Journal of the Association for Information Systems

Research Paper ISSN: 1536-9323

Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-generated Content

Roman Lukyanenko
University of Saskatchewan, Canada
lukyanenko@edwards.usask.ca

Jeffrey Parsons
Memorial University of Newfoundland, Canada

Yolanda Wiersma
Memorial University of Newfoundland, Canada

Gisela Wachinger
DIALOGIK, Non Profit Institute for Communication and Cooperation Research, Germany

Benjamin Huber
University of Stuttgart, Germany

Robert Meldt
University of Stuttgart, Germany

Abstract:

Organizations' increasing reliance on externally produced information, such as online user-generated content (UGC) and crowdsourcing, challenges common assumptions about conceptual modeling in information systems (IS) development. We demonstrate UGC's societal importance, analyze its distinguishing characteristics, identify specific conceptual modeling challenges in this setting, evaluate traditional and recently proposed approaches to modeling UGC, propose a set of conceptual modeling guidelines for developing IS that harness structured UGC, and demonstrate how to implement and evaluate the proposed guidelines using a case of development of a real crowdsourcing (citizen science) IS. We conclude by considering implications for conceptual modeling research and practice.

Keywords: Conceptual Modeling, Information Systems Development, User-Generated Content, Crowdsourcing, Citizen Science, Case Study, Design Science Research.

Roger Chiang was the accepting senior editor. This paper was submitted on September 24, 2015, and went through two revisions.

1 Introduction

Traditionally, information systems (IS) have been developed and primarily used within organizational boundaries (Hirschheim & Klein, 2012; Winter, Berente, Howison, & Butler, 2014). Consequently, conceptual modeling research and practice have generally assumed that organizations develop models to support well-defined and stable internal requirements (Abdel-Hamid, 1988; Eckerson, 2002; English, 2009). In this setting, conceptual modeling plays an important role (Grossman, Aronson, & McCarthy, 2005; Hirschheim, Klein, & Lyytinen, 1995; Kummer, Recker, & Mendling, 2016; Kung & Solvberg, 1986; Recker, 2015; Wand & Weber, 2002) because it is one of the first phases of IS development during which analysts elicit and represent requirements. Conceptual models help in designing and developing key IS components, including database schema and tables, user interfaces, and code. They also help parties involved (e.g., users, developers) understand the domain and communicate with each other during the development. Finally, conceptual models support maintenance and querying/reporting.

Increasingly, new forms of information production have given rise to an important class of applications to which traditional conceptual modeling may not apply. In contrast to information that employees or others closely associated with an organization produce, members of the general public are increasingly creating (often casually) digital information and, thus, contributing to a proliferation of user-generated content (UGC). Major sources of UGC include social media (Culnan, McHugh, & Zubillaga, 2010; Johnson, Safadi, & Faraj, 2015; Susarla, Oh, & Tan, 2012; Whelan, Teigland, Vaast, & Butler, 2016) and crowdsourcing (Brabham, 2013; Brynjolfsson & McAfee, 2014; Doan, Ramakrishnan, & Halevy, 2011; Jagadish et al., 2014; Prpić, Shukla, Kietzmann, & McCarthy, 2015). UGC can take diverse forms such as comments, blogs, tags, product reviews, videos, maps, or contest solutions (Gao, Greenwood, Agarwal, & McCullough, 2015; Krumm, Davies, & Narayanaswami, 2008; Palacios, Martinez-Corral, Nisar, & Grijalvo, 2016; Wattal, Schuff, Mandviwalla, & Williams, 2010; Zhao & Han, 2016).

While organizations can mine existing user-generated data sources (e.g., Twitter, Facebook, Flickr, Youtube, Wikipedia, forums, blogs) (Byrum & Bingham, 2016; Chen, Chiang, & Storey, 2012; Culnan et al., 2010; Wattal et al., 2010; Whelan et al., 2016), we focus on cases in which organizations develop IS to collect specific information to satisfy organizational information needs. We refer to such activities as "organization-directed" or "crowdsourced" UGC (we refer to these activities as UGC for brevity)[1] (Lukyanenko, Parsons, Wiersma, Sieber, & Maddah, 2016b). Such applications target interested and motivated online users and, thereby, make the resulting data more focused, easier to interpret, and more immediately usable.

To illustrate how one might apply organization-directed UGC, consider a recent announcement by the Chinese smartphone maker ZTE to crowdsource the design of its new smartphone (Vincent, 2016)[2]. Turning to the crowd makes it possible to harness the creativity and ingenuity of people who actually use the devices. At the same time, as the company admits, there are many unknowns about the process of seeking contributions from the crowd. For ZTE to fulfill its desire to revolutionize smartphone designs, it faces having to decide how best to design the crowdsourcing process, which includes deciding what kind of information it needs to collect from people. To address this issue, analysts would traditionally turn to conceptual modeling in order to understand and capture all pertinent requirements.

User-generated content differs considerably from traditional organizational settings in which many researchers have conducted their conceptual modeling research. While UGC opens many opportunities to engage with customers, citizens, and anonymous Internet users, it also creates challenges. In this paper, we address the question of how to perform conceptual modeling in the UGC setting.

Motivated by UGC's ongoing diffusion, we:

1) Demonstrate the societal importance of UGC and highlight the need to create technological solutions for this domain
2) Analyze UGC's distinguishing characteristics and identify specific conceptual modeling challenges in this setting
3) Evaluate traditional and recently proposed approaches to modeling structured UGC and identify their strengths and limitations

[1] Crowdsourcing is a more general term than UGC because it also includes contests, crowdfunding, product development and co-creation, and information generation (Brabham, 2013; Doan, Ramakrishnan, & Halevy, 2011).
[2] We thank our second reviewer for suggesting this example.

4) Propose a set of conceptual modeling guidelines that others can use when developing IS that harness structured UGC, and

5) Demonstrate how to implement and evaluate the proposed using a case of developing a real crowdsourcing (citizen science) IS.

Because we propose a novel artifact—design guidelines—we structure the paper in a way that follows the general guidelines for conducting design science research (DSR), including identifying a practical need, selecting kernel theories, deriving design guidelines based on kernel theories, and demonstrating/evaluating the application of guidelines in a case study (Gregor & Hevner, 2013; Hevner, March, Park, & Ram, 2004; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Prat, Comyn-Wattiau, & Akoka, 2015). We conclude our paper by outlining implications of this work for theory and practice and suggesting directions for future research.

2 The Rise of Organization-led User-generated Content

Organization-directed UGC plays an important role in the overall UGC landscape because it promises to support organizational decision making and operations by delivering focused data from target audiences (see Appendix A for examples of organization-directed UGC projects). Companies look to UGC to better understand their customers, develop new products, gain market insights, or better manage corporate assets (Brabham, 2013; Brynjolfsson & McAfee, 2014; Byrum & Bingham, 2016; Hill & Ready-Campbell, 2011; Whitla, 2009). Gartner (2013) has projected that "by 2017, more than half of all consumer goods manufacturers will receive 75% of their consumer innovation and R&D capabilities from crowdsourcing solutions". In healthcare, hospitals, governments, and other agencies encourage patient feedback to improve the quality of care. Scientists seek contributions from ordinary people and build novel citizen science information systems to obtain it (He & Wiggins, 2015; Nov, Arazy, & Anderson, 2014; Parsons, Lukyanenko, & Wiersma, 2011; Prestopnik & Tang, 2015; Simpson, Page, & De Roure, 2014).

Another emerging trend is the proliferation of platforms that organizations can use to rapidly establish a UGC application. For example, Amazon's Mechanical Turk and CrowdFlower.com each maintain a massive pool of "crowdworkers" for hire (Chittilappilly, Chen, & Amer-Yahia, 2016; Garcia-Molina, Joglekar, Marcus, Parameswaran, & Verroios, 2016; Li, Wang, Zheng, & Franklin, 2016; Paolacci, Chandler, & Ipeirotis, 2010; Stewart et al., 2015). EpiCollect.net provides a Web and mobile toolkit for the point-and-click generation of data collection interfaces on mobile platforms. Projects powered by EpiCollect engage crowds in such diverse activities as cataloguing archaeological sites, determining animal and plant distributions, and monitoring and mapping locations of street graffiti. Generally, developers use their own intuition to drive these efforts given that no established principles for harnessing UGC exist.

UGC's potential has prompted economists to suggest its inclusion in future calculations of the national gross domestic product (Brynjolfsson & McAfee, 2014). Yet, while incorporating UGC in organizational decision making can be beneficial, these environments pose fundamental conceptual modeling challenges.

3 Modeling Challenges in User-generated Content Settings

3.1 Review of Traditional Modeling Assumptions

Systems analysts have traditionally used conceptual models to capture information requirements during the earliest stages of IS development (Hirschheim et al., 1995; Wand & Weber, 2002). Typically, data's eventual consumers (e.g., managers or other employees who require data to perform some tasks) provide information requirements. Consequently, a conceptual model reflects the data's intended uses, which modelers assume to be established in advance and stable over time (Chen, 2006; Liddle & Embley, 2007; Lukyanenko & Parsons, 2013a).

Organizational settings with innate governance structures made it possible to reconcile conflicting perspectives and promote common understanding about how to collect and interpret data. Indeed, a final conceptual model often represents an integrated global view but may not represent the view of any individual user (Parsons, 2003). The fundamental approach to conveying domain semantics in traditional conceptual modeling is representation by abstraction (Mylopoulos, 1998; Smith & Smith, 1977). Abstraction makes it possible to deliberately ignore the many individual differences among phenomena and represent only relevant information (where those who consume data determine its relevance based on its known uses). Abstraction lies at the heart of popular conceptual modeling grammars. For example, a typical script made using the entity-relationship (ER) or Unified Modeling Language (UML) grammars depicts classes, attributes

of classes, and relationships between classes. Classes (e.g., student) abstract from differences among instances (e.g., particular individuals who happen to be students).

Researchers have traditionally assumed that representation by abstraction enables one to completely and accurately represent relevant domain semantics (Olivé, 2007). Close contact with users (typical to traditional organizational settings) makes it feasible to capture all relevant perspectives. To then ensure that data collection satisfies the information requirements, information contributors are trained to provide data in the desired format. As Lee and Strong (2003) contend, “[a]t minimum data collectors must know what, how, and why to collect the data" (p. 33).

3.2 Conceptual Modeling Challenges in User-generated Content Settings

In contrast to settings where information creation is (assumed to be) well understood and controlled, UGC projects typically impose no constraints on who can contribute information because engaging broad and diverse audiences is frequently their raison d'être. Due to the unrestricted, open, and democratic nature of UGC projects, it may be impossible to reach every relevant and representative stakeholder, which makes it difficult to determine appropriate and adequate conceptual structures (e.g., classes, relationship types) that would be congruent with views of every potential user (Lukyanenko et al., 2016b). Moreover, in UGC, it is particularly important to capture views of information contributors because they are the ones actually contributing data about the phenomena of interest to organizational decision makers (Gao et al., 2015; Kallinikos & Tempini, 2014).

Recent empirical work in UGC settings suggests that information contributors have extremely diverse views, which results in vastly different input on the same underlying phenomena. Among other things, diversity manifests in long-tail distributions of user-generated datasets (Clow & Makriyannis, 2011; Cooper et al., 2011; Dewan & Ramaprasad, 2012; Johnson, Faraj, & Kudaravalli, 2014; Lukyanenko, Parsons, & Wiersma, 2014a; Meijer, Burger, & Ebbers, 2009). Long-tail distributions of UGC suggest that non-experts conceptualize domain phenomena in terms of very different classes and attributes (Kallinikos & Tempini, 2014; Lukyanenko & Parsons, 2013a). Modeling must account for the possibility that some legitimate users are domain novices and may not fully understand or be able to conform to others’ domain views. Yet, the incongruence between a model embedded in information systems and the one natural for a particular data contributor may dissuade them from effectively engaging and contributing (Kleek, Styke, Schraefel, & Karger, 2011; Lukyanenko et al., 2014a; Stevens et al., 2014) or could result in dirty data due to guessing or sabotage.

The need to be sensitive to the views of extremely diverse potential information contributors is specific to UGC. Traditional modeling relies extensively on the availability of representative users to elicit requirements (Browne & Ramesh, 2002; Parsons, 2003). Traditional systems use predefined abstractions to promote a “consistent and concise” (Clarke, Burton-Jones, & Weber, 2016) “consensus view” (Parsons, 2003) among various parties that emphasizes the commonality of perspectives rather than their diversity. Further, traditional research does not concern itself with representing views of information contributors; instead, it focuses primarily on data consumers based on the assumption that information contributors consort with consumers or that one can train them to provide data with the desired content and form (Hirschheim et al., 1995; Lee, Pipino, Funk, & Wang, 2006; Olivé, 2007; Wang, Reddy, & Kon, 1995). Finally, the kind of extreme view diversity typical of UGC makes it infeasible to apply traditional modeling techniques such as roles and table views because they adopt the premise that one can discover all the relevant views (and pragmatically assume a finite, manageable number of views). Discovering all relevant views is quite infeasible in the context of UGC. Therefore, a novel modeling challenge in UGC concerns how to represent and even encourage via modeling views that would be congruent with every potential contributor of UGC. Therefore, we propose our first modeling challenge (MC):

MC1: Representing diverse (individual) information contributor views.

Importantly, in UGC settings, information contributors are inherently close to objects or events that organizations may be interested in learning about. In contrast, in typical organizational settings, information contributors, such as data entry operators or customers, have a well-defined affiliation with the organization. Thus, information contributors, much like data consumers, are inside organizational boundaries and effectively share the same reality. In contrast, in UGC projects, data consumers exist outside these boundaries and, critically, are exposed to events and objects to which members of the organization may not be exposed. Thus, through UGC, organizations expand their ability to sense and monitor their environments (Culnan et al., 2010; Goodchild, 2007). For example, many citizen science projects hope to harness individuals’ unique local knowledge to detect signs of climate change (Theobald et al., 2015). Thus, UGC

has become a vehicle for discovery (Cardamone et al., 2009; Khoury et al., 2014; Lintott et al., 2009). Researchers have argued that citizens are ideally situated to make discoveries due to their lack of scientific training, diverse interests, and proximity to phenomena (Kennett, Danielsen, & Silvius, 2015; Lukyanenko et al., 2016a). As a result, a particular contribution may involve previously unidentified phenomena (instances) or novel characteristics of known things, which creates a need to model the unknown. Translating this challenge into typical constructs of conceptual modeling, we propose:

MC2a: Representing instances of unknown (to the IS) classes.

MC2b: Representing unknown (to the IS) attributes of instances of known classes.

Traditional IS serve specific uses expressed through predefined abstractions. In contrast, because online contributors may see novel things (MC2a), researchers have actively mined UGC for unanticipated insights (Baur, 2016; Byrum & Bingham, 2016; Hara, Sun, Moore, Jacobs, & Froehlich, 2014; Sheng, Provost, & Ipeirotis, 2008). The rapid improvement in artificial intelligence and machine learning technologies that enable one to discover and extract patterns from data sets have facilitated this practice (Castillo, Castellanos, & Tremblay, 2014; Chen et al., 2012; Davenport, 2013). Machine learning, for example, enables one to repurpose contributions to discover new patterns or predict targets that data contributors do not explicitly provide. Predefined and fixed structures (as in traditional modeling) limit the extent to which the attributes and relationships in the structure may support unanticipated uses. While contemporary machine learning techniques provide powerful facilities for imputing missing values, it is significantly more challenging to recognize that a certain relevant attribute is missing (Tremblay, Dutta, & Vandermeer, 2010), which may hamper wider discoveries in UGC (Lukyanenko et al., 2016a). In contrast, while allowing users more flexibility in user input does not guarantee consistency and uniformity of data, there is a greater likelihood that users may (eventually, with large enough samples) report on certain potentially relevant attributes that research thus far has neglected or does not know about (Davis, Fuller, Tremblay, & Berndt, 2006; Raymond, 2001). Still, the question remains: how should one foster this ability to discover unknown unknowns (Davis et al., 2006) while keeping information, to the best extent possible, consistent, manageable and useful to the organization? This question is a novel conceptual modeling challenge in UGC. Therefore, we propose:

MC3: Supporting unanticipated uses of data.

In traditional organizational settings, when developing conceptual models, analysts assume that all relevant users understand any representations (abstractions) in the model (indeed, users are supposed to be intensely involved in requirements elicitation) and, if needed, that they can train users to provide data according to these conceptual structures. In contrast, in UGC settings, users are only loosely connected to the organization and are under no obligation to undergo training and comply with the organization's information needs. Since UGC is intended for organizational decision making, one needs to ensure that the data the project generates is useful to the organization sponsoring the project. However, because the information contributors are often non-experts, they may be unable to naturally generate useful data. The challenge, therefore, concerns how to leverage conceptual modeling to ensure one obtains useful data.

MC4: Ensuring that information contributors can provide data that is immediately or potentially useful to the sponsoring organization.

The modeling challenges presented above (summarized in Table 1) demonstrate the key differences between traditional and UGC settings. The comparison between the two settings shows that traditional solutions and assumptions do not align with the challenges of modeling in UGC. Unlike traditional corporate environments, UGC projects are inherently open: thus, it may not be possible in many cases to develop stable conceptual structures in advance (e.g., classes, attributes relationships) congruent with every potential user (stakeholder) in this setting. Next, we consider approaches to conceptual modeling that hold potential to address the modeling challenges in UGC.

