Querying Instances – A Protocol Analysis Study

Full Paper

Arash Saghafi
Sauder School of Business
University of British Columbia
Vancouver, BC, Canada
arash.saghafi@sauder.ubc.ca

Yair Wand
Sauder School of Business
University of British Columbia
Vancouver, BC, Canada
yair.wand@sauder.ubc.ca

Jeffrey Parsons
Faculty of Business Administration
Memorial University of Newfoundland
St. John's, NL, Canada
jeffreyp@mun.ca

Abstract

The instance-based paradigm – introduced as an alternative to traditional class-based database management methods – does not require imposing a well-defined schema over data, nor does it entail central control and planning. As a consequence, it supports information requirements agility, enables collection of higher quality data, and reduces the schema and database operation problems associated with traditional methods. This study investigates the ability of content-consumers to use instance-based representations effectively for information retrieval purposes. A visual representation of the instance-based data was created and empirically evaluated with 12 subjects using protocol analysis. Results show that instance-based users were able to retrieve the required information more accurately compared to users of the traditional representation. From a cognitive point of view, instance-based users were more efficient than class-based users – they experienced fewer breakdowns in their problem solving process and, when breakdowns occurred, were more successful in recovering from them.

Keywords: Database Design; Instance-based data model; Ontology; Laboratory Experiment

Introduction

Information systems are representations or models of real world domains (Wand and Weber 1989). Thus, success of an information system is contingent on how effectively and faithfully the representations are generated and interpreted by analysts and designers (Wand and Weber 2002). Prior research has suggested using ontology – a branch of philosophy that deals with the structure of reality in the broadest sense (Angeles 1981) – as guidance for the modeling process. Various ontological theories exist, but according to Fonseca (2007), the most widely used ontology in systems analysis and design and conceptual modeling research is that of Mario Bunge (Bunge 1977). Wand and Weber (1989) adapted this approach to the Information System domain.

A large body of work has focused on developing ontological guidelines – based on Bunge's ontology - for different conceptual modelling grammars (e.g., Recker et al. 2006, Evermann and Wand 2006, Bera et al. 2011). The approach has also been used for evaluating the effectiveness of such guidelines on users' performance of tasks that require understanding conceptual models (Bodart et al. 2001, Burton-Jones and Meso 2006). A meta-analysis of prior work indicates a strong effect of ontological guidance on improving the users' understanding of the *conceptual domain models* (Saghafi and Wand 2014). However, the results related to the effectiveness of ontologically guided *data models* have been inconclusive (Allen and March 2006, Bowen et al. 2009).

This paper explores the application of ontological and cognitive principles to data modelling via an *instance-based* data modelling approach (Parsons and Wand 2000). Unlike traditional data modelling methods that describe a domain in terms of a shared model of classes of entities and relationships among them, the instance-based approach neither requires imposing well-defined structure (classes) over the data, nor entails central control and planning. Instead, the instance-based approach gives users the ability to dynamically organize the data in classes useful for their purposes. Research done so far on the instance-based indicates that this approach supports information requirements agility (Parsons and Wand 2013), provides flexibility (Parsons and Wand 2000), and can improve the quality of user-generated content (Lukyanenko et al. 2014).

Specifically, our focus is on use of instance-based data models (compared to class-based data) by human users within organizational settings. Based on schemata theory (Derry 1996), we predict that users' assimilation of information (or cognitive performance) improves when they have the option to construct mental frameworks of concepts (i.e. schemas) based on their prior knowledge. The ability to view information through one's prior mental models is granted by the instance-based paradigm. In class-based representations, on the other hand, users need to view the data based on someone else's (e.g. a database designer's) model.

The objective of this research is two-fold: (1) propose a formalized data structure to guide the implementation of instance-based systems; and (2) empirically evaluate the ability of the users to retrieve the required information effectively from an instance-based system.

