The Impact of Conceptual Modeling on Dataset Completeness: A Field Experiment

Completed Research Paper

Roman Lukyanenko
Florida International University
roman@nlnature.com

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Yolanda F. Wiersma
Memorial University of Newfoundland
ywiersma@mun.ca

Abstract

This paper investigates the impact of conceptual modeling on the information completeness dimension of information quality in the context of user-generated content. We propose a theoretical relationship between conceptual modeling approaches and information completeness and hypothesize that traditional class-based conceptual modeling negatively affects information completeness. We conducted a field experiment in the context of citizen science in biology. The empirical evidence demonstrates that users assigned to an instantiation that is based on class-based conceptual modeling provide fewer observations than users assigned to an instance-based condition. Users in the instance-based condition also provided a greater number of new classes of organisms. The findings support the proposed hypotheses and establish that conceptual modeling is an important factor in evaluating and increasing information completeness in user-generated content.

Keywords: Information quality, Conceptual modeling, Information modeling, Dataset completeness, Field experiment, Citizen science, User-generated content, Crowdsourcing

Introduction

Organizations are increasingly looking to make decisions driven by the data they collect or acquire. Following a recent survey of more than 3,000 business executives, LaValle et al. (2014) concluded that global business leaders "now want businesses run on data-driven decisions" (p. 22). With the rise of data-driven enterprises, a prominent concern for organizations is ensuring that data (information) is of acceptable quality. Traditionally information was created by the organization itself – using processes designed to provide data needed for operations and analysis. This allowed the organization to have well-understood and tightly controlled mechanisms of information production (Zuboff 1988). Traditional information quality (IQ) management required all parties related to data (e.g., producers and users) be in contact and agreement (e.g., Lee and Strong 2003). However, this view of IQ management is challenged by the growing practice of using information produced outside organizational boundaries.

Among various sources of externally produced data, of special interest is user-generated content (UGC) (Cha et al. 2007, Daugherty et al. 2008, Krumm et al. 2008), defined here as information produced by members of the general public – who often are casual content contributors (the crowd) – rather than by employees or others closely associated with an organization. To harness UGC, organizations create

1 Following a widely-accepted convention in the research community we use the terms information and data interchangeably (Lee 2006; Parssian et al. 2004; Wang 1998).
specially-built information systems (known as crowdsourcing IS, see Barbier et al. 2012; de Boer et al. 2012; Doan et al. 2011; Whitla 2009) or mine large pools of data from forums, blogs, social media, comments or product reviews (Andriole 2010; Chau and Xu 2012; Susarla et al. 2012). Companies that sell consumer products or services, for example, may use UGC to learn what customers think about their products and use this information to modify their offerings (Barwise and Meehan 2010; Culnan et al. 2010). Public sector organizations increasingly turn to the crowd to reduce operation costs and improve services. For instance, CitySourced (www.citysourced.com) is a US-wide project that encourages people to report civic issues (e.g., crime, public safety, environmental issues) and makes this data available to participating municipalities for analysis and action. OpenStreetMap (www.openstreetmap.org) constructs user-generated maps, thereby providing affordable geographical information to individuals, non-profit organizations and small businesses. Projects such as Amazon’s Mechanical Turk (www.mturk.com) and CrowdFlower (www.crowdflower.com) maintain a virtual workforce and lease it to clients for specific projects (e.g., to classify products in an e-commerce catalog). One of the best-known examples of crowdsourcing is Wikipedia (www.wikipedia.org) that leverages crowds to create a world-wide compendium of knowledge. Scientists and monitoring agencies sponsor online systems - citizen science IS - that allow ordinary users to provide observations of local wildlife, report on weather conditions, track earthquakes and wildfires, or map their neighborhoods (Planagin and Metzger 2008; Haklay 2010; Hand 2010).

Despite the growing reliance on UGC, a pervasive concern is the quality of data produced by ordinary people. The online environment differs substantially from traditional organizational settings. Online users are typically volunteers (Arazy et al. 2011). Participation in projects that collect UGC is often open, resulting in highly variable user motivation and domain expertise (Coleman et al. 2009). The controls over information creation available to the host organization are limited and casual online users may choose not to comply with the organization’s standards for data collection (Luksyanenko and Parsons 2013). These characteristics of UGC settings bring significant challenges to ensuring quality data.2

Among various data quality concerns related to UGC, accuracy and completeness of data are of particular interest. As organizations seek to make decisions based on UGC, it is important for data to be as error-free as possible. In this paper, we focus on information completeness - a less-examined dimension. Information completeness, or the extent to which valid states of reality are represented in IS (see Wand and Wang 1996), is an important requirement for decisions to be well-informed and unbiased.

In this paper, we advance an alternative perspective on information completeness in UGC. We argue that conventional wisdom in IQ under-represents a critical role of information contributors in information production. In a flexible and open UGC environment, with weak controls over information production and an explicit objective of harnessing the unknown, information contributors should be given more freedom to determine what and how much information to provide.

Adopting a contributor-oriented perspective to information completeness, however, requires rethinking fundamental approaches to IS design. In the traditional development paradigm, IS data structure reflects intended uses of information as defined by information consumers (Parsons and Wand 2000; Wand and Weber 1995). These views shape design of user interfaces, application logic, and data structures defined in databases. As the structure and behavior of IS objects are informed by conceptual modeling (Wand and Weber 1990), we focus on conceptual modeling as a factor affecting IQ. Little is known about how conceptual modeling decisions affect information completeness. Although research has stressed the relationship between systems analysis/design and the quality of stored data (Hoxmeier 1998; Levitin and Redman 1995; Wand and Wang 1996), the prevailing modeling approaches (e.g., Entity Relationship diagrams) are typically viewed as unproblematic for data quality (Ravichandran and Rai 1999; Wang et al. 1995) rather than as a potentially contributing factor to data incompleteness. In contrast, we argue that employing traditional conceptual modeling in UGC settings may have detrimental impact on IQ and call for alternative approaches to modeling UGC.

Recent IS research considers the relationship between conceptual modeling and information quality, but the focus is mainly on the quality dimensions of accuracy and attribute (information) loss (Luksyanenko et

---

2 While we believe that motivation of crowd users could impact the quality of observations we focus solely on the impact of modeling and consider motivation to be outside the scope of this paper.
al. 2011; Parsons et al. 2011). In this paper, we aim to increase understanding of the impact of conceptual modeling on information completeness in the context of UGC. We present a novel definition of information completeness informed by the pivotal role of information contributors in UGC projects. We then develop a theoretical model of the relationship between conceptual modeling approaches and information completeness. We evaluate the proposed theoretical model in a field experiment using a real IS. The paper concludes by outlining an approach to resolving IQ problems arising from the use of traditional class-based conceptual models and discussing implications of our work for future research and practice.