Table 1. Comparison of Traditional Modeling Assumptions with Modeling Challenges in UGC

	Traditional modeling approach / assumption	Modeling challenge in UGC
Challenge 1	Modeling unified, consensus view based on abstractions (classes, attributes, and relationships) that users provide and can be further trained to understand; ability to discover all relevant domain views; focus on information consumers	Representing diverse (individual) information contributor views
Challenge 2	Modeling based on well-understood abstractions (classes, attributes, and relationships)	Representing instances of unknown (to the IS) classes (2a)
Challenge 3	Modeling assumes known uses and embeds them in pre-specified abstractions (classes, attributes, and relationships)	Encouraging unanticipated uses of data
Challenge 4	Modeling assumes stakeholders provide abstractions that reflect information needs, or can be trained to understand the abstractions	Ensuring that information contributors can provide data that is immediately or potentially useful to the sponsoring organization

4 Current Approaches to Conceptual Modeling for UGC

Although UGC is rapidly growing in importance, no established principles for conceptual modeling in this setting exist (Lukyanenko & Parsons, 2012). Based on our review of existing research and practice (below), we divide all extant approaches into three categories:

1) Traditional (based on pre-defined abstractions designed to satisfy consumers' information needs).
2) Emerging (based on pre-defined abstractions that are most appropriate to information contributors).
3) Emerging (based on flexible or no structure).

We briefly describe each approach and its advantages and limitations in addressing the modeling challenges we identify above.

4.1 Traditional Modeling Approach in UGC

In the absence of proven alternatives and guidelines, major projects continue to implement traditional abstraction-based modeling approaches that rely on such modeling techniques as entity-relationship diagrams and relational database storage (Alabri & Hunter, 2010; Bubnicki, Churski, & Kuijper, 2016; Connors, Lei, & Kelly, 2012; Hunter & Hsu, 2015; Kelling, Yu, Gerbracht, & Wong, 2011; Michener & Jones, 2012; Oberhauser & Prysby, 2008; Wiggins et al., 2013). Many researchers advocate these as "best practice" for conceptual modeling in UGC (Dickinson, Zuckerberg, & Bonter, 2010; Wiggins et al., 2013). The traditional development appears to be an implicit assumption for this type of UGC. Accordingly, Dickinson (2010, p. 150) describe a typical process of citizen science as: "Participants in ecological projects, typically outdoor hobbyists, …gather data, and enter them online into *centralized, relational databases*" (emphasis added).

To illustrate the application of traditional modeling in UGC, consider eBird (www.ebird.org). The project adopts entity-relationship diagrams and relational database technology (Wiggins et al., 2013) and organizes its conceptual structures in terms of classes and attributes useful to scientists. For example, after observing birds in the wild, eBird contributors report observations by indicating the biological species of the observed birds. These reports generate structured UGC that scientists can readily analyze and aggregate (thereby addressing MC4). Online volunteers across the world use eBird and submit millions of bird sightings each month (Sheppard, Wiggins, & Terveen, 2014).

However, members of the general population are not biology experts and may not be able to correctly identify organisms at the species level (Gura, 2013; Lewandowski & Specht, 2015; Lukyanenko, Parsons, & Wiersma, 2014b). Requiring contributors to provide data in terms of classes and attributes of interest to scientists (or, more broadly, organizational data consumers) may lead to guessing or project abandonment (Lukyanenko et al., 2014a, 2014b). Traditional modeling, as the eBird project exemplifies, clearly struggles to address MC1.

One may potentially address MC1 by applying machine learning techniques to classify objects (e.g., birds) based on the data provided. Indeed, eBird recently launched MerlinApp, which uses machine learning to

automatically detect species based on an uploaded photograph (He & Wiggins, 2015). Currently, such techniques work with only few target classes[3] (and may not work in most general purpose projects) and only when the photos are available and have sufficient resolution. In addition, such approaches do not realize the value of capturing additional, unanticipated information about individual observations (e.g., novel classes or novel attributes of existing classes as per MC2) or support unanticipated uses of information (as per MC3). Finally, addressing the limitations of conceptual modeling using post hoc data manipulation (e.g., via machine learning) offers practical value but does not focus on the limitations of conceptual modeling grammars or methods and, thus, does not contribute directly to advancements in conceptual modeling research.

Despite its limitations, traditional abstraction-based conceptual modeling remains dominant for collecting UGC for organizational uses (see Appendix D for examples of major projects that rely on traditional modeling). Notably, some projects implement flexible (e.g., noSQL) database technologies (e.g., for reasons of scalability and performance). Thus, the Zooniverse project, powered by the schema-less MongoDB database, engages over 900,000 online volunteers. Despite the flexible schema, however, the project collects data based on pre-specified abstractions (Simpson et al., 2014). Likewise, many generic platforms (e.g., Amazon Mechanical Turk) increasingly support a variety of data collection options but suggest creating project-specific pre-specified abstractions in the default templates (see Appendix D).

Recognizing the shortcomings of traditional conceptual modeling for UGC, several alternatives to modeling dynamic, heterogeneous or distributed information have emerged.

4.2 Emerging Approaches to Conceptual Modeling of UGC

Facing the challenge of creating "complete and accurate" specifications with diverse or unstable user views, researchers have examined the possibility of focusing on a narrow aspect of the domain that would become a domain "core" and, thus, could be shared (in principle) among heterogeneous users. For example, models may employ only very basic concepts (e.g., general level classes such as "bird", "tree", and "fish" rather than more specific ones such as "American robin", "white birch", and "rainbow trout") (Castellanos, Lukyanenko, Samuel, & Tremblay, 2016; Lukyanenko et al., 2014b; McGinnes, 2011). Recent experimental research in conceptual modeling in UGC settings offers strong evidence that, despite diverse user views, heterogeneous users readily understand basic level categories and can use them to collect data (Lukyanenko et al., 2014b). Thus, it is possible to use basic level categories to address MC1 (representing diverse user views).

One challenge of basic-level specifications, however, concerns how to convey essential semantics while keeping models simple and lean (Anwar & Parsons, 2010). Among possible solutions, researchers have proposed domain ontologies to "bridge" different user views (McGinnes, 2011). Experts or, more appropriately in UGC, the crowd (via outsourcing) can construct these ontologies to generate more intuitive representations (Braun, Schmidt, Walter, Nagypal, & Zacharias, 2007; Robal, Haav, & Kalja, 2007). Such approaches tend to encapsulate diverse user perspectives and are increasingly popular.

Another promising emerging approach that analysts can use independently or to support basic level categories involves putting the onus of modeling on users by allowing them to dynamically change models (Krogstie et al., 2004; Roussopoulos & Karagiannis, 2009). Analysts can combine this approach with modeling based on core or basic classes in which they develop only a basic-level model with the expectation that users update the model. However, such an approach leads to unresolved issues of cooperative schema evolution and concurrent access and modification of schemas (Roussopoulos & Karagiannis, 2009). It is also unclear if this approach is scalable online because some users may lack the skill and motivation to create and alter models.

Despite their promise, a concern about emerging approaches based on using predefined abstractions most appropriate to information contributors is that these generic classes are "information poor" and do not support specific inferences often needed in organizational decision making. For example, in natural history-based citizen science projects, biological species is the most commonly used classification level (Cottman-Fields, Brereton, & Roe, 2013, Figure 1; Lukyanenko et al., 2014b, Table 1) because it represents the primary unit of analysis and conservation in biology (Lewandowski & Specht, 2015; Mayden, 2002). In this case, merely modeling "generic" concepts (e.g., bird) would not address MC4 (ensuring information is immediately useful) and would also likely fail to meet the objectives of MC2 and MC3 (being unable to capture unknown classes or unknown attributes of instances and to support unanticipated uses of data). Empirical research further

[3] As of the time of writing, Merlin App allows one to identify 400 species (out of over 10,000 bird species); see http://merlin.allaboutbirds.org/help-and-faqs.

suggests that predefined abstractions bias online users toward existing classification structures, which anchors users to these classes and fails to capture novel classes or instances that users see as members of classes different from those defined in the IS (Lukyanenko et al., 2014a). Finally, because real-world objects may in some ways differ from other objects of the same class (i.e., birds of the same species may differ from one another due to their unique history), relying on modelling by abstraction may make it impossible to capture objects individuality reported by users, which further fails to address MC2.

In response to the limitations of traditional class-based information modeling, researchers and practitioners have investigated and developed flexible modeling, data collection, and storage technologies (Abiteboul, 1997; Decker et al., 2000; Fayoumi & Loucopoulos, 2016; Heath & Bizer, 2011; Krogh, Levy, Dutoit, & Subrahmanian, 1996; Parsons & Wand, 2000). An active area of research involves models that evolve along with changing enterprise requirements (Chen, 2006; for recent review, see Fayoumi & Loucopoulos, 2016). The most notable of these are noSQL databases, which often implement flexible (e.g., schema-less) data models (Chang et al., 2008; DeCandia et al., 2007; Pokorny, 2013), and semantic Web technologies that assume flexible data formats (Decker et al., 2000; Ding, Fensel, Klein, & Omelayenko, 2002; Patel-Schneider & Horrocks, 2007).

Consider one approach that does not rely on a priori structures but stores information in a flexible key-value pair or entity-attribute-value (EAV) format. This model is common in both noSQL (e.g., DynamoDB (DeCandia et al., 2007)) and semantic Web infrastructure including, the resource description framework (RDF); one can also find it in the Datalog logic programming language (Patel-Schneider & Horrocks, 2007). The RDF framework supports semantic Web applications by which one can describe things and concepts on the Web using triples of subject-predicate-object (Heath & Bizer, 2011). In Datalog, one can declare individuals without reference to a class. One can use Datalog to declare and store facts about individuals, such as “married (Mary, John)” to describe the relationship between the individuals Mary and John.

Flexible storage technologies suggest a potential conceptual modeling approach in which one does not conduct conceptual modeling as commonly understood and simply stores data in an unstructured form (e.g., as loosely structured free-form text, open tags, key-value pairs) (DeCandia et al., 2007; Kaur & Rani, 2013; Lukyanenko & Parsons, 2013a). Free-form text is dominant in social networking and social media applications (e.g., Facebook, Twitter) and is also the primary means of collecting and storing data in online forums, chats, wikis, and knowledge sharing communities (Bifet & Frank, 2010; Haklay & Weber, 2008; Kallinikos & Tempini, 2014; Russell, 2013; Wattal et al., 2010). A clear advantage of these technologies is their ability to accommodate diverse user perspectives and seamlessly capture novel, unanticipated information—fully addressing MC1 and MC2.

These approaches lead to the question: is there a role for conceptual modeling when using flexible schema-less or text-based storage solutions? For example, Kaur and Rani (2013) argue that, when implementing noSQL database technologies, conceptual modeling is unnecessary. Lukyanenko and Parsons (2013a) make a similar argument by saying that conceptual modeling in these settings may be “becoming obsolete”.

Without any conceptual modeling, however, it is unclear how effective flexible database technologies are for organizations looking to collect UGC for specific purposes (MC4). There is a booming practice of inferring structure from unstructured sources, but without some structured data production, users may enter any data they want (relevant or not), and making sense of this data is a major challenge (Abiteboul, 1997; Buneman, Davidson, Fernandez, & Suciu, 1997; Larsen, Monarchi, Hovorka, & Bailey, 2008). Indeed, it appears that, even when the technology is flexible, some conceptual modeling still occurs. Many projects (e.g., eBird, GalaxyZoo, see Appendix D) employ a hybrid solution: they have some predefined fields (e.g., type of galaxy; bird species observed) driven by traditional conceptual modeling assumptions, but the projects also ask for additional information via unstructured comment fields, emails, and discussion forums. However, this practice lacks principles and theoretical justifications for many design decisions. Furthermore, even in hybrid cases, traditional conceptual modeling prevails: projects appear to capture what is important using predefined structures that serve intended informational uses and relegate any extraneous information to unstructured sources. However, without strong theoretical grounding, one cannot determine the best allocation of data between structured and unstructured forms and what flexible approaches to storage to implement.

In summary, many projects continue to implement traditional approaches to conceptual modeling despite mounting evidence that it may be limiting in UGC settings (see Table 2). Modeling UGC using abstraction-driven modeling (both traditional and emerging based on “core” classes) is premised on the a priori availability of specifications of the kinds of data users might contribute. Abstraction-based conceptual models depict stylized (i.e., generalized and simplified) representations of actual complex user experiences

and beliefs (Kaldor, 1961). Abstraction is a mental mechanism essential for humans to survive in a diverse and changing world (Harnad, 2005; Lakoff, 1987; Parsons & Wand, 2008). In conceptual modeling, a key benefit of abstraction is that a predetermined and consistent structure makes it easier to collect data and use it for analysis and decision making (Mylopoulos, 1998). At the same time, conceptual modeling based on representation by abstraction assumes that users understand and share domain abstractions and that pre-defined abstractions can sufficiently represent real-world objects. In UGC settings, neither assumption may hold: users may have highly idiosyncratic views of a domain that do not fit into predefined abstractions; further, since abstractions are based on similarity among objects, abstractions are inherently limited for capturing unique, novel information about objects—a major promise of UGC.

Table 2. Analysis of Existing Conceptual Modeling Approaches against Challenges of Modeling in UGC Settings (We Bold Fully Addressed Challenges)

	Traditional	**Emerging approaches**	
	Specialized abstractions useful to data consumers	Pre-defined "core" or "basic" abstractions appropriate to information contributors	Flexible approaches with no pre-defined structure
Challenge 1	Struggles to address	**Fully addresses**	**Fully addresses**
Challenge 2	Struggles to address	Does not address	**Fully addresses**
Challenge 3	Partially addresses	Struggles to address	Partially addresses
Challenge 4	**Fully addresses**	Struggles to address	Struggles to address

Likewise, emerging flexible approaches carry other limitations. In particular, they struggle to provide the kind of consistency that abstraction-based modeling offers, which is problematic in cases where organizations sponsor platforms to obtain specific UGC required for organizational decision making and operations. Further, flexible approaches, such as noSQL or tagging, focus on collecting and storing data and do not specifically consider how to conduct conceptual modeling, which leaves a notable conceptual and practical gap.

In Section 5, we propose a set of theoretically motivated modeling guidelines for collecting and managing UGC and demonstrate how to develop IS using these guidelines.

5 Conceptual Modeling Guidelines for User-generated Content

While both traditional and emerging approaches have limitations, they also appear to complement each other, which Table 2 shows. The table suggests that one can address the shortcomings of one perspective (e.g., traditional) by adopting the other (e.g., emerging flexible). However, one must then address how to combine the two fundamentally different conceptual modeling philosophies to leverage both of their advantages while avoiding their deficiencies. We propose theoretically grounded guidelines for conducting conceptual modeling in UGC that combine the advantages of traditional and emerging flexible approaches.

Because conceptual modeling deals with representing the world as humans understand it, researchers have considered two theoretical foundations as useful for understanding conceptual modeling grammars: ontology and cognition (Clarke et al., 2016; Henderson-Sellers, 2015; Hirschheim et al., 1995; Sabegh, Recker, & Green, 2016; Wand, Monarchi, Parsons, & Woo, 1995). Because we intend our guidelines to represent reality, we base them primarily on ontology. At the same time, we follow recent work in conceptual modeling that argues that ontology alone is insufficient (Lukyanenko et al., 2014b). Thus, we build on findings from psychology to ensure that the ontological postulates are consistent with how people conceptualize and perceive reality.

Philosophers have long debated about what exists: views range include those that posit only mental (e.g., idealism) or physical reality (e.g., materialism) exists, or that both do (e.g., dualism). Further, philosophers have proposed differing views on the principal constituents of reality, including the primacy of forms and ideas (e.g., idealism), individual physical objects (e.g., individualism), and interconnected systems (e.g., holism) (Bunge, 2006; Herbert, 1987; Hylton, 1990; Mattessich, 2013; Walker, 1989). Guidance for our choice of ontology came from the four modeling challenges we identify above. Thus, we adopted the ontological individualism view that both has a strong philosophical tradition (Bunge, 2006; Clarke et al., 2016; Mattessich, 2013) and is consistent with cognitive theories on how humans perceive the world (Carey, 2009; Kahneman, 1992).

Ontologically, one can argue that the physical world comprises unique material objects (Bunge, 1977; Rosch, 1978). The physical world also provides an analogy for humans to think about social reality (e.g., reality of corporations, laws, social structures), and, thus, humans conceive of the social world in terms of individual entities as well (e.g., The European Union, this university) (Searle, 1995). Humans create abstractions, such as classes, to capture some equivalence among existing objects (e.g., birds, trees) (Murphy, 2004; Smith & Medin, 1981) or to create templates from which social entities can be instantiated (e.g., a corporation).

Psychology research has shown that, because of varying prior experience, domain expertise, conceptualization, and ad hoc utility, people construct different abstractions of the same domain (McCloskey & Glucksberg, 1978; Smith, 2005). Differences in perceiving and thinking about the same objects are natural unless a strong mechanism for enforcing a unified view of reality exists (generally not the case in UGC settings). For example, a citizen scientist may create a class of oiled birds to refer to distinct objects (birds) that are covered in oil; this class helps the citizen scientist communicate vital cues about a potential environmental disaster. A group of tourists or scientists could have classified the same birds a few days earlier as “beautiful wildlife” or “double-crested cormorants”, respectively.

