary (DaVe) and documents our design process, Section 5 evaluates the vocabulary with respect to objective criteria for knowledge sharing and through a case study, and finally Section 6 presents our conclusions.

2 USE CASES

In this section we identify a set of use cases that illustrate scenarios where the data value vocabulary can be applied. The information gathered from the use cases is then used to identify requirements for the vocabulary. In general, a use case will be described and will demonstrate some of the main challenges to be addressed by the data value model. According to the challenges, a set of requirements for the data value vocabulary are abstracted, usually as competency questions (Ren et al., 2014).

2.1 Data Value Monitoring

In Brennan et al. we identified the data value monitoring capability as a fundamental part of any control mechanism in an organisation or information system that seeks to maximise data value, and hence data-driven innovation (Brennan et al., 2018). Data monitoring focuses on assessing and reporting data value throughout the value chain by gathering metrics on datasets, the data infrastructure, data users, costs and operational processes, and it provides us with the following challenges:

- Integration of diverse metrics that span many data value dimensions and which most likely pertain to a range of different tools in different formats. The goal here is to be able to build unified views of value from many data sources.
- Intelligent methods for identification of the appropriate metric for a given data asset could be supported by a knowledge model of the available metrics, the tools available to collect them, and how metrics are related to differing value dimensions.
- Providing explanations about the context and measurement of a metric when reporting on data value assessment results, for example in data governance applications.
- Accommodating new metrics - since data value is a new domain and the scope of tools and metrics is evolving it is necessary to be able to define new metrics and relate them to specific data value dimensions.

A data value vocabulary will help with these tasks by providing a common vocabulary for data value metric metadata that could be used to annotate the results of diverse tools and thus support data integration. If the vocabulary identifies links between metrics and tools, it will be possible to query a knowledge base using the data value vocabulary in order to select appropriate tools. By encoding the context and metric definitions it would be possible to support users in interpreting metric measurements of data value.

2.2 Curating Data

In Attard et al. we identify curation as a role that stakeholders can undertake whilst participating within a data value network (Attard et al., 2017). Fundamentally data curation is still a labour-intensive process and often requires human input from expensive and time-poor domain experts (Francois et al., 2016). Hence optimisation of the data curation process by using data value estimates as a lens with which to focus

human effort is a possible application area. This has the following challenges:

- Monitoring data value in a curation environment (see above use case).
- Using data value estimates to identify which data value dimensions of a dataset are both scoring poorly and are suitable for remediation through data curation processes, e.g. increasing data quality.
- Enabling a data curator to identify which value dimensions for a dataset are relevant to a specific data value chain, and to incorporate them in a dataset description. This is to support targeting the most significant data value dimensions during the curating process and throughout the value chain.

2.3 Data Management Automation

Several authors have already applied data value metrics to drive automated data management processes such as file migration (Turczyk et al., 2007), data quality assessment (Even et al., 2010), and information lifecycle management (Chen, 2005). However all these initiatives represent discrete value-driven systems that use heterogeneous data value metrics and estimates for a single application or purpose. A more generalised application of data value-driven automation calls for integrated tool-chains of applications whereby the impacts or reports of one tool can be consumed by others in order to execute follow-on activities, such as dataset repair after value assessment. This use case has the following challenges:

- Existing tools contain diverse value metrics and lack a common representation semantics. This results in a challenge to enable diverse tools to be able to relate them to a coherent view of relevant value dimensions and value calculations.
- No common format to express data value metric thresholding or targets.
- Capture of the relationships between data value, data assets, dataset metadata, data quality metrics, and data quality engineering methods, tools and processes. This would enable the application of probabilistic or semantic reasoning to be applied to goal-setting, monitoring and control of the automated data management control loop.

2.4 Data Governance based on Data Value

According to Tallon, data governance must become a facilitator of value creation as well as managing risk (Tallon, 2013). However, organisations are fundamentally challenged to understand how big data can create value (Demirkan and Delen, 2013). This means that creating links between data assets and organisational value as a basis for data governance is the most direct way to map between corporate strategy and data operations. This is a multi-faceted problem though; access to information and its interpretation through analytics to extract insights is at the core of decision-making. But more importantly, big data governance could drive business model innovation (Davenport, 2014), i.e. the appropriate deployment of data to develop new products and services based on the data, or the exploitation of data to transform how key organisational functions operate. The challenges of this use case are as follows:

- Flexibly representing data value so that it can be related to other business domain models such as data assets, business goals, key employees, and organisational knowledge.
- Existing data value chains are not optimally executed, in part due to a lack of data value estimates.
- Supporting operational decision making processes by informing them of high relevance and high value data assets and organisational information channels or processes.
- Identification of value faults or issues within data value chains over time in order to initiate mitigating actions.
- Estimating data value for data aquisition decisions to ensure its utility and "worth" in a specific context.