Theoretical Foundations and Research Propositions

It is widely assumed that IS are developed to support specific and predefined uses. This is reflected in IS analysis that, among other things, seeks to discover entities in a domain about which an organization wants to keep information (Parsons and Wand 2008; Parsons and Wand 1997). Such information often originates in use cases that communicate at a high level the data and activities to model (Jacobson et al. 1999). Once the system is designed, its quality is assessed insofar as it provides functionality and information necessary to fulfill the needs of its users (DeLone and McLean 1992; Petter et al. 2013).

In IQ research, quality is typically understood as fitness for use of data by information consumers (Lee 2006; Lee 2003; Strong et al. 1997; Wang and Strong 1996). Completeness is a prominent dimension of IQ and the fitness for use perspective is reflected in widely-cited definitions. Parssian et al. (2004) define completeness as "availability of all relevant data to satisfy the user requirement" (p. 968). Lee et al. (2002) developed measurement items to evaluate completeness, asking whether "information is sufficiently complete for our needs" and "covers the needs of our tasks" (p. 143). Completeness has been classified as context-dependent (Wang and Strong 1996). Nelson et al. (2005) explain (p. 203):

> It is important to recognize that the assessment of completeness only can be made relative to the contextual demands of the user and that the system may be complete as far as one user is concerned, but incomplete in the eyes of another.

From this perspective, all parties to IQ management (e.g., data creators, data consumers) should share a common understanding of what data is relevant, how to capture it and why it is important (Lee 2003, p. 33):

> To process organizational data, a firm’s data production process is conceptually divided into three distinct areas: data collection, data storage, and data utilization...

> To solve data quality problems effectively, the members in all three processes must hold and use sufficient knowledge about solving data quality problems appropriate for their process domains. At minimum, data collectors must know what, how, and why to collect the data.

However, this process breaks down in UGC. As discussed earlier, the organizational controls exercised over data production are minimal in UGC settings. Users vary in domain expertise and levels of motivation (Coleman et al. 2009). Many online contributors (e.g., amateur citizen scientists) may know very little about how data are going to be used by the sponsoring organizations (e.g., scientists). Furthermore, as online settings are inherently open and democratic, expecting a shared understanding on how much data is needed and why (Lee and Strong 2003) is not feasible for some or even most users.

Completeness, therefore, becomes a function of the judgment and abilities of independent online information contributors. However, since the data is intended for organizational use, it is important to consider data consumers as well. Thus, we argue for an alternative approach: information must be of potential interest to the host organization, but the amount of information should be up to contributors to decide. For example, suppose a mobile phone manufacturer is interested in learning what customers think about their products and create a UGC platform for this purpose. In this case, any information on a relevant product (mobile phone) provided by an online contributor might be of value to the company but, since user views may vary, the objective is to capture these as fully as possible.

An important issue that arises out of the new approach to completeness is how to judge when information is complete. Psychology research suggests that there are many unique ways to describe reality (Murphy 2004). To demonstrate our approach to completeness, we focus on a special case, dataset
**completeness** - the extent to which a database captures all records of phenomena of potential interest to data consumers that data contributors are willing to provide. In the example above, if the online contributor attempts to provide some information about the mobile phone, but the system rejects the entire attempt (for various reasons, as discussed later) resulting in failure to store any information about the phone, the resulting dataset is not as complete as would otherwise be. Dataset completeness is of critical concern to organizations, but is one of the most overlooked dimensions in IQ research. Fan and Geerts (2012) warn, "not only attribute values but also tuples are often missing from our databases" (pp. 93–94).

The implications of defining completeness from the contributor perspective differ from ‘fitness-for-use’ approaches to IQ management. Whereas fitness-for-use suggests the need to discover and support intended uses of data, the contributor-oriented perspective on completeness implies a need to design systems sensitive to contributor points of view. In the next section, we suggest how different approaches to IS development (focusing on conceptual modeling) can be used to improve completeness of UGC datasets.

Informing our approach to completeness is the perspective taken by Wand and Wang (1996) who argued that "completeness is the ability of an information system to represent every meaningful state of the represented real world system" (p. 93). Although their analysis is premised on IQ that reflects "the intended use of information" (p. 87), it suggests that completeness maybe undermined if an IS is incapable of representing every relevant state of the world. Approaching information completeness from the representational perspective suggests to investigate the impact of the process by which an IS representation of reality is created, or conceptual modeling, on IQ.

Research on conceptual modeling and information quality are both well-established in IS, but the relationship between conceptual modeling (e.g., conceptual modeling grammars) and information quality (including information completeness) remains poorly understood. Typically, discussions of the impact of modeling on IQ examine the importance of high quality requirements translated into good database design, such as implementing integrity constraints and business rules (Hoxmeier 1998; Olivé 2007; Redman 1996). Wand and Weber (1990) further stipulate that "[i]f an information system is intended to be a representation of a real world system, the grammars used during the design and implementation process must be capable of fully describing the structure (statics) and behavior (dynamics) of the real world" (p. 63). Here we examine the impact of two perspectives in conceptual modeling: traditional – as typified by such popular approaches as the Entity-Relationship and UML; and emerging – modeling based on instances rather than classes. Using these two approaches, we demonstrate the role of conceptual modeling grammars in improving completeness of UGC datasets.

The prevailing class-based conceptual modeling grammars, such as the Entity-Relationship model and UML class diagrams, involve *a priori* specification of conceptual entities (classes, entity types), attributes (or properties) of these entities, and relationships between entities assumed to exist in a domain (Chen 1976). Based on these models, developers produce IS design objects such as database schema, user interface, and code. Data creation is then mediated by these objects: all individual entities in a domain must be recorded in terms of predefined classes specified in the conceptual model.

We argue that class-based modeling may negatively impact completeness of UGC due to the requirement to comply with the constraints specified in class-based conceptual models. For example, a class that is not defined in the conceptual model will be rejected by an IS. This places unnecessary limitations in domains, such as UGC, in which completely specifying the relevant data structures in advance is unrealistic. Furthermore, a mismatch between models of a contributor and those defined in the IS may dissuade data contributors from reporting information. For example, users may be apprehensive about submitting potentially incorrect data (e.g., an instance of a product for which no class is found), or even be disappointed with the gulf between their own model and an IS and thus avoid using it.