Multiple and unique perspectives are consistent with the underlying reality and human conceptualization of reality. Thus, rather than attempting to find predefined unifying abstractions, we propose the *assumption of representational uniqueness*: that each representation of the same instance may be unique (e.g., expressed using different attributes and classes), including representations by the same user at different times[4]. This assumption addresses the modeling challenge of representing and encouraging diverse views (MC1) and, as we show later, enables one to seamlessly capture novel classes and attributes of objects (MC2). Further, recognizing the growing importance of information reuse, representational uniqueness does not assume any specific use of information and, thereby, provides for more information reuse opportunities (MC3).

Guideline 1: Adopt a representational uniqueness conceptual modeling assumption.

To enable representational uniqueness, conceptual modeling should proceed *without* prior specification of domain-specific abstractions[5]. Instead, IS developers should provide flexible logical and physical database structures and a flexible user interface to accommodate potentially unique user input. Because representational uniqueness results in potentially redundant, heterogeneous data, it departs fundamentally from the four decades of the conceptual modeling tradition premised on grammars that promote consistency and precision (Chen, 1976; Clarke et al., 2016; Peckham & Maryanski, 1988).

Under this approach, users can provide information according to how they conceptualize reality and based on unique properties of objects without having to conform to a particular pre-defined structure. One can store such information using a flexible data model, such as most noSQL (Cattell, 2011; Grolinger, Higashino, Tiwari, & Capretz, 2013; Kaur & Rani, 2013) or semi-structured (Abiteboul, 1997) data models.

Having adopted the representational uniqueness assumption, the question arises: what kind of logical structures support this assumption? To derive guidelines for structuring flexible user input, we extend a particular materialistic ontology (that of Mario Bunge) as guidance to generate specific statements about reality that we use as the foundation for modeling UGC.

Bunge (1977) postulates that the world comprises “things” (which one can also refer to as instances, objects, or entities). We apply the notion of instances to things in the physical, social, and mental worlds (Perszyk, 2013). Examples of instances include specific objects that individuals can sense in the physical world (e.g., this chair, bird sitting on a tree, Barack Obama) and any mental objects humans conceive (e.g., specific promise, rule of algebra, Hamlet). They can also include social objects (Bergholtz & Eriksson, 2015; Eriksson & Agerfalk, 2010; March & Allen, 2012, 2014; Searle, 1995) (e.g., European Union, a specific bank account). As such, we define an instance as any material, social, or mental phenomenon to which someone ascribes an individual, unique identity.

The psychology literature and traditional conceptual modeling grammars (Carey, 2009; Kahneman, 1992; Scholl, 2002) support instances’ fundamental role. We argue that the instance is an elementary and

[4] Representational uniqueness does not imply that every stored representation is unique because two different users may independently provide the same set of attributes and classes for the same instance.

[5] This does not suggest that modeling is completely absent from IS development—it merely emphasizes the absence of a traditional specification of the classes of information that an IS is designed to manage. We recognize, however, that any development inherently involves some degree of modeling, a point we consider later.

fundamental construct, and a key objective of modeling is to represent instances as fully and faithfully as possible, which leads to the second guideline:

Guideline 2: Represent UGC based on unique instances and represent instances independently of any other construct.

We now consider how to represent instances—a challenging task. According to Bunge (1977), every instance is unique in some way by virtue of having distinct properties. Properties are always attached to things and cannot exist without them: the materiality of properties is directly derived from the materiality of things. According to Bunge (1977), people cannot observe properties directly and perceive them instead as "attributes" (or sense impressions termed in philosophy "qualia" or "secondary properties") (Curley, 1972; Loar, 2003). Several attributes can potentially refer to the same property. The existence of an attribute does not imply that a particular property exists (e.g., the attribute "name" is an abstraction of an undifferentiated bundle of properties). While material things exist independently of an observer, individual observers may consider different attributes of things at different points in time. Indeed, attributes are basic abstractions of reality insofar as any attribute (e.g., the color red, a texture's roughness, a building's height) is a generalization formed by compressing diverse sensorimotor input (or memory) into a mentally stable coherent element[6]. Attributes are fundamental building blocks of representation to the extent that one can use them to identify instances and form higher-level abstractions (e.g., one can group things with similar attributes into classes). Thus, the third guideline states:

Guideline 3: Use attributes to represent individual instances.

We now consider how to model classes consistent with the representational uniqueness assumption. According to Bunge (1977), ontologically classes are based on shared properties of instances, which is consistent with research in psychology that individuals use classes to group instances they perceive (e.g., based on attribute similarity) as equivalent (Fodor, 1998; Medin, Lynch, & Solomon, 2000; Rosch, 1978). Classification allows humans to abstract from differences among instances (i.e., unique attributes of instances) and, thereby, gain cognitive economy and the ability to infer unobservable properties of instances (Medin et al., 2000). Using classes improves communication efficiency by reducing the effort of having to provide an exhaustive list of attributes for each instance. Classes are also intuitive when reasoning about instances. It is unnatural for users to refer to instance x in terms of its attributes alone. It is likely that users refer to x using some class (e.g., dog, employee, bank, account). Finally, knowing what classes users assign to instances reveals any biases in the kinds of attributes users attach to instances. The classes known to a person influence human perception as illustrated by stereotype effects (Jussim, Nelson, Manis, & Soffin, 1995) and categorical perception (Harnad, 1990). More broadly, cognitive penetration of perception posits that higher-level mental mechanisms such as classification affect lower-level perceptual functions (Elman & McClelland, 1988; Vetter & Newen, 2014). Thus, knowing the classes users attach to instances illuminates gaps and biases in the attributes they provide. Finally, Searle (1995) suggests that classes may precede instances in social worlds (i.e., humans first create a class of corporation, define its attributes, and then assign/declare its instances) (March & Allen, 2015).

In summary, classes are a critical, convenient, and natural mechanism by which users can reason about instances and describe their properties of interest. They also help to understand the attributes they provide, which helps to address MC3 and MC4 (making data amenable for reuse and making data useful for organizations). Finally, as Lukyanenko et al. (2014b) demonstrate, when given freedom to classify in an open-ended manner, non-expert users tend to provide generic classes (typically "basic-level" classes) with high accuracy. Therefore, we conceptualize classes as constructs that can be attached to instances but that do not constrain their attributes:

Guideline 4: Use classes to represent individual instances.

A corollary to the claim that all instances are unique is the fact that it is often challenging to fully and precisely represent instances using linguistic attributes and classes that humans maintain to deal with typical and recurring phenomena (see Lukyanenko, Parsons, & Samuel, 2015b). We argue that instances are informationally infinite because no collection of classes and attributes can fully exhaust the representation of instances (see Lukyanenko et al., 2014b; Murphy, 2004). Further, representing some properties of instances via attributes and classes can be unnatural and cumbersome in some situations. For example, it may be easier to state "this bird

[6] When considering visual modality, with every input interruption or environment change, such as the movement of eyes (saccades) or of the object of interest, the retina senses the focal object (stationary or moving) differently, but maintains operational constancy and equivalence of attributes, such as shape, color, length, texture, size (see, e.g., Harnad 1990).

was flying faster than any other ones I've seen before" using an unstructured format rather than forcing the user to convert the statement into a set of attributes and classes. Thus, we encourage modelers to consider additional mechanisms to represent individual instances. For example, providing data contributors the ability to describe observed objects using unstructured text, photos, videos, or virtual reality simulations may increase the overall representational fidelity of instances, which leads to the following guideline:

Guideline 5: Use additional mechanisms (e.g., unstructured text, videos, photos) to represent individual instances.

Finally, because we focus on modeling UGC intended for specific organizational purposes, we propose that, independently of the guidelines above, analysts may elicit and represent specific predefined abstractions useful for organizational decision makers. When developing user-generated content projects that organizations sponsor (e.g., Cornell University in the case of eBird), analysts have access to organizational actors, including data consumers (e.g., scientists). To leverage this opportunity, we suggest analysts develop a traditional conceptual model, termed here a *target organizational model* (TOM), to capture an organizational view of phenomena. Analysts should develop this conceptual model using traditional methods and approaches (see Borgida, Chaudhri, Giorgini, & Yu, 2009; Hirschheim et al., 1995; Jacobson, Booch, & Rumbaugh, 1999; Olivé, 2007) in conceptual modeling and may use any traditional conceptual modeling grammar (e.g., ERD, UML). This organizational model serves two objectives:

1. It models intended organizational uses that may otherwise be lost if one follows only the above guidelines. Since, according to the guidelines above, users can freely define their own attributes and classes, one can leverage the organizational model to define the project's scope and provide examples of the kind of instances the organization is interested in. One can then use this information to develop TOM cues, defined here as examples, instructions, and general informational content that the project will include, which serves as a frame of reference for users as they enter the data without explicitly constraining their input.
2. It creates a target for aligning data collected through the instance-based guidelines above and the informational view the organization needs. For example, as the guidelines above may result in highly heterogeneous data about birds, the organizational view may suggest that data consumers (e.g., scientists) are focusing on the species level of classification. One can then leverage this information to construct machine-learning processes that take the heterogeneous data that users provide and infer species based on it. Thus, the organizational model serves as a prediction target can be used in automating post hoc data transformations over sparse and heterogeneous instance-based data. Therefore, having an organizational view of the data addresses MC4 (making the resulting data useful for data consumers).

Thus, we propose:

Guideline 6: Develop a target organizational model using traditional conceptual modeling methods and the input elicited from data consumers to capture organizational information requirements and provide TOM cues and a target for automatically reconciling instance-based data obtained using previous guidelines.

Table 3 summarizes the benefits of the proposed guidelines for addressing the modeling challenge. One can also consider these benefits that arise from addressing the modeling challenges as "outcomes" or "consequences" of applying the guidelines. The guidelines have an implicit causal mechanism: we propose that implementing them results in attaining the corresponding modeling challenge. For example, we propose that using attributes, classes, text and TOM make the resulting data more useful to data consumers. Thus, our work contributes a novel design theory because it "gives prescription for design and action: it says how to do something" (Gregor & Hevner, 2013, p. 339). In providing ontological and psychological rationale for each guideline, we imbue our theory with explanatory power because one can explain the effectiveness and the underlying mechanisms of the guidelines by referencing these theories. Finally, the rationale underlying the challenges (e.g., providing organizations with high-quality relevant data, allowing users to participate and express themselves) are dependent variables of the design theory, which one can evaluate empirically. In Section 4, we provide a demonstration and evaluation of the proposed guidelines using a case study of development of an information system artifact—a real system designed to capture UGC.

Table 3. Summary of the Challenges Our Guidelines Address

Modeling challenge	Guideline(s) that address the challenge
Challenge 1: (information contributor view diversity)	Guideline 1: representational uniqueness
Challenge 2: (unknown classes and or attributes)	Guidelines 2, 3, 4, 5: flexible entry of attributes, classes
Challenge 3: (unanticipated uses)	Guidelines 3, 4, 5: new attributes and classes, text
Challenge 4: (making information useful)	Guidelines 3, 4, 5, 6: attributes, classes, text, TOM

6 Evaluation of the Proposed Guidelines

Because our contribution has both theoretical and practical components, we evaluate the proposed guidelines from three perspectives: 1) we used a case study to provide evidence of the proposed guidelines' utility (compared with traditional approaches), which also allowed us to provide a concrete instantiation example for practitioners; 2) we conducted interviews with the users of the system we developed to evaluate the extent to which the system addresses the modeling challenges above; and 3) we followed an emerging design science strategy (e.g., Alter, 2013) to consider its relevance, novelty, clarity, and usefulness (see Appendix C).

6.1 Case Study of the Application and Utility of the Guidelines

To better understand the advantages of the proposed guidelines over traditional approaches, we conducted a case study in which we created an information system based on the proposed guidelines. Case methodology is particularly powerful for addressing "why" and "how" questions in a complex, real-world setting (Baxter & Jack, 2008; Dubé & Paré, 2003; Lee, 1989; Yin, 2013). Since design science research aims to support practice (Hevner & Chatterjee, 2010), practitioners in particular need to understand how to apply the proposed guidelines. Given the inherent ambiguity in implementing abstract design guidelines, researchers recommend offering rich contextual descriptions of the implementations based on the principles (Chandra, Seidel, & Gregor, 2015; Chandra Kruse, Seidel, & Purao, 2016; Lukyanenko & Parsons, 2013b). The rich contextual descriptions also allow researchers to demonstrate how to trace specific features of the artifact to the proposed guideline and, thereby, establish instantiation validity of the implementation (Lukyanenko, Evermann, & Parsons, 2015).

The case study adopted the first author's experience with developing a UGC system following, first, traditional guidelines and, second, the proposed guidelines. This approach afforded detailed insights into the development of UGC projects based on the proposed guidelines. At the same time, only the first, second, and third authors were involved in the case study the initial stage and, of these authors, the third did so indirectly.

The first author implemented the guidelines by redesigning a citizen science IS, NLNature (http://www.nlnature.com). The third author (an ecology professor) founded the NLNature project in 2009 to map the biodiversity of a region in North America using amateur sightings of nature (e.g., plants, animals). Typical to other design science research, the third author involved IS researchers in the project due to a real-world problem (Hevner et al. 2004): allowing diverse users (information contributors) to report observations and to promote discovery of new phenomena with high veracity while generating useful data for scientific analysis.

One can break down the project into two distinct phases: a class-based approach (2009 to May 2013) and an instance-based approach (May 2013 to present). In the first phase, because the development team had no available guidelines for conceptual modeling in citizen science applications, it developed the project using a traditional class-based approach to conceptual modeling. We redesigned the project in 2013 to implement the proposed modeling guidelines.

6.1.1 Phase 1 Design: Traditional Modeling in UGC

Traditionally, in conceptual modeling, one first identifies a set of concepts (entity types, classes) that describe the domain (Parsons & Wand, 1997). Consistent with similar projects (e.g., www.eBird.org, www.iSpot.org.uk, see Appendix D), we choose biological species as the focal classes relevant to the domain. We focused on species-level classes due to the scientific team's information requirements: data consumers and sponsors of the project required biological observations be reported as species (to address

Challenge 4). Species are widely established units for monitoring, protecting, and conserving plants and animals (Mayden, 2002). This level of classification has been in the focus of broader citizen science research and practice (Crall et al., 2010; Hochachka et al., 2012; Wiersma, 2010).

To address MC1, the project sponsors suggested a mixed convention of biological nomenclature and general knowledge ("folksonomy") to conceptually organize the entities for which the project collected information. In this approach, species-level classes became lower-level classes in a generalization-specialization hierarchy in which higher-level classes were intuitive ones (see Figure B1). Hence, if a user selected the top-level class first (e.g., "sea bird"), it could limit the species-level options (e.g., to only sea birds) and, thus, help the user to locate the intended one.

We performed conceptual modeling using the popular UML class diagram notation (Dobing & Parsons, 2008; Grossman et al., 2005). We designed a relational database based on the conceptual model (Teorey, Yang, & Fry, 1986); the conceptual model also informed menu items and the options in the data collection interface.

Once the project launched, we assessed the extent to which the IS addressed the modeling challenges (including analyzing contributions, comments from users, and benchmark comparisons with parallel scientific sampling). The project team determined that the quality and level of participation were below expectations. Based on the arguments above, we identified the class-based approach to conceptual modeling that supported the system as a detriment to both quality and participation (Parsons et al., 2011). The analysis of user comments suggested that some users, when unsure how to classify unfamiliar organisms, made guesses (to satisfy the requirement to classify organisms). Figure B2 shows a vignette with an observation classified as a merlin (*Falco columbarius*) in which the observation creator admits to guessing. Notably, in this example, eight months passed before another member reported that the original example had an incorrect classification.

Additionally, in several cases, individuals could not fully describe the organisms using attributes of the correctly chosen species-level class (e.g., morph foxes had additional attributes not deducible from the class red fox). Finally, there was evidence that individuals did not report many observations because of the incongruence between the conceptual model and user views. For example, in contrast to biological nomenclature shown in Figure 1, non-experts may consider double-crested cormorants as shore birds rather than sea birds due to the strong association with shore areas; as a result, a user may not be able to locate a double-crested cormorant option under the shore bird level—a consequence of the rigid classification that drove the design (causing it to fail to address MC2). The identified issues motivated an effort to implement the proposed conceptual modeling guidelines on NLNature.

6.1.2 Phase 2 Design

The development guided by the proposed guidelines represented a fundamental shift from the previous approach. Whereas traditional IS development begins with eliciting and analyzing user requirements (Appan & Browne, 2012; Browne & Ramesh, 2002; Jacobson et al., 1999), the new guidelines suggest representing individual (unique) instances. Since the project had access to a stable cohort of data consumers (the scientists), we began by interviewing them to derive a TOM (Guideline 6) but did not develop database tables based on this model (unlike in phase 1).

Since information requirements of scientists remained the same as during phase 1, the TOM was effectively the same conceptual model created in phase 1. This model became the basis for formulating the intended project objectives, which included monitoring species distributions, informing conservation policy (which was at the species level), protecting endangered species, and educating students and the general public.

