Journal of the International Society for the Systems Sciences | 67th Meeting of the International Society for the Systems Sciences
J.M. Wilby, Editor.

FRAMING SEMANTIC DATA WAREHOUSING FROM A SYSTEMS PERSPECTIVE

Susan Campher
Roelien Goede

Abstract

The challenge associated with data warehousing has escalated in the era of big data with masses of fast-moving heterogeneous data sources. As organisations attempt to exploit an ever-growing complex and dynamic datasphere, traditional data warehousing practices seem to produce systems that are inflexible and unable to scale. In a dynamic world of flux and change, systemic process thinking provides an alternative paradigm from which to approach the data warehousing challenge. This paper provides a framing of semantic data warehousing from such a systems perspective. Semantic data warehousing involves data semantification – enriching data with its context and meaning – to achieve higher levels of automation and adaptability. The framing elucidates the inherent systems approach of incorporating semantic technologies and automated dimensionalisation in data warehousing. It provides a case for the data management community to appreciate and accept complexity and multiple perspectives, and to incorporate systemic process thinking and semantic relativism into data management practices.

Keywords

Data warehousing, business intelligence, data management, automation, dimensional modelling, systems approach, systemic process thinking, semantic relativism, Semantic Web, ontologies.

1 | Introduction

Data warehousing comprises methods and practices for collecting, integrating, managing, interpreting, and using data in support of data analytics and business intelligence. Managing the content of a *data warehouse* (DW) has always been challenging, but has become more so in the era of big data. First, DW content management is a complex function since enterprise information systems that constitute the main sources of data are designed for specific purposes (or *use cases*) and are semantically and structurally heterogeneous. Secondly, data architectures have become highly complex and dynamic with large volumes and a variety of data moving at high speeds through the datasphere. Manual development of DW content management services cannot scale with big data and traditional data warehousing practices are not adaptable enough. Addressing the problem requires a systematic and holistic approach (Koehler et al., 2017).

Current research by the authors involves investigating the use of enhanced data semantics for intelligent data warehousing, specifically in support of automated content management. The use of Semantic Web technologies is considered since they improve the level of automated processing of information on the Web without the need for sophisticated artificial intelligence (Berners-Lee, 2009; Bizer, Heath, & Berners-Lee, 2009). It involves semantification of data – enriching data with metadata about its context and meaning – so that machines and humans may collaborate effectively within a large and complex digital information space, or *datasphere*.

Since there are many different perspectives from which the same enterprise information may be viewed as well as various analytical requirements that would not be known in advance, determining the required semantic metadata and a suitable data model for semantic data warehousing proves to be a difficult problem. One could argue that it is a *wicked problem* as described by Rittel and Webber (1973) because (among other characteristics) solution objectives are not clear and it is impossible to know the full consequences of a particular implementation. Even the overarching problem situation – of how to better manage big data for meaningful and effective use – is defined in terms of a possible resolution or action towards improvement, i.e., to semantify data for intelligent machine processing of information. Such a proposed action came about because of a particular explanation of the problem: a lack of suitable data semantics in data management practices. However, fundamental concepts of systems approaches, including *systemic process thinking*, are considered appropriate to provide a frame from which the problem situation may be approached.

This paper contains an exposition of a systems approach for semantic data warehousing. The aim is to elucidate the systems concepts behind a proposed resolution to the data warehouse content management challenge. Section 2 provides background to the problem of content management in data warehousing in the era of big data. Section 3 contains an overview of several fundamental concepts of a systems approach. In Section 4, these ideas are related to the philosophy of information, conceptual modelling, a dimensional view of data, and semantic data warehousing using ontologies. In conclusion, Section 5 provides a summary of this systems framing for semantic data warehousing.

2 | Background

2.1 | Designing and Modelling in Data Warehouses

Traditionally, up to 70% of the DW development effort is attributed to designing and developing the *extract, transform, and load* (ETL) subsystem of a DW (Kimball, Ross, Thornthwaite, Mundy, & Becker, 2008). The purpose of the ETL subsystem is to gather, integrate, and consolidate data and loading it into a DW. The purpose of the DW itself is to present enterprise-wide information to users and analytics applications for business intelligence and decision making. The overall system is therefore often referred to as the *data warehouse/business intelligence* (DW/BI) system.

The information in a DW is typically presented in multidimensional schemas (or dimensional models). Dimensional models are popular because they provide a predictable, standard framework for analytics, are business-focused, intuitive and easy to understand, provide fast query performance, and are flexible to analytic needs and unexpected changes (Kimball, 1997). Data modelling is performed when DWs are designed. The modelling process involves analysing business processes and the associated data sources to understand both the information (or analytics) requirements and the available data. The available data may be in a variety of formats, including structured and unstructured data. Enterprise information system experts must be consulted to gain understanding of the conceptualisations within the source data systems, their associated data models, and the meaning of the data within those systems. Once the source data is understood, consensus must be reached on how the data will be consolidated, and transformation processes or mappings are subsequently developed to present data to users and analytics applications in multidimensional schemas.

This modelling process, as well as the ETL system development, is tedious and requires continuous revision as business processes and data sources evolve. However, Kimball and Ross (2013) consider automated dimensionalisation the best approach to integrate heterogeneous data for meaningful use in big data environments, but it would require relevant semantic metadata and a suitable data model.