In contrast, instance-based conceptual modeling (e.g., Lukyanenko and Parsons 2012) is based on the ontological assumption that instances "exist" independent of classifications (Bunge 1977; Parsons and Wand 2000). Ontologically, every "thing" possesses a unique set of properties (or attributes). Instances can be grouped into classes when they share some relevant properties. As an instance can possess many properties, it can belong to a very large number of classes, depending on the context.

In practice, instance-based conceptual modeling means that, when reporting information, users are free to describe observed instances in terms of attributes or, when possible, classes of interest. If an instance is
reported solely in terms of attributes, these attributes can be used to infer classes. The classes of interest to decision makers may be discovered after information is recorded, provided the user reports enough attributes for positive class identification. This removes the constraint of traditional modeling to understand and comply with an a priori defined schema.

As instance-based data collection is grounded in theories of ontology and psychology, it should also be an intuitive and ontologically consistent way to report data. For example, in projects that map biodiversity, the objects of interest (e.g., birds, animals) may be fleeting with an extremely short exposure time. This makes it unrealistic to expect consistent classification of these instances at some predefined levels. For example, even if an online volunteer is very familiar with the needs of scientists, short exposure or poor visibility may preclude the person from definitively identifying an instance at very specific (i.e., useful to scientists) classification levels. Unlike IS based on the prevalent abstraction-based models, systems that implement instance-based conceptual modeling do not require users to “force-fit” their individual conceptualizations into predefined abstraction-based models. We therefore expect that class-based conceptual models result in lower dataset completeness when the classes defined in an information system do not match those of the information contributors. In the next section we examine these questions in a field experiment in a real-world UGC project.

**Field Experiment**

To provide empirical evidence of the impact of conceptual modeling on information completeness, we conducted a field experiment in the context of citizen science in biology. Citizen science is a type of crowdsourcing in which scientists enlist ordinary people to generate data to be used in scientific research (Louv et al. 2012; Silvertown 2009). Citizen science promises to reduce information acquisition costs and facilitate discoveries (see, for example, Hand 2010). Recognizing UGC as a source of unanticipated insights, scientists consider collecting citizen data in a hypothesis-free manner (for discussion, see Wiersma 2010). Recently, IS researchers argued that giving more flexibility to citizen scientists should improve accuracy of observations and promote broader user engagement (Lukyanenko et al. 2011; Parsons et al. 2011).

Citizen science in biology is a convenient ground for research in IQ: it has established standards for information quality (e.g., biological nomenclature) and a well-defined cohort of data consumers (scientists). This makes it easier to evaluate the impact of modeling approaches on real decision making. Further, citizen science has an immutable requirement for high-quality data - a requisite for valid research. Citizen science is a voluntary endeavor and the challenge is to induce data of acceptable quality while keeping participation democratic and open (Louv et al. 2012).

Biology has a well-established conceptual schema. Specifically, species is considered the focal classification level into which instances in this domain are commonly organized. Species are units of research, international protection, and conservation (Mayden 2002). Major citizen science projects (e.g., eBird.org, which collects millions of bird sightings) implement prevailing modeling approaches (e.g., Entity-Relationship) and collect observations of instances as biological species (Parsons et al. 2011; Wiggins et al. 2013). Class-based conceptual modeling based on species-level classes therefore constitutes a natural benchmark against which performance of the novel instance-based approach can be gauged.

As an alternative to class-based models, we argue observations collected from citizen scientists should be collected and stored in terms of instances, their attributes and any classes contributors deem relevant. This approach represents a realistic compromise. Non-experts do not always know (or may not be willing to identify down to the level required by scientists) the instance that was observed. It is more realistic to expect a volunteer to remember some features of unknown species than to expect a precise classification.

The completeness of information stored in the instance-based IS can be compared with that stored in a traditional IS due to the fact that both systems represent instances. For example, records in traditional IS (e.g., such as those in popular citizen science projects), are about instances of interest to biologists, but reported in terms of the classes useful to them (e.g., species). Similarly, instance-based IS can have records about the same instances, but their attributes and classes would naturally vary (reflecting varying levels of domain expertise, motivation, and other contextual factors inherent in UGC settings). While in the latter case some classes and attributes relevant to data consumers may be missing, information is still relevant insofar as it pertains to the instances of interest to the organization, satisfying our the definition of information completeness. We thus hypothesize:
Hypothesis 1: Number of instances stored. Contributors will report significantly more instances of plants and animals and other taxa in the instance-based IS compared with an IS based on classes useful to data consumers (biological species).

Hypothesis 1 is primarily motivated by our contention that the prevailing modeling grammars may have inherent barriers to describing instances of interest. This leads to less complete datasets due to the mismatch between the conceptualizations of online contributors and the class-based models embedded in the IS. We now turn to another important use of crowdsourcing – enlisting crowds to gather information about novel classes of things. We believe an instance-based IS has a greater propensity to capture instances of novel classes due to: (1) the difficulty in UGC settings to determine in advance a comprehensive set of classes for a domain, (2) the greater propensity of the instance-based IS to capture novel classes, and (3) the anchoring effect that might potentially prevent participants from providing instances of classes that are not present in the original schema.

In rich and complex domains (e.g., science, healthcare, consumer markets) it may be difficult to determine in advance all relevant classes of things (regardless of whether participants are previously familiar with them or not). For example, projects maybe local in scope (Sheppard et al. 2014) and be concerned with monitoring and conservation in a small geographic area. Since distributions of plants and animals are not static, it may be impossible to develop a comprehensive classification suitable to describe everything that might be observed in a given local area. Indeed, finding anomalies and outliers might be the raison d’être for some UGC projects. Even a single valid data point would spell success for a project like SETI@Home that leverages distributed crowd computing in search of extraterrestrial intelligence (Korpela 2012).

Organizations increasingly hope to harness UGC to learn something new about their target domains. One approach may be to encourage participants to contact the organizers when they encounter something unusual or not fitting into the predefined structure. Anecdotally, a discovery of a previously unknown to astronomy object, Hanny’s Voorwerp, occurred when an online contributor, Hanny van Arkel discovered a huge blob of green-glowing gas while performing a task of classifying galaxies in a project GalaxyZoo (Lintott et al. 2009). A project schema was not designed to accommodate this class of instances. van Arkel sensibly posted this information in a forum created to support the project and this post was eventually noticed by scientists. While online contributors may find workarounds to record information they believe to be important, class-based IS lack inherent affordances (or design features that correspond to human abilities, see Norman 1983) to capturing unanticipated attributes and classes. In contrast, instance-based information management is naturally suitable for representation of any unanticipated insights by allowing users to record novel classes and attributes users wish to provide.