During the interviews with scientists, we identified the domain of the project to be all of natural history (i.e., plants, animals, and other taxa). Instance-based modeling according to the guidelines above has no mechanisms to set domain boundaries: a user may report an instance of a rock along with an instance of a bird. However, one can leverage the TOM in generating instructions to guide data collection to the potentially relevant (for the sponsoring organization) instances. Since NLNature's mandate was to provide data to satisfy the sponsoring organization's information needs, the IS design should remain sensitive to these views. However, embedding these views in the deep structure of the IS (Wand & Weber, 1990), such as the conceptual models and, consequently, database tables, would violate the representational uniqueness assumption. Consequently, we embedded TOM cues such as the data collection instructions to accompany data collection fields and descriptions and explanations of the project's objectives and purposes (e.g., see Figure 5, Appendix B).

Implementing the organization's information requirements as TOM cues constitutes a reasonable compromise between flexible modeling and the pragmatics of projects driven by specific interests and agenda (MC1, MC2, MC3, and MC4). The TOM cues do not stand in the way of user expression (in contrast to traditional class-based structures when they are incongruent with information contributors' views)—particularly if they make an explicit call for unanticipated kinds of instances (MC2 and MC3) and still promote organizational needs.

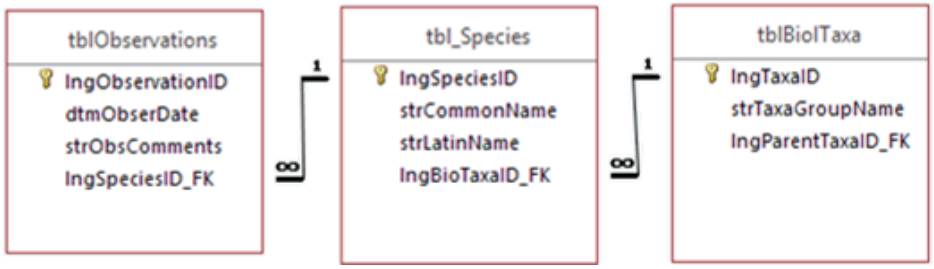
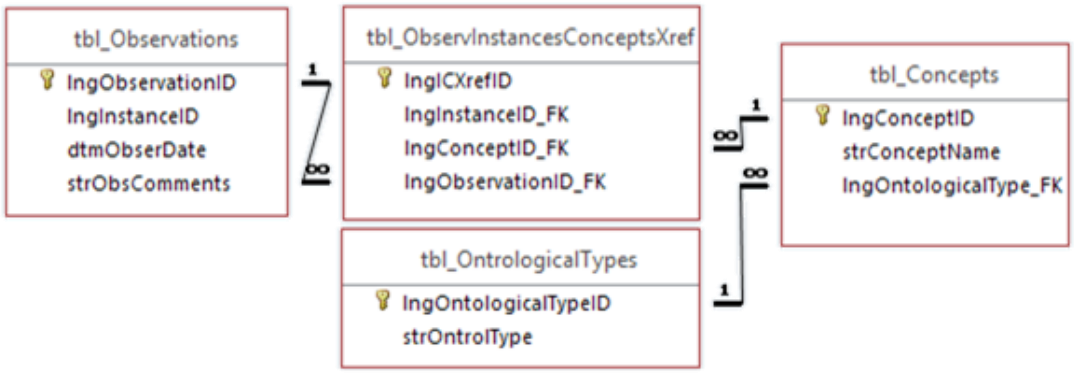
Having established the TOM, we turned to implementing the remaining guidelines. The representational uniqueness (Guideline 1) suggests providing structures that can capture attributes and classes without constraints (Guidelines 2, 3, and 4). To do that, we selected a flexible database technology with corresponding functionality. There is a booming market of flexible noSQL databases that provide several suitable schema-less databases (Cattell, 2011). Potential candidate data models included key-value pair (DeCandia et al., 2007), document-focused (Chang et al., 2008) instance-based (Parsons & Wand, 2000) and graph (Angles & Gutierrez, 2008) data models. Of these, the closest model to what we needed was the instance-based model (Parsons & Wand, 2000) because it shares the ontological and cognitive foundations that underlie this research and includes the relevant modeling constructs. Consequently, NLNature adopted the instance-based data model to store UGC.

One can deploy an instance-based data architecture on top of relational database management software. Table 4 compares database architecture used in phases 1 and 2 of the project using Microsoft Access, a common database management system. Since the focal class in phase 1 was biological species, an observation contained a required SpeciesID field that referenced the species table (in turn, species were grouped into high-level biological taxa; e.g., the species “common tern” belonged to the group “sea birds”). Failure to classify an observed phenomenon as an instance of the species table resulted in a failure to report an observation. In contrast, in phase 2, we focused on representing individual (unique instances). To capture observations of instances, we created the observations table. The table contained date and time of the instance observation (guided by the assumption that individuals observes instances at some moment in time)[7]. NLNature stored attributes and classes in a generic concepts table that contained a unique identifier, a concept name, and a flag that distinguished classes (e.g., bird) from attributes (e.g., is red). In this implementation, we chose to equate instances with observations under the assumption that, in a wide-scope natural science project, it may be infeasible to uniquely identify instances (i.e., to know that observation x and y report on the same instance). Nevertheless, we included InstanceID field in the observations table to permit data consumers to probabilistically (e.g., based on common descriptions, coordinates, time) link observations of the same instance.

We then proceeded with developing the user interface. As Guidelines 2, 3, 4 and 5 suggest, we focused on how to collect attributes and classes that describe instances. The representational uniqueness assumption suggests that data collection interfaces should be, to the extent possible, open and flexible. Following popular practice on social media websites (e.g., Facebook, Twitter, PatientsLikeMe.com), on search websites (e.g., Google), and in citizen science projects (e.g., www.iSpot.org.uk), we decided to use a prompt-assisted (“autocomplete”) text field (Kallinikos & Tempini, 2014). This feature allows a prompt to dynamically show recommendations based on the strings participants type (e.g., see Figure B3). This approach has advantages over a traditional constrained-choice mode (such as in Figure B1). Because a text field is always initially empty, it mitigates any adverse ordering and priming effects (and, thereby, satisfies the independence clause of Guideline 2). When reporting attributes or classes, the interface instructed users to begin typing in the textbox and click “add” or press “enter” when finished. As soon as someone entered more than two characters, a suggestion box appeared with the classes or attributes that contained the string entered. Users could select an item from the list or provide novel attributes and classes.

[7] This implementation is simplified for the purposes of comparison and general design guidance. In real projects such as NLNature, each table could include additional attributes, including a time stamp, system ID of the record creator, multiple comments by different users, and any security, validation, and monitoring keys.

Table 4. Comparison of Database Architecture used in Phase 1 and 2 of the Project

Phase	Table schema
1	
2	

Since we now collected data on instances in a novel manner, we provided instructions for participants about how to report observed instances. Specifically, next to the dynamic text field, we defined attributes (e.g., "Attributes (or features) are words that describe the organism you observed, including its properties, behavior and the environment."). The system would dynamically remove these definitions once participants added attributes. We converted Guidelines 3, 4 and 5 into simple questions: "What was it?" (for classes; where it denotes any instance), "Please describe it" (for attributes), "Other sighting information" (for unstructured text, see Figure B4). Following Guideline 5, we allowed users to describe instances using unstructured text but only after asking them to provide attributes and classes (to encourage greater structure) and any photos if available. Once a user finished the process (by clicking a button), the observation became public. The website also contained a dynamic map on the front page of the project that showed the most recent sightings (see Figure B6).

After introducing the TOM, we could align the instance-based data with organizational needs through a series of transformations. Thus, knowing (based on the TOM) that scientists were interested in certain species allowed us to establish an extract-transform-load (ETL) (Simitsis & Vassiliadis, 2008) process that automatically prepared the raw instance-based information for the scientists to visualize and analyze (e.g., transforming sparse attribute data into structured records organized by species). One can make this process more effective by leveraging recent advances in artificial intelligence, including natural language processing. In particular, we showed that one can use the attributes collected from non-expert volunteers who observe generally unfamiliar plants and animals in following the proposed guidelines to train machines to predict species of interest to scientists (captured in the TOM) (Lukyanenko, Wiersma, & Parsons, 2016c). For example, using common text mining methods (Larsen et al., 2008; Provost & Fawcett, 2013; Weiss, Indurkhya, & Zhang, 2010), we could obtain classification accuracy of species identification a high as 74.4 percent (Lukyanenko et al., 2016c) based on the attributes that the non-experts provided (a percentage notably higher than the percentage of correct species-level classifications by non-experts reported in the literature (Lukyanenko et al., 2014b)). One can use a similar approach to obtain other classification targets of interest (e.g., nocturnal vs. diurnal, marine vs. terrestrial, poisonous vs. edible) that are missing from the original data input via inferring them based on a training sample.

Furthermore, we have evidence of the effectiveness of the TOM cues for promoting organizational uses. In particular, while the interface was inherently flexible, over 50 percent of the observations continued to report species-level information due to the presence of multiple cues, which suggests the importance of this level for the project.

6.1.3 Comparison between Phase 1 and Phase 2 Implementations

To better understand how the proposed guidelines address the modeling challenges, we now compare the phase 1 (no guidelines, traditional modeling) and phase 2 (with proposed guidelines) implementations.

MC1: First, the phase 2 implementation addressed MC1 in a more comprehensive way than phase 1 version. In particular, representational uniqueness (Guideline 1) allowed the project to capture individual data contributors' perspectives no matter how diverse they were because individuals gained much greater flexibility to attach any attributes and classes to instances and describe them using unstructured text. Because the phase 2 interface assumed no predefined schema, different users could supply different attributes (or classes) for the same instance based on their knowledge, interests, ability, or motivation. Failure to agree on classes or even attributes was not a problem because the implementation accommodated novel classes and attributes. In contrast, in the phase 1 implementation, to meaningfully engage with the project, online volunteers had to be able to identify species and otherwise agree with scientists—a requirement that severely restricted the project to a few lay experts in the crowd.

Phase 2 further simplified the process of modeling compared with phase 1. Following the assumption of representational uniqueness, we did not engage in comprehensive requirements elicitation except for when developing the TOM. A major part of IS development (i.e., creating a formal representation of knowledge in a domain) was a relatively minor phase in which we mostly aimed to understand organizational needs to construct NLNature's TOM.

MC2: Representing instances via classes, attributes, and other means (here, unstructured text) following Guidelines 2, 3, 4 and 5 addressed MC2 (capturing novel classes and novel attributes of objects). Phase 1 offered no direct way of capturing classes and attributes that did not exist in the original schema (e.g., the project encouraged users to send project administrators an email with suggestions, but that relied on the diligence and motivation of online volunteers to actually do so). In contrast, the phase 2 implementation offered a direct and immediate mechanism in the main data collection process to report any concept not present in the schema. Since the launch of the redesigned project, users have reported hundreds of novel classes and attributes, including novel observations of biological significance (e.g., Fielden et al., 2015). These reports would have been impossible to capture if the choices were predefined and based on known distributions of organisms in the area.

MC3: In the phase 1 implementation, the resulting data were in a consistent and predictable format. While the data's consistency did not preclude one from drawing unexpected inferences from it, the phase 2 implementation opened additional opportunities to use data in a novel way (MC3). Due to Guidelines 3 and 5 (attributes and unstructured text), users provided a vast amount of specific textual descriptions and many low-level specific attributes. Thanks to Guideline 4 (classes), users put these attributes into context (e.g., one can only interpret some attributes when one knows what class of objects is being described) by reporting many generic classes (e.g., bird, tree, fish). This abundance of specific data collected without the predefined structure, but enriched with context, opened opportunities for discoveries. Novel configurations of classes, attributes, and additional textual information provided ample opportunities for finding unanticipated and unknown patterns. If phase 1 enabled one to discover "unknown knowns" (e.g., new locations of known species), phase 2 enabled one to capture "unknown unknowns" (Davis et al., 2006; Recker, 2015) (e.g., new facts about unanticipated phenomena; see Section 6.2).

MC4: The only challenge where phase 2 struggled compared with phase 1 was in ensuring that the project delivered data traditionally useful to scientists (MC4). Phase 1 of the project embedded these uses in its core specification (e.g., by focusing the conceptual model on select biological species of interest). In contrast, the phase 2 implementation did not guarantee data suitable for traditional uses, and we had to proactively foster the collection of such data. Here, Guideline 6 played an essential role, but Guidelines 3, 4, and 5 were also useful. First, as we mention above, the TOM strongly suggested that contributors report certain species, and they voluntarily classified over half of the instances reported at that level. Second, when contributors reported no biological species, one could often identify a species from an observation using machine learning provided they reported enough attributes and textual description to produce a positive identification (Lukyanenko et al., 2016c). When required, scientists could assemble

a dynamic classification based on the collection of attributes that are of interest at a given moment. For example, if scientists were interested in an attribute such as “active period”, then they could construct at least two classes based on values: nocturnal and diurnal animals. The same system could also use attributes that connect each species with a biological taxonomy to reproduce scientific biological classification. Thus, in principle, NLNature could achieve the objectives of a traditional classification without the inherent limitations of traditional approaches in addressing the previous three challenges.

In sum, despite their flexibility, the guidelines can deliver data similar to the traditional IS as long as the project provides clear instructions and examples of desirable contributions to give volunteers a general sense of the kinds of data the organization requires. Still, we need more work in this direction to claim that the guidelines fully address MC4. It is not clear if data consumers (who are experts) can work meaningfully with the sparse attributes that non-experts provide. Similarly, because domain expertise is a scarce resource and does not scale in large datasets, another question concerns the potential for using machine learning techniques to obtain fine-grained classifications from datasets based on the proposed guidelines. By proposing and demonstrating the guidelines for modeling UGC in this paper, we hope to catalyze more work in the area of applying machine learning to non-expert datasets in UGC.

6.2 Interview with NLNature's Users

Because we implemented the guidelines in a real-world application, we could obtain direct feedback from the data contributors who worked with NLNature. As such, we conducted a series of interviews and focus groups. These interviews provide additional evidence about the ability of these guidelines to address the modeling challenges in UGC. We developed a suite of questions around three themes (motivations for participating, perceptions of the website and data quality, and problem framing and agenda setting for future NLNature-like projects).

We sent invitations to participate via email (using participant supplied email addresses on NLNature). We held one focus group in the same city as the university that sponsored NLNature and one in a smaller community approximately two hours away (to try to compare users from urban vs. rural areas). We conducted a total of eight interviews (five face-to-face and three via phone) and two focus groups (five people attended the city one and two attended the smaller community one). The fourth author (an expert in participatory research and facilitation) facilitated each interview/focus group. Interviews lasted 45-60 minutes and focus groups lasted 120 minutes (see Appendix E for the detailed protocol). The conversations were semi-structured, fully recorded, transcribed, and analyzed by qualitative methods. We chose the semi-structured interview format to avoid biasing participants to specific answers (King & Horrocks, 2010; Loftus, 1975; Wengraf, 2001) and because it allowed the participants to express their attitudes towards NLNature more freely. Below, we provide evidence from the transcripts of the texts of the 15 participants (in total) of the ability of the new version of NLNature to address the modeling challenges.

MC1: During phase 1, the requirement to classify phenomena at the species level restricted participation to those data contributors capable or willing to provide information at this classification level. However, having adopted a representational uniqueness assumption (Guideline 1), we have strong evidence that the system no longer contained participation barriers related to the underlying conceptual model. Indeed, the interviews offered rich insight into the diversity of the data contributors and the ability of the system to accommodate them.

Our interview subjects ranged from people who worked in jobs that brought them outdoors (e.g., parks and recreation, fishing, bee keeping, forestry) to those who worked in an indoor setting (e.g., deputy mayor, computer programmer, retail), and most did not have a professional background in biology or ecology (Table 5 summarizes interview/focus group participants' backgrounds and Appendix F provides details on each participant). Many listed outdoor pursuits (e.g., hiking, fishing, mountain biking, birding) as key hobbies and saw the NLNature website as fitting with them. However, not all saw themselves as avid outdoors people. One noted: “I'm not a super-super nature-fit hiking kind of person”, but added “if I'm outside it's often because I'm observing things or just enjoying nature, going for a walk with my dog, that kind of stuff”. Two participants saw the website as a way to enhance intergenerational relationships. One woman who contributed sightings with her father said that participating in NLNature “definitely improved and enhanced our relationship, because it gives us something to do [together]”. Another participant noted that: “We have four granddaughters. And we want to provide a nurturing environment for them socially and physically.”. He felt that a website such as NLNature helped him “nurture … curiosity about the world” in his grandchildren.

Table 5. Summary of Interview/Focus Group Participants' Backgrounds

Gender	Female	5
	Male	10
Age category	Adult: 15	
Profession / industry	Non-biology / ecology:	8
	Related to biology/ecology	5
	No information provided	3
Expertise / background in biology/ecology	No professional background in biology/ecology	9
	Some ecology / biology background / expertise	5
	No information provided	1
Hometown	No information provided	9
	Small town / rural area	6

Several of the participants highlighted an interest in science but identified themselves mainly as "non-experts" or "lay persons". Many of them appreciated that the website allowed them to post things that they could not identify; indeed, they often posted such things and hoped the online community would assist with identification and expressed disappointment when some were never identified. Thus, the phase 2 design appears to accommodate diverse non-expert crowds—people without deep subject matter knowledge—a feature that the phase 1 design struggled with together, as have similar projects that have pursued phase 1-style conceptual modeling philosophy (Lukyanenko et al., 2016a).