2.5 Requirements for a Data Value Vocabulary

By examining the use cases and challenges described above we have established the following requirements for the data value vocabulary. Each requirement has been validated according to three criteria: (1) Is the requirement specifically relevant to data value representation and reasoning? (2) Does the requirement encourage reuse or publication of data value meta data as (enterprise) linked data? (3) Is the requirement testable? Only requirements meeting those three criteria have been included.

1. The vocabulary should be able to represent data value comprehensively through a common representation.
2. It must be possible to extend the vocabulary with new metrics and assign them to specific data quality dimensions;

3. Data value metrics should enable the association to a set of measurements that are distributed over time;
4. It should be possible to associate a data asset (dataset) to a set of documented, and, if available, standardised value metrics;
5. It must be possible to associate a metric with a specific tool or toolset that supports generation of that metric; and
6. It must be possible to define the meaning of data value in the context of a specific data asset in terms of a number of dimensions, metrics and metric groups.

In addition we adopt the general requirements for data vocabularies from the W3C Data on the Web Best Practices Use Cases and Requirements working group note[1] to guide us on vocabulary engineering requirements:

- Vocabularies should be clearly documented;
- Vocabularies should be shared in an open way;
- Existing reference vocabularies should be reused where possible; and
- Vocabularies should include versioning information.

3 RELATED WORK

Data value is recognised as a key issue in information systems management (Viscusi and Batini, 2014). Data value is not a new concept; it has been extensively explored in the context of data value chains (Lee and Yang, 2000; Crié and Micheaux, 2006; Peppard and Rylander, 2006; Miller and Mork, 2013; Latif et al., 2009). The rationale of these data value chains is to extract the value from data by modifying, processing and re-using it. Yet, to date, the literature on data value chains only provides varying sequences and/or descriptions of the processes required to create value on a data product. This makes it challenging for stakeholders to easily identify what **characterises** data value. Hence methods and metrics to measure it are still immature (Tallon, 2013).

The existing literature offers varying definitions of data value. For example, Jin et al. define the value of data as a commodity to be determined by its use-value (Jin et al., 2008), Al-Saffar and Heileman define information value to be a function of trust in the source, and the impact of a specific piece of information on its recipient (Al-Saffar and Heileman, 2008), whilst Castelfranchi identifies the value of knowledge to be derived from its use and utility, and also from its necessity and reliability (Castelfranchi, 2016).

Despite this lack in literature, formal methods for establishing the value of data or information (which are typically used interchangeably in the literature) have been studied at least since the 1950s in the field of information economics (or infonomics). Moody and Walsh define seven laws of information that explain its unique behaviour and relation to business value (Moody and Walsh, 1999). They highlight the importance of metadata, saying that "[f]or decision-making purposes just knowing the accuracy of information is just as important as the information being accurate". They also identify three methods of data valuation: utility, market price, and cost (of collection), and conclude that utility is in theory the best option, but yet impractical, and thus cost-based estimation is the most effective method.

Data value in literature is also depicted or modelled through different dimensions, matching the definition of data value that is being followed. Many of these dimensions overlap with data quality dimensions. For example, Ahituv suggests timeliness, contents, format, and cost (Ahituv, 1980), which clearly parallel modern research on data quality dimensions (Zaveri et al., 2015). This large variety of dimensions results in an equally large number of domain-specific models that singularly are not adequate to provide a domain-independent, comprehensive, and versatile view of data value. Other existing models, while representing a valid data value dimension, do not (yet) adequately model all aspects. For instance, the Dataset Usage Vocabulary (DUV) (Lóscio et al., 2016) fails to model usage statistics, such as number of users, frequency of use, etc. The W3C Dataset Quality Vocabulary (daQ) (Debattista et al., 2014) is relevant but is specialised for capturing data quality metrics rather than data value metrics. Since these may overlap it sets an important requirement for the data value vocabulary that its metric definitions are compatible with those of the data quality vocabulary. In fact, Otto has also recently argued that research efforts should be directed towards determining the functional relationship between the quality and the value of data (Otto, 2015).

To date, there has been no attempt to specify a formal data value knowledge model. Moreover, existing models cannot be considered for providing complete answers to the queries and scenarios as identified in the use cases in Section 2. However one advantage of adopting a linked data approach is that our model can be interlinked with existing W3C standard models of usage, quality and dataset descriptions to form

[1] https://www.w3.org/TR/dwbp-ucr/

a complete solution for use cases like data governance driven by data value.

4 DATA VALUE VOCABULARY - DaVe

In this section we use ontology engineering techniques and standard vocabularies in order to define a vocabulary that enables the comprehensive representation of data value. In turn, this will enable the quantification of data value in a concrete and standardised manner. The Data Value Vocabulary (DaVe) is a lightweight core vocabulary for enabling the representation of data value quantification results as linked data. This will allow stakeholders to easily re-use and manipulate data value metadata, whilst also representing information on the dataset in question in other suitable vocabularies such as the W3C DCAT vocabulary for metadata describing datasets.