Finally, the predefined classes commonly used to structure information collection may act as an anchor – an initial starting condition that affects the subsequent information users provide. According to cognitive psychology, exposure to some information relevant to the task (i.e., the initial value or an anchor) increases the likelihood of this information being considered or assimilated into the subsequent response (Strack and Mussweiler 1997; Tversky and Kahneman 1974). In IS research anchoring effects have been observed during IS development (e.g., information requirements determination, query writing, programming; see, for example, Allen and Parsons 2010; Browne and Ramesh 2002; Parsons and Saunders 2004) and IS use (e.g., in the context of decision support systems, see, for example, George et al. 2000). To the best of our knowledge, anchoring effect has not been considered as a factor that may negatively impact information quality.

We argue traditional IS necessarily exposes users to existing classes (the anchor) in order to capture information. This may lead to assimilation of the predefined classes by the users (adjustment bias) and affect users’ perceptions of the kinds of classes organizations sponsoring UGC projects might require. Thus, while users may encounter a variety of instances from the project domain in the physical world, they may form a judgment that the instances of classes defined in the IS are those of sole or particular relevance and fail to communicate instances of classes not present in the schema.

Here we compare the ability of two modeling approaches to capture unanticipated kinds of instances. To ensure equitable and conservative comparison we focus on new (i.e., not previously included in the project schema) species-level classes (as opposed to, for example, new attributes of instances). This comparison is conservative insofar as species-level identification is the explicit focus of class-based IS and
is arguably de-emphasized in the instance-based IS where the focus is on attributes and classes. Following theoretical arguments presented above, we hypothesize:

**Hypothesis 2: Number of instances of new species stored.** Contributors will report significantly more instances of new biological species in the instance-based IS compared with an IS based on classes useful to data consumers.

To evaluate the proposed hypotheses we used data from an existing web-based citizen science project, NLikeNature.\(^3\) Using field experimentation offers several advantages. The project allowed us to track real user behavior (as opposed to behavioral intentions) as well as evaluate the impact of user-generated data on data consumers and scientific decision making. Studying behavior directly is a growing trend in a variety of disciplines from economics to psychology, where scholars argue that deeper understanding of actual behavior and its circumstances affords unique insights about unobservable states of human mind (Bargh and Chartrand 1999). Conducting research in a real setting further increases external validity compared to similar studies conducted in laboratory environments.

The project was launched in 2009 by researchers at a North American university as part of a national initiative to investigate how the public uses the Internet to communicate environmental change. The scientific objective of the project is mapping biodiversity in a region of North America based on amateur sightings of nature (e.g., plants and animals). During the four years of its existence, assessments of IQ were performed (including analysis of contributions, comments from users, and benchmark comparison with parallel scientific sampling). The authors of this paper, together with the biologists (i.e., project sponsors), determined that quality and level of participation were below expectations. Consistent with the arguments above, we identified the prevailing class-based approach to information modeling as a detriment to IQ. 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). Additionally, in several cases, the organisms could not be fully described using attributes of the correctly chosen species-level class (e.g., morph foxes had additional attributes not deducible from the class Red fox). Finally, the authors believed that many observations were not reported because of the incongruence between the conceptual model and user views.

The decision to conduct the experiment was made one year prior to its commencement in May 2013. The 12 months preceding the launch of the experiment were spent in planning and development. Importantly, any promotional activities were halted during this time to avoid attracting attention of the public to the project and ensuring a fresh start. Preceding the launch of the redesigned NLNature, activity on the website was low. This allowed us to rebrand NLNature to the community as a fresh start.

During 2012 we substantially redesigned NLNature, changing its appearance and behavior. The data collection interfaces were completely changed. We also timed the launch of the experiment to coincide with the end of spring - a time when wildlife becomes accessible as people spend more time outdoors. NLNature was promoted among the members of the general public through a tour across the region. The project was advertised/featured on radio, television, newspapers, and online (e.g., through Google Adsense and Search network, Facebook, Twitter, through website partnerships). All demographic groups were targeted in the promotional activities to ensure a representative sample of users. The promotional activities produced substantial traffic on the project - peaking at 10,000 visitors per month at the height of the experiment in mid-summer of 2013.

To compare prevailing and alternative approaches to modeling, we developed two data collection interfaces, each corresponding to different conceptual modeling assumptions: a class-based interface using species-level data entry, and an instance-based interface. The interfaces were designed to be visually similar and were dynamically generated from the same master template (differing only in the aspects relevant to the underlying conceptual modeling approaches).

Potential information contributors (citizen scientists) were randomly assigned to one of two data collection interfaces upon registration and remained in the originally assigned conditions for the duration of the experiment. The data entry form required authentication to ensure that users were not exposed to different conditions. Regardless of the assigned condition, all users received equal access to other areas of

---

\(^3\) See, [http://www.nlnature.com](http://www.nlnature.com)
the project (e.g., internal messaging system, forum) and equal support from the project sponsors. This ensured equivalent facilitating conditions (Venkatesh et al. 2003) across the three groups.

In the class-based condition, users were required to report sightings by selecting from a predefined list of species (no attributes of species were collected – to ensure equivalent treatments and consistent with existing real IS projects). Since it is entirely possible that a contributor may not know or be confident in the species-level identification, we provided an explicit option (with clear instructions) to bypass the species-level classification by clicking on the "Unknown or uncertain species” checkbox below the data entry field (see Figure 1, left panel). In the second condition we instructed participants to provide attributes and, if possible, classes (see Figure 1, right panel). This allowed users to report sightings even if they could not determine a class for the instance observed.

In both conditions, to see a list of options (classes or attributes) users were instructed to begin typing in the textbox and click "Add” or press "Enter” when finished. As soon as more than two characters are entered, a suggestion box appeared with the classes or attributes that contain the string entered. In the species-based condition participants were required to select an item from the list (or supply a new class in the comments, as per instructions). In the instance-based condition, participant could select an item from the list or provide novel attributes and classes via direct entry.