Next, we present the theoretical background behind the instance-based approach. Then, we present the representation grammar proposed for modelling instance-based data. We subsequently propose the research model, and describe the experiment and the procedure used for the evaluation. We present our results and conclude the paper with a discussion of the implications of our findings.

Background

Recent advances in information technologies have facilitated collection and storage of massive amounts of diverse data within and across organizations (Brown et al. 2011). Moreover, users have also become content providers as exemplified by phenomena such as social networks, citizen science and the idea of "Open Data". Data have become available in huge volumes, with variety in data types (e.g. text and image) and might have high velocity in rate of data generation (e.g., streams, which may not be permanently kept) - these are termed "Volume-Variety-Velocity" (Zikopoulos and Eaton, 2011). Specifically, this new information paradigm is characterized by three features: (i) data may come from multiple sources with varying structure, (ii) information is used by multiple users (who might require their own customized views of the information), and (iii) the data are used for various applications, which might not be fully known in advance. All this requires flexibility in managing and using information.

A more flexible information management system can provide several advantages. First, it can enable integration of data from multiple sources with different structures. Second, users can have their information needs met more effectively by allowing them to access the data according to their own views. Third, it can support emerging applications that were not originally anticipated. However, traditional methods are typically based on a well-defined data schema structured to accommodate anticipated applications, and the assumption that data sources and users are well-known. Hence, these methods do not provide the flexibility required to attain the above advantages (Parsons and Wand 2000; Groves et al. 2013).

The instance-based paradigm – proposed by Parsons and Wand (2000) – is an alternative approach that provides flexibility by not binding data to predefined classes. From an ontological point of view, things in the world and their properties exist independently of any classification defined by humans (Parsons and Wand 2000, p. 239). Cognitively, classes are abstractions over instances that highlight useful similarity among groups of instances (Parsons and Wand, 2008). These principles are followed in the instance-based model using a two-layered architecture: (i) the instance layer that includes all the instances along with their properties, and (ii) the class layer, which includes class definitions (based on properties) that can be added or removed from the second layer as needs arise.

To represent instance-based data, we developed a simple grammar in which every *thing*[1] in the domain is represented as a node (Figure 1) and the properties that it possesses by itself (i.e., intrinsic properties) are grouped and modelled with the node. If a thing shares a property with another thing (i.e., mutual property), we model it as a link that connects two things (each represented by a node). The information related to the shared property is grouped and modeled along with the link. Figure 1 shows a generic model of our proposed grammar. This representation – which models things in the world without assigning them to predefined classes – is aligned with the ontological importance of "things". The relationships (links) modeled in this representation simply indicate what connections *might* exist. We do not consider this to be a fixed classification, as none of the property bundles or links are considered to be fixed and predetermined.

