Citizen Science 2.0: Data Management Principles to Harness the Power of the Crowd

Roman Lukyanenko[1], Jeffrey Parsons[1], and Yolanda Wiersma[2]

[1] Faculty of Business Administration, Memorial University, St. John's Canada
[2] Department of Biology, Memorial University, St. John's Canada
{roman.lukyanenko,jeffreyp,ywiersma}@mun.ca

Abstract. Citizen science refers to voluntary participation by the general public in scientific endeavors. Although citizen science has a long tradition, the rise of online communities and user-generated web content has the potential to greatly expand its scope and contributions. Citizens spread across a large area will collect more information than an individual researcher can. Because citizen scientists tend to make observations about areas they know well, data are likely to be very detailed. Although the potential for engaging citizen scientists is extensive, there are challenges as well. In this paper we consider one such challenge – creating an environment in which non-experts in a scientific domain can provide appropriate and accurate data regarding their observations. We describe the problem in the context of a research project that includes the development of a website to collect citizen-generated data on the distribution of plants and animals in a geographic region. We propose an approach that can improve the quantity and quality of data collected in such projects by organizing data using instance-based data structures. Potential implications of this approach are discussed and plans for future research to validate the design are described.

Keywords: design, citizen science, management, database design, conceptual modeling, data quality.

1 Introduction

Citizen science is a term used to describe the voluntary participation of amateur scientists in scientific endeavors [1]. Humans are increasingly regarded as effective sensors of their environment [2] and the potential for using information collected by individuals is continuously expanding [3]. Citizen science has a long tradition. During the Victorian era many wealthy individuals engaged in natural history as a hobby, and made contributions to the understanding of species distributions and behavior as a result. With the development of the Internet, it has become easier for ordinary people to participate and contribute large amounts of information. Yet, given the expertise and language gap between scientists and ordinary people, information transfer in citizen science projects is not straightforward. While citizen scientists can offer insights and generate new ideas [4], their lack of training and expertise results in inconsistent and incorrect data [5,6,7]. In particular, where direct elicitation of people's opinions is required we can expect lower scientific accuracy of data as wider

H. Jain, A.P. Sinha, and P. Vitharana (Eds.): DESRIST 2011, LNCS 6629, pp. 465–473, 2011.

audiences with lesser expertise get engaged. This research attempts to address this problem by suggesting data management principles that maximize the quantity and quality of information collected from non-experts.

There are many advantages to harnessing citizen scientists. Participants spread across a large area will collect more information than an individual researcher can. Because citizen scientists tend to make observations about areas they know well, data are likely to be very detailed. An additional advantage is the potential longevity of such data; some citizen science programs (e.g., the Audubon Christmas Bird Count) have been in existence for over 100 years, resulting in data sets extending over long periods, thus enabling analysis of trends. Coupled with the availability of relatively inexpensive photo and video equipment, harnessing the power of ordinary people to provide data and observations about the natural world can lead to major advances in the natural sciences, as well as assist in vital areas of wildlife conservation and emergency management in the event of natural disasters (such as the Gulf of Mexico oil spill).

Although the potential for engaging citizen scientists is extensive, there are challenges as well. In this paper we describe one such challenge – creating an online environment in which non-experts in a scientific domain can provide appropriate and accurate data regarding their observations. We describe the problem in the context of a research project that includes the development of a website and database to collect citizen-generated data on the distribution of plants and animals in a geographic region. We propose an approach to improving the quantity and quality of data collected in such projects by using instance-based data structures [8]. Potential implications of this approach are discussed and plans for future research to validate the design are described.