2.2 | The Complexity Challenge

The idea of *VUCA*, short for volatility, uncertainty, complexity, and ambiguity, is relevant to the world of big data and affects the development and evolution of analytics environments like DW/BI systems. For example, the concept of a *data lake* as an architectural approach to big data analytics has gained popularity with affordable and scalable cloud storage. However, it was not long before the term *data swamp* was used to refer to these large repositories of raw data that became difficult to maintain and use without contextual metadata and proper data governance (Halper & Kobielus, 2021; Rao, 2018). Since larger, hybrid data architectures lack application of traditional (relational) database principles, data integrity, consistency, and clear meaning are compromised (Helland, 2011). Rao (2018) notes that developing the data lake concept must be a process of continuous evolution, and areas of further development include architectural standardisation, interoperability, and proper data governance and curation.

Other emerging approaches to enterprise data management include the *data fabric* and the *data mesh*. A data fabric incorporates active metadata, knowledge graphs, semantics, and machine learning to support augmented data integration and services. In a data mesh, data are organised into domains according to usage patterns, and the domains are then contextualised with business context descriptors. However, any attempt to provide a universal platform for data analytics must incorporate a wide variety of technologies with some level of standardisation.

2.3 | The Challenge of Standardisation

Because of rapidly developing data technologies and unforeseeable data use cases, attempting standardisation in data analytics environments remains a challenge. In practise, standardisation for data interoperability suffers from adoption hesitance as well as a lack of semantic standardisation. Although many data interchange standards exist (e.g., CSV, XML, and JSON) and many applications support them, interoperability remains limited since the focus is on structural interoperability, not semantic interoperability. Efforts toward semantic standardisation often fail due to insufficient feedback from practical use into the standards development process (Drath & Barth, 2012). A deadlock situation occurs when adoption of a standard is postponed until it reaches higher maturity while widespread application with feedback is required to reach that level of maturity.

2.4 | Semantic Web Technologies and Ontologies

The Semantic Web discipline is considered rich in knowledge about effective data management. Mainstream adoption in industry is evident with a wide range of applications, such as those using knowledge graphs and Industry 4.0 technologies (Hitzler, 2021). The idea of the Semantic Web involves distributing self-describing data via the Web for automated information processing that is meaningful but not dependent on artificial intelligence (Berners-Lee, Hendler, & Lassila, 2001). Initially, the Web developed mostly for human use of information distributed in document format, and *weaving* the Semantic Web involves *semantic markup* of an information source with metadata describing the meaning of its content. The metadata is generally expressed using *controlled vocabularies* or *ontologies*.

The concepts of *linked data* and the *web of data* are closely associated with the Semantic Web. Linked data provides *connections* for information discovery and exploration by humans and computers: ‘With linked data, when you have some of it, you can find other, related, data’ (Berners-Lee, 2009, para. 1). The web of data is considered an additional layer on top of the general architecture of the Web, a global data graph built using linked data principles and technologies such as the *Resource Description Framework* (RDF) and RDF ontologies (Bizer et al., 2009). The idea of *openness* underpins development of the web of data. Openness in this context refers to anyone publishing any type of data without constraints on the ontologies used, complexity, or possible links between data as well as being freely available in non-proprietary formats (Bizer et al., 2009).

3 | A Systems Approach

Although what constitutes a systems approach depends on a social dynamic involving a connection between systems practitioners and the history of systems thinking, there are generally accepted concepts that provide a common grounding (Ison, 2017). A systems approach involves *systemic thinking*: ‘the understanding of a phenomenon within the context of a larger whole’ (Ison, 2017, p. 24), or ‘looking at each component part in terms of the role it plays in the larger system’ (Churchman, 1979, p. 103). Three features (or purposeful orientations) of systemic thinking are: understanding of *interrelationships*, commitment to *multiple perspectives,* and awareness of *boundaries* (Williams & Hummelbrunner, 2010). Surfacing these elements brings understanding to the whole. Systemic thinking is particularly helpful when considering complex problems or situations. A *systemic inquiry* is regarded as a systems approach that is adaptive to change since it purposefully acknowledges and addresses uncertainty and facilitates social learning within an action-oriented cyclic process (Ison, 2017).

A *system* is an integrated whole, distinguished by an observer in a situation for having a certain *boundary* and *purpose*. The boundary determines what is relevant in the context of a particular systemic inquiry (Williams & Hummelbrunner, 2010) and the purpose describes ‘what the system does or exists for from the perspective of the observer’ (Ison, 2017). Churchman (1979) describes a system as ‘a *set of parts* coordinated to accomplish a *set of goals*' (p. 29) and presents five elements to consider when thinking about a system: (1) its overall objectives – relating to its true purpose; (2) its components (parts or subsystems), their activities, objectives and performance measures; (3) its environment – what is *outside* of the system relative to its objectives and control; (4) its resources – everything *inside* the system, used to perform its activities; and (5) management of the system. These elements are not trivial to determine and must be continuously revisited.

A system is furthermore regarded as a *process,* rather than a thing (Hammond, 2017). Asby (2023) argues that this process view is 'fundamental to developing models and understanding of the stability we find in our world' (p. 131). It differs from the traditional (Western) analytic view of the world focused on objects and their attributes detached from their context and relations to their environment. In Asby's *systemic process thinking*, a learning process is essential in a dynamic world, and *out there* is regarded as 'a network of interacting systems that I perceive, but models are developed to the point that systems can be recognised from multiple perspectives and now thought of as objects' (Asby, 2023, p. 144). Object thinking thus emerges from process thinking and modelling.