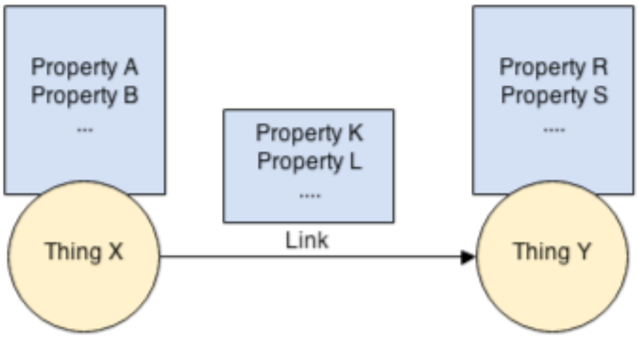
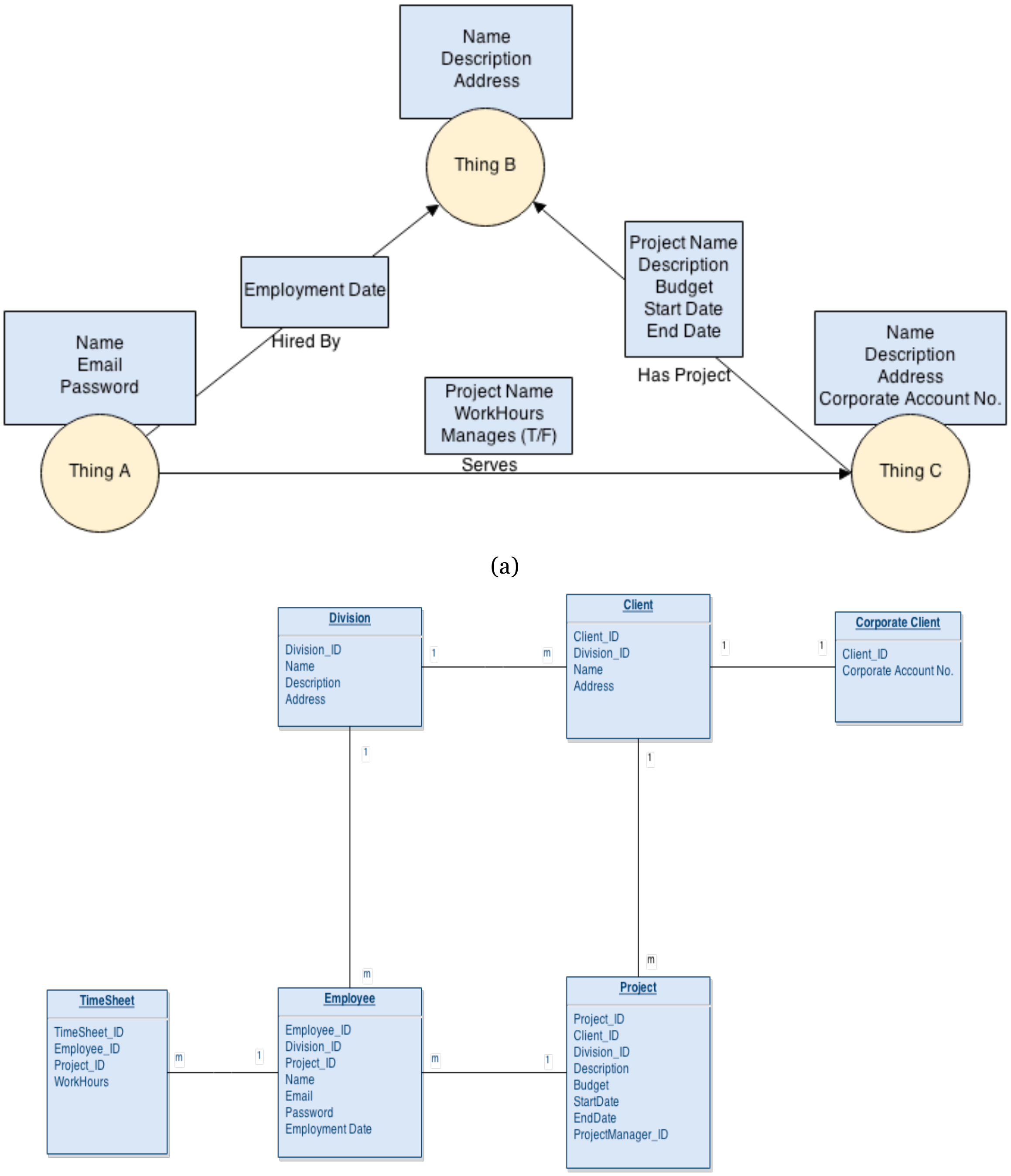


Figure 1. Instance-based representation of two things along with their properties

An example of our proposed representation grammar is provided in Figure 2(a) modeling a fictional consulting firm. For comparison, the same domain is modeled using a class diagram in Figure 2(b). In addition, Table 1 presents a query and shows the steps required for running that query in each approach.