4.1 Vocabulary Design

Data value is not only subjective, but also depends on the context where the data is being used. Due to this specific nature of data value, the definition of a generic data value vocabulary is quite challenging. In fact, varying contexts of use will require the quantification of different value dimensions, and therefore the use of the relevant metrics. In Figure 1, we present DaVe, an abstract metadata model that, through extending the vocabulary, enables a comprehensive representation of Data Value. This representation will also be fluid in that it will allow the use of custom data value dimensions that are relevant to the context in question, whilst also maintaining interoperability. For DaVe we follow the Architectural Ontology Design Pattern[2] which affects the overall shape of the ontology and aims to constrain how the ontology should look like. This pattern is shared with the Dataset Quality Vocabulary (daQ) for its structure, and thus increases interoperability between the vocabularies and easily allows reuse of data quality metrics as metrics for data value dimensions when deemed appropriate.

Essentially, the *DataValue* concept is the central concept within DaVe, and will contain all data value metadata. As shown in Figure 1, in DaVe, we distinguish between three layers of abstraction. A DataValue concept consists of a number of different *Dimensions*, which in turn contain a number of *MetricGroups*. Each Metric Group then has one or more *Metrics* that quantify the Dimension that is being assessed. This relationship is formalised as follows:

Definition 1.

$$V \subseteq D,$$
$$D \subset G,$$
$$G \subset M;$$

where V is the DataValue concept (`dave:DataValue`), $D = \{d_1, d_2, ..., d_x\}$ is the set of all possible data value dimensions (`dave:Dimension`), $G = \{g_1, g_2, ..., g_y\}$ is the set of all possible data value metric groups (`dave:MetricGroup`), $M = \{m_1, m_2, ..., m_z\}$ is the set of all possible data value metrics (`dave:Metric`), and $x, y, z \in \mathbb{N}$.

These three abstract classes are not intended to be used directly in a DataValue instance. Rather, they should be used as parent classes to define a more specific data value characterisation. We describe the abstract classes as follows:

- **`dave:Dimension`** - This represents the highest level of the characterisation of data value. A Dimension contains a number of data value Metric Groups. It is a subclass of `qb:DataSet`; the W3C Data Cube DataSet. This enables rich metadata to be attached describing both the structure of the data collected in this dimension, and conceptual descriptions of the dimensions through W3C Simple Knowledge Organisation System (SKOS) models[3].
- **`dave:MetricGroup`** - A metric group is the second level of characterisation of data value, and represents a group of metrics that are related to each other, e.g. by being a recognised set of independent proxies for a given data value dimension.
- **`dave:Metric`** - This is the smallest unit of characterisation of data value. This concept represents metrics that are heuristics designed to fit a specific assessment situation. The `dave:ValueMeasurement` class is used to represent an instance of an actual measurement of a data value analysis.

In DaVe we reuse two W3C standard vocabularies, namely the RDF Data Cube Vocabulary (Cyganiak et al., 2014), and the Data Catalog Vocabulary (DCAT) (Maali et al., 2014). The latter, through `dcat:Dataset`, has the purpose of identifying and describing the dataset which is analysed with the intention of measuring its value. On the other hand, the