Common fundamental concepts encountered in popular systems approaches discussed by Reynolds and Holwell (2020) are as follows: *emergence* – 'the property of a "whole" that arises from the interaction of the parts' (p. 308); *hierarchy* – referring to levels or layers within the whole; *communication* – including development of mutual understandings as well as the exchange of information; and *control* – relating to corrective actions. The concept of hierarchy relates to a multilayered structure of systems within systems – a *holarchy* (Hammond, 2017). Communication also involves *feedback,* an interconnection or circularity commonly found in systems, so that systemic awareness is also associated with understanding *cycles*, *counterintuitive effects,* and *unintended consequences* in systems or situations (Ison, 2017).

An *iterative* process to address a complex situation or problem provides an opportunity *to learn*. Each attempt at resolution results in better understanding of a situation: new insights are gained and often involve *sweeping in* other perspectives or aspects of the larger system not considered before (Churchman, 1982; Ulrich, 1994). Meaningful learning requires continuous and critical reflection by the practitioner, but also interaction with a wider community of practice, systems, and beyond (Asby, 2021; Reynolds & Holwell, 2020). A systems approach acknowledges that there is no ultimate solution to be found. In a dynamic environment, it is a matter of evolution: 'the processes of feedback and learning play a much more pivotal role, and the reductionist paradigm [of traditional science] becomes increasingly inadequate' (Hammond, 2017, p. 4).

In a systems approach, the use of models (and diagrams) facilitates the learning process, and the models remain conceptual constructs not to be confused with either representations of actual reality or the end-goal of the exercise (Reynolds & Holwell, 2020). A frequent analogy used to describe this notion is 'the map is not the territory', whether the map refers to a data model (Kent & Hoberman, 2012, p. 22) or any other type of conceptual model (Bawden, 2010, p. 53). Models are developed from a particular perspective and for a specific (subjective) purpose that determines the aspects of the real world that are represented (Asby, 2021).

4 | A Systems Framing for Semantic Data Warehousing

4.1 | Philosophy of Information when Modelling Databases

In 1975, the ANSI/SPARC Study Group on Database Management Systems published a report on their three-level approach to modelling databases. They introduced a framework with a *conceptual information schema* as a bridge between internal schemas (database-level representations) and external schemas (user or application views of the information): 'The framework is designed to support change and evolution in an enterprise by providing a suitable basis for generating and controlling mappings between internal and external schemas' (Bachman et al., 1975, p. 1). The incentive for the three-level approach was achieving data independence and the foreseen benefits of the conceptual schema were (1) to provide a relatively stable and formalised central frame of reference of an enterprise's information model that (2) serves as a starting point for a data dictionary describing information assets and how they are used, (3) serves as a central authority for security and integrity management, (4) reduces the number of mappings between internal and external schemas, and (5) acts as an insulator between changes in internal and external schemas.

In Chapter 5 of their report, the study group lays out their notion of how data models are devised. Three realms of interest in the philosophy of information are depicted (Exhibit 1), *reality*, the *conceptual realm*, and the *data realm(s)*. The conceptual realm entails how reality is perceived and converted into mental and symbolic models. Models are continuously refined through scientific processes and abstraction. The *best* model of reality relates to our total conceptualisation of reality. Through engineering abstractions, limited models are developed and formally described for a particular purpose (relating to a *universe of discourse*) while irrelevant aspects are deliberately ignored. Symbolic abstraction involves using a standard language to formally describe the conceptualisation, resulting in a conceptual schema from which internal and external schemas may be derived.

Exhibit 1. Realms of Interest when Modelling Information (Bachman et al., 1975).