When designing the two data collection interfaces, it was important to make sure they were appropriate for the target class- and instance-based conceptual modeling. As discussed earlier, the class-based version implements traditional approaches to systems analysis and design. Therefore, when making specific design decisions (e.g., the design of data entry forms), it was important to have high ecological validity. Consistent with similar projects (e.g., www.eBird.org, www.iSpot.org.uk), we instructed participants to provide a positive identification of species (Lukyanenko et al. 2011) where the actual classes were predefined based on the list provided by scientists. Following popular practice on social media websites (e.g., Facebook) and citizen science projects (e.g., www.iSpot.org.uk) we decided to provide options via a prompt-assisted text field. This allowed a participant to begin typing a class and a prompt would dynamically show potential matches based on the string being typed. As more people become engaged with social media, the dynamic text field is becoming a norm for data entry. It also appeared as a superior alternative to a dropdown list as it mitigated any potential adverse ordering and priming effects.

We used the traditional class-based condition as a template for the instance-based one as it was more important to ensure equivalence across conditions than produce the most effective implementation of the instance-based IS. Specifically, we reused every design element included in the class-based condition that was not pertinent to instance-based modeling. As with the class-based data entry, we began with the instructions but asked users to describe instances and attributes rather than classes (as shown in Figure 1).

---

4 As we expected the instance-based IS to outperform the class-based IS we were not looking to design the most effective instance-based IS, but rather an implementation that would make the comparison more equitable.
right panel). The same dynamic text box was used for data entry. Unlike in the class-based condition, users in the instance-based condition were not constrained to the predefined choices and were free to provide their own attributes and classes.

**Results**

The results are based on data from six months of usage between June and December 2013. This period spanned high and low tourism seasons and also allowed participants to observe major changes in ecology due to seasonal changes.

In designing the project we followed established practices of engaging citizen science participants in scientific research including voluntary and anonymous participation (Robson et al. 2013; Snäll et al. 2011). In order to use NLNature, participants were required to accept a consent form that outlined the nature of their interaction with the website. Failure to accept the consent disallowed people from using any data-collecting features. No incentives for participation were provided and no personally-identifying information was collected. The existence of the experimental manipulation was not disclosed.

The results of the study are based on information provided by the website members who accepted the consent form after June 1st 2013 when the two manipulations outlined above took effect. From June to December 2013, 230 members accepted the consent form and began to participate. Upon accepting the consent form users were randomly assigned into the study conditions. Some participants registered, but never landed on the observation collection page and, hence, were not actually exposed to manipulation (this was determined by analyzing server logs). The final number of participants who at least once visited one of the two observation collection interfaces was 42 in the species-based condition and 39 in the instance-based condition. The remaining analysis is based on the information provided by these users.

While we did not require users to provide demographic data, some volunteered this information by completing an optional form. Fifteen participants indicated their age (50.87 avg., 15.54 st. dev.). Seventeen participants indicated how many years they lived in the target region (18.85 avg., 17.30 st. dev.). Fourteen participants provided the number of hours per week they spent outdoors (19.14 avg., 15.54 st. dev.). While the majority of participants abstained from contributing demographic information, those who provided information were mature and had considerable local experience.

To evaluate H1, we analyzed observations (i.e., sightings of instances that could be classified or described using attributes) provided by 81 participants exposed to manipulation in the two conditions. As, in the class-based condition, it was possible that a contributor would not know or be confident in species-level identification, we provided an explicit option (with clear instructions) to bypass the species-level classification by clicking "Unknown or uncertain species" checkbox below the data entry field (see Figure 1). We further instructed participants to write, in the comments, any class to which they believed the instance belonged. As in this case a user could provide classes at levels other than species, we removed such non-species observations from the count for users in the class-based condition. Table 1 reports the number of contributions in each condition: they included sightings made in the instance-based condition and species-level classifications in the species-based condition.

Finally, because we defined completeness as the extent to which stored information represents the phenomena of potential interest to data consumers, we counted an observation as valid if it described an instance from the domain of biology, including plants, animals and other taxa.6

---

6 Approximately 30% of the members who accepted the consent during the study period were randomly excluded from manipulation and were assigned to a third condition to test hypotheses that are outside the scope of this paper.

6 This led to the removal of one observation of ‘island’.
Table 1. Number of observations by condition

| Experimental Condition | No of users in condition | Observations |          |            |            |            |            |
|------------------------|--------------------------|--------------|----------|------------|------------|------------|
|                        |                          |              | Total    | Mean       | St. dev.   | Skewness   | Kurtosis   |
| Class-based            | 42                       | 87           | 2.07     | 2.56       | 2.08       | 4.23       |
| Instance-based         | 39                       | 390          | 10.00    | 37.83      | 5.47       | 29.66      |

Before hypothesis testing we verified the assumption of normality in the data using the Shapiro-Wilks test. In each condition, the distribution of observations per user significantly deviated from normal (with W=0.690 and p-value < 0.000 for the class-based and W= 0.244 and p<0.000 for the instance-based condition). We also note the presence of outliers in each condition. As seen from Table 1, in both cases the distributions are skewed and leptokurtic. This was confirmed using Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit statistics, where the best fitting distributions were power-law, lognormal and exponential. Commonly, these are referred to as "long tail" distributions. Indeed, the top 4 contributors (or 10% of the user sample) in the instance-based condition produced 80.8% of the observations in that condition. These results are not surprising: long-tail distributions have been observed consistently in other user-generated datasets including in citizen science projects (Lukyanenko and Parsons 2013). However, the instance-based condition has greater mean, variance, skewness and kurtosis (see Table 1). Figure 2 further illustrates this by showing that users in the instance-based condition tend to contribute a higher number of observations and few users in this condition contributed one or zero observations.

![Figure 2. Non-zero observation frequencies (y-axis) per user (x-axis) in the two conditions](image)

To determine if the difference in the number of observations per user is significantly different across the conditions we used the exact permutation test (Gibbons and Chakraborti 1992; Good 2001; Hayes 1996). The exact permutation test is suitable when data are not normally distributed, sample sizes are low and medium, outliers and ties (i.e., same values in two samples, as in Figure 2) are present. Based on the exact permutation test of observations per user between the two conditions, the p-value is 0.033, indicating that users in the instance-based condition provided significantly more observations than those in the species-based condition. This supports Hypothesis 1 and is consistent with our contention that different conceptual modeling approaches may result in significantly different numbers of instances of interest represented in IS.

To gain a deeper insight into the impact of different conceptual modeling approaches on information completeness, we further analyzed the categories and attributes provided to identify specific causes of lower performance by the users in the class-based group. Specifically, we note three (observable)
behavioral patterns of users in the class-based condition that we predicted would contribute to lower information completeness. Below we elaborate on each pattern.