Acknowledging the distinction between information representation and information management, the focus of the current paper is only on representation of information – however, information representation is a key element of information management and as discussed in the literature, more faithful representations of the domain lead to better information systems (Wand and Weber 1989).

[1] In Bunge's ontology (1970) the world is made of things that possess properties. In other words, *things* are the building blocks of reality.



(a)

(b)

Figure 2. Sample Consulting Firm; (a) instance-based; (b) class-based

Table 1. Identifying the steps to retrieved desired information using instance-based and class-based models

Edward McKay, one of the employees, has asked for overtime pay for his effort in completing the market research project for the EZLink company. Describe the procedure to identify the average hours worked per day by Edward on the market research project – for the sake of simplicity, weekends and holidays are also included.	
Instance-based	**Class-based**
a) Locate the thing that has "Edward McKay" as the value of its "Name" property. b) Follow the "Serves" link that goes out from Edward's node. Note the "WorkHours" property on the "Serves" link as well as "Project Name". c) From the thing at the end of the "Serves" link from step b (that should have "Name: EZLink"), look for "Has Project" links that have the same "Project Name" value. Note "Start Date" and "End Date". d) Divide the work hours of the employee (e.g. Edward) by the number of days it took to complete the project (for sake of simplicity, weekends and holidays are also included).	a) Locate Edward McKay in the "Employee" table. Note his "Employee_ID". b) From the "Timesheet" table, find Edward's timesheet using his "Employee_ID". Note his "WorkHours". c) From the "Project" table, look up EZLink's market research project. d) Using the start and finish date of the project, calculate the number of days it took to complete the project. e) Divide the work hours of the employee (e.g. Edward) by the number of days it took to complete the project (for sake of simplicity, weekends and holidays are also included).

Research Model

The proposition tested in this paper is that *users of the instance-based representation will be able to retrieve data from information systems more accurately than users of class-based systems*. By accuracy, we mean successful execution of the steps required to perform a task.

We limited the scope of the current work to studying business users – that is, users who are not database experts and are not involved in the design of the databases. Moreover, we developed a task focused on retrieval of information; in other words, describing the procedure for answering queries that have a well-defined solution.

This proposition is backed by *schema theory* (Derry 1996), which predicts that assimilation of information by human subjects will be more effective when they can construct schemas or frameworks that are congruent with the mental models in their minds. We believe that the instance-based approach allows users to form models of concepts in their minds informed by and consistent with their prior knowledge. Class-based representations, on the other hand, are typically based on a model created by a database designer, and the users of these representations need to understand the model that was created based on someone else's prior knowledge. Based on these arguments, we predict that the cognitive assimilation and processing of information will be more efficient using the instance-based approach (i.e., when data are not bound to predefined classes).

We expect that first-time users (whether instance-based or class-based) will require some effort to understand a domain representation. In comparison, users of the class-based representations may incur additional cognitive load (on top of the initial learning effort) to understand and adapt to a class structure defined by a database designer.

Experiment

We use a control-treatment design (class-based vs. instance-based) for this study. In the control group, subjects access data that are classified by a database designer. In the treatment group, however, users retrieve information that is not bound to any classification, and they can use their own mental frameworks of prior knowledge to think about the data.

To gain a deeper understanding of the possible advantages of an instance-based representation with respect to information retrieval accuracy (compared to a class-based representation), we chose to conduct a protocol analysis experiment. As mentioned in the previous section, we predict that users of instance-based representation are able to assimilate information more effectively using their own mental schema (based on their prior knowledge). Our protocol analysis investigates the effectiveness of the cognitive process of users in solving problems using both representations.

Design, Participants, Experimental Material, and Procedure

Following Burton-Jones and Meso (2006) and Bera et al. (2011), we designed a process tracing study with a small sample of 12 subjects. Six were randomly assigned to the control group (receiving a class-based representation) and six to the treatment group (receiving an instance-based representation). We recruited students of an undergraduate-level course titled "Information Systems Technology and Development", who had learned basic database concepts class diagrams, relationship types, queries, and forms[2].

