

Conceptual Modeling Principles for Crowdsourcing

Roman Lukyanenko
Faculty of Business Administration
Memorial University
St. John's, Canada
roman.lukyanenko@mun.ca

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

ABSTRACT

Traditionally, the research and practice of conceptual modeling assumed all relevant information about a domain could be discovered through user-analyst communication. The increasing ubiquity of crowdsourcing challenges a number of long-held propositions about conceptual modeling. Given significant differences in levels of domain expertise among contributors in crowdsourcing projects, it is often impossible to predict all valid conceptualizations of a domain by potential users. Approaching conceptual modeling in crowdsourcing using traditional principles of modeling is highly constraining. This paper explores fundamental conceptual modeling challenges in crowdsourcing domains. We then use theoretical foundations in philosophy (ontology) to offer potential solutions.

Categories and Subject Descriptors

D.3.3 [Information Systems]: Models and Principles – *general*.

General Terms

Management, Design

Keywords

Conceptual modeling, User-generated content, Crowdsourcing

1. INTRODUCTION

Crowdsourcing refers to the engagement of the general public in specific tasks defined by project owners or their clients [1, 2]. Perhaps the single largest product of crowdsourcing is user-generated information - various forms of digital content (e.g., product reviews, tags, bird observations, wikis), *typically created for a predefined purpose* by participants of crowdsourcing projects. Crowdsourcing user-generated information promises to expand organizational “sensor” networks, making it possible to collect large amounts of data from highly diverse audiences. For example, Cornell University’s eBird (www.ebird.com) collects bird sightings from online contributors around the world [3] and Amazon Mechanical Turk (www.mturk.com) enlists users to perform tasks for their clients (e.g., classify products). Crowdsourcing increasingly supports traditional business operations, emergency management, healthcare, politics/governance and other activities [1, 4-6].

Despite growing social and economic importance, little is known about how to conduct analysis and create *conceptual models* of crowdsourcing systems. Conceptual models of a domain –

informal or formal representations of knowledge about the kinds of phenomena in the domain – are central to information systems (IS) development. Conceptual models support communication between developers and users, guide database and application design, formally document system requirements and promote domain understanding [7-9]. It is widely contended that conceptual models increase effectiveness and efficiency of IS development and lead to design of better systems [7, 10].

We argue traditional conceptual modeling approaches are inadequate in crowdsourcing environments. Crowdsourcing reveals an important gap in conceptual modeling research and motivates a search for novel forms of domain representations potentially useful in other (including corporate) domains.

A fundamental tenet in conceptual modeling is reliance on *users* (and other stakeholders) for elicitation and verification of domain semantics to be captured by conceptual models. Maintaining close contact with users and project stakeholders is a commonly prescribed guideline [8, 10, 11]. Indeed, one of the roles of conceptual models is to facilitate communication between users and analysts [12]. Accordingly, it has been suggested that “poor execution of [user requirement elicitation] will almost guarantee that the final project is a complete failure” [13, p. 66].

User-driven conceptual modeling originated in corporate IS development, where potential users of a system were presumed to be known in advance and accessible for intensive developer–user dialogs and conceptual model verification [14, 15].

Unlike the traditional corporate setting, in crowdsourcing projects there are typically no constraints on who can participate and engaging broader audiences is often highly desirable. Given the diversity of users, many of them may *lack domain expertise* and may have incongruent views or conceptualizations of issues in a domain. In this context, many questions arise, including: (how) can conceptual models faithfully represent the resulting diversity of views and accommodate variable level of expertise? How do crowdsourcing domains change the role of conceptual models as a facilitator of user-designer communication? What type of information should a conceptual model specify? It is widely believed that poorly executed analysis leads to unnecessary costs and project failures [7, 16]. Thus, addressing these questions can have a significant impact on crowdsourcing systems and the organizations they increasingly support.

This paper explores fundamental conceptual modeling challenges in crowdsourcing domains. We then use theoretical foundations in philosophy (ontology) to identify the root cause of these challenges and derive potential solutions.

2. CONCEPTUAL MODELING CHALLENGES IN CROWDSOURCING IS

Traditionally research in systems analysis was conducted under the assumption that relevant information about a domain is known

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CrowdSens '12, November 2, 2012, Maui, Hawaii, USA.