MC2: As we saw from the case study, the traditional implementation struggled to represent instances of unknown (to the IS) classes and unknown attributes of instances of known classes (i.e., MC2). With the flexibility in reporting attributes and classes (Guidelines 3 and 4), data contributors now gained the ability to report on novel phenomena. One active member with an enthusiasm for photographing insects at the micro-scale probably contributed the most significant novel sighting. When we initially designed NLNature, we presented it to local natural history groups who suggested that the main focus should be on birds and wildflowers. Our original pre-specified categories focused on these "charismatic" taxa. By allowing individuals to directly enter new classes and attributes, one person used this opportunity to provide a host of insect sightings. Notably, two professional entomologists followed his posts quite closely. In one case, he reported a particular banding pattern on the legs of a mosquito, which he had not seen before. His online research led him to believe it was a new species to the province; consultations with entomologists confirmed as much, and the sighting was subsequently published in a scientific journal, with the volunteer being a co-author (Fielden et al., 2015).

Although many people are not interested in (and may be fearful of) insects, other more charismatic species captured interest among participants on NLNature in ways that we did not anticipate. In November 2014, a flock of domestic pheasants escaped from a backyard. These striking birds were quickly noted and posted to the website. User feedback indicated that people were curious where they came from (prompting the owner to go online and explain), and many expressed concern for their wellbeing. Over the following weeks, repeated geographically dispersed sightings showed that NLNature had the ability to document movement of organisms (including completely novel ones) and, thus, offer unique data on the distribution of domesticated animals in the wild.

While the website communicates that it encourages participants to post sightings of nature, some people expressed interest in animal behavior that they had observed directly and explained that such observations had motivated them to find out more about the animal and then post it. For example, one interviewee said that:

> *One day [I] saw something I'd never seen before, and it was a bee. And it was doing something I'd never seen a bee do, which was to cut off a piece of a leaf and fly away with the piece. So I was fascinated by this, and went on to discover what it was and discovered it was a leafcutter bee, again, something I'd never heard of before.*

MC3: The uses of information that are specified in advance drive traditional IS development. In contrast, having representational uniqueness opens an IS to more diverse potential uses (i.e., MC3), such as those identified by users themselves.

While MC3 suggests that data consumers will use a system in unanticipated ways, in the age of social media, data contributors also consume data, which makes them data consumers as well (Coleman, Georgiadou, & Labonte, 2009; Zwass, 2010). The interviews show how our participants used the website's data for different purposes than what we anticipated. Interviews provided evidence that Guidelines 3 and 4 (i.e., flexibility in entering new attributes and classes) fostered unanticipated uses. One participant said that she used it as an "outlet" for her curiosity, and another mentioned that he liked it "just to go out there and see… what was around, that was odd or unusual". The fact that, due to Guidelines 3 and 4, the classes could be at any level (not necessarily species) resulted in the data being used as a self-education tool. One participant said: "it is actually fun… to go in the pictures of others and identify birds they know. It's like a game.".

Our participants identified several potential ways in which one could use the data beyond the stated project's objectives. For example, they suggested that one could correlate data from the website with abiotic data (such as weather data) to see "patterns over time", use the data as indicators for water pollution, or use the data to track changes in abundance over time, including the emergence of non-native species. One participant noted non-experts' ability to detect such changes: "my mother-in-law who is 89—there are species of things there that did not exist in that community when she was a young person, and so with climate change and all that, it is important to track what new stuff there is". This comment again underscores the importance of Guidelines 3 and 4, in that they enable individuals to report on unanticipated things. Several people saw the website's main scientific use as an inventory or baseline of what is in the province; one said: "You never know, maybe later we'll say 'thank goodness we had this site, because now we know—ever since 2009—this animal's been in this area'.". Participants mentioned NLNature several times as a useful monitoring tool: "NLNature is potentially used to support questions we haven't dreamed of". Finally, one said "I may never know the results [of what I post] but in 20 years [it may have an impact]. You do not know the valuable stuff.".

MC4: Finally, we have evidence of the new system's ability to continue to support focal organizational needs via TOM (owing to Guideline 6). In particular, users continue to conceptualize the needs of the system in terms of species: the focal units of analysis for scientists. This ability is notable because the data collection interface no longer includes references to species, nor does it require one to identify instances at this level. Still, the abundance of TOM cues that point to this level provide soft guidance for contributors to generate data consistent with the traditional needs of the organizational decision makers.

Some of the participants stressed a species-specific focus and suggested that NLNature be harnessed for specific projects, such as monitoring backyard birds, assessing pesticide impacts on honey bees, or monitoring bats or coyotes. Other participants realized that NLNature might have more potential in the future, including for purposes the project sponsors could not identify. One said: "[it] is playing more of the 'big data' role. It is probably not collecting enough observations yet to call it big data…[but] I see that as an important role that NLNature plays that a common citizen science project would not play.". One participant of the first focus group emphasized the importance of mapping species: "We need to monitor what is going on with climate change.... It is not a priority [of government] so if they don't do it, maybe we have to organize it ourselves. And this seems like one of the best tools to do that.".

Of course, the evidence we provide from interviewing 15 participants does not permit statistical analyses or strong inferences of causality; however, it is consistent with the evidence from the case study and offers additional support for the utility of the proposed guidelines and their ability to address the major conceptual modeling challenges in UGC. Furthermore, interviewing real users was particularly important because UGC is inherently participatory and the feedback, perceptions, and attitudes of the people who provide UGC about the system used to collect it is particularly important (Hagen & Robertson, 2010; Nov et al., 2014).

7 Contributions and Future Work

With the growing importance of UGC, a vital question concerns whether and how one should perform conceptual modeling in this environment. Traditional approaches to conceptual modeling cannot address the challenges of modeling UGC. To address the emerging challenges, we propose conceptual modeling guidelines intended to help one develop UGC projects. We provide the guidelines to support the ever-

increasing organizational efforts to harness information outside organizational boundaries and contribute broadly to conceptual modelling research and practice.

7.1 Contributions to Practice

This work offers guidance for projects that implement both traditional and flexible technologies. While new projects continue to sprout worldwide, consistent and theory-grounded design guidelines presently do not guide such developments. Many projects continue to implement traditional approaches to modeling UGC. For example, a growing number of businesses now "rent" UGC, and many tools automatically generate data collection forms that one can rapidly deploy in desktop or mobile environments. CitySourced organizes the domain in terms of predefined classes related to crime and public safety. Similarly, Amazon's Mechanical Turk, CrowdFlower, and EpiCollect provide tools (e.g., API methods, point-and-click interfaces) for rapidly generating data collection forms with predefined choices. With the availability of easy-to-use tools, UGC applications are growing at an even faster rate. By motivating, formulating, and demonstrating the application of the guidelines for conceptual modeling in UGC, we hope to inform platform providers, API developers, and project owners of the value of alternative approaches to modeling. Similarly, organizations can design new (or redesign existing) systems following the guidelines. This work demonstrates the application of the proposed guidelines in redesigning a real IS. The NLNature design attests to the feasibility of the guidelines and provides a blueprint that practitioners can follow when developing UGC projects.

This work contributes broadly to the infrastructures to support the booming interest in big data. As digital information grows in size, velocity, and heterogeneity, the need to design appropriate data collection processes and extract information from data sources and integrate them becomes ever more pressing (Abbasi, Sarker, & Chiang, 2016; Chen et al., 2012; Jagadish et al., 2014). In particular, when designing big data infrastructures, Rai (2016) argues that a major challenge involves "address[ing] the tension between the stability of pre-existing categorization schemas that may have worked well for historical data and…challeng[ing] and revis[ing] ontological assumptions underlying the schemas when anomalies are detected in new observational data" (p. vi). In this paper, we propose an approach that can become a general solution for data-intensive architectures and support the burgeoning practice of big data.

This work contributes to noSQL database development. Because conceptual models typically inform database design, the proposed guidelines based on representation of instances introduce theoretical and practical perspectives on how to design and model databases. Specifically, this work provides theoretical justification and demonstrates how to conduct analyses with flexible data models. The NoSQL database market has grown significantly in recent years, which has led to development of numerous commercial packages such as MongoDB, DynamoDB, Apache CouchDB, Neo4J, Virtuoso, Allegro, OrientDB, FoundationDB, (see, Cattell, 2011; Grolinger et al., 2013; Pokorny, 2013)[8]. Recent research on noSQL database design admits that the setting lacks a conceptual approach applicable to it (Kaur & Rani, 2013). With this work, we introduce a conceptual layer on top of noSQL databases, which fills in important development gap and may lead to wider adoption of these database technologies.

Similarly, technical considerations such as scalability, latency, and redundancy have driven major developments in the noSQL databases area (Cattell, 2011; Pokorny, 2013). Research has dedicated considerably less attention to semantic issues—an area where noSQL databases differ from traditional data models in several important ways, which may lead to potentially significant differences in the ability of noSQL databases to represent reality. We at least partially address this deficit in this paper.

7.2 Contributions to Theory

With this research, we contribute to conceptual modeling theory by introducing several theoretical concepts, including representational uniqueness, target organizational model, target organizational model cues, and abstraction-free conceptual modeling. This work establishes a new paradigm in conceptual modeling research: representational uniqueness based on concrete instances. This paradigm constitutes a significant change to the way one would normally understand and use conceptual modeling, which includes its role in IS development and the functions that analysts and users perform. Conceptual modeling has been central to IS development since the field's early days (Checkland & Holwell, 1998; Mumford & Henshall, 1979; Rossi & Siau, 2000; Wand & Weber, 2002). In the past 40 years, researchers have proposed scores of modeling notations, some of which have become popular and widely used, including the entity-relationship

[8] http://www.oracle.com/technetwork/database/database-technologies/nosqldb/overview/index.html

(ER) model (Chen, 1976) and the Unified Modeling Language (UML) (Evermann & Wand, 2001; Jacobson et al., 1999). Notwithstanding this proliferation, most grammars share a common principle of representation by abstraction (Mylopoulos, 1998; Peckham & Maryanski, 1988; Smith & Smith, 1977). Even recent theoretical work on conceptual modeling assumes "consistent and concise" grammars (Clarke et al., 2016). In contrast, we propose modeling IS based on minimal abstraction.

One can convert the modeling guidelines we propose here into testable propositions. For example, research can measure the impact of these guidelines on dependent variables of interest (e.g., domain understanding, problem solving, or information quality) (Burton-Jones & Meso, 2006; Recker, 2015; Samuel, Watkins, Ehle, & Khatri, 2015; Topi & Ramesh, 2002; Wand & Weber, 2002). Given that conceptual models are central to user and analyst domain understanding, comprehension, verification, design of IS objects such as database schema, user interface, programming, and even the quality of data stored in IS (Lukyanenko et al., 2014b) and system use (Burton-Jones & Grange, 2012), this change opens significant opportunities for rethinking other aspects of IS development. One can conduct such research by deriving IS objects based on the guidelines we propose and comparing them with those based on traditional conceptual modeling. One can further use the guidelines to design IS or their components in a real or laboratory settings. One can also use the guidelines to evaluate existing conceptual modeling grammars or even suggest ways to develop graphic notations that could support communication and interaction during UGC IS development and, thus, pave the way to significant future research in conceptual modeling.

Acknowledgements

We thank the Social Sciences and Humanities Research Council of Canada for funding that supported this research. We also thank the senior editor and the reviewers for detailed feedback that substantially improved the paper.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), i-xxxii.

Abdel-Hamid, T. K. (1988). The economics of software quality assurance: A simulation-based case study. *MIS Quarterly*, *12*(3), 395-411.

Abiteboul, S. (1997). Querying semi-structured data. In *Proceedings of the 6th International Conference on Database Theory* (pp. 1-18).

Alabri, A., & Hunter, J. (2010). Enhancing the quality and trust of citizen science data. In *IEEE eScience 2010* (pp. 81-88).

Alter, S. (2013). Work system theory: Overview of core concepts, extensions, and challenges for the future. *Journal of the Association for Information Systems, 14*(2), 72-121.

Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys, 40*(1), 1-39.

Anwar, S., & Parsons, J. (2010). An ontological foundation for agile modeling with UML. In *Proceedings of the Americas Conference on Information Systems* (pp. 1-12).

Appan, R., & Browne, G. J. (2012). The impact of analyst-induced misinformation on the requirements elicitation process. *MIS Quarterly, 36*(1), 85-106.

Baur, A. W. (2016). Harnessing the social Web to enhance insights into people's opinions in business, government and public administration. *Information Systems Frontiers, 18*(2), 1-21.

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report, 13*(4), 544-559.

Bergholtz, M., & Eriksson, O. (2015). Towards a socio-institutional ontology for conceptual modelling of information systems. In Jeusfeld, M. A., & Karlapalem, K. (Eds.), *Advances in conceptual modeling* (pp. 225-235). Berlin: Springer.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data discovery science. In B. Pfahringer, G. Holmes, & A. Hoffmann (Eds.), *Discovery science* (LNCS, vol. 6332, pp. 1-15). Berlin: Springer.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience, 59*(11), 977-984.

Borgida, A. T., Chaudhri, V., Giorgini, P., & Yu, E. (Eds.). (2009). *Conceptual modeling: Foundations and applications: Essays in honor of John Mylopoulos* (vol. 5600). Berlin: Springer.

Borne, K., & Team, Z. (2011). *The Zooniverse: A framework for knowledge discovery from citizen science data.* Paper presented at the fall meeting of the American Geophysical Union.

Brabham, D. C. (2013). *Crowdsourcing*. Cambridge, MA: MIT Press.

Braun, S., Schmidt, A., Walter, A., Nagypal, G., & Zacharias, V. (2007). Ontology maturing: A collaborative Web 2.0 approach to ontology engineering. In *Proceedings of the 16th International World Wide Web Conference.*

Browne, G. J., & Ramesh, V. (2002). Improving information requirements determination: A cognitive perspective. *Information & Management, 39*(8), 625–645.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: work, progress, and prosperity in a time of brilliant technologies.* New York, NY: WW Norton & Company.

Bubnicki, J. W., Churski, M., & Kuijper, D. P. (2016). TRAPPER: An open source Web-based application to manage camera trapping projects. *Methods in Ecology and Evolution, 7*(10), 1209-1216.

Buneman, P., Davidson, S., Fernandez, M., & Suciu, D. (1997). Adding structure to unstructured data. In *Proceedings of the 6th International Conference on Database Theory* (pp. 336-350).

Bunge, M. (1977). *Treatise on basic philosophy: Ontology I: The furniture of the world.* Boston, MA: Reidel.

Bunge, M. (2006). *Chasing reality: strife over realism*. Toronto: University of Toronto Press.

Burton-Jones, A., & Grange, C. (2012). From use to effective use: A representation theory perspective. *Information Systems Research, 24*(3), 632-658.

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.

Byrum, J., & Bingham, A. (2016). Improving analytics capabilities through crowdsourcing. *MIT Sloan Management Review, 57*(4), 43-48.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C., Lintott, C., Keel, W. C., Parejko, J., Nichol, R. C., Thomas, D., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., & VandenBerg, J. (2009). Galaxy zoo green peas: Discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society, 399*(3), 1191-1205.

Carey, S. (2009). *The origin of concepts.* New York, USA: Oxford University Press.

Castellanos, A., Lukyanenko, R., Samuel, B. M., & Tremblay, M. C. (2016). Conceptual modeling in open information environments (pp. 1-7). In *Proceedings of the AIS SIGSAND Symposium.*

Castillo, A., Castellanos, A., & Tremblay, M. C. (2014). Improving case management via statistical text mining in a foster care organization. In M. C. Tremblay, D. VanderMeer, M. Rothenberger, A. Gupta, & V. Yoon (Eds.), *Advancing the impact of design science: Moving from theory to practice* (LNCS, vol. 4863, pp. 312-320). Berlin: Springer.

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record, 39*(4), 12-27.

Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. In *Proceedings of the Hawaii International Conference on System Sciences* (pp. 4039-4047).

Chandra Kruse, L., Seidel, S., & Purao, S. (2016). Making use of design principles (pp. 37-51). In *Proceedings of the 11th International Conference on Tackling Society's Grand Challenges with Design Science* (LNCS, vol. 9661, pp. 37-51). Berlin: Springer.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems, 26*(2), 4-23.

Checkland, P., & Holwell, S. (1998). *Information, systems, and information systems: Making sense of the field*. Hoboken, NJ: John Wiley & Sons.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Chen, P. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems, 1*(1), 9-36.

Chen, P. (2006). Suggested research directions for a new frontier—active conceptual modeling. In In D. W. Embley, A. Olive, & S. Ram (Eds.), *Conceptual modeling* (LNCS, vol. 4215, pp. 1-4). Berlin: Springer.

Chittilappilly, A. I., Chen, L., & Amer-Yahia, S. (2016). A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering, 28*(9), 2246-2266.