We argued that, since the class-based models constrain user input to predefined classes and attributes, users may not be able to record instances unless they provide classes that are congruent with the predefined structure in an IS. Evidence for this comes from the analysis of classes users entered in the dynamic textbox. The use of a dynamic textbox for data entry allows comparing words and phrases users attempted to submit against the classes defined in the IS. While in the instance-based condition entering directly new attributes and classes was allowed, in the class-based condition the entries were vetted against the active species list and only matching entries were allowed (unless a user explicitly bypassed this step to report and an unknown or new species).

The analysis of user input reveals instances of mismatch between the intended classification and the active class base. While we specifically instructed users to provide species-level responses and identification at that level is the prevailing practice in natural history citizen science, users still attempted to provide classes at other levels. These were generally at higher levels in the classification hierarchy (e.g., dolphin, toad, slug) potentially reflecting classification uncertainty (e.g., due to conditions of observation) or lower levels of domain expertise (non-experts are generally more comfortable with more general taxonomic levels).

Each case in which the provided class did not match the target species level was not included in the analysis above, contributing to the lower number of observations in the class-based condition. The existence of cases where users attempted to enter data at levels above the species-level provides evidence for the mismatch between the model of the contributor and the data consumer-oriented view embedded in the IS.

The second pattern observed showed that, when facing a structure incongruent with their own, some users changed the original submission. In several cases this resulted in loss of instances. For example, in one case a user began with typing “otter” (non-species level) - the entry was rejected by the system. The user then proceeded to record "Little Brown Bat (Myotis lucifugus)" under the same observation ID. In another case a user typed "grackle" (non-species level) 5 times before finally selecting "Common Grackle (Quiscalus quiscula)". Another user began with "black bear scat", and after two attempts to record it, typed "Black Bear (Ursus americanus)". In all examples above the original input had to be changed by users to comply with the model. In the case of "otter" the instance was not stored as the user simply abandoned data entry.

We predicted that class-based modeling has weak affordances for capturing new classes. Our results offer some evidence for this. In 12 cases, users in the class-based condition selected to by-pass species identification, but then failed to provide any species-level labels. These cases were also excluded from the final count, further expanding the difference in the number of observations between the conditions.

Another source of difference between the conditions is the prevalence of non-species-level classification in the instance-based sample. Many classes provided in the instance-based condition were at levels higher-than the species. Of 390 observations in the instance-based condition, 179 (45.9%) were not classified at the species level. For these observations, participants provided 583 classes and 69 attributes (222 distinct classes and 43 unique attributes). Among the classes provided, 110 are what are known in psychology research as basic-level categories (e.g., fly, bird, tree, fish, spider). Basic-level categories are widely accepted in cognitive psychology as the generally preferred classification level for non-experts and are typically in the middle of the taxonomy (e.g., “bird” is a level higher than “American Robin”, and lower than “animal”) (Corter and Gluck 1992; Eimas and Quinn 1994; Markman and Wisniewski 1997; Rosch et al. 1976; Tanaka and Taylor 1991). The use of basic-level categories by participants can stem from at least three sources (which are not mutually exclusive):

- low level of domain expertise of some users, as argued in the psychology literature;
- conditions of observation (e.g., too dark, fleeting, at a distance), when positive identification at more specific levels could not be made;
- attempts to provide additional evidence in cases when confidence in lower-level identification is low.
The evidence demonstrates that the mismatch between the conceptualization of users (situational or expertise-related) and those embedded in the IS contributed to the lower number of observations in the class-based condition.

We now evaluate Hypothesis 2, which posits that a greater number of new species would be reported in the instance-based condition. Users in both conditions provided 997 attributes and classes including 87 in the class-based and 910 in the instance-based condition. Of these 701 attributes and classes were new - they did not exist in the schema or data and were suggested by users as additions. This was done directly by users in the instance-based condition and indirectly (via comments to an observation) by users in the class-based condition.

During the experiment 126 new species-level classes were suggested by the participants, including 119 in the instance-based and 7 in the class-based condition (see Table 2). In each condition the distribution of new species per user significantly deviates from normal (W=0.430 and p-value<0.000 for the class-based and W = 0.232 and p<0.000 for the instance-based condition). The distribution is long-tailed in the instance-based condition (fitted using Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit) and uniform (Chi-squared = 47, Monte Carlo p=0.424) in the class-based condition. Based on the exact permutation test the number of new species is significantly greater in the instance-based condition (p=0.007) providing support for Hypothesis 2. This suggests that instance-based approach to modeling may be more effective for harnessing unanticipated phenomena of interest.

<table>
<thead>
<tr>
<th>Experimental Condition</th>
<th>No of users in condition</th>
<th>New Species Total</th>
<th>Mean</th>
<th>St. dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class-based</td>
<td>42</td>
<td>7</td>
<td>0.17</td>
<td>0.44</td>
<td>2.53</td>
<td>5.96</td>
</tr>
<tr>
<td>Instance-based</td>
<td>39</td>
<td>119</td>
<td>3.05</td>
<td>13.17</td>
<td>5.35</td>
<td>28.51</td>
</tr>
</tbody>
</table>

Users also provided interesting attributes for some sightings. Many appeared to augment the classes provided and offered additional information not inferable from the classification labels:

- attributes describing situational behavior of the instances observed (e.g., mating, hopping, fluttering together);
- attributes describing something unusual about an instance (e.g., tagged, only has one antler, has identifying marks);
- attributes describing the environment / location of the instance (e.g., near highway, 10 feet away from highway, near bike trail).

As these attributes cannot be deterministically established from knowing the species and thus constitute information beyond what could be collected in a traditional class-based model.

Several sightings of biological significance were reported during the experiment. These included unanticipated distribution of species (e.g., vagrant birds, fish and insects), a mosquito alien to the geographic area of the study, and a discovery of a possibly new species of wasp presently pending verification. All these occurred in the instance-based condition. It is also notable that some of the new things suggested by the instance-based users belonged to groups that were poorly represented in the project schema, including insects (e.g., 29 sightings of flies, 10 sightings of moth, 8 sightings of mosquitoes).