Copyright 2012 ACM 978-1-4503-1715-3/12/11...\$15.00.

in advance and typically comes from users, who are also subject-matter experts [14]. These assumptions reflect a corporate environment with common training, instructions, and shared corporate culture. In contrast, the characteristics of rapidly proliferating crowdsourcing domains challenge existing modeling practices. Below, we outline premises of traditional systems analysis that we argue are problematic for crowdsourcing systems.

Direct involvement of subject-matter experts who are typically end-users. Traditionally, analysts rely on users (or, more generally, stakeholders) for subject-matter expertise and system requirements. The information is typically elicited through direct contact with end-users or their representatives (e.g., supervisors, team leaders). Based on user requirements, domain semantics is represented in a conceptual model [12].

In contrast, in crowdsourcing there may be no constraints on who can participate and engaging broader audiences is often desirable. While crowdsourcing projects are established primarily to serve the needs of project owners or their clients (subject matter experts), the actual users or contributors are ordinary people, often lacking subject matter expertise and possessing diverse domain views [3, 17]. Indeed, such broad participation raises many concerns about the quality of crowdsourced data [3, 18-22]. In addition to the difficulty in reaching distributed online users, it is unclear how to determine which user requirements are appropriate to consider and how to engage users with varying levels of expertise in model verification activities.

Unified global schema. Traditionally, close contact with users provided an opportunity to resolve conflicts in individual views and conceptualizations: any “conflict must be solved through communication among people” [23, p. 250]. A final conceptual model, therefore, typically represents a global, integrated view of a domain but often does not represent the view of any individual user [24]. This parallels an organizational process of reaching a collective judgment through dialog, negotiation or specialized techniques [e.g., 25]. The global schema is then “the basis for understanding by all users and applications” [26, p. 140].

In a crowdsourcing environment, reaching all potential users and reconciling their views is not generally possible. Consider a sample conceptual model of a natural history citizen science project (see Figure 1). Citizen science is a type of crowdsourcing which aims to use contributions of online users for scientific research [27-29]. Figure 1 depicts potential generalization-specialization relationships among entities in this domain based on a mixture of biological nomenclature and general knowledge.

While Figure 1 appears to capture faithfully a typical domain view [see 30, 31], it is deficient as it may not represent or even be congruent with the views of many users (citizen scientists). Different potential contributors may organize perceived reality using alternative criteria. For example, some may group birds by sound or appearance. Depending upon the criteria chosen, issues of reconciling different taxonomies may arise. For example, *polar bears* may be classified by some non-experts as *land*, rather than *marine* mammals. Similarly, some users may advocate different, but valid, conceptualizations of the instances represented by classes in Figure 1. For example, given the same instance, an *osprey*, some users may classify it as *bird*, *predator*, *fish hawk*, or *osprey*. Each classification may reflect varying individual domain expertise (e.g., non-expert contributors may not be familiar with *osprey* and be more comfortable with the more general concept of *bird*). Traditional approaches to modeling do not provide sufficient guidance to the problem of reconciling extreme diversity of user views.

Representation by abstraction. The fundamental approach to conveying domain semantics in a conceptual model is representation by abstraction [8, 32]. Abstraction enables analysts to deliberately ignore the many individual differences among phenomena and represent only relevant generalized information. A typical script made using the popular entity-relationship (ER) grammar, for example, may depict classes (entity types, categories), attributes of classes, and relationships (e.g., Figure 1). Classes (e.g., *student*, *tree*, *bird*) abstract from differences among instances (e.g., a *particular student*, or a *specific bird*), instead capturing perceived commonalities of instances.

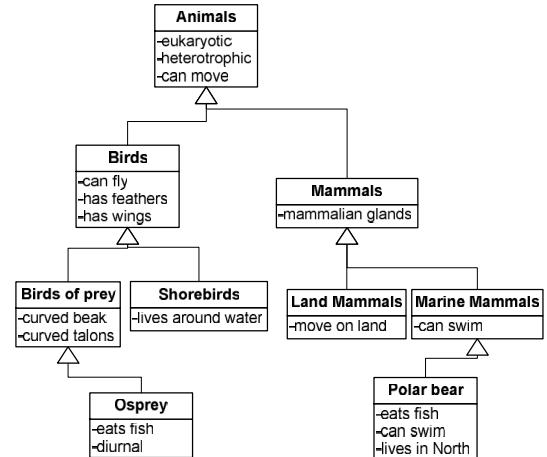


Figure 1. Hypothetical conceptual model in citizen science

As seen from the citizen science discussion, in crowdsourcing domains a conceptual model representing information deemed relevant by some users may marginalize or exclude potentially valuable conceptualizations of other users. As a result the final information system may preclude users whose views are not adequately accommodated from effectively engaging and contributing [33]. In the next section we identify the root cause of these challenges and explore a potential solution.