Clarke, R., Burton-Jones, A., & Weber, R. (2016). On the ontological quality and logical quality of conceptual-modeling grammars: The need for a dual perspective. *Information Systems Research, 27*(2), 365-382.

Clow, D., & Makriyannis, E. (2011). iSpot analysed: Participatory learning and reputation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 34-43).

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research, 4*(1), 332-358.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers, 102*(6), 1267-1289.

Cooper, A. K., Rapant, P., Hjelmager, J., Laurent, D., Iwaniak, A., Coetzee, S., Moellering, H., & Düren, U. (2011). Extending the formal model of a spatial data infrastructure to include volunteered geographical information. In *Proceedings of the 25th International Cartographic Conference*.

Cottman-Fields, M., Brereton, M., & Roe, P. (2013). Virtual birding: Extending an environmental pastime into the virtual world for citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2029-2032).

Craigle, V. (2007). MedWatch: The FDA safety information and adverse event reporting program. *Journal of the Medical Library Association, 95*(2), 224-225.

Crall, A., Newman, G., Jarnevich, C., Stohlgren, T., Waller, D., & Graham, J. (2010). Improving and integrating data on invasive species collected by citizen scientists. *Biological Invasions, 12*(10), 3419-3428.

Culnan, M. J., McHugh, P. J., & Zubillaga, J. I. (2010). How large U.S. companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive, 9*(4), 243-259.

Curley, E. M. (1972). Locke, Boyle, and the distinction between primary and secondary qualities. *The Philosophical Review*, *81*(4), 438-464.

Davenport, T. H. (2013). *Enterprise analytics: Optimize performance, process, and decisions through big data*. New York, NY: Pearson Education.

Davis, C. J., Fuller, R. M., Tremblay, M. C., & Berndt, D. J. (2006). Communication challenges in requirements elicitation and the use of the repertory grid technique. *Journal of Computer Information Systems, 46*(5), 78-86.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. In *Proceedings of 21st ACM Symposium on Operating Systems Principles* (pp. 205-220).

Decker, S., Melnik, S., Harmelen, F. van, Fensel, D., Klein, M., Broekstra, J., Erdmann, M., & Horrocks, I. (2000). The semantic Web: The roles of XML and RDF. *IEEE Internet Computing, 4*(5), 63-73.

DeMeritt, M. (2011). Simplifying citizen reporting. *ArcUser, Magazine for ESRI Software User, 14*(1), 26-27.

Dewan, S., & Ramaprasad, J. (2012). Research note-music blogging, online sampling, and the long tail. *Information Systems Research, 23*(3), 1056-1067.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics, 41*(1), 112-149.

Ding, Y., Fensel, D., Klein, M., & Omelayenko, B. (2002). The semantic Web: Yet another hip? *Data & Knowledge Engineering, 41*(2-3), 205-227.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM, 54*(4), 86-96.

Dobing, B., & Parsons, J. (2008). Dimensions of UML diagram use: A survey of practitioners. *Journal of Database Management*, 19(1), 1-18.

Dubé, L., & Paré, G. (2003). Rigor in information systems positivist case research: Current practices, trends, and recommendations. *MIS Quarterly*, *27*(4), 597-636.

Eckerson, W. W. (2002). *Data quality and the bottom line*. The Data Warehouse Institute.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*(2), 143-165.

English, L. P. (2009). *Information quality applied: Best practices for improving business information, processes and systems*. New York: Wiley.

Eriksson, O., & Agerfalk, P. J. (2010). Rethinking the meaning of identifiers in information infrastructures. *Journal of the Association for Information Systems, 11*(8), 433-454.

Evermann, J., & Wand, Y. (2001). Towards ontologically based semantics for UML constructs. In H. S. Kunii, S. Jajodia, & A. Sølvberg (Eds.), *Conceptual modeling* (LNCS, vol. 2224, pp. 354-367). Berlin: Springer.

Fayoumi, A., & Loucopoulos, P. (2016). Conceptual modeling for the design of intelligent and emergent information systems. *Expert Systems with Applications, 59*, 174-194.

Fielden, M. A., Chaulk, A. C., Bassett, K., Wiersma, Y. F., Erbland, M., Whitney, H., & Chapman, T. W. (2015). Aedes japonicus japonicus (Diptera: Culicidae) arrives at the most easterly point in North America. *The Canadian Entomologist, 147*(6), 737-740.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong.* Oxford, UK: Clarendon Press.

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, J., & Wallin, J. (2011). Galaxy zoo: Morphological classification and citizen science. In M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (Eds.), *Advances in machine learning and data mining for astronomy* (pp. 213-236). Boca Raton, FL: CRC Press.

Gao, G. G., Greenwood, B. N., Agarwal, R., & McCullough, J. S. (2015). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *MIS Quarterly, 39*(3), 565-589.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, *26*(3), 10-14.

Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering, 28*(4), 901-911.

Gartner. (2013). *Gartner reveals top predictions for IT organizations and users for 2014 and beyond*. Retrieved from http://www.gartner.com/newsroom/id/2603215

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211-221.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*(2), 337-355.

Gregor, S., & Jones, D. (2007). The anatomy of design theory. *Journal of the Association for Information Systems, 8*(5), 312-335.

Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and Applications, 2*(22).

Grossman, M., Aronson, J. E., & McCarthy, R. V. (2005). Does UML make the grade? Insights from the software development community. *Information and Software Technology, 47*(6), 383-397.

Gura, T. (2013). Citizen science: Amateur experts. *Nature, 496*(7444), 259-261.

Hagen, P., & Robertson, T. (2010). Social technologies: Challenges and opportunities for participation. In *Proceedings of the 11th Biennial Participatory Design Conference* (pp. 31-40).

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing, 7*(4), 12-18.

Hara, K., Sun, J., Moore, R., Jacobs, D., & Froehlich, J. (2014). Tohme: Detecting curb ramps in Google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 189-204).

Harnad, S. (2005). To cognize is to categorize: Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 20-45). Amsterdam: Elsevier Science.

Harnad, S. R. (1990). *Categorical perception: The groundwork of cognition*. Cambridge, MA: Cambridge University Press.

He, Y., & Wiggins, A. (2015). Community-as-a-service: Data validation in citizen science. In *Proceedings of the 4th International Workshop on Methods for Establishing Trust of (Open) Data.*

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the Web into a global data space* (vol. 1). San Rafael, CA: Morgan & Claypool Publishers.

Henderson-Sellers, B. (2015). Why philosophize; why not just model? In P. Johannesson, M. Lee, S. Liddle, A. Opdahl, & O. Pastor López (Eds.), *Conceptual modeling* (LNCS, vol. 9381). Berlin: Springer.

Herbert, N. (1987). *Quantum reality: Beyond the new physics*. New York: Anchor.

Hevner, A., & Chatterjee, S. (2010). *Design research in information systems: Theory and practice* (vol. 22). Berlin: Springer.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Hill, S., & Ready-Campbell, N. (2011). Expert stock picker: The wisdom of (the experts in) crowds. *International Journal of Electronic Commerce, 15*(3), 73-101.

Hirschheim, R., & Klein, H. K. (2012). A glorious and not-so-short history of the information systems field. *Journal of the Association for Information Systems, 13*(4), 188-235.

Hirschheim, R., Klein, H. K., & Lyytinen, K. (1995). *Information systems development and data modeling: Conceptual and philosophical foundations*. Cambridge, UK: Cambridge University Press.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution, 27*(2), 130-137.

Hunter, J., & Hsu, C.-H. (2015). Formal acknowledgement of citizen scientists' contributions via dynamic data citations. In R. B. Allen, J. Hunter, & M. L. Zheng (Eds.), *Proceedings of the 17th International Conference on Asia-Pacific Digital Libraries* (pp. 64-75). Berlin: Springer.

Hylton, P. (1990). Russell, idealism, and the emergence of analytic philosophy. New York: Oxford University Press.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 64-67).

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process* (vol. 1). Reading, MA: Addison-Wesley.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86-94.

Johnson, S. L., Faraj, S., & Kudaravalli, S. (2014). Emergence of power laws in online communities: The role of social mechanisms and preferential attachment. *MIS Quarterly, 38*(3), 795-808.

Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. *Information Systems Research, 26*(1), 165-187.

Jussim, L., Nelson, T. E., Manis, M., & Soffin, S. (1995). Prejudice, stereotypes, and labeling effects: Sources of bias in person perception. *Journal of Personality and Social Psychology, 68*(2), 228-246.

Kadry, B., Chu, L. F., Kadry, B., Gammas, D., & Macario, A. (2011). Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. *Journal of medical Internet Research, 13*(4), e95.

Kahneman, D. D. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive psychology, 24*(2), 175-219.

Kaldor, N. (1961). Capital accumulation and economic growth. In F. A. Lutz & D. C. Hague (Eds.), The theory of capital (pp. 177-222). London: Macmillan.

Kallinikos, J., & Tempini, N. (2014). Patient data as medical facts: Social media practices as a foundation for medical knowledge creation. *Information Systems Research, 25*(4), 817-833.

Kaur, K., & Rani, R. (2013). Modeling and querying data in NoSQL databases. In *Proceedings of IEEE International Conference on Big Data*.

Kelling, S., Yu, J., Gerbracht, J., & Wong, W.-K. (2011). Emergent filters: Automated data verification in a large-scale citizen science project. In *Proceedings of the e-Science Computing for Citizen Science Workshop* (pp. 20-27).

Kennett, R., Danielsen, F., & Silvius, K. M. (2015). Conservation management: Citizen science is not enough on its own. *Nature, 521*(7551), 161-161.

Kessler, D. A., Natanblut, S., Kennedy, D., Lazar, E., Rheinstein, P., Anello, C., Barash, D., Bernstein, I., Bolger, R., Cook, K., Couig, M. P., Donlon, J., Johnson, J., Lorraine, C., McGinnis, T., Nazario, J., Nightingale, S., Peck, C., Pendergast, M., Rastogi, S., Reynolds, C., Schapiro, R., Tollefson, L., & Wion, A. (1993). Introducing MEDWatch: A new approach to reporting medication and device adverse effects and product problems. *JAMA, 269*(21), 2765-2768.

Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popovic, Z., Jaskolski, M., & Baker, D. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural & Molecular Biology, 18*(10), 1175-1177.

Khoury, G. A., Liwo, A., Khatib, F., Zhou, H., Chopra, G., Bacardit, J., Bortot, L. O., Faccioli, R. A., Deng, X., He, Y., Krupa, P., Li, J., Mozolewska, M. A., Sieradzan, A. K., Smadbeck, J., Wirecki, T., Cooper, S., Flatten, J., Xu, K., Baker, D., Cheng, J., Delbern, A. C. B., Floudas, C. A., Keaser, C., Levitt, M., Popovic, Z., Scheraga, H. A., Skolnick, J., & Crivelli, S. N. (2014). WeFold: A coopetition for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics, 82*(9), 1850-1868.

King, N., & Horrocks, C. (2010). *Interviews in qualitative research*. Thousand Oaks, CA: Sage.

Kleek, M. G. V., Styke, W., Schraefel, M., & Karger, D. (2011). Finders/keepers: A longitudinal study of people managing information scraps in a micro-note tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2907-2916).

Krogh, B., Levy, S., Dutoit, A., & Subrahmanian, E. (1996). Strictly class-based modelling considered harmful. In Proceedings of the *29th Hawaii International Conference on System Sciences* (pp. 242-250).

Krogstie, J., Lyytinen, K., Opdahl, A. L., Pernici, B., Siau, K., & Smolander, K. (2004). Research areas and challenges for mobile information systems. *International Journal of Mobile Communications, 2*(3), 220-234.

Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing, 7*(4), 10-11.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension*. Chicago, IL: Chicago Press.

Kummer, T.-F., Recker, J., & Mendling, J. (2016). Enhancing understandability of process models through cultural-dependent color adjustments. *Decision Support Systems, 87*, 1-12.

Kung, C. H., & Solvberg, A. (1986). Activity modeling and behavior modeling. In *Proceedings of the IFIP WG 8.1 Working Conference on Information Systems Design Methodologies* (pp. 145-171).

Lakoff, G. (1987). *Women, fire, and dangerous things : What categories reveal about the mind.* Chicago: University of Chicago Press.

Larsen, K. R., Monarchi, D. E., Hovorka, D. S., & Bailey, C. N. (2008). Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems. *Decision Support Systems, 45*(4), 884-896.

Lee, A. S. (1989). A scientific methodology for MIS case studies. *MIS Quarterly*, *13*(1), 33-50.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to data quality*. Cambridge, MA: MIT Press.

Lee, Y. W., & Strong, D. M. (2003). Knowing-why about data processes and data quality. *Journal of Management Information Systems, 20*(3), 13-39.

Lewandowski, E., & Specht, H. (2015). Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology, 29*(3), 713-723.

Li, G., Wang, J., Zheng, Y., & Franklin, M. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2296-2319.

Liddle, S. W., & Embley, D. W. (2007). A common core for active conceptual modeling for learning from surprises. In P. C. Peter & Y. W. Leah (Eds.), *Active conceptual modeling of learning: Next generation learning based system development* (pp. 47-56). Berlin: Springer.

Lintott, C. J., Schawinski, K., Keel, W., Arkel, H. V., Bennert, N., Edmondson, E., Thomas, D., Smith, D. J. B., Herbert, P. D., Jarvis, M. J., Virani, S., Andreescu, D, Bamford, S. P., Land, K., Murray, P, Nichol, R. C., Raddick, M. J., Slosar, A., Szalay, A., & Vandenberg, J. (2009). Galaxy zoo: Hanny's Voorwerp, a quasar light echo? *Monthly Notices of the Royal Astronomical Society, 399*(1), 129-140.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 389*(3), 1179-1189.

Loar, B. (2003). Qualia, properties, modality. *Philosophical Issues, 13*(1), 113-129.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive psychology, 7*(4), 560-572.

Lukyanenko, R., Evermann, J., & Parsons, J. (2015). Guidelines for establishing instantiation validity in IT artifacts: A survey of IS research. In *DESRIST 2015* (LNCS, vol. 9073). Berlin: Springer.

Lukyanenko, R., & Parsons, J. (2012). Conceptual modeling principles for crowdsourcing. In *Proceedings of the International Workshop on Multimodal Crowdsensing* (pp. 3-6).

Lukyanenko, R., & Parsons, J. (2013a). Is traditional conceptual modeling becoming obsolete? In W. Ng V. C. Storey & J. C. Trujillo (Eds.), *Conceptual modeling* (LNCS, vol. 8217). Berlin: Springer.

Lukyanenko, R., & Parsons, J. (2013b). Reconciling theories with design choices in design science research. In *DESRIST 2013* (LNCS, vol. 7939, pp. 165-180). Berlin: Springer.

Lukyanenko, R., & Parsons, J. (2015). Information quality research challenge: Adapting information quality principles to user-generated content. *ACM Journal of Data and Information Quality, 6*(1), 1-3.

Lukyanenko, R., Parsons, J., & Samuel, B. (2015b). Do we need an instance-based conceptual modeling grammar? In *Proceedings of the Symposium on Research in Systems Analysis and Design*.

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014a). The impact of conceptual modeling on dataset completeness: A field experiment. In *Proceedings of the International Conference on Information Systems* (pp. 1-18).

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014b). The IQ of the crowd: Understanding and improving information quality in structured user-generated content. *Information Systems Research, 25*(4), 669-689.

Lukyanenko, R., Parsons, J., & Wiersma, Y. (2016a). Editorial: Emerging problems of data quality in citizen science. *Conservation Biology, 30*(3), 447-449.

Lukyanenko, R., Parsons, J., Wiersma, Y., Sieber, R., & Maddah, M. (2016b). Participatory design for user-generated content: Understanding the challenges and moving forward. *Scandinavian Journal of Information Systems, 28*(1), 37-70.

Lukyanenko, R., Wiersma, Y., & Parsons, J. (2016c). *Is crowdsourced attribute data useful in citizen science? A study of experts and machines.* Paper presented at the Collective Intelligence.

March, S., & Allen, G. (2012). Toward a social ontology for conceptual modeling. In *Proceedings of the 11th Symposium on Research in Systems Analysis and Design* (pp. 57-62).

March, S., & Allen, G. (2015). Classification with a purpose. In *Proceedings of the Symposium on Research in Systems Analysis and Design* (pp. 1-10).

March, S. T., & Allen, G. N. (2014). Toward a social ontology for conceptual modeling. *Communications of the AIS, 34,* 1347-1358.

Mattessich, R. (2013). *Reality and accounting: Ontological explorations in the economic and social sciences.* New York: Routledge.

Mayden, R. L. (2002). On biological species, species concepts and individuation in the natural world. *Fish and Fisheries, 3*(3), 171-196.

McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*(4), 462-472.

McGinnes, S. (2011). Conceptual modelling for Web information systems: What semantics can be shared? In O. De Troyer, C. Bauzer Medeiros, R. Billen, P. Hallot, A. Simitsis, & H. Van Mingroot (Eds.), *Proceedings of the 30th International Conference on Conceptual Modeling* (vol. 6999, pp. 4-13). Berlin: Springer.

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology, 51*(1), 121-147.