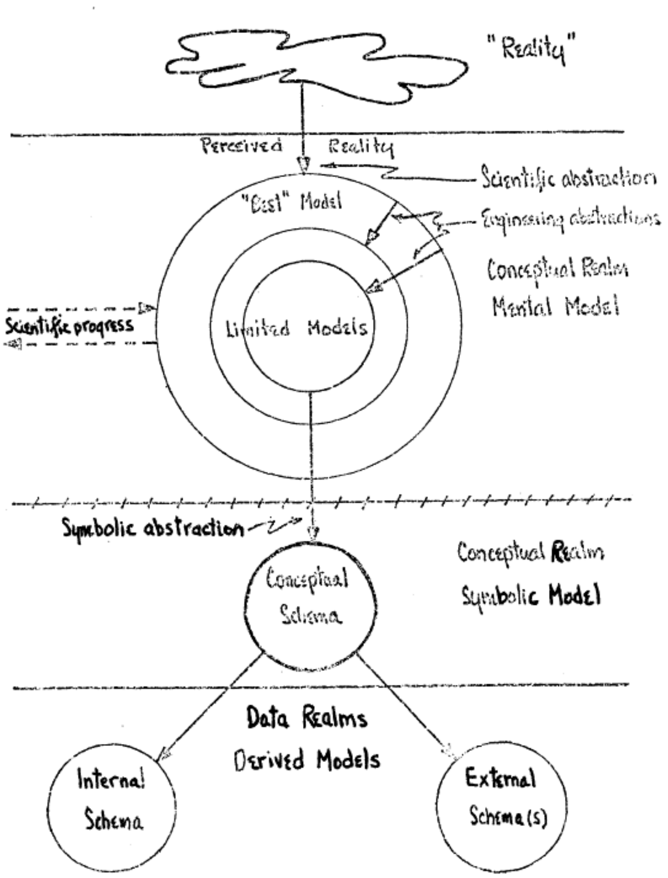
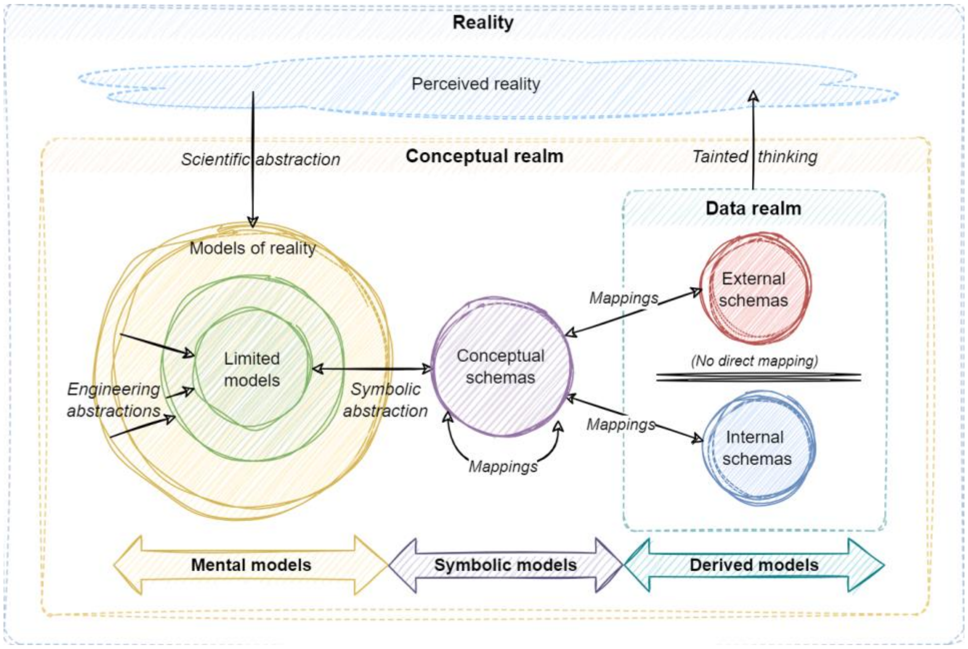


Figure VIII - 5.1

Not just Reality - Data

Exhibit 2 presents an adaptation of these realms of interest. First, the three realms (reality, conceptual, and data) may be considered nested: our conceptualisations and implementations into symbolic and derived models are part of reality and further influence our perceived reality. Besides there being multiple mental models (or conceptualisations) of both reality and a universe of discourse, there are many conceptual schemas, including different logical data models (e.g., entity-relational, multidimensional, object-oriented, object-role, etc.) and modelling languages (e.g., XML, UML, ORM 2.0) so that mappings between different conceptual schemas are required.

Exhibit 2. Realms of Interest when Modelling Information: An Adaptation.



Regarding our mental models of the data realm, Kent and Hoberman (2012) explain that 'people in the data processing community have gotten used to viewing things in a highly simplistic way, dictated by the kind of tools they have at their disposal' (p. 25). Our mental models about data and information are influenced by the industrialised data realm. Kent is furthermore of opinion that concerns about the semantics of data are aggravated when computerised systems are incorporated into record keeping facilities and suggests that it is due to a larger data realm and a loss of mutual understanding of what the data means among a larger community. Hoberman subsequently suggests that it is due to three aspects: system complexity, role specialisation, and *tainted thinking*. System complexity refers to having multiple applications with diverse purposes, and thus diverse views of interrelated information. Role specialisation happens when fewer people in a growing organisation understand the larger picture. Tainted thinking involves thinking about a business activity in terms of the software used (even though the software tends to change more often than the actual activity). Tainted thinking is akin to *WEIRD thinking* (Western, Educated, Industrialised, Rich, and Democratic) that tends to be more analytical, with a focus on objects and their attributes detached from their context, and associated with abstract, symbolic representational systems (Henrich, Heine, & Norenzayan, 2010). With WEIRD thinking models are 'assumed to be a description of the thing as it is, rather than what we would perceive in the particular circumstances in which we were looking' (Asby, 2023, p. 132).

When considering these realms as a holarchy with multiple symbolic models and conceptual schemas, several questions remain: What other holons and relationships may be present? What exactly is the role of tainted thinking within our models? Should we aim for creating a single global conceptual model (a canonical conceptual schema) and modelling language?

4.2 | Conceptual Modelling in the Era of Big Data and Distributed Computing

Ptiček, Vrdoljak, and Gulić (2019) argue that the big data integration challenge needs to be addressed on a higher level of abstraction, i.e., the conceptual data model, that this modelling should be evolutionary and iterative, and that it can be performed using Semantic Web technologies. Many other attempts to integrate data from heterogeneous sources involve schema extraction and matching techniques that focus on creating mappings between lower-level data representations. However, big data sources typically lack clear schema information and direct mappings between derived data models was never the intend with the three-level approach to modelling databases (cf. Exhibit 2). Moreover, schema matching (manual or automated) is complex and has a high risk of being inaccurate, as it is always incomplete and cannot cater for all possible (future) schema representations (Miller et al., 2001). This view clearly relates to a systems perspective – appreciating the functions of different levels of abstraction and approaching a problem situation in an evolutionary manner.