[2] `http://ontologydesignpatterns.org/wiki/Category:ArchitecturalOP`

[3] `https://www.w3.org/2004/02/skos/`

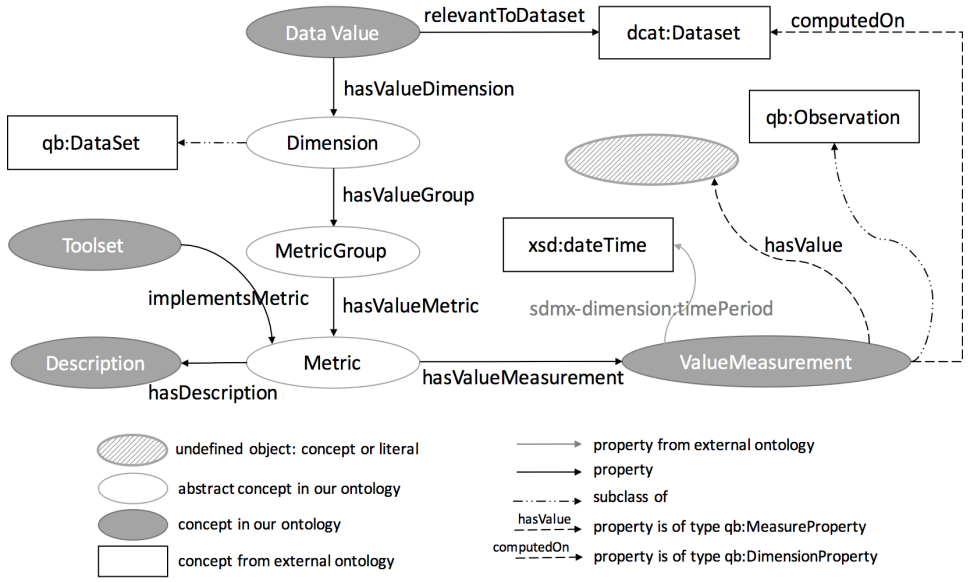

Figure 1: The Data Value Vocabulary - DaVe

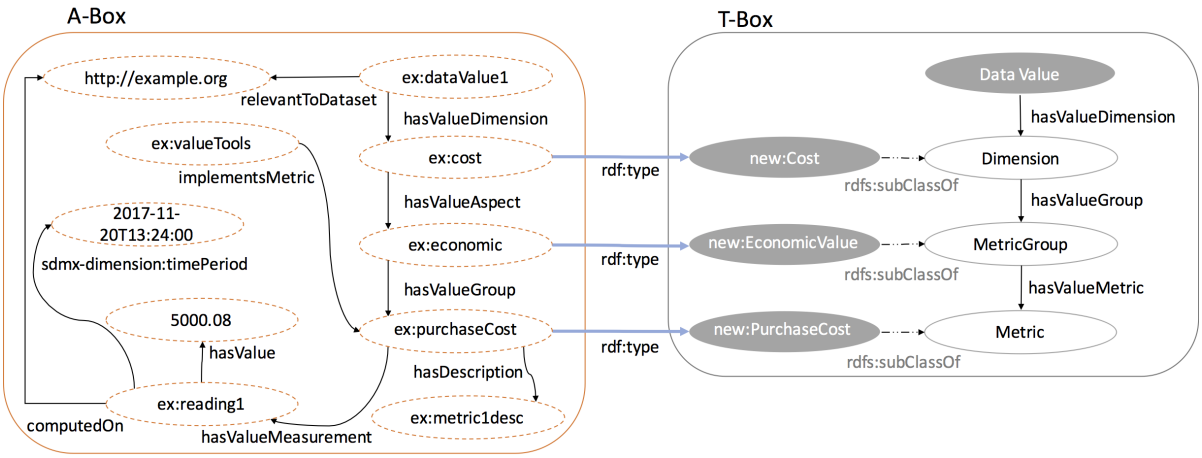

Figure 2: Extending DaVe - A-Box and T-Box

Data Cube Vocabulary enables us to represent data value metadata of a dataset as a collection of readings. This is essential to provide for the requirements as identified in the use cases in Section 2. Therefore, through the use of the Data Cube Vocabulary, users of DaVe will be able to:

- view all the metrics and their respective value measurements, grouped by dimension;
- view the various available value measurements for a specific metric (typically collected at different points in time as the dataset evolves);

We describe the remaining concepts within DaVe as follows:

- **`dave:ValueMeasurement`** - As a subclass of `qb:Observation`, this concept enables the representation of multiple readings of a single metric, as they occur, for example, on different points in time, or otherwise for different revisions of the same dataset. `dave:ValueMeasurement` also provides links to the dataset that the metric was computed on through the `dave:computedOn` property, a timestamp when the metric was computed through the `sdmx-dimension:timePeriod` property, and the resulting value of the metric through the `dave:hasValue` property. The latter value is multi-typed since results might vary amongst different types, including boolean, floating point numbers, integers, etc.

- **dave:Toolset** - This concept provides a link to a toolset or framework that provides functionality for a specific metric, therefore enabling users to easily identify the toolsets supporting the value metrics they require.
- **dave:Description** - This concept provides an overview of the metric and the context in which it is used.

4.2 Extending and Instantiating the Ontology

In order to comprehensively model data value, a user will need to extend the DaVe vocabulary with new data value measures that inherit the defined abstract concepts `dave:Dimension`, `dave:MetricGroup`, and `dave:Metric`. This will enable a user to represent data value in the specific domain at hand. Figure 2 portrays how DaVe can be extended with specific data value measures (T-Box). These measures can then be used to represent actual data value metadata (A-Box). In Figure 2 we extend DaVe with Cost as an example of the `dave:Dimension` concept, Economic Value as an example of `dave:MetricGroup`, and PurchaseCost as an example `dave:Metric`. According to LOD best practices, such extensions should not be included in DaVe's own namespace. For this reason we recommend users to extend DaVe in their own namespaces. In future work we plan to provide sample dimension and metric specifications using DaVe that will be refined via community feedback and serve as a catalog of examples that DaVe users can reuse directly or draw upon to build their own specifications.

5 EVALUATION

In this section we provide preliminary evaluation of the DaVE vocabulary in two ways; by leading out a structured analysis on the features of the ontology, and by applying the vocabulary to a use case in order to validate its usability and capability of modelling data value in context.

5.1 Design-Oriented Evaluation

Table 1 presents the evaluation of the DaVe vocabulary in accordance to the desired qualities expected from a well designed ontology. The methodology we use here follows the structured analysis approach laid out in (Solanki et al., 2016). We here define a number of generic and specific criteria, and evaluate our ontology according to how it fares with regard to these criteria.

We have also evaluated the ontologies in accordance to one of the most widely adapted, objective criteria for the design of ontologies for knowledge sharing; the principles proposed by Gruber (Gruber, 1995).