**Discussion**

The results demonstrate that conceptual modeling approaches affect information completeness. Using a real IS, we found that participants provided on average more observations when assigned to the version that implements instance-based, rather than traditional class-based, information modeling. Similarly, participants in the instance-based condition provided greater number of novel classes of organisms. The results indicate that traditional modeling can be a barrier to providing information, one that can be mitigated by instance-based modeling.
It is also notable that of the top 5 contributors, 4 belonged to the instance-based condition - collectively producing 315 sightings - 80.8% of the observations in the instance-based condition and 66.0% of all the observations collected during the study period. In contrast, the top 4 contributors in the class-based condition created 33 observations - or 37.9% of the observations in their condition and 6.9% of all observations. This hints that instance-based modeling might encourage the emergence of "superstars" - people who contribute a disproportionately large share of the projects' content. Given that the typical distribution of user activity in UGC projects is long-tailed, superstars constitute a stable core of the project - a group of regular and potentially most loyal users. Nurturing the growth of superstar users may be central to a project's success, as they play a key role in content production, dissemination of ideas and influence on others (Chau and Xu 2012; Zhang et al. 2013).

Instance-based conceptual modeling appears to be more effective at capturing unanticipated phenomena. Users in the instance-based condition reported 17 times more observations of new species than in the class-based condition. One concern about the definition of information completeness from the perspective of data creators is that this may result in information that is irrelevant and of no value to the sponsoring organization. Our findings appear to point to the contrary. Instance-based users outperform the users in the class-based condition in the traditional classification task. A potential explanation for this paradox has to do with the increased flexibility and freedom afforded by the instance-based model. While the class-based users were given a mechanism to report new species it was not direct and seamless. In several instances users in this condition appeared on the path to provide new classes (by clicking on the bypass identification button), but contrary to the instructions provided no valid descriptions in the comments. Another reason for the lower number of new species in the species-based condition might be related to the fact that users in this condition were directly exposed to the schema of the project - and thus could have formed a preconceived notion of the kinds of things that were of interest to the project sponsors. Indeed, users in that condition were required to select from predefined options. In contrast, users in the instance-based condition were not required to comply with any predefined options. Indeed, some of the new instances logged by the instance-based users belonged to groups that were originally poorly represented in the project schema, including spiders, flies, and mosquitoes. These readily observable organisms were rarely reported in the four years preceding the experiment. A widely-held assumption in citizen science holds that non-experts mostly report "charismatic" organisms, fueling concerns that citizen science produces a distorted view of biodiversity (Boakes et al. 2010; Galloway et al. 2006). The results obtained indicate that the imposition of schema may bias participants toward predefined options and the bias may be potentially overcome using instance-based modeling.

Despite finding significant differences between the two conditions, it is notable that among the information provided by participants in the instance-based condition, many classes were at the species-level. This level of granularity is natural for domain experts, whereas novices are typically more comfortable with more generic classes (Tanaka and Taylor 1991). The results suggest that, despite efforts to attract members of the general population, many participants on NLNature had higher-than average levels of domain expertise. This may be explained by the fact that, being unaware of the novel experimental condition, prospective participants assumed that getting engaged in the project required some level of domain expertise. This could have dissuaded non-expert participants from joining and discovering the instance-based condition.

While participants in the instance-based condition provided more observations than participants in the class-based condition, a natural question arises as to the extent to which the instances in the instance-based condition belonged to classes (species) provided on NLNature before the start of the experiment. This question is important as these classes typically support intended uses of citizen science information by the scientists. In the instance-based condition, participants provided 51 of the 343 (14.9%) species that were in the schema of the NLNature before the start of the experiment. By comparison, in the class-based condition participants provided only 36 (10.5%) of the original species. While this may be in part due to the overall larger number of observations in the instance-based condition (there were 390 observations in that condition and only 87 observations in the class-based condition), and the overall message of the project that implicitly oriented participants toward providing species-level classes (a point we consider later when we introduce the concept of soft factors in guiding user input), it illustrates that the use-
agnostic instance-based approach does not necessarily result in failure to capture information known to be of immediate relevance and usefulness to data consumers.  

### Implications for research and practice

Our findings have important implications for the theory and practice of data management, information quality, conceptual modeling and user-generated content. Having reasonably complete information is important for supporting informed and unbiased decisions and is a core concern in a data-driven society. Traditionally, information completeness has been understood as a contextual dimension (Wang and Strong 1996). This, for example, suggested that before data is collected, the uses need to be established and understood. We advocate an alternative perspective on information completeness: rather than tying completeness to intended uses of data, we consider it to be a function of the extent to which various factors involved in data production inhibit or encourage the provision of information. In this paper we explore one such factor - the conceptual modeling approaches to IS development. Our results demonstrate that different approaches to conceptual modeling result in different levels of dataset completeness. This finding establishes a novel connection between conceptual modeling and dataset completeness and suggests a new mechanism for increasing IQ. This work increases our understanding of, and provides empirical evidence for, the relationship between conceptual modeling and information quality building on conceptual work by Lukyanenko and Parsons (2011).

This paper contributes to the growing body of work that aims to establish antecedents of information quality dimensions and discover mechanisms for improving quality. Despite extensive research on IQ and its centrality to organizational decision making, relatively little is known about what causes low quality data - resulting in "a significant gap in the IS research" (Petter et al. 2013, p. 30). Furthermore, of various factors that influence quality (e.g., training, instructions, process controls), issues related to modeling of IS have been surprisingly neglected (Levitin and Redman 1995; Lukyanenko and Parsons 2011). Even fewer papers consider this issue theoretically. This makes it difficult to effectively leverage conceptual modeling in improving this quality dimension. Likewise, a better understanding of data quality implications promises to inform the theory and practice of conceptual modeling and suggest directions for improving modeling methods and grammars.

The results indicate that different conceptual modeling approaches can have a profound impact on the completeness of datasets. Specifically, prevailing conceptual modeling grammars can have a detrimental impact on dataset completeness. In many cases, valid information that users wanted to provide in a live citizen science project was not captured by the system due to the core assumptions of traditional conceptual modeling: "complete specification"; the requirement to comply with predefined class-based structures, and the tenet that an instance must belong to some class in order to be stored (Parsons and Wand 2000). These assumptions of traditional modeling appear to be severely limiting in environments where complete specification is infeasible and users may be exposed to hitherto unknown instances. This is an important finding given that traditional modeling techniques have been considered unproblematic for IQ (Ravichandran and Rai 1999; Wang et al. 1995).

Exposing the limitations of prevailing approaches to conceptual modeling is of significant practical value. First, it suggests that before proceeding with systems analysis using traditional approaches, analysts should examine the extent to which the assumptions of modeling methods and grammars are congruent with the nature of the project and characteristics of potential information contributors. For example, the objective of a project might involve discovery of new classes of things - as may be the case in scientific endeavors - leading analysts to reconsider using traditional grammars. Second, the potential of class-based models to stifle representation of valid instances may warrant evaluating completeness of existing datasets. Any significant gaps in representing relevant instances be could inform further quality improvement strategies. Knowing the gaps can help organizations better qualify datasets for decision making, suggest changes to class structures to bridge the identified gaps, or encourage adopting alternative modeling approaches to information collection.