Meijer, A., Burger, N., & Ebbers, W. (2009). Citizens4citizens: Mapping participatory practices on the internet. *Electronic Journal of e-Government, 7*(1), 99-112.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution, 27*(2), 85-93.

Mumford, E., & Henshall, D. (1979). *Designing participatively: A participative approach to computer systems design: A case study of the introduction of a new computer system*. Manchester, UK: Manchester Business School.

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Mylopoulos, J. (1998). Information modeling in the time of the revolution. *Information Systems, 23*(3-4), 127-155.

Nov, O., Arazy, O., & Anderson, D. (2014). Scientists@ home: What drives the quantity and quality of online citizen science participation. *PloS One, 9*(4), 1-11.

Oberhauser, K., & Prysby, M. D. (2008). Citizen science: Creating a research army for conservation. *American Entomologist, 54*(2), 103-104.

Olivé, A. (2007). *Conceptual modeling of information systems*. Berlin: Springer.

Palacios, M., Martinez-Corral, A., Nisar, A., & Grijalvo, M. (2016). Crowdsourcing and organizational forms: Emerging trends and research implications. *Journal of Business Research, 69*(5), 1834-1839.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419.

Parsons, J. (2003). Effects of Local Versus Global Schema Diagrams on Verification and Communication in Conceptual Data Modeling. *Journal of Management Information Systems, 19*(3), 155–184.

Parsons, J., Lukyanenko, R., & Wiersma, Y. (2011). Easier citizen science is better. *Nature, 471*(7336), 37-37.

Parsons, J., & Wand, Y. (1997). Using objects for systems analysis. *Communications of the ACM, 40*(12), 104-110.

Parsons, J., & Wand, Y. (2000). Emancipating instances from the tyranny of classes in information modeling. *ACM Transactions on Database Systems, 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.

Patel-Schneider, P. F., & Horrocks, I. (2007). A comparison of two modelling paradigms in the semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*(4), 240-250.

Peckham, J., & Maryanski, F. (1988). Semantic data models. *ACM Computing Surveys, 20*(3), 153-189.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Perszyk, K. J. (2013). *Nonexistent objects: Meinong and contemporary philosophy*. Berlin: Springer.

Phan, M., Thomas, R., & Heine, K. (2011). Social media and luxury brand management: The case of Burberry. *Journal of Global Fashion Marketing, 2*(4), 213-222.

Pokorny, J. (2013). NoSQL databases: A step to database scalability in Web environment. *International Journal of Web Information Systems, 9*(1), 69-82.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems, 32*(3), 229-267.

Prestopnik, N. R., & Tang, J. (2015). Points, stories, worlds, and diegesis: Comparing player experiences in two citizen science games. *Computers in Human Behavior, 52*, 492-506.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.

Prpić, J., Shukla, P. P., Kietzmann, J. H., & McCarthy, I. P. (2015). How to work a crowd: Developing crowd capital through crowdsourcing. *Business Horizons, 58*(1), 77-85.

Rai, A. (2016). Editor's comments: Synergies between big data and theory. *MIS Quarterly, 40*(1), iii-ix.

Rao, L. (2011). Crowdflower raises $7M, launches e-commerce tool for data categorization. *Techcrunch*, 1-4.

Raymond, E. S. (2001). *The cathedral & the bazaar: Musings on Linux and open source by an accidental revolutionary*. Sebastopol, CA: O'Reilly Media.

Recker, J. (2015). Research on conceptual modelling: Less known knowns and more unknown unknowns, please (vol. 165, pp. 3-8). In *Proceedings of the 11th Asia-Pacific Conference on Conceptual Modelling*.

Robal, T., Haav, H.-M., & Kalja, A. (2007). Making Web users' domain models explicit by applying ontologies. In J.-L. Hainaut (Ed.), *Advances in conceptual modeling-foundations and applications* (pp. 170-179). Berlin: Springer.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hoboken, NJ: John Wiley & Sons.

Rossi, M., & Siau, K. (2000). *Information modeling in the new millennium*. Hershey, PA: IGI Global.

Roussopoulos, N., & Karagiannis, D. (2009). Conceptual modeling: Past, present and the continuum of the future. In A. Borgida, C. V, P. Giorgini, & E. Yu (Eds.), *Conceptual modeling: Foundations and applications* (Vol. 5600, pp. 139-152). Berlin: Springer.

Russell, M. A. (2013). *Mining the social Web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. Sebastopol, CA: O'Reilly Media.

Sabegh, M. A. J., Recker, J., & Green, P. (2016). Designing experiments to test the theory of combined ontological coverage. In *Proceedings of the International Conference on Information Systems.*

Samuel, B. M., Watkins, L., Ehle, A., & Khatri, V. (2015). Customizing the representation capabilities of process models: Understanding the effects of perceived modeling impediments. *Software Engineering, IEEE Transactions on, 41*(1), 19-39.

Scholl, B. J. (Ed.). (2002). *Objects and attention*. Cambridge, MA: MIT Press.

Searle, J. R. (1995). *The construction of social reality*. New York: Simon and Schuster.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614-622).

Sheppard, S., Wiggins, A., & Terveen, L. (2014). Capturing quality: Retaining provenance for curated volunteer monitoring data. In *Proceedings of the ACM Conference on Computer Supported Cooperative work & Social Computing* (pp. 1234-1245).

Simitsis, A., & Vassiliadis, P. (2008). A method for the mapping of conceptual designs to logical blueprints for ETL processes. *Decision Support Systems, 45*(1), 22-40.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web Companion* (pp. 1049-1054).

Smith, A. M., Lynn, S., & Lintott, C. J. (2013). An introduction to the Zooniverse. In *Proceedings of the 1stt AAAI Conference on Human Computation and Crowdsourcing*.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, J. M., & Smith, D. C. P. (1977). Database abstractions: Aggregation and generalization. *ACM Transactions on Database Systems, 2*(2), 105-133.

Smith, L. B. (2005). Emerging ideas about categories. In L. Gershkoff-Stowe & D. H. Rakison (Eds.), *Building object categories in developmental time* (pp. 159-173). Mahwah, NJ: L. Erlbaum Associates.

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1-8).

Stevens, M., Vitos, M., Altenbuchner, J., Conquest, G., Lewis, J., & Haklay, M. (2014). Taking participatory citizen science to extremes. *Pervasive Computing, IEEE, 13*(2), 20-29.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*(5), 479-491.

Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research, 23*(1), 23-41.

Teorey, T. J., Yang, D., & Fry, J. P. (1986). A logical design methodology for relational databases using the extended entity-relationship model. *ACM Computing Surveys, 18*(2), 197-222.

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E.., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A., & Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation, 181*, 236-244.

Topi, H., & Ramesh, V. (2002). *Human factors research on data modeling: A review of prior research, an extended framework and future research directions*. *Journal of Database Management*, *13*(2), 3-19.

Tremblay, M. C., Dutta, K., & Vandermeer, D. (2010). Using data mining techniques to discover bias patterns in missing data. *Journal of Data and Information Quality, 2*(1), 1-19.

Van Pelt, C. R., Cox, R., & Sorokin, A. (2012). *Dynamic optimization for data quality control in crowd sourcing tasks to crowd labor* (United States patent application, US 13/428,708).

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition, 27*, 62-75.

Vincent, J. (2016). ZTE is ready to hear your ideas for its crowdsourced smartphone. *The Verge*. Retrieved from http://www.theverge.com/circuitbreaker/2016/8/3/12366474/zte-crowdsourced-smartphone-project-csx

Walker, R. C. S. (1989). *The coherence theory of truth: Realism, anti-realism, idealism*. New York: Routledge

Wand, Y., Monarchi, D. E., Parsons, J., & Woo, C. C. (1995). Theoretical foundations for conceptual modelling in information systems development. *Decision Support Systems, 15*(4), 285-304.

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

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. *Decision Support Systems, 13*(3-4), 349-372.

Wattal, S., Schuff, D., Mandviwalla, M., & Williams, C. B. (2010). Web 2.0 and politics: The 2008 U.S. presidential election and an e-politics research agenda. *MIS Quarterly, 34*(4), 669-688.

Weber, R. (2012). Evaluating and developing theories in the information systems discipline. *Journal of the Association for Information Systems, 13*(1), 1-30.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of predictive text mining* (vol. 41). Berlin: Springer.

Wengraf, T. (2001). *Qualitative research interviewing: Biographic narrative and semi-structured methods*. Thousand Oaks, CA: Sage.

Whelan, E., Teigland, R., Vaast, E., & Butler, B. (2016). Expanding the horizons of digital social networks: Mixing big trace datasets with qualitative approaches. *Information and Organization, 26*(1), 1-12.

Whitla, P. (2009). Crowdsourcing and its application in marketing activities. *Contemporary Management Research, 5*(1), 15-28.

Wiersma, Y. F. (2010). Birding 2.0: citizen science and effective monitoring in the Web 2.0 world. *Avian Conservation and Ecology, 5*(2), 13.

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., LeBuhn, G., Litauer, R., Lots, K., Michener, W., & Newman, G. (2013). Data management guide for public participation in scientific research. *DataOne Working Group*. Retrieved from http://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf

Winter, S., Berente, N., Howison, J., & Butler, B. (2014). Beyond the organizational "container": Conceptualizing 21st century sociotechnical work. *Information and Organization, 24*(4), 250-269.

Yin, R. K. (2013). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.

Zhao, Y., & Han, Q. (2016). Spatial crowdsourcing: Current state and future directions. *IEEE Communications Magazine, 54*(7), 102-107.

Zwass, V. (2010). Co-Creation: Toward a taxonomy and an integrated research perspective. *International Journal of Electronic Commerce, 15*(1), 11-48.

Appendix A

Table A1. Examples of Organization-directed UGC Projects

Sector	Typical crowd tasks	Projects sample (including sponsoring organizations)
Private / commercial	Product improvement, ideation, Classification of product items	• The New York Times (identification of ads in old newspapers, http://madison.mytimes.com) • Burberry Inc. (product ideation and development) (Phan, Thomas, & Heine, 2011) • Procter & Gamble Inc. (feedback on product experiences, www.beinggirl.com); Lumenogic, LLC (crowd-based predictions, www.lumenogic.com)
Community management, public policy	Reporting on civic issues, help in disaster response	• CitySourced.com (USA) (citizen reports of local crime, graffiti, potholes, broken street lights, (DeMeritt, 2011)) • Fix My Community (report the need of local maintenance and repairs), including FixMyStreet.com (UK), Fixmystreet.org.au (Australia), FixMyCommunity.ug (Uganda), Aduanku.my (Malaysia), Cuida Alcalá (http://cuida.alcala.org), Fixamingata.se (Sweeden) • Ushahidi.com (crisis information, social activism and public accountability) (Gao, Barbier, & Goolsby, 2011)
Healthcare	Patient symptom reports, doctor and hospital ratings	• RateMD.com (patient ratings of physicians) (Gao et al., 2015; Kadry, Chu, Kadry, Gammas, & Macario, 2011) • US Food and Drug Administration's MedWatch (patient reports on drugs and medical devices) (Craigle, 2007; Kessler et al., 1993)
Science	Describing and classifying, reporting on observations, transcribing data	• Cornell University (www.eBird.org) – reporting bird sightings • Oxford University and others: GalaxyZoo (www.galaxyzoo.org) – identifying galaxies from photos (Lintott et al., 2008) • University of Washington and others: Fold.It (protein folding to develop new drugs) (Khatib et al., 2011)

Appendix B: Select NLNature Screenshots

Below we provide the screenshots of the NLNature relevant to the case discussion above.

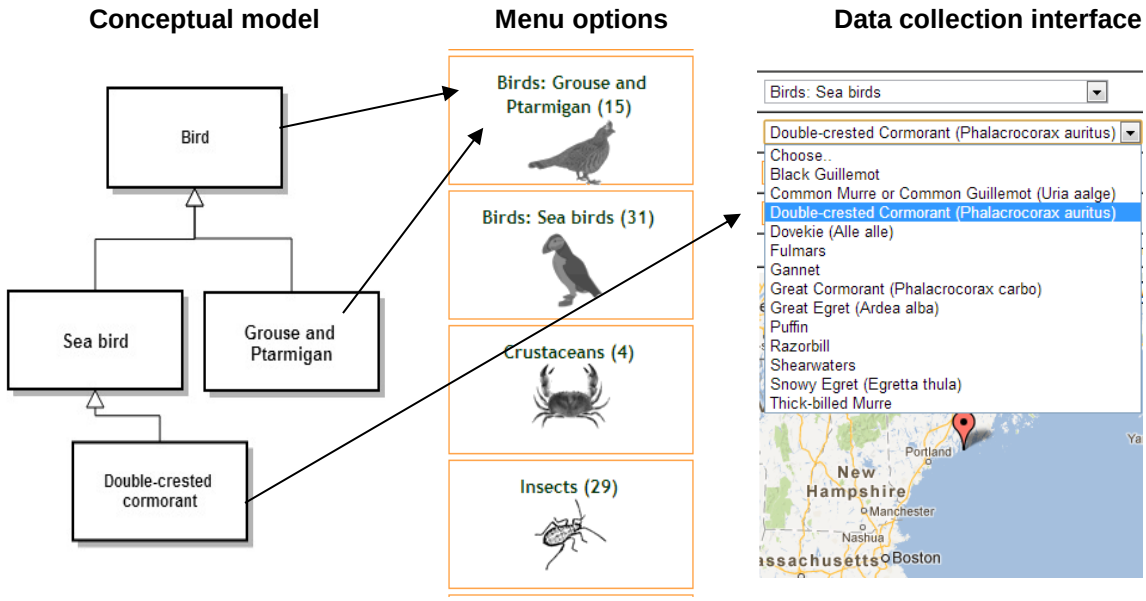


Figure B1. Conceptual Model Fragment and User Interface Elements Based on the Model in Phase 1 NLNature

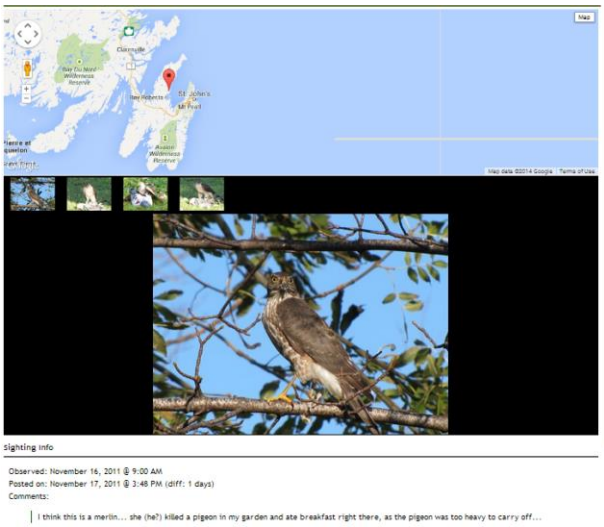
Screenshot of the observation	Public correspondence between the observation creator, Lynette, and another user, Timothy.	
	Lynette Nov. 17 2011	I think this is a merlin... she (he?) killed a pigeon in my garden and ate breakfast right there, as the pigeon was too heavy to carry off...
	Timothy July 28 2012	Actually an accipiter. Sharpshinned hawk
	Lynette July 28 2012	Thank-you, Timothy! I'm an amateur, I Was guessing as to what it was!