- **Clarity** - DaVe meets two of Gruber's three criteria for clarity in ontological definitions as follows:
  1. Conceptualisation in DaVe focuses solely on modelling the requirements for recording data value metric measurements and their grouping into data value dimensions, irrespective of the computational framework in which these will be implemented (Gruber's "independence from social and computational contexts");
  2. Definitions in DaVe (such as the definition of `dave:Metric`) have not been asserted in every case using necessary and sufficient conditions, due to the additional complexity this definition style places on the interpretation of the vocabulary (Gruber's recommendation of providing logical axioms); and
  3. Finally, DaVe has been very well documented with labels and comments (Gruber's requirement for natural language documentation).
- **Coherence** - There are two aspects to coherence according to Gruber:
  1. Definitions in an ontology must be logically consistent with the inferences that can be derived from it; and
  2. The logical axioms of the ontology and its natural language documentation should be consistent. DaVe has been checked using popular reasoners for logical consistency, although further work will have to be done on applications and field trials to explore the range of the inferences possible and to validate them. DaVe has been extensively documented using inline comments, labels and metadata using the LODE [6] documentation generation framework. This process ensures that ontology engineers working on DaVe can easily update the documentation when updating the vocabulary and that documentation generation is automatic and nearly instantaneous, which facilities validation and consistency checking.
- **Extendibility** - Gruber states that to ensure extendibility, a vocabulary should allow for monotonic extensions of the ontology. For DaVe we

[6] `http://www.essepuntato.it/lode`

Generic criteria	Evaluation
Value Addition	(1) The vocabulary adds data value specific metadata to the processes of data management / data governance / data value chain management, and enriches information about datasets to include data value metrics and their collection context. Tools can then use this context dependent information for automation and automatic generation purposes. (2) DaVe is used to provide details about the data value assessment process outcomes. (3) It links together related concepts in data value, data quality, data usage and data catalogs. (4) DaVe can also help inform governance decision-making or reasoning about data value dimensions, metrics and tools in a governance knowledge base, for example to enable metric selection or combination.
Reuse	(1) Potential reuse across a wider community of data producers, data value chain managers, dataset managers, ontology engineers of new or related vocabularies. (2) Potential users and uses of DaVe include developers of data profiling/assessment tools, data governance platforms, decision support systems and business intelligence systems. (3) The vocabulary is easy to reuse and published on the Web together with detailed documentation. It defines a general abstraction of value dimensions and metrics that can be extended for specific use cases or domains. Furthermore, the models are extendable and can be inherited by specialised domain ontologies for specific data governance platforms.
Design and Technical quality	(1) All ontologies have been designed as OWL DL ontologies, in accordance to ontology engineering principles (Noy and Mcguinness, 2001). (2) Axiomatisations in the ontologies have been defined based on the competency questions identified during requirements scoping. (3) The vocabulary has been validated by the OOPs! ontology pitfall scanner[4]. (4) The ontology contains descriptive, licensing, and versioning metadata.
Availability	The ontology has been made publicly available at `http://theme-e.adaptcentre.ie/dave`. Further, it has been given persistent w3id URIs, deployed on public facing servers and is content negotiable. The vocabulary is licensed under a Creative Commons Attribution License. DaVe has also been registered in LOV[5].
Sustainability	The ontology is deployed on a public Github repository. It is supported by the ADAPT Centre, a long-running Irish government funded research centre. Long term sustainability has been assured by the ontology engineers involved in the design.
Specific criteria	
Design suitability	The vocabulary has been developed in close association with the requirements emerging from potentially exploiting applications, as presented in the use cases section of this paper. Thus they closely conform to the suitability of the tasks for which they have been designed.
Design elegance and quality	Axiomatisation in the ontologies have been developed following Gruber's principles of clarity, coherence, extendability, minimum encoding bias, and minimum ontological commitment (Gruber, 1995). These ontologies are based on the ADAPT Centre's past history of vocabulary standards development with the W3C.
Logical correctness	The ontologies have been verified using DL reasoners for satisfiability, incoherency and inconsistencies. The OOPs! model checker has been deployed to validate the ontologies.
External resources reuse	Concepts from external ontologies such as W3C's Data Cube and the DCAT vocabulary have been used in DaVe. Moreover, other ontologies such as the Data Quality Ontology daQ and the Dataset Usage Vocabulary DUV can be used in instances of the ontology as required by the user and the context of data use.
Documentation	The vocabulary have been well documented using rdfs:label, rdfs:comment and author metadata. HTML documentation via the LODE service (`http://www.essepuntato.it/lode`) has also been enabled. All ontologies have been graphically illustrated. This paper also documents the vocabulary, its use cases and provides example instances.

Table 1: Evaluating the DaVe Vocabulary

have reused the structural pattern of the Data Quality ontology (DaQ), where we define an abstract metric framework designed to be extended with new data value concepts as required, whilst still maintaining the defined structure and existing definitions.