2 The Challenge – Facilitating Participation

The success of a citizen science project depends on the willingness and ability of members of the general public to voluntarily observe and report information. In many cases, this in turn requires some level of scientific knowledge by participants. For example, the website of the Cornell Ornithology Lab, eBird (www.ebird.com), draws on the enthusiasm of avid birders to provide detailed information about bird sightings. The Cornell Lab is an international leader in ornithological research, and eBird is an exemplar of a successful online citizen science project. However, engagement of the lay public with eBird may be limited by the application domain. Citizen scientists who wish to upload bird sightings need to be familiar with bird taxonomy and identification. The bird checklist provided in the online interface assumes the user has already made a positive identification (i.e., identified the species) and knows to which taxonomic group the bird belongs. This is acceptable for a reasonably experienced citizen scientist, but the rank beginner ([7] provides a taxonomy of "expertise levels" among citizen scientists) may not be able to participate, or may provide data of poor quality as a result of his/her inability to make a positive identification [9]. Thus, useful participation may be limited to more experienced amateurs.

The issue of quality and reliability of user-supplied data in citizen science projects has attracted much attention in recent research [5]. Although the literature is limited (given the relative recency of Web 2.0 applications), the implied assumption of much of the work to date is that there exists an inherent trade-off between data quality and the level of participation (data quantity). Experts are considered to be the source of the most accurate volunteered information [7], but there are fewer "expert amateurs" than "beginners" available to participate.

The common method of increasing data quality considered in the literature is training and educating the volunteers. For example, data inconsistency may result from volunteers' lack of experience, inadequate guidelines and insufficient training [4], "rolled up" into larger monitoring projects [6]. Training, while generally desirable, may not always be possible, especially for low budget projects.

A typical way to increase quality is through expert verification, an approach that has been used for by-catch and beached bird observation [12] and for unusual observations on eBird [19]. However, with the size of data sets increasing [6], individual verification becomes unrealistic and in many ways is contrary to the spirit of citizen science.

Another line of research suggests social networking as key to increasing data quality. Some research has proposed a trust and reputation model for classifying knowledge using the social networking practice of peer evaluation of content [13]. This approach is the basis for iSpot, a website that exploits a user reputation mechanism to determine accuracy of observations [14]. The reputation-trust model adapted from well-developed trust research in e-commerce [e.g., 15,16,17] has been applied to the context of citizen science [18]. While the social networking approach appears promising, it has a number of limitations. Although it has been compared to the "scientific peer review process" [13], social networking is useful only for popular citizen science projects with large numbers of users. Web sites with a small number of users may not have sufficient user activity per observation to ensure rigorous peer review. In addition, as even a very popular website cannot guarantee that every observation will receive equal scrutiny, this metaphor of scientific review is not fully justified. Furthermore, users with high reputation who are considered experts in some domain may still provide inaccurate data in other domains. Most importantly, social networking may fail to harness the potential of an individual non-expert, as in the absence of domain knowledge such volunteers may feel too intimidated to express their opinion (consider the description of a type 'neophyte' [7]). Finally, the social networking approach lacks generality, as it relies on a particular technology, and may exclude many citizen science projects that do not currently employ a social networking model.

Notwithstanding the value of the above approaches, we argue that it is possible to increase the quality of data generated by of an individual volunteer by minimizing subject information that has a high likelihood of being inaccurate. Requiring volunteers to make a (potentially inaccurate) positive identification of natural history phenomena implies that the observer has some knowledge of traditional scientific taxonomy. We argue that an alternative to classifying observations according to a fixed taxonomy is to allow volunteers to provide information about observations and that this will increase the general success of citizen-scientist projects.

3 A Proposed Solution – Attribute-Based Data Collection

A traditional approach to citizen participation in scientific data collection works well (i.e., makes it possible to collect accurate data from a broad constituency) only if the participants are capable of classifying observed phenomena accurately. For example, accurate classification of observed plants and animals by species requires that participants understand the distinguishing characteristics of species. We contend that imposing this requirement on participation, as in projects such as eBird, imposes a severe and unnecessary restriction on the level of participation that can be realized in citizen science projects.

To combat this limitation, we propose an approach to data collection and storage that does not require users to classify observed phenomena. Instead, they record any attributes associated with the observation. We illustrate the approach in the context of NLNature – an ongoing citizen scientist-based project to collect data about the flora and fauna of Newfoundland & Labrador (www.nlnature.com). Our proposal is based on the instance-based data model (IBDM) [8] and our application of the model has implications both for interface design and for database design. Working within the framework of the IBDM, we extend the model to address issues of identifying phenomena, and suggest how the model offers a solution to the challenges of a typical citizen scientist project.