3. TOWARD CONCEPTUAL MODELING PRINCIPLES OF CROWDSOURCING IS

We contend that the challenges of reaching user consensus and developing a unified global schema in crowdsourcing domains point to *the fundamental deficiency of abstraction-based models to represent semantics in crowdsourcing systems*. Abstraction-based conceptual models depict generalized or *stylized* [34, p. 178] representations of perceived concrete phenomena. For example, some subject matter experts in biology may specify a *bird* class, assume all instances of this class possess attributes *has feathers*, *has wings*, and *can fly*, and contend that other attributes are irrelevant (see Figure 1). They may further stipulate constraints on the data types for attributes and prescribe particular relationships for all *bird* instances (e.g., generalization relationships as in Figure 1).

In crowdsourcing systems each contributor may have very different ways of abstracting reality. According to research in human psychology, differences in prior experience, domain expertise, conceptualization, and ad hoc utility, may result in different abstractions of the same domain between contributors and for the same contributor over time [35-37]. Unlike in more homogeneous corporate environments, however, in crowdsourcing domains *discovering and reconciling* individual models is unrealistic.

To address these issues some consider folksonomies or collaboratively-created domain-specific ontologies [38]. Such approaches tend to encapsulate diverse user perspectives, yet even these adaptive approaches may neglect all valid views and inadvertently inhibit domain understanding [see 39]. Another promising approach is “*schema during*”, in which users are free to alter the data model [26]. This, however, invites unresolved issues of cooperative schema evolution and concurrent access and modification of schemas. It is also unclear if “*schema during*” is reliable online, as some users may lack skills and motivation to create and alter abstract data models [17].

Given that anticipating and reconciling individual models of domain users appears unrealistic in crowdsourcing systems, we argue for **representations that are, to the extent possible, concrete and observer-independent**. To derive principles for such representation, we turn to ontology.

Ontology, a branch of philosophy that studies what exists in reality, is actively used in conceptual modeling research to establish a substantive basis for creating a shared understanding among parties involved in conceptual modeling [40, 41]. To illustrate an ontological approach, we use the general ontology of Mario Bunge [42] that specifies fundamental elements of existence assumed to be observer independent [43, p. 4].

According to Bunge [42] the world consists of “things” (or *instances*), elementary ontological constructs. Every instance is unique in some way as it has distinct properties. Properties are always attached to instances and cannot exist without them. Properties can be *intrinsic* if they are inherent in things (e.g., *height* or *mass*) or *mutual* if they belong to more than one thing (e.g., *date hired* is a joint property of a person and a company). People perceive properties as *attributes*, which are basic abstractions of reality [see 44]. Attributes can be used to form higher-level derived abstractions (e.g., things with similar attributes can be grouped in *classes*).

Bunge’s ontology can be used to develop conceptual modeling principles and grammars to model crowdsourcing systems. Instances can be modeled directly by allowing different potential users to report *attributes* of instances in a domain. The constancy and identifiability of instances (e.g., same public figures or spatial objects) provides common ground for productive user-designer communication.

Conceptual models depicting instances and attributes can be used to guide database development, as they are compatible with the instance-based data model [45]. Instance-based approaches recognize the primacy of instances and attributes over classes and allow instances to be stored independent of classes. This enables seamless storage of heterogeneous information, and thus is a suitable structure for crowdsourcing systems [33].

Inspecting user-provided attributes in a conceptual model affords an early glimpse into what actual user contributions may look like. This may reveal potential data conflicts (e.g., *white breast* vs. *grey breast* - reported by two users for the same instance) or suggest the extent of information-richness in application domains (e.g., typical *birds* contain a greater number of perceivable attributes than typical *chairs*). Such early insights can inform developers and project owners so preemptive measures can be taken to properly collect and manage future data. Attributes can provide inferences to construct higher-level abstractions based on ad hoc needs (e.g., classes, hierarchies, relationship types).

4. DISCUSSION AND CONCLUSION

With the rising importance of crowdsourcing, the question of how to create conceptual models of crowdsourcing systems is a pressing one. Traditional conceptual modeling approaches appear ill-equipped to address emerging challenges of representing unpredictable user input.