- **Minimal encoding bias** - For wider adoption of the ontology, Gruber states that the ontology should use a conceptualisation mechanism that minimises the dependencies on encoding formats. DaVe has been formalised in OWL 2, which is a W3C standard for representing ontologies on the Web. It has its foundations in Description Logics. Multiple serialisation formats are available for the ontology. The axiomatisation in DaVe is therefore accessible to all tools and frameworks that support these serialisations. There are limits to the expressivity of OWL (Grau et al., 2008) and it has modelling quirks that impact on any conceptualisations it captures, but nonetheless it has been designed specifically for knowledge capture and to minimise the impact on models.
- **Minimum ontological commitment** - Gruber's final test requires that an ontology should only make assertions that require only a minimum commitment from implementing agents, providing them the flexibility to extend and enrich the ontology, albeit in a monotonic way. DaVe meets this criteria in at least two ways:
  1. It minimises the number of imported ontologies. Each imported ontology or referenced term has been assessed for the impact it has on the overall model and incomplete, inconsistent or overly wide ontologies have not been included.
  2. Rather than providing a static model of the data value domain based on our current understanding, DaVe provides a framework of value dimensions, metrics and measurements with their relationships which is designed to be extended to incorporate new metrics, dimensions and tools.

5.2 Use Case Driven Evaluation

In this section we describe a deployment scenario for DaVe in MyVolts Ltd.[7]; an Irish data-driven online retailer, that wishes to assess data value to drive internal business process optimisation.

MyVolts is a successful SME with a 15 year track record that develops and operates a highly automated internet retail and business intelligence system. They have served over 1 million customers and are a leading source for consumer device power supplies in the markets where they operate: the USA, Ireland, the UK, France, and Germany. In addition to importing and designing standard power supplies, MyVolts has its own power products. MyVolts collect, manage and analyse data on their customers, the evolving market of power supply device specifications, and the power supply needs of all consumer electronics. This involves monitoring social media, web sales data such as Amazon top seller lists, customer queries and complaints, and device manufacturer homepages. New consumer electronic devices must be discovered, categorised, profiled for potential sales value and have their power supply technical specifications (voltage, polarity, tip type and dimensions) mined from open web data. There are an estimated 5.5 million consumer electronics devices on sale today and the number of powered devices is growing rapidly. The lack of standardised machine-readable repositories means that PDF is the dominant data publication format. Integrating this data while maintaining strict quality control is a major challenge for MyVolts' semi-automated data collection system (which may be modelled as a data value chain).

Our aim here is to identify how to model data value in this context in order to optimise this data value chain. This requires five specific steps:

1. Identify data value as it occurs within the value chain (data value creation/consumption);
2. Identify the data value dimensions that are relevant in this context;
3. Model data value using DaVe;
4. Implement model and metrics to quantify data value; and
5. Adapt data value chain accordingly.

In Figure 3 we portray an example of a data value chain within MyVolts that shows various value creating processes as well as decision-making processes. Through this figure we can identify the following as relevant data value dimensions (not exhaustive):

- Quality - Data must be accurate, timely, accessible, complete, etc.
- Cost - Data must have manageable costs, including production, maintenance, or purchasing costs.
- Usage - Data with more uses (actual or planned) will be more valuable to MyVolts, as it will have more impacts on the data value chain.

Based on these data value dimensions, in Figure 4 we provide a T-Box example using DaVe for the MyVolts data value chain scenario. Once this model is