Using a pre-experiment questionnaire, we asked whether subjects had ever written a database query before, and also measured their prior database knowledge and domain knowledge. The measures of prior knowledge are shown in Table 2.

Table 2. Measures of Prior Knowledge

	Written Queries Before (Y/N)	**Database Knowledge**	**Travel Knowledge**	**Consulting Knowledge**
Class-based (N=6)	66%	2.66 / 7	4.33 / 7	3.16 / 7
Instance-based (N=6)	50%	2.83 / 7	4.66 / 7	3.66 / 7

We used two cases: a fictional travel agency and a consulting firm. First, subjects were trained for 20 minutes in the instance-based or class-based method, according to the group to which they were assigned. Then participants were provided with descriptions, a general schema, and actual data (in hard copy format) from the travel agency and consulting cases (the order of case assignment was varied among subjects). Figure 2 illustrates the general schema of the consulting case.

We asked the subjects to verbalize their mental process (i.e., think out loud) and describe the steps required for answering queries related to each case. Audio recordings were taken from each subject, and analyzed by two coders.

Dependent Variables, Data Analysis and Results

To investigate the effectiveness of users' cognitive processes in solving problems using instance-based versus class-based representations, we measured three dependent variables: *performance, breakdowns, and recovery*. *Performance* was measured on a five-point scale: subjects who had the correct answer received the full mark (out of 1); when most of the steps were correct, we awarded the subject 0.75; for half, and fewer than half, correct steps we awarded 0.50 and 0.25 respectively; if the answer was completely wrong, the subject received 0. Each case had four questions, thus, subjects received a mark out of 4 for each case.

The *breakdown* variable – first defined by Newell and Simon (1972), and also used by Burton-Jones and Meso (2006) and Bera et al. (2011) - is defined as a *failure in the line of thought of an individual when he/she is searching their problem space*. *Recovery* (from a breakdown) occurs when a subject returns to an earlier step in their line of problem solving process to continue solving the problem.

[2] The participants received $20 for their time. The top two performers were awarded an additional $20 gift card (as motivation for better performance).

Based on our proposition justification (section 3), we expect that users of the instance-based representation will have fewer breakdowns since they are able to assimilate information based on their own mental schemas – rather than a schema defined by a database designer (as in class-based methods). We also expect that users of the instance-based representation will be more successful in recovering from a breakdown due to the flexibility this approach provides in viewing and organizing information according to one's mental model (Derry 1996).

To demonstrate our application of breakdown and recovery concepts, consider three hypothetical scenarios: (1) A subject starts answering a question and follows the steps in a metaphorical flow chart (or a business process), but gets to a dead-end. If he/she abandons that line of thought (i.e., the metaphorical flow chart), we consider that a breakdown with no recovery. The subject may start approaching the problem from a different angle (i.e., new line of thought), or give up on answering that question. (2) A subject gets to a dead-end in his/her solution, but goes a few steps back and continues on the same line of thought (i.e., metaphorical flow-chart). This is considered a breakdown with a recovery. (3) Without any breaks or failures, a subject goes from the start to end of the solution flow chart. While the answer may not be correct, we consider this a solution with no breakdowns (and hence no recoveries). Table 3 summarizes the results from our protocol analysis for the travel agency and the consulting domains.

As mentioned earlier, two coders evaluated the protocol analysis results. They calibrated their rating scheme using the data from a pilot (of the protocol analysis) with two subjects. The intra-class correlations (ICC) between the two coders are shown in Table 4. These numbers indicate high levels of agreement between the two coders on the measured variables of the experiment.