In this paper we draw on theories of ontology to propose a solution that avoids the negative consequences of abstraction-based models. Instance-based conceptual models enable one to capture basic elements of a domain as perceived by individual users. Instances and attributes promise to provide early insights into the domain and inform subsequent data management. To this extent, the proposed solution reinforces the traditionally critical role of conceptual models in crowdsourcing system development.

At the same time, the proposed approach leaves a number of important questions unresolved. Despite flexibility in representing individual perspectives, conceptual models of concrete instances and attributes are ineffective in capturing the full scope of a domain. Using specific instances and attributes allows depicting a sample of perceived reality, but can never provide a complete view: it is infeasible to show all instances and attributes past and future. Instances and attributes are also unintuitive to use in communication. It is unnatural for users to refer to instance *x* in terms of its attributes alone. Instead, it is more likely that users would intuitively refer to *x* using some *class* (e.g., *bird*, *osprey*, *shorebird*, *predator*, *object in the sky*). Finally, higher-level abstractions are fundamental to human beings and completely excluding them may impoverish instance-based conceptual models. Indeed, existing classes known to an organism may work as *filters* through which humans abstract properties of instances in the world, biasing perception [see 44]. This suggests that employing higher-level abstractions in a way that does not undermine domain representation and understanding may prove complementary to instance-based models. In future work, we intend to investigate how different approaches to domain representation may be combined to maximize their benefits and hope to develop a comprehensive suite of principles and grammar.

With this paper we hope to initiate a dialog in the research community about the issue of alternative forms of domain representation. As we show in this paper, with the continued proliferation of IS, it is becoming vital to ensure a proper fit between conceptual modeling approaches and the requirements and nature of the modeled domains.

5. REFERENCES

- [1] Doan, A., Ramakrishnan, R., and Halevy, A.Y. 2011. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM* 54, 4, 86-96.
- [2] Howe, J. 2006. The Rise of Crowdsourcing. In *Wired*.
- [3] Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K., and Kelling, S. 2012. Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution* 27, 2, 130-137.
- [4] Majchrzak, A.N.N. and More, P.H.B. 2011. Emergency! Web 2.0 to the Rescue! *Communications of the ACM* 54, 4, 125-132.
- [5] Andriole, S.J. 2010. Business Impact of Web 2.0 Technologies. *Communications of the ACM* 53, 12, 67-79.
- [6] Behrend, T., Sharek, D., Meade, A., and Wiebe, E. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43, 3, 800-813.