---

8 Unfortunately, the field setting precluded us from knowing what was observed limiting our ability to compare accuracy of the information provided by the two approaches. There could be a potential tension between the volume of observations and their accuracy and we hope to investigate this issue in future research.
In this paper, we extended prevailing conceptualizations of completeness to emphasize the critical role of data contributors in UGC settings. We argue information completeness becomes a function of the judgment and abilities of independent online information contributors – qualities that can be extremely diverse and unpredictable. It is important, however, to remain cognizant of the organizational needs - which led us to define completeness as the depth and breadth of information as determined by online contributors insofar as it pertains to the phenomena of potential interest to data consumers.

Defining quality from the contributor point of view raises concerns whether the resulting data is relevant to the organizations. Our results can be used to address some of these concerns. Despite the flexibility afforded by the instance-based condition, all contributions in this condition were of phenomena that belonged to the target biology domain. This appears to suggest that users assimilated the general domain of the project - this could come from the advertisements they saw, any information they read about the project, their ideas and beliefs about its goals. This shows that organizations can leverage these "soft factors" for guiding input in the desired direction without imposing rigid constraints on the kinds of data users provide thereby opening door for user expression, creativity and discovery.

The contributor perspective suggests additional mechanisms for quality improvement. For example, whereas the fitness for use paradigm requires discovering the uses of data, the contributor-oriented approach to IS motivates efforts to design systems sensitive to contributor points of view. In this paper, we explicitly advanced instance-based modeling as one design strategy sensitive to data contributors. Future work can extend this by exploring other design approaches. Here, of potential value are the soft approaches to guiding user input in the desired direction without imposing rigid constraints (e.g., general instructions, allowing users to interact with each other). This promises exciting opportunities for future research.

We believe our arguments and findings extend beyond the current context of crowdsourcing citizen science. The findings presented here have important implications for information systems development and data management in environments with weak controls over information production or evolving uses of information. Specifically, contributor-oriented IQ can inform emergent practice of UGC development including crowdsourcing (Doan et al. 2011) and social media (Culnan et al. 2010; Hemsley and Mason 2012). In these domains, traditional quality improvement strategies (e.g., training, instructions) may be limited. Here we establish conceptual modeling as a promising new strategy for quality improvement - arguably better suited for the nature of UGC.

There is a growing market of platforms that make available UGC upon request. Yet, despite the growing reliance of organizations on UGC, little is known about how to effectively harness the knowledge of online contributors and existing platforms implement traditional approaches to modeling. For example, CitySourced organizes the domain in terms of predefined classes related to crime and public safety. Similarly, Amazon’s Mechanical Turk and CrowdFlower provide tools (e.g., API methods) for rapid generation of data collection forms with predefined choices. We hope the findings presented here will caution organizations looking to rely on such datasets, as they could be potentially incomplete as well as under-representative of the complete real-world phenomena. We further hope to inform platform providers of the advantages of the use-agnostic approach to UGC and hope to catalyze development of new use-agnostic APIs.

The use-agnostic approach to data collection also can be promising for scientific gathering of data. As our study suggests, participants can provide more information on the phenomena of interest, and report unanticipated phenomena, when assumptions about the kinds of information needed for later analysis, are relaxed. This echoes the concern raised by Hufnagel and Conca (1994) that “closed format questions” may lead to “instability in response data” especially when participants have low knowledge of the subject area. Following our findings we caution researchers when asking participants to fill predefined data collection fields (a common method in survey research). The empirical evidence in our study suggests, for example, that anchoring and insufficient adjustment may occur when predefined choices are made available. This could prevent research participants from communicating additional and unanticipated views potentially hindering researchers’ ability to discover novel perspective about phenomena. The findings further indicate that a potential cause for incomplete responses stems from the misalignment between the model of phenomena as perceived by the scientist and the model of phenomena as understood by the potential research participants. By identifying the potential for and causes of incomplete data, we hope to catalyze future research on more effective ways to gather scientific data.
An important question that remains open is the impact of the contributor-oriented IQ on data consumers. Here, one issue is whether organizations can take advantage of the novel affordances of contributor-oriented IQ. For example, data that is more faithful to the crowd’s perspective can be leveraged in designing better customer-facing products or services or redesigning internal processes to make them more agile and flexible (see Kharabe and Lytinen 2013). Future research can also examine challenges that data consumers (e.g., scientists) may face when interpreting and analyzing instance-based data as well as opportunity this data presents.

Limitations and Conclusions
We note several limitations of the present work. One general concern relates to the nature of empirical evidence obtained as a result of field experimentation. While using field experimentation offers advantages (discussed earlier), the results should be interpreted with caution. Working in a field setting raises common concerns about experimental control. One issue is ensuring that users in one condition were not experiencing treatments in different conditions. We tried to address this by using password authentication before any manipulation could be experienced. Having to enter (and remember) a user name and password, however, potentially deterred some (e.g., less determined) users from engaging. Another issue is whether the users of NLNature were representative of the broader population. We expended considerable effort to reach as many different segments of the population as possible (as expounded above). At the same time, the analysis of observations revealed an unexpectedly large proportion of species-level identifications - indicative of domain experts (Lukyanenko et al. 2011; Tanaka and Taylor 1991). This can be potentially explained by the volitional nature of the project where users with domain knowledge or interest in biology would be more likely to participate. While the presence of experts makes the evaluation of our theories more conservative, we caution readers when interpreting our findings, particularly those related to effect sizes, and ratios of generic classes and attributes. As our theory assumes a context where information contributors are non-experts with respect to the intended information uses by project sponsors (in this case, biologists), we expect the impact of modeling on completeness be even greater in purely novice populations. We are addressing these concerns in a follow-up laboratory experiment. Similarly, the issue of how instance-based data can be used in scientific analysis was beyond the scope of the current work. We hope to investigate this issue in future studies.

As organizations invite diverse and unpredictable user-generated content into internal decision making, they face the challenge of managing the quality of such datasets. Applications such as citizen science create opportunities to collect and analyze data in ways that are not otherwise possible. Despite the potential for online engagement with citizen science and online users in general, the prevailing assumptions and practices underlying data collection in these projects limit the amount of relevant information that organizations are able to harness. In this paper, we outline the root of this problem and suggest a solution designed to increase quantity of collected information in UGC settings.

References