When modelling for integration on a higher level of abstraction, considerations for conceptual models and canonical conceptual schemas become relevant. The benefits of having a canonical conceptual schema include a reduced number of mappings between different representations. Saltor, Castellanos, and García-Solaco (1991) believe that 'databases are to represent conceptualisations that people have about reality, and to be interpreted as such conceptualisations' (p. 44). Also, 'a conceptualisation is never absolute, it is relative to the point of view of a user or group of people, because different persons perceive and conceive reality in different ways' (p. 45). Apart from being a subjective process, creation of a conceptual data model is furthermore regarded as a highly arbitrary process (Fisher & Sheth, 2004; Kent & Hoberman, 2012). The same universe of discourse may be modelled differently by different data modellers. Because of these reasons, any data model that must play the role of a canonical data model for database interoperability must incorporate *semantic relativism*. Semantic relativism is described as 'the degree to which [a database] can accommodate all these different conceptualisations (of the same real world)' (Saltor et al., 1991, p. 45). The systems orientation of multiple perspectives is highly relevant to this view. It furthermore incorporates the core systems concept of communication (information exchange and mutual understanding) as well as consideration of models as conceptual constructs that are not absolute.

Another argument for moving the focus away from low-level data integration and technology-specific attempts to address the problem of data semantics is provided by Borgida and Mylopoulos (2004):

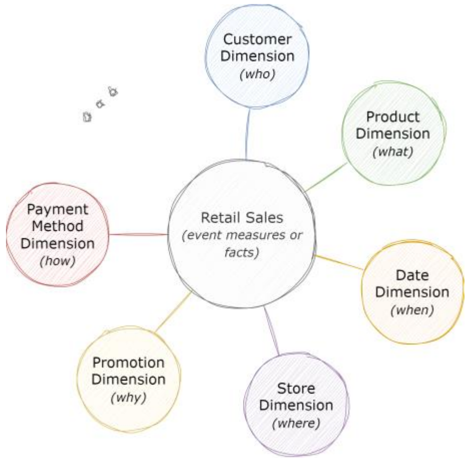
> *The problem [of data semantics] arises from general trends towards open, distributed computing, where it is no longer possible to assume that the operational environment of an information source is closed and stable. Accordingly, we should be looking for solutions that are general, i.e., not Web technology-specific. (p. 24)*

This argument relates to the proposal of the authors for a general solution that incorporates *intentional semantics* to truly understand the meaning of data. While research into Web technologies at the time focused on providing semantic mappings *between* different conceptualisations (e.g., two ontologies or two database schemas) describing the *what* and *when* aspects of a subject area, they neglected aspects of the *how* relating to the purpose and stakeholder goals within the conceptualisations. Borgida and Mylopoulos (2004) are of opinion that ever more expressive modelling languages developed for Web technologies are not enough; semantics describing the intent of a design must be incorporated into the modelling process, i.e., its purpose and perspective, to ensure trust and understanding.

4.3 | A Dimensional View of Data with the Kimball Approach

In addition to providing a predictable and standard framework for analytics, business process dimensional models are considered intuitive and flexible to analytic needs and unexpected changes (Kimball, 1997). When data is *dimensionalised*, it is conceptually divided into facts (business process measures), and dimensions describing the context of the facts. This context may include any attribute or characteristic associated with a measurement event and describes the *who*, *what*, *when*, *where*, *why*, and *how* that were true at the moment the fact was recorded (Kimball et al., 2008). An example of a dimensional model for retail sales is provided in Exhibit 3. In this case, the dimensional depiction is star shaped. It is also common to think of (and represent in OLAP databases) these models as multidimensional *cubes*.

Exhibit 3. A Dimensional Model for a Retail Sales Business Process.



Facts are typically generated (and recorded in an operational information system) when some event occurs and are conceptually grouped together in a *business process* related to the event. For example, when a product is sold, the quantity, sales price, and sales amount are facts associated with the Retail Sales business process.

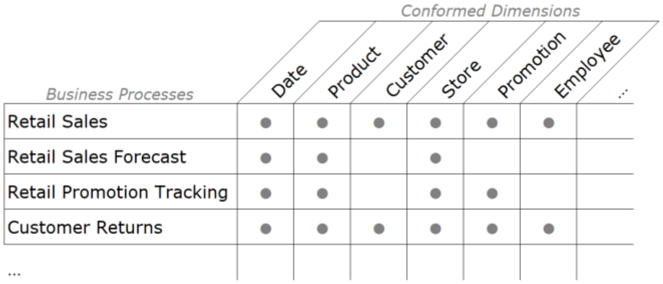
The dimensions, like Date, Customer, Product, Store, etc., are mutually understood concepts related to the business process, and typically contain one or more natural hierarchies. For example, the Store dimension may contain a hierarchy in which the location of the store aggregates from Street, City, Region, and Country all the way up to All Stores. The Date dimension rolls up from a Calendar Date to Month, Quarter, and Year.

The case for dimensional modelling in data warehousing are discussed in *A dimensional modeling manifesto* (Kimball, 1997). It is also core to the *Business Dimensional Lifecycle* (also known as the *Kimball Lifecycle*) methodology for data warehousing that has the following three core principles (Kimball et al., 2008):

- Focus on the business. Design and development centres around understanding what drives the enterprise; it relates to the overall objectives and performance measures.
- Dimensionally structure the data that is delivered for BI and decision-making (using business process dimensional models as described above).
- Iteratively develop the overall DW/BI system, one business process dimensional model at a time.