- [7] Wand, Y. and Weber, R. 2002. Research commentary: Information systems and conceptual modeling - A research agenda. *Information Systems Research* 13, 4, 363-376.
- [8] Mylopoulos, J. 1998. Information modeling in the time of the revolution. *Information Systems* 23, 3-4, 127-155.
- [9] Olivé, A. 2007. *Conceptual modeling of information systems*. Springer, Berlin Heidelberg New York.
- [10] Moody, D.L. 2005. Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering* 55, 3, 243-276.
- [11] Gould, J.D. and Lewis, C. 1985. Designing for usability: key principles and what designers think. *Communications of the ACM* 28, 3, 300-311.
- [12] Parsons, J. and Cole, L. 2005. What do the pictures mean? Guidelines for experimental evaluation of representation fidelity in diagrammatical conceptual modeling techniques. *Data & Knowledge Engineering* 55, 3, 327-342.
- [13] Hickey, A.M. and Davis, A.M. 2004. A Unified Model of Requirements Elicitation. *Journal of Management Information Systems* 20, 4, 65-84.
- [14] Lukyanenko, R. and Parsons, J., 2012. Conceptual modeling in open-world domains. In *Proceedings of the 11th Symposium on Research in Systems Analysis and Design* (Vancouver, Canada).
- [15] Browne, G.J. and Ramesh, V. 2002. Improving information requirements determination: a cognitive perspective. *Information & Management* 39, 8, 625-645.
- [16] Boehm, B.W. 1981. *Software engineering economics*. Prentice-Hall.
- [17] Coleman, D.J., Georgiadou, Y., and Labonte, J. 2009. Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research* 4, 332-358.
- [18] Mackechnie, C., Maskell, L., Norton, L., and Roy, D. 2011. The role of 'Big Society' in monitoring the state of the natural environment. *Journal of Environmental Monitoring* 13, 10, 2687-2691.
- [19] Arazy, O., Nov, O., Patterson, R., and Yeo, L. 2011. Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *Journal of Management Information Systems* 27, 4, 71-98.
- [20] Flanagan, A. and Metzger, M. 2008. The credibility of volunteered geographic information. *GeoJournal* 72, 3, 137-148.
- [21] Rowland, K. 2012. Citizen science goes 'extreme'. *Nature News* 17 February 2012.
- [22] Wiggins, A., Newman, G., Stevenson, R.D., and Crowston, K., 2011. Mechanisms for Data Quality and Validation in Citizen Science. In *Proceedings of the "Computing for Citizen Science" workshop* (Stockholm, SE).
- [23] Pohl, K. 1994. The three dimensions of requirements engineering: a framework and its applications. *Information Systems* 19, 3, 243-258.
- [24] 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.
- [25] Eden, C. and Ackermann, F. 1998. *Making Strategy: The Journey of Strategic Management*. Sage Publications.
- [26] Roussopoulos, N. and Karagiannis, D., 2009. Conceptual Modeling: Past, Present and the Continuum of the Future. In *Conceptual Modeling: Foundations and Applications*, A. Borgida, V. Chaudhri, P. Giorgini and E. Yu, Eds. Springer Berlin / Heidelberg, 139-152.
- [27] Silvertown, J. 2009. A new dawn for citizen science. *Trends in Ecology & Evolution* 24, 9, 467-471.
- [28] Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4, 211-221.
- [29] Wiggins, A. and Crowston, K., 2011. From Conservation to Crowdsourcing: A Typology of Citizen Science. In *44th Hawaii International Conference on System Sciences* (Hawaii, US).
- [30] Kaufman, K. and Peterson, R.T. 1999. *A Peterson Field Guide to Advanced Birding: Birding Challenges and How to Approach Them*. Houghton Mifflin.
- [31] Peterson, R.T. and Peterson, V.M. 2002. *Birds of eastern and Central North America*. Houghton Mifflin Co.
- [32] Smith, J.M. and Smith, D.C.P. 1977. Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems* 2, 2, 105-133.
- [33] Parsons, J., Lukyanenko, R., and Wiersma, Y. 2011. Easier citizen science is better. *Nature* 471, 7336, 37-37.
- [34] Kaldor, N. 1961. Capital Accumulation and Economic Growth. In *The Theory of Capital*, F.A. Lutz and D.C. Hague, Eds. Macmillan, London, 177-222.
- [35] Murphy, G.L., 2004. *The big book of concepts*. MIT Press, Cambridge, Mass.
- [36] Smith, L.B. 2005. Emerging ideas about categories. In *Building object categories in developmental time*, L. Gershkoff-Stowe and D.H. Rakison, Eds. L. Erlbaum Associates, 159-175.
- [37] McCloskey, M. and Glucksberg, S. 1978. Natural categories: Well defined or fuzzy sets? *Memory & Cognition* 6, 4, 462-472.
- [38] Braun, S., Schmidt, A., Walter, A., Nagypal, G., and Zacharias, V. 2007. Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. In *Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Web Conference* (Banff, Canada).
- [39] Lukyanenko, R. and Parsons, J., 2011. Unintended Consequences of Class-Based Ontological Commitment. In *Advances in Conceptual Modeling. Recent Developments and New Directions*, O. De Troyer, C. Bauzer Medeiros, R. Billen, P. Hallot, A. Simitsis and H. Van Mingroot, Eds. Springer Berlin / Heidelberg, 220-229.
- [40] March, S. and Allen, G. 2012. Toward a social ontology for conceptual modeling. In *Proceedings of the 11th Symposium on Research in Systems Analysis and Design* (Vancouver, Canada), 57-62.
- [41] Wand, Y., Monarchi, D.E., Parsons, J., and Woo, C.C. 1995. Theoretical foundations for conceptual modelling in information systems development. *Decision Support Systems* 15, 4, 285-304.
- [42] Bunge, M. 1977. *Treatise on basic philosophy: Ontology I: the furniture of the world*. Reidel, Boston, MA.
- [43] Wand, Y. and Weber, R. 2006. On Ontological Foundations of Conceptual Modeling: A Response to Wyssusek. *Scandinavian Journal of Information Systems* 18, 1, 1-11.
- [44] Harnad, S. 2005. To Cognize is to Categorize: Cognition is Categorization. In *Handbook of Categorization in Cognitive Science*, H. Cohen and C. Lefebvre, Eds. Elsevier Science, 20-45.
- [45] Parsons, J. and Wand, Y. 2000. Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Transactions on Database Systems* 25, 2, 228-268.