[7] http://myvolts.com/

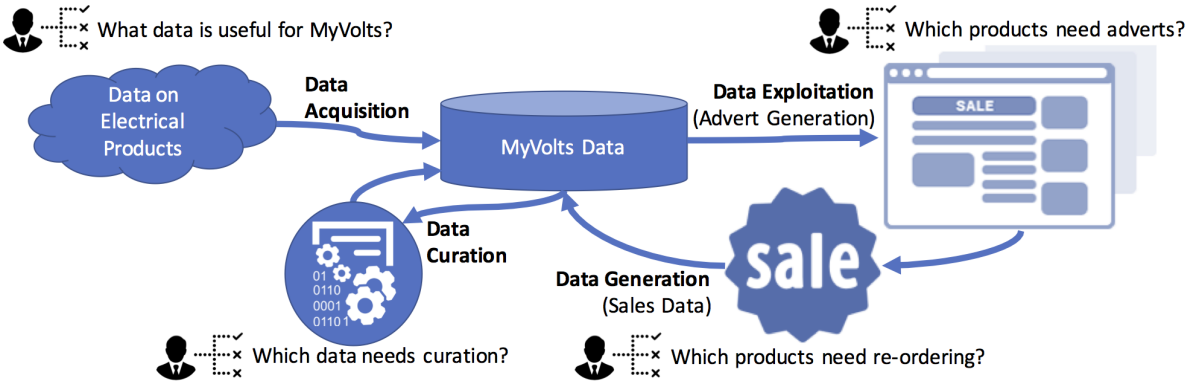

Figure 3: MyVolts Data Value Chain Example

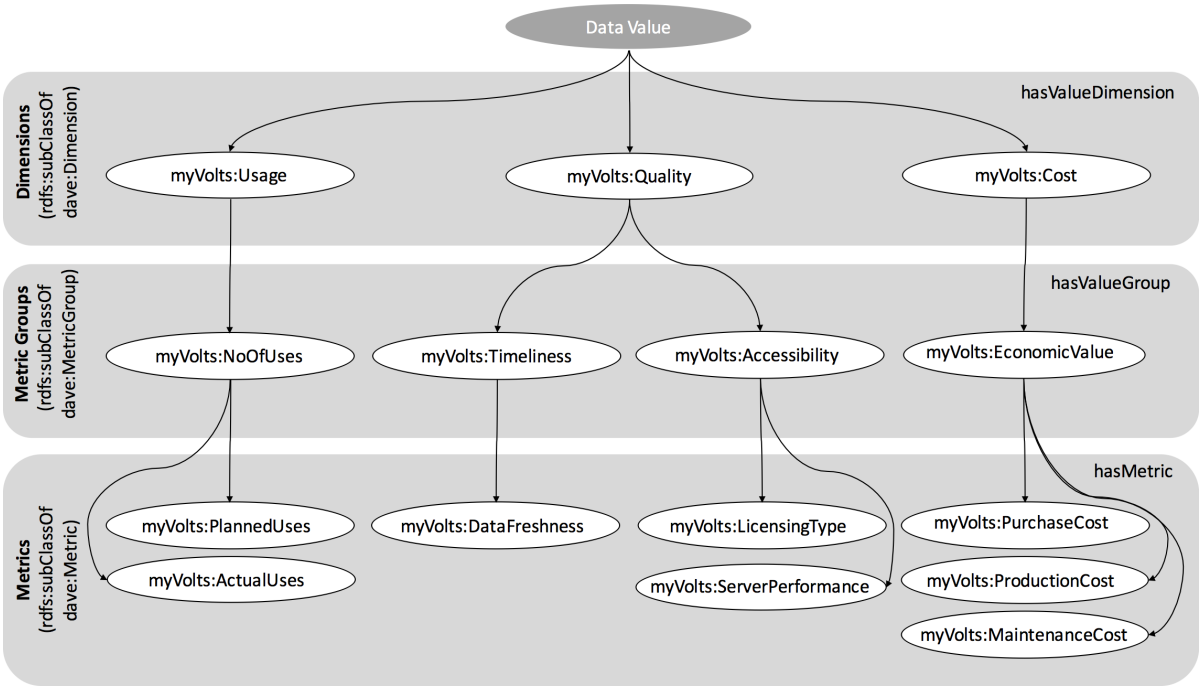

Figure 4: MyVolts Data Value Model based on DaVe

applied and the data value is quantified using the relevant metrics, a stakeholder from MyVolts can then analyse how to exploit this data value monitoring information in order to optimise their data value chain. For instance, the data acquisition process can be optimised by first analysing the quality of the data to be acquired, and also its purchasing cost. This will ensure that an optimal decision is made when acquiring the data, and that the data will provide maximal benefits for its intended use. It also demonstrates the ease of definition of a consistent schema for all assessment tools to have their data uplifted. For example, the R2RML mapping language can be used to map usage data stored in a relational database into a semantic format using DaVe's structure, which will allow for easy integration and unified querying.

Through this use case driven evaluation we have a preliminary validation of the DaVe vocabulary. We demonstrate its flexibility in enabling the comprehensive modelling of data value, as well as its potential impact on data exploitation.

6 CONCLUSION

Data is increasingly being considered as an asset with social and commercial value. The exploitation of data is ongoing in many dimensions of society, and data value has been extensively explored in the context of data value chains. Yet, due both to the multi-

dimensionality of data value and to the relevance of context in quantifying it, there is no consensus of what characterises data value or how to model it.

In this paper we identify a set of use cases with the aim of illustrating scenarios where a data value model can be applied. From these use cases we also extract a number of requirements that such a vocabulary should cater for. We therefore define DaVe; a lightweight data value vocabulary that enables the representation of data value quantification results as linked data. This vocabulary can be extended with custom data value dimensions that characterise data value in a specific context. It also allows for the integration of diverse metrics that span many data value dimensions and which most likely pertain to a range of different tools in different formats. We lead out a preliminary evaluation by (1) leading out a structured analysis on the features of the ontology, and (2) by applying the vocabulary to a use case to validate its usability and capability of modelling data value in context.

By enabling the comprehensive representation of data value, DaVe allows users to monitor and assess the value of data as it occurs within any data value chain, as data is being exploited. This will in turn enable the effective management of value, and hence efficient exploitation of data.

ACKNOWLEDGEMENTS

This research has received funding from the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund. If any, should be placed before the references section without numbering. To do so please use the following command: *\section*{ACKNOWLEDGEMENTS}*

REFERENCES

Ahituv, N. (1980). A Systematic Approach toward Assessing the Value of an Information System. *MIS Quarterly*, 4(4):61.