Since many *big bang* attempts to develop DW/BI systems fail, the Kimball approach prescribes incremental development based on a data warehouse bus architecture, or matrix, as depicted in Exhibit 4. The bus matrix provides an overview of the whole enterprise: its constituting parts, i.e., the business processes, are listed in rows while relevant resources and aspects of the environment are listed as shared or *conformed* dimensions on the columns; the dots on the matrix intersections indicate how they are related. This architecture is considered vital for communicating, planning, and designing a DW/BI system (Kimball & Ross, 2013) and provides a framework, or blueprint, for incremental, adaptive designs and development over distributed systems (Kimball & Ross, 2016). It supports a divide-and-conquer approach that incorporates a learning process: each iteration of a development programme involves analysing a single business process in detail to determine which facts (or event measures) and dimensions are relevant from the perspective of the subject area stakeholders. Each business process may also involve more than one internal activity that is monitored in the *value chain* of the enterprise so that it may spawn several groups of related facts (or fact tables), each such group having a certain *grain* (Kimball et al., 2008). The grain correlates with one step in an activity for which data is recorded in the value chain, e.g., *when a sales forecast is made for a product at a store*, or *when a customer returns an item*, and determines the dimensions and dimension levels that naturally falls out of the activity. In any subsequent iteration for another business process, what was learnt and developed during previous iterations, including the business vocabulary and dimensions, is reused and refined to incorporate the new perspective into the overall architecture. This process naturally involves identifying commonalities and differences of various perspectives.

Exhibit 4. An Enterprise Data Warehouse Bus Matrix, Adapted from Kimball and Ross (2013).



	Conformed Dimensions						
Business Processes	Date	Product	Customer	Store	Promotion	Employee	...
Retail Sales	●	●	●	●	●	●	
Retail Sales Forecast	●	●		●			
Retail Promotion Tracking	●	●		●	●		
Customer Returns	●	●	●	●	●	●	
...							

The bus architecture and dimensional models are independent of any technology, or the way data is modelled in operational information systems. Dimensional models are also highly flexible to change and gracefully extensible. Because of the standard, symmetrical way dimensions are associated with facts, addition of new facts, dimension attributes, and even completely new dimensions, has no effect on how existing data is modelled or accessed by users and applications. Therefore, the use of dimensional models anticipates the continuous change and evolution of analytics needs and available data (Kimball & Ross, 2016).

Essential to the (multi-) dimensional view of data is the hierarchical structure of cubes, as well as within dimensions themselves. This hierarchical nature of dimensional models allows for *drilling up*, *drilling down*, and *slicing* or *dicing* (adding or removing various dimensions) to provide different perspectives on the data.

4.4 | Semantic Data Warehousing and the Use of Ontologies

The primary motivation behind the Semantic Web is to create a highly *interconnected* web of data – the more connected, the better the context and discoverability of new data. Realising such a high level of integration and interoperability requires a *complete* language with universal expressive power capable of representing all kinds of data and views of information. The (conventional) approach to organise information based on global consistency requirements, such as in traditional knowledge representation models, is considered a hinderance to knowledge becoming *webised*; Berners-Lee (1998) argued that 'centralised concepts of absolute truth, total knowledge, and total provability' must first be eliminated (Knowledge Representation goes Global section). When considering an appropriate language to be used, Heflin, Hendler, and Luke (2005) also describe the nature of the Web as being *free from centralised authority* (i.e., distributed), highly *dynamic* (including the ontologies used), very large, and hence also an *open world* in which complete reasoning algorithms are not feasible and the assumption that all knowledge is available (the closed-world assumption) cannot be made. The ideas behind the Semantic Web are in correlation with both holism and pluralism. Linked data principles and the aim towards highly interconnected data promote holistic views of data, while rejection of global consistency requirements recognise the pluralistic nature of information and its uses.

Semantic Web technologies and ontologies are considered vehicles for unifying different conceptualisations, and thus also data models and sources. Ontologies are suitable for data interchange and integration, and domain ontologies are considered the only artefacts 'to reconcile, at the semantic level, heterogeneous data source models' (Jean, Pierra, & Ait-Ameur, 2005, 2006, p. 242); to fully automate integration, ontologies must be explicitly represented in databases. Baclawski et al. (2018) furthermore assert:

> *The use of ontology for context is a unifying conceptual model: a common language across the enterprise. In order to enable interoperability among applications or re-use of data across the enterprise, what were implicit contexts for each set of data must become explicit ontological classes and relations within the ontology. … If ontologies are able to formally specify their contextual assumptions, then logic can be applied to cross the contextual boundaries. (pp. 5,7)*

The contextual assumptions include not only the terminology used in a system or data set but also information related to the granularity and provenance of the data, as well as the way time and spatial aspects are represented. Ontologies are furthermore purposefully structured with hierarchies – another fundamental systems concept – to lay bare the intrinsic relationships between constructs in a conceptualisation.

Different ways in which to incorporate Semantic Web technologies and ontologies into databases and data warehousing have been proposed. Fankam, Jean, Bellatreche, and Aït-Ameur (2008), for example, proposed an extended ANSI/SPARC architecture with an ontological layer containing explicit semantic information of the data in terms of three types of ontologies: *conceptual canonical ontologies* (representing shared conceptualisations), *nonconceptual canonical ontologies* (representing different views of the data), and *linguistic ontologies* (representing word definitions or contextual usages). Gacitua, Mazon, and Cravero (2019) explored and summarised the use of Semantic Web technologies in data warehousing. They concluded that adoption in industry is inhibited by a technological gap and query the applicability and transferability of current research findings to industry. Antunes, Cardoso, and Barateiro (2022) also provide a review of approaches to incorporate ontologies into DW/BI systems. They found that ontologies can 'support, simplify and help automate design and development tasks and processes in DW/BI systems' (p. 12), but that they are primarily used for intermediary support in DW development and design, and not typically for data enrichment per se. The authors further argue that data enrichment must be considered to cope with the growing complexity associated with big data.