Figure B2. A Vignette of an Observation Classified as Merlin (*Falco columbarius*) where the Observation Creator Admits to Guessing

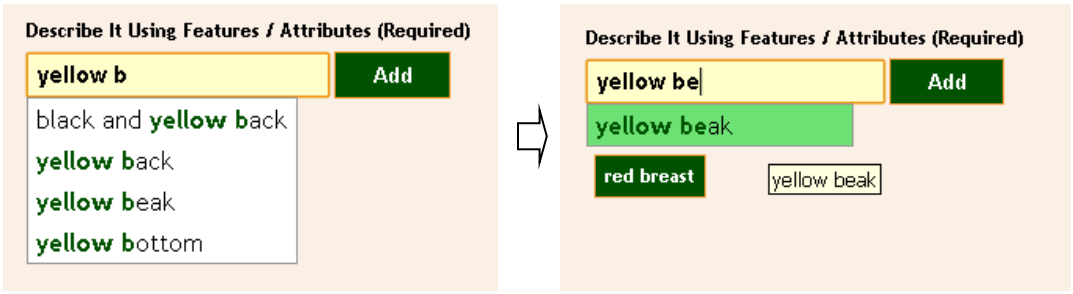


Figure B3. Example of Data Collection in Phase 2

Figure B4. NLNature Phase 2 Data-entry Interface

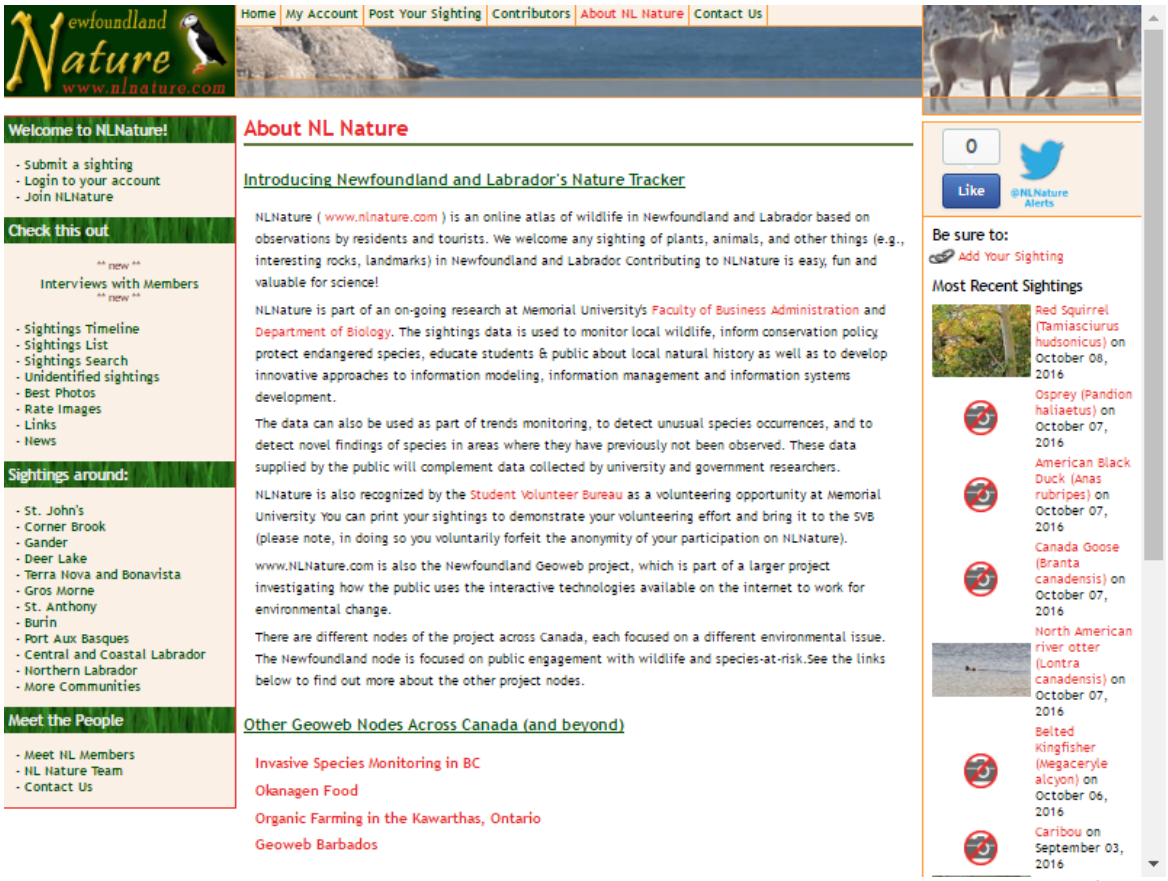


Figure B5. The "About Us" Page on NLNature Phase 2 Describing the Focus of the Project[9]

[9] Arrow and rectangle highlight the TOM cues, including indications of the project's scope and its objectives.

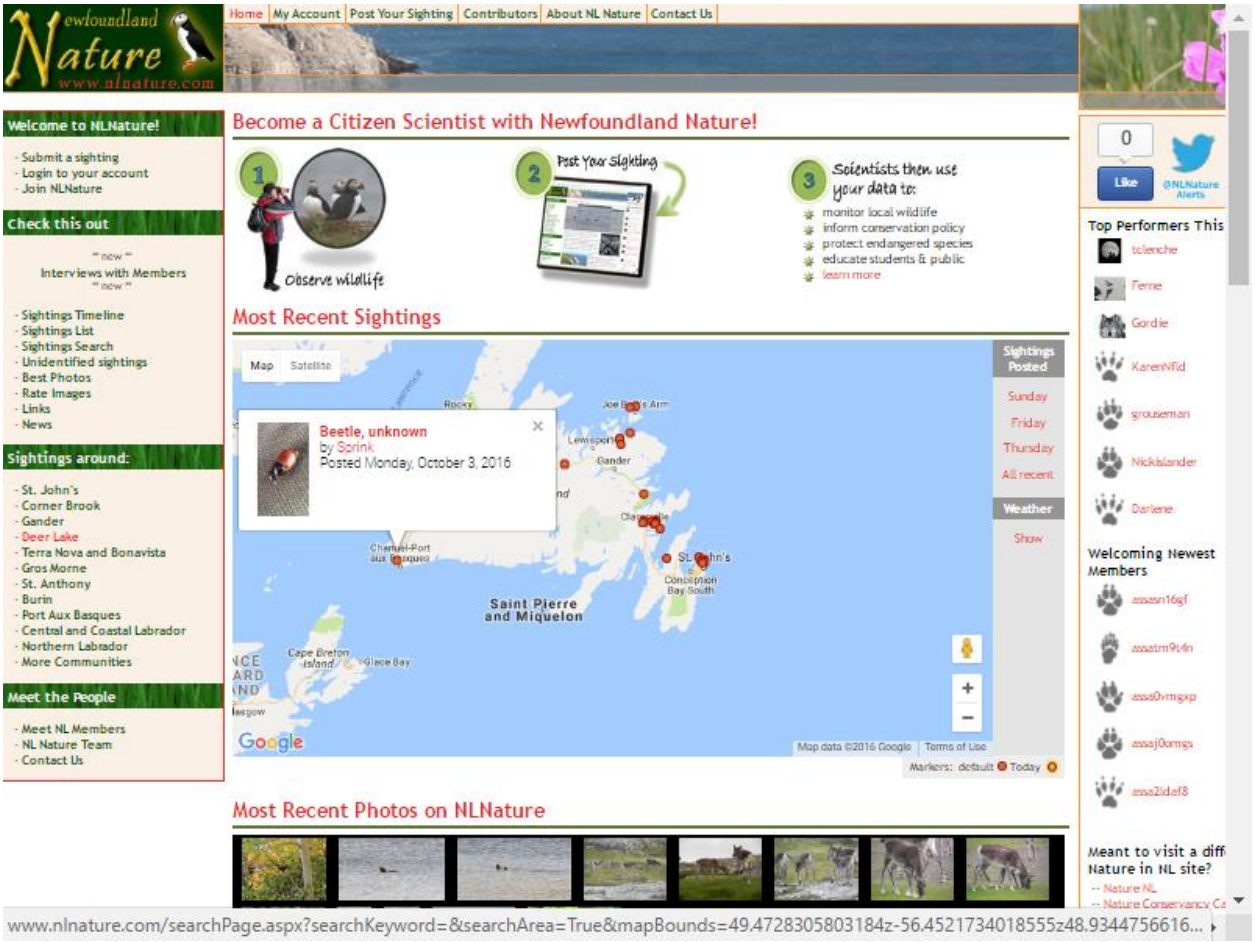


Figure B6. Redesigned Front Page of NLNature (Public View)

Appendix C: Evaluation of the Design Theoretic Contribution

Because the proposed guidelines constitute a nascent design theory (Gregor & Hevner, 2013), we adopt the four most common criteria used in evaluating theories in IS: relevance, novelty, clarity, and usefulness (Alter, 2013; Gregor & Jones, 2007; Kuhn, 1977; Weber, 2012).

Relevance: as we discuss the introduction, our guidelines address an important emerging challenge in conceptual modeling: how to model information systems used to collect UGC for organizational purposes. As people are becoming more comfortable in creating content, UGC is turning into a new capital that organizations can leverage (Brynjolfsson & McAfee, 2014; Prpić et al., 2015). Yet, no established principles for harnessing this socially important resource exist. As such, we believe this paper addresses an important and socially relevant issue.

Novelty: according to Gregor and Hevner (2013), one can classify design science contributions into four categories based on solution and application domain maturity: improvement, invention, routine design, and exaptation. Our work falls into the category of inventions as it provides "new solutions for new problems" (Gregor & Hevner, 2013, p. 345). User-generated content is a new and radically different setting compared to traditional settings in which information was commonly produced (Lukyanenko & Parsons, 2015). Working in this new setting, we proposed an approach to conceptual modeling radically different from the traditional approaches premised on a priori specification of domain specific abstractions. Doing so requires rethinking the fundamental assumptions behind conceptual modeling and introduction of several new conceptual modeling concepts, including representational uniqueness, hybrid structured and unstructured modeling, target organizational models, and TOM cues. The high novelty of our work paves the way for a promising new direction in future conceptual modeling research and practice.

Clarity: we articulate the design guidelines carefully and systematically. In particular, we start by identifying the four challenges faced when modeling UGC. We then adopt ontology and cognition as to reference fields to address the challenges. We strictly derived each proposed challenge from the underlying theoretical foundations, and the guidelines are complete and sufficient in so far as they both exhaust the main theoretical basis and the four target challenges. To further clarify the guidelines, below, we provide a detailed demonstration of the application of the proposed guidelines when developing a real UGC IS.

Usefulness/utility: to provide evidence of the usefulness of the proposed guidelines in addressing the modeling challenges in UGC, we draw on a case study of real IS development based on these guidelines and interviews with real users after two years of using the project derived from these guidelines. Both the case in the interviews provide strong evidence of the utility of the proposed guidelines.

Appendix D: Major Projects that Implement or Support Traditional Conceptual Modeling

Project	Information collection objective	Pre-specified abstractions used for data collection	Means of capturing additional information	References containing discussion of the modeling architecture
eBird www.ebird.org	Sightings of birds across the world	List of bird species to select from	Free-form comment box, email	(Bonney et al., 2009; He & Wiggins, 2015; Hochachka et al., 2012; Kelling et al., 2011)
CitySourced www.citysourced.com/	Reports on urban and civic issues	List of categories (e.g., crime, graffiti, broken street light) to select from	Free-form comment box, email	(DeMeritt, 2011)
GalaxyZoo www.galaxyzoo.org/	Classification of galaxies from digital images	List of pre-specified galaxy images to match form, size of galaxies	List of additional pre-specified galaxy images to report any "odd features"; free-form comment box, discussion forum, email	(Fortson et al., 2011; Lintott et al., 2008; Lukyanenko et al., 2016a)
Zooniverse https://www.zooniverse.org	Platform for volunteer-based data collection projects	Allows defining project-specific pre-specified abstractions	Features as in GalaxyZoo	(Borne & Team, 2011; Fortson et al., 2011; Simpson et al., 2014; Smith, Lynn, & Lintott, 2013)
Amazon Mechanical Turk www.mturk.com	Platform for paid on-demand data collection	Default templates suggest creating project-specific pre-specified abstractions but also supports free-form tagging, free textboxes, and custom scripting of data collection elements		(Ipeirotis, Provost, & Wang, 2010; Sorokin & Forsyth, 2008)
CrowdFlower www.crowdflower.com	Platform for paid on-demand data collection	Default templates suggest creating project-specific pre-specified abstractions but also supports free-form tagging, free textboxes, and custom scripting of data collection elements		(Rao, 2011; Van Pelt, Cox, & Sorokin, 2012)

Appendix E: Interviews and Focus Groups Protocol Details

Introduction (30 minutes or less). Welcoming, organization, background of the facilitators (two of the six co-authors of the paper), names of the group members, confidentiality, asking for agreement for note taking, audio record and transcription (names are coded by numbers).

Free discussion about the participants, their motivation and general experience on NLNature (approx. 30 minutes):

Leading question:

1. Why do you participate on NLNature?

Possible additional questions:

1. How did you discover the website?
2. What is your main reason of using the website?
3. What are the most useful aspects of the website?
4. How do you enjoy being in the outdoors?
5. How much time do you spent being in the outdoors?
6. If you look for information on the website, did you get it?
7. If you asked questions, did somebody answer?
8. What are the groups of animals/plants you are interested in?
9. Do you report sightings from hiking, fishing, your profession?
10. Do you target a specific taxonomic group?
11. Why did you choose this group?

Questions about NLNature to ascertain the utility of the Phase 2 design (approx. 30 minutes):

1. What do you mean, when you think of quality of information? What kind of quality would be needed?
2. How do you participate? (How often? Are you systematic? Do you log sightings, or participate mainly through commenting?)
3. What features of the website do you use?
4. What information do you get out of the website?
5. Is it easy to search/ to find stuff?
6. What is your perception of the quality of the provided information?
7. Do you think you have to report the species?
8. How do you feel about the fact that you don't have to identify the species?
9. Are attributes useful?
10. Why do you not contribute the attributes?
11. What do you think is important to describe in your favorite group?
12. Do you have suggestions for making the website better?
13. When we think about redesigning the website, what can we do?
14. What else can we improve?

Free discussion about problem framing and agenda setting of future of projects like NLNature (approx. 30 minutes):

Leading question:

1. What are the main topics you think science should address in context of NLNature?

Possible additional questions:

1. What do you think how the data on NLNature are used or might be used?
2. What is your definition of a "natural area'?

3. What is the name “NLNature” about? What is your understanding of the term “nature” within the name of the website that you are contributing to?
4. What is your impression of “wilderness”?
5. What would you think are the main sources of environmental problems?
6. What do you think is the most pressing issue facing plants and animals in your area/ in the province?
7. What is your background/profession?
8. Do you participate in other online citizen science projects, and if so, how does their participation and experience compare to that on NLNature?
9. How does the nature of participation in NLNature compare to participation by citizens in other projects?
10. Do you perceive yourself as scientists?

Wrap-up, feedback, thanks and goodbye

Appendix F: Details on Interview and Focus Group Participants

Interviewee or participant no.	Gender	Age category	Profession / industry	Background / expertise in biology/ecology	Hometown
1	Male	Adult	Computer programmer	No professional background in biology/ecology	Small town / rural area
2	Female	Adult	Retail	No professional background in biology/ecology	Small town / rural area
3	Male	Adult	GIS-systems and geography	Biology minor at college, but mainly self-trained	No information
4	Female	Adult	Stay-home mom	No professional background in biology/ecology	Small town / rural area
5	Male	Adult	No information	Involved with the bee keeping association	No information
6	Female	Adult	No information	No information	No information
7	Male	Adult	Software developer	No professional background in biology/ecology	No information
8	Male	Adult	Social Scientist	Studied Biology before he switched to Social Science	Small town / rural area
9	Female	Adult	Marketing	Gardener	No information
10	Male	Adult	Marketing	No professional background in biology/ecology	No information
11	Male	Adult	Forester	Forester, working for a natural resource management government agency	No information
12	Male	Adult	Fishery	No professional background in biology/ecology	Small town / rural area
13	Male	Adult	Fishery	No professional background in biology/ecology	No information
14	Female	Adult	Communications, deputy mayor	No professional background in biology/ecology	No information
15	Male	Adult, Senior Citizen	No information	No professional background in biology/ecology	Small town / rural area

About the Authors

Roman Lukyanenko is an assistant professor of information systems at Edwards School of Business, University of Saskatchewan. He received his PhD in information systems from Memorial University of Newfoundland. His research interests include citizen science, crowdsourcing, information (data) quality, conceptual modeling, and business analytics. His research has been published in *Nature, Information Systems Research, Scandinavian Journal of Information Systems, Conservation Biology, ACM Journal of Data and Information Quality* as well as leading conferences in information systems and computer science.

Jeffrey Parsons is University Research Professor and Professor of Information Systems in the Faculty of Business Administration at Memorial University of Newfoundland. He holds a Ph.D. in Information Systems from the University of British Columbia. His research interests include conceptual modeling, data management, crowdsourcing, and recommender systems; he is especially interested in classification issues in these and other domains. His research has been published in journals such as *Nature*, *Management Science*, *MIS Quarterly*, *Information Systems Research*, *Journal of Management Information Systems*, *Communications of the ACM*, *ACM Transactions on Database Systems*, and *IEEE Transactions on Software Engineering*. He is a senior editor at *MIS Quarterly* and has served as a senior editor at the *Journal of the Association for Information* Systems and an associate editor at *Information Systems Research*

Yolanda Wiersma is an Associate Professor in Landscape Ecology in the Department of Biology, Memorial University of Newfoundland. Her research interests include applications of landscape ecology to forest and wildlife management, developing metrics to measure landscape connectivity, spatial pattern analysis and citizen science. Her research has been published in a variety of journals including *Landscape Ecology*, *Conservation Biology*, *BioScience*, and *Biological Conservation*. Her research germane to citizen science is published in *Nature*, *Information Systems Research,* and the *Scandinavian Journal of Information Systems*.

Gisela Wachinger holds a PhD in Biology (HIV modeling) from the Ludwig Maximilians University of Munich. She was a co-founder of, and a teacher in, the Master Program Urban Planning and Citizen Participation (M.Sc.). She is working as a project leader at the DIALOGIK non-profit institute for communication and cooperation research, leading a number of public participation and conflict resolution projects. Her research has appeared in international and German journals (e.g., *FEBS letters*, *Soil Biology and Biochemistry*, and *Risk Analysis*). Her main research interests include mediation and participation in the areas of science, health and the environment, risk analysis for natural hazards, and citizen science in biodiversity and climate change research.

Robert Meldt is a Master's (Urban Planning and Citizen Participation, MSc) student at the University of Stuttgart. He worked as a graduate assistant for ZIRIUS, the Stuttgart Research Center for Interdisciplinary Risk and Innovation Studies on topics of citizen participation in Politics and Science. He is currently writing his master's thesis on motivation in citizen scientists.

Benjamin Huber is a Master's (Urban Planning and Citizen Participation, MSc) student at the University of Stuttgart.