Table 3. Results of the protocol analysis

Domain	**Condition**	**Performance**	**Breakdowns**	**Recovery**	**Recovery %**
Travel Agency	Class-based (n=6)	2.88	5.83	2.17	37%
	Instance-based (n=6)	3.50	4.17	2.33	56%
Consulting	Class-based (n=6)	2.13	7.00	2.67	38%
	Instance-based (n=6)	3.63	7.00	3.50	50%

Performance is out of 4, n: Sample Size, Recovery Percentage = Recovery/Breakdowns.

Table 4. Agreement between coders

Domain	**Variable**	**Intra-Class Correlation**
Travel Agency	Performance	82%
	Breakdowns	98%
	Recoveries	88%
Consulting	Performance	85%
	Breakdowns	97%
	Recoveries	94%

The protocol analysis results are consistent with our proposition that users of an instance-based representation were able to assimilate the information more effectively than users of a class-based representation. This is evidenced by the higher performance, fewer breakdowns, and higher success rate in recovering from a breakdown (i.e., recovery percentage), by instance-based users in both cases. Notably, participants in the instance-based condition appeared to perform better on the tasks than those in the class-based condition (however, we do not provide statistical support for this due to the low number of subjects in each condition).

After they completed the experimental task, we asked subjects to describe the biggest challenge they faced while interacting with the representation that was assigned to them. Table 5 lists the subjects' self-reported challenges. We categorized these challenges based on similarity of the reasons.

Discussion

We provide a preliminary evaluation of the ability of novices to query an instance-based data structure in comparison to a traditional class-based approach. Using cognitive schema theory, we predicted that users of instance-based representations assimilate information more effectively than users of class-based representations. A protocol analysis involving 12 participants corroborated this proposition by showing that users of the instance-based representation provided more accurate answers and had fewer breakdowns than users of the class-based representation. Moreover, in case of a breakdown, users of the instance-based representation were more likely to recover. We should note that the subjects recruited for this experiment had received some training in the traditional class-based methods. Despite this (i.e., even though the "deck was stacked" in favor of class-based representations), subjects who worked with instance-based representations performed better.

Previous research on the instance-based paradigm has argued for its value in supporting information requirements agility (Parsons and Wand 2013), demonstrated its flexibility (Parsons and Wand 2000), and provided evidence of its ability to improve the quality of user-generated data collected in a citizen science setting (Lukyanenko et al. 2014). The current study focused on users who act as consumers of content, and demonstrates the practical usability of the instance-based approach.

We believe that the flexibility afforded by the instance-based approach caters to current trends within organizations to incorporate data from various sources into their decision-making process. Although integration of various sources of data is probably the responsibility of a database administrator (and beyond the scope of the current paper), we believe advantages afforded by the instance-based paradigm to content consumers should entice organizations to consider adopting this approach. Overall, the benefits of the instance-based paradigm warrant further exploration, both in research and in industry.

Note that the current work was limited to novice users of information systems, and routine information retrieval (query) tasks. As a laboratory experiment, our study has relatively high internal validity, but it is also limited due to the fact that we used student subjects. The choice of grammar and decisions made in creating the experimental material may also be considered to be subjective. We tried to mitigate this threat by having the material validated by two modeling experts.

Future work will perform a larger scale experiment, and evaluate users' performance in both routine retrieval tasks, as well as open-ended exploratory tasks. In addition, future research could investigate adoption of the instance-based paradigm by users with expertise in databases as well as in the business. Business experts might look for varied usage settings (e.g. e-commerce, health care, and financial data) where instance-based representation will be advantageous.