5 | Conclusion

The challenge associated with data warehousing – attempting to integrate and consolidate heterogeneous data sources – has escalated in the era of big data. Traditional data warehousing practices cannot scale and adapt in such a *messy* situation involving multiple, often conflicting perspectives, as well as high levels of complexity and uncertainty. This paper contributes to the field of data warehousing by providing a systems frame for an approach to semantic data management that explicitly incorporates systemic awareness. This framing can be summarised as follows.

The field of data warehousing and business intelligence is inherently antireductionist (or holistic) and systemic. The purpose of a DW is to provide a platform for analytics with as many integrated views of information as possible – a 360-degree view of the enterprise – to enable discovery of interconnection between phenomena and emergent properties, analysis of root causes, and insight into ways beneficial change may be employed. The business intelligence analytics cycle itself is, in essence, a feedback loop, described by Kimball (2016) as a cycle of (1) monitoring what is happening, (2) identifying exceptional performance situations, (3) determining the causal factors, (4) modelling different decision or action alternatives, and (5) tracking actions taken by feeding them back into the DW and returning to step (1) to evaluate the outcomes. The DW/BI system *is* the management information system proposed by Churchman (1979) for recording 'the relevant information for decision-making purposes and specifically [for telling] the richest story about the use of resources' (p. 39). The DW provides the necessary information regarding the enterprise components, resources, environment, and performance, and the BI component supports the management of the enterprise through the analytics function as a purposeful intervention.

A dimensional view of data is process-oriented: objects such as facts (event measures) and dimensions emerge naturally as mutual understandings within an enterprise. It furthermore provides a view of this information that is naturally hierarchical in structure, enabling high-level and detail-level analysis. Dimensional models are highly flexible to change and the Kimball approach using the bus architecture is designed to be iterative and adaptable. From a systems perspective the bus architecture model is a demonstration of wholistic planning. It assists systems designers to understand the enterprise as well as the consequences of any change. A bus matrix provides a high-level view of an enterprise as a system in which its value chain is visualised, with the subsystems represented as business processes, and resources and environmental factors represented as dimensions. The bus matrix not only illustrates the shared information and interaction between the different subsystems, but also their boundaries.

A single, relatively stable, conceptual database schema as a mediator between different derived data models (as proposed by the SPARC Study Group) can no longer be assumed. Different symbolic data representations require adaptive mappings that may be achieved using semantic technologies and ontologies. However, mere mappings between various data models are not enough. Conceptual models should include intentional semantics and contextual assumptions about the subject area, as well as semantics about its constructs. Conceptual modelling must, furthermore, be approached as an evolutionary process that remains independent of specific technologies. Automated dimensionalisation of data based on a bus architecture model is considered appropriate but would also require semantification of data for that purpose.

Any universal model used for data semantification must have a high level of *semantic expressiveness* and support for *semantic relativism*. Ontologies provide a vehicle for unification of conceptual models across different system boundaries. Ontologies are naturally hierarchical in structure, and it is possible to define canonical, noncanonical, and linguistic ontologies to integrate conceptualisations between various systems on various levels of abstraction.

This systems framing for semantic data warehousing provides the following potential benefits for the data management community. The community would appreciate and accept complexity, multiple perspectives, and the need for semantic relativism within metadata. Because it provides both stability and flexibility, business process dimensional modelling and the bus architecture model would be supported as a means for data management that is sustainable, even *future responsive*, in the era of big data.

Future research opportunities include the development of a formal system for semantic content management in support of automated dimensionalisation of data.

6 | References

Antunes, A. L., Cardoso, E., & Barateiro, J. (2022). Incorporation of ontologies in data warehouse/business intelligence systems – a systematic literature review. *International Journal of Information Management Data Insights, 2*(2), Article 100131.

Asby, R. (2021). *Thinking systems: An organic language of harmony for human survival*. Triarchy Press.

Asby, R. (2023). On the framing of systems and cybernetic models. *Philosophy Kitchen-Rivista di filosofia contemporanea, 18*(2023), 131–146.

Bachman, C. W., Cohn, L., Florance, W. E., Kirshenbaum, F., Kunecke, H., Lavin, M., . . . Yormark, B. (1975). Interim Report: ANSI/X3/SPARC Study Group on Data Base Management Systems 75-02-08. *ACM SIGMod Record, 7*(2), 1–140.

Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Luciano, J., . . . Sowa, J. (2018). Ontology summit 2018 communiqué: Contexts in context. *Applied Ontology, 13*(3), 181–200.

Bawden, R. (2010). The community challenge: The learning response. In C. Blackmore (Ed.), *Social learning systems and communities of practice* (pp. 39–56). Springer.

Berners-Lee, T. (1998, September 17). What the Semantic Web can represent. Retrieved from https://www.w3.org/DesignIssues/RDFnot.html

Berners-Lee, T. (2009, June 18). Linked data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34–43.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data – the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Borgida, A., & Mylopoulos, J. (2005). Data semantics revisited. In C. Bussler, V. Tannen, & I. Fundulaki (Eds.), *Semantic Web and databases* (pp. 9–26). Second International Workshop (SWDB 2004), Toronto, Canada. Springer.