Al-Saffar, S. and Heileman, G. L. (2008). Semantic Impact Graphs for Information Valuation. In *Proceedings of the Eighth ACM Symposium on Document Engineering*, DocEng '08, pages 209–212, New York, NY, USA. ACM.

Attard, J., Orlandi, F., and Auer, S. (2017). Exploiting the value of data through data value networks. In *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance ICEGOV'17*, pages 475–484.

Brennan, R., Attard, J., and Helfert, M. (2018). Management of data value chains, a value monitoring capability maturity model. In *Proceedings of 20th International Conference on Enterprise Information Systems (ICEIS 2018). Under Review*.

Castelfranchi, C. (2016). In search of a principled theory of the value' of knowledge. *SpringerPlus*, 5(1):1617.

Chen, Y. (2005). Information Valuation for Information Lifecycle Management. In *Second International Conference on Autonomic Computing (ICAC'05)*, pages 135–146. IEEE.

Crié, D. and Micheaux, A. (2006). From customer data to value: What is lacking in the information chain? *Journal of Database Marketing & Customer Strategy Management*, 13(4):282–299.

Cyganiak, R., Reynolds, D., and Tennison, J. (2014). The rdf data cube vocabulary. W3c recommendation, World Wide Web Consortium (W3C).

Davenport, T. H. (2014). How strategists use big data to support internal business decisions, discovery and production. *Strategy & Leadership*, 42(4):45–50.

Debattista, J., Lange, C., and Auer, S. (2014). Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems - SEM '14*, pages 92–99, New York, New York, USA. ACM Press.

Demirkan, H. and Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, 55(1):412 – 421.

Even, A., Shankaranarayanan, G., and Berger, P. D. (2010). Evaluating a model for cost-effective data quality management in a real-world CRM setting.

Francois, P., Manning, J. G., Whitehouse, H., Brennan, R., Currie, T., Feeney, K., and Turchin, P. (2016). A Macroscope for Global History. Seshat Global History Databank: a methodological overview. *Digital Humanities Quarterly*, 10(4).

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. (2008). Owl 2: The next step for owl. *Web Semant.*, 6(4):309–322.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928.

Jin, H., Xiong, M., and Wu, S. (2008). Information Value Evaluation Model for ILM. In *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, pages 543–548. IEEE.

Klann, J. G. and Schadow, G. (2010). Modeling the Information-value Decay of Medical Problems for Problem List Maintenance. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 371–375, New York, NY, USA. ACM.

Latif, A., Us Saeed, A., Hoefler, P., Stocker, A., and Wagner, C. (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In *Proceedings of International Conference on Semantic Systems*, pages 568–576.

Lee, C. C. and Yang, J. (2000). Knowledge value chain. *Journal of Management Development*, 19(9):783–794.

Lóscio, B. F., Stephan, E. G., and Purohit, S. (2016). Data usage vocabulary (duv). Technical report, World Wide Web Consortium.

Maali, F., Erickson, J., and Archer, P. (2014). Data catalog vocabulary (dcat). W3c recommendation, World Wide Web Consortium.

Miller, H. G. and Mork, P. (2013). From Data to Decisions: A Value Chain for Big Data. *IT Professional*, 15(1):57–59.

Moody, D. and Walsh, P. (1999). Measuring The Value Of Information: An Asset Valuation Approach. *Seventh European Conference on Information Systems (ECIS'99)*, pages 1–17.

Noy, N. F. and Mcguinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical report.

Otto, B. (2015). Quality and Value of the Data Resource in Large Enterprises. *Information Systems Management*, 32(3):234–251.

Peppard, J. and Rylander, A. (2006). From Value Chain to Value Network:. *European Management Journal*, 24(2-3):128–141.

Ren, Y., Parvizi, A., Mellish, C., Pan, J. Z., van Deemter, K., and Stevens, R. (2014). *Towards Competency Question-Driven Ontology Authoring*, pages 752–767. Springer International Publishing, Cham.

Sajko, M., Rabuzin, K., and Ba??a, M. (2006). How to calculate information value for effective security risk assessment. *Journal of Information and Organizational Sciences*, 30(2):263–278.

Solanki, M., Božić, B., Freudenberg, M., Kontokostas, D., Dirschl, C., and Brennan, R. (2016). *Enabling Combined Software and Data Engineering at Web-Scale: The ALIGNED Suite of Ontologies*, pages 195–203. Springer International Publishing, Cham.

Tallon, P. P. (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, 46(6):32–38.

Turczyk, L. A., Heckmann, O., and Steinmetz, R. (2007). File Valuation in Information Lifecycle Management. *Managing Worldwide Operations & Communications with Information Technology*, pages 347–351.

Viscusi, G. and Batini, C. (2014). Digital Information Asset Evaluation: Characteristics and Dimensions. pages 77–86. Springer, Cham.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2015). Quality assessment for linked data: A survey. *Semantic Web Journal*, 7.