Table 5. Challenges reported by the subjects in the protocol analysis

Condition	Challenge	Subject Statements
Class-based (n=6)	Finding the required information attributes	"Pinpointing the location of attributes was difficult" (Subject #3) "Couldn't find the information in the tables" (Subject #4) "Finding information and tables, [...] was challenging" (Subject #11)
	Understanding the relationships between classes	"Questions were challenging. Understanding the relationships between classes were difficult, in particular one-to-many relationships" (Subject #6) "It takes time. Accessing and searching for [relevant] information were difficult" (Subject #10) "[...] understanding their relationships was challenging" (Subject #11) "Seeing the connection between records was difficult. Connecting foreign keys to primary keys adds overhead in data retrieval" (Subject #12)
Instance-based (n=6)	Confusion due to prior familiarity with class-based approach, or lack of familiarity with instance-based approach	"I found the approach easy, however, concepts from the course confused me" (Subject #1) "Understood the system, but describing the sequence of actions threw me off guard" (Subject #2) "Not knowing what operations were allowed in this view was challenging" (Subject #5) "I was so used to class-based that switching to this view became difficult for me" (Subject #9)
	Finding the required information attributes	"Also locating what property is on what thing was difficult" (Subject #5) "Visualizing what properties to look at based the question was challenging, but graphics and links make finding connections between objects easier" (Subject #8)
	Understanding the relationships between instances	"Too many links become confusing. However, actual data made understanding the question easier by providing example" (Subject #7)

References

Allen, G. N., & March, S. T. 2006. "The effects of state-based and event-based data representation on user performance in query formulation tasks". *MIS Quarterly*, 269-290.

Bera, P., Burton-Jones, A., & Wand, Y. 2011. "Guidelines for Designing Visual Ontologies to Support Knowledge Identification." *MIS Quarterly*, *35*(4).

Bodart, F., Patel, A., Sim, M., & Weber, R. 2001. "Should optional properties be used in conceptual modelling? A theory and three empirical tests." *Information Systems Research*, *12*(4), 384-405.

Bowen, P. L., O'Farrell, R. A., & Rohde, F. H. 2009. “An empirical investigation of end-user query development: the effects of improved model expressiveness vs. complexity.” *Information Systems Research*, *20*(4), 565-584.

Brown, B., Chui, M., & Manyika, J. 2011. Are you ready for the era of ‘big data’?. *McKinsey Quarterly*, *4*, 24-35.

Bunge, M. 1977. *Treatise on basic philosophy: Ontology I: the furniture of the world* (Vol. 1). Springer.

Burton-Jones, A., & Meso, P. N. 2006. “Conceptualizing systems for understanding: An empirical test of decomposition principles in object-oriented analysis.” *Information Systems Research*, *17*(1), 38-60.

Derry, S. J. 1996. “Cognitive schema theory in the constructivist debate.” *Educational Psychologist*, *31*(3-4), 163-174.

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. 2013. The big data revolution in healthcare: accelerating value and innovation. *New York (NY): McKinsey Global Institute*.

Lukyanenko, R., Parsons, J., & Wiersma, Y. 2014. “The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content.” *Information Systems Research*, 25(4), 669-689.

Newell, A., & Simon, H. A. 1972. *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Parsons, J., & Wand, Y. 2000. “Emancipating instances from the tyranny of classes in information modeling.” *ACM Transactions on Database Systems (TODS)*, *25*(2), 228-268.

Parsons, J., & Wand, Y. 2008. “Using cognitive principles to guide classification in information systems modeling.” *MIS Quarterly*, *32*(4), 839-868.

Parsons, J., & Wand, Y. 2013. “Cognitive Principles to Support Information Requirements Agility.” In *Advanced Information Systems Engineering Workshops* (pp. 192-197). Springer Berlin Heidelberg.

Saghafi, A., & Wand, Y. 2014. “Do Ontological Guidelines Improve Understandability of Conceptual Models? A Meta-analysis of Empirical Work”. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 4609-4618). IEEE.

Wand, Y., & Weber, R. 1989. “An ontological evaluation of systems analysis and design methods. *Information System Concepts: An In-Depth Analysis.” Elsevier Science Publishers BV, North-Holland.*

Wand, Y., & Weber, R. 2002. “Research commentary: information systems and conceptual modeling—a research agenda.” *Information Systems Research*,*13*(4), 363-376.

Zikopoulos, P. and Eaton, C., 2011. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.