Churchman, C. W. (1979). *The systems approach: Revised and updated*. Dell.

Churchman, C. W. (1982). *Thought and wisdom*. Intersystems.

Drath, R., & Barth, M. (2012). Concept for managing multiple semantics with AutomationML – Maturity level concept of semantic standardization. In M. Miskowicz & R. Zurawski (Eds.), *Proceedings* (pp. 1–8). 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation (ETFA 2012), Kraków, Poland. IEEE.

Fankam, C., Jean, S., Bellatreche, L., & Aït-Ameur, Y. (2008). Extending the ANSI/SPARC architecture database with explicit data semantics: An ontology-based approach. In R. Morrison, D. Balasubramaniam, & K. Falkner (Eds.), *Software Architecture* (pp. 318–321). Second International Conference (ECSA 2008), Paphos, Cyprus. Springer.

Fisher, M., & Sheth, A. P. (2004). Semantic enterprise content management. In M. P. Singh (Ed.), *The practical handbook of internet computing* (pp. 9-1–9-21). Chapman & Hall.

Gacitua, R., Mazon, J. N., & Cravero, A. (2019). Using Semantic Web technologies in the development of data warehouses: A systematic mapping. *WIREs Data Mining and Knowledge Discovery, 9*(3), Article e1293.

Halper, F., & Kobielus, J. (2021). *TDWI Best Practices Report: Building the unified data warehouse and data lake*. 1105 Media. Retrieved from https://tdwi.org/research/2021/05/arch-all-best-practices-report-unified-data-warehouse-and-data-lake/asset.aspx

Hammond, D. (2017). Philosophical foundations of systems research. In M. Edson, P. Buckle Henning, & S. Sankaran (Eds.), *A guide to systems research: Philosophy, processes and practice* (pp. 1–19). Springer.

Heflin, J., Hendler, J. A., & Luke, S. (2005). SHOE: A blueprint for the Semantic Web. In D. Fensel, J. A. Hendler, H. Lieberman, & W. Wahlster (Eds.), *Spinning the Semantic Web: Bringing the World Wide Web to its full potential* (pp. 29–63). MIT Press.

Helland, P. (2011). If you have too much data, then 'good enough' is good enough. *Communications of the ACM, 54*(6), 40–47.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences, 33*(2–3), 61–83.

Hitzler, P. (2021). A review of the semantic web field. *Communications of the ACM, 64*(2), 76–83.

Ison, R. L. (2017). *Systems Practice: How to act: In situations of uncertainty and complexity in a climate-change world* (2nd ed.). Springer.

Jean, S., Pierra, G., & Ait-Ameur, Y. (2007). Domain ontologies: A database-oriented analysis. In J. Filipe, J. Cordeiro, & V. Pedrosa (Eds.), *Web Information Systems and Technologies* (pp. 238–254). International Conferences, WEBIST 2005 and WEBIST 2006. Springer.

Kent, W., & Hoberman, S. (2012). *Data and reality: A timeless perspective on perceiving and managing information in our imprecise world* (3rd ed.). Technics.

Kimball, R. (1997). A dimensional modeling manifesto. *Database Management Systems (DBMS), 10*(9), 58–70.

Kimball, R. (2016). Drill down to ask why. In R. Kimball & M. Ross (Eds.), *The Kimball Group reader: Relentlessly practical tools for data warehousing and business intelligence* (2nd ed., pp. 22–24). Wiley.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.

Kimball, R., & Ross, M. (2016). *The Kimball Group reader: Relentlessly practical tools for data warehousing and business intelligence* (2nd ed.). Wiley.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., & Becker, B. (2008). *The data warehouse lifecycle toolkit* (2nd ed.). Wiley.

Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A. A., . . . Paton, N. W. (2017). Data context informed data wrangling. In J.-Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, & M. Toyoda (Eds.), *Proceedings* (pp. 956–963). 2017 IEEE International Conference on Big Data (BIGDATA), Boston, MA, USA. IEEE.

Krishnan, K. (2013). *Data warehousing in the age of big data*. Morgan Kaufmann.

Miller, R. J., Hernández, M. A., Haas, L. M., Yan, L., Howard Ho, C., Fagin, R., & Popa, L. (2001). The Clio project: Managing heterogeneity. *ACM SIGMod Record*, *30*(1), 78–83.

Ptiček, M., Vrdoljak, B., & Gulić, M. (2019). The potential of semantic paradigm in warehousing of big data. *Automatika*, *60*(4), 393–403.

Rao, V. (2018, March 6). Data lakes and data swamps. International Business Machines. Retrieved from https://developer.ibm.com/articles/ba-data-becomes-knowledge-2/

Reynolds, M., & Holwell, S. (2020). Epilogue: Systems approaches and systems practice. In M. Reynolds & S. Holwell (Eds.), *Systems approaches to making change: A practical guide* (2nd ed., pp. 307–321). Springer.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*, 155–169.

Saltor, F., Castellanos, M., & García-Solaco, M. (1991). Suitability of datamodels as canonical models for federated databases. *ACM SIGMod Record*, *20*(4), 44–48.

Ulrich, W. (1994). Can we secure future-responsive management through systems thinking and design? *Interfaces*, *24*(4), 26–37.

Williams, B., & Hummelbrunner, R. (2010). *Systems concepts in action: A practitioner's toolkit*. Stanford University Press.