The IBDM is based on ontological and cognitive principles [8, 20]. Ontologically, every "thing" possesses a unique set of properties. Classes are formed based on the principle that one can classify things based on a subset of their observed properties, and make inferences about unobserved properties the instance possesses by virtue of belonging to the class [21]. Since an instance can possess very many properties, it can belong to a very large number of potential classes, depending on the context.

By shifting the focus from a predefined classification to the thing (instance) and its attributes (see Fig. 1) we do not need to model a domain *a priori* in terms of the classes of interest. It is sufficient to ensure that the application has a comprehensive collection of instances, and each instance contains a set of well-defined attributes. When required, a user can assemble a dynamic classification based on the collection of attributes that are of interest at a given moment. For example, if an attribute such as "behavior" is of interest, then at least two classes can be constructed based on values: animals that are nocturnal (active at night) vs. diurnal (active during the day). The same system can also use attributes that connect each species with a biological taxonomy to reproduce scientific biological classification. Thus, the instance-based model is capable of achieving the objectives of a traditional classification without the inherit limitations.

We posit that attribute-based design will enable potential citizen scientists to provide data efficiently and effectively, thereby increasing their participation in data gathering. We propose a data collection interface designed based on the primacy of a phenomenon and its attributes over classification of the phenomenon. A user is asked to identify those attributes (e.g., size, color, appearance, behavior, location, sound) of an observed plant or animal. In principle, the primary scientific object of an observation (the species observed) can be identified by an expert after the observation is recorded, provided that the user reports enough attributes to produce a positive

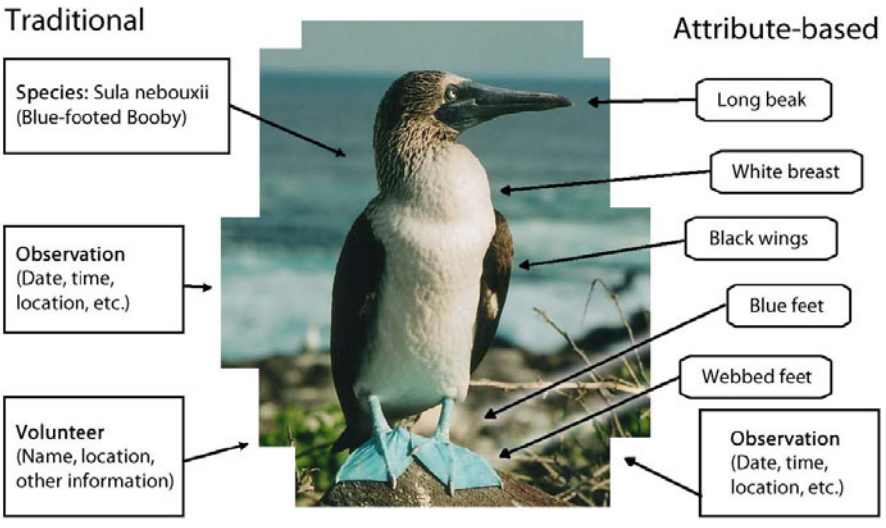

Fig. 1. Traditional vs. attribute-based information (Image source: Wikimedia Commons)

identification. This contrasts with traditional approaches requiring *a priori* classification (e.g., requiring users to select from a checklist of species), which are usable only by more expert volunteers. Once several attributes are selected, the system will match them with pre-existing sets of identifying attributes for species, and either infer a species or ask for additional attributes that could also be automatically inferred from those previously supplied.

Although the final attribute set resulting from an observation can potentially match multiple species, this proposed solution offers a realistic compromise. Non-experts do not always know the phenomenon that was observed. It is more realistic to expect a volunteer to remember some features of unknown species then to expect a precise classification and identification. The key activity of identification therefore shifts from designing a perfect classification to facilitating effective attribute management. The more the system can guide the choice of attributes, the higher inferential value such records hold, and the easier it is to classify observations.

4 Attribute-Based Database Design

Database structure can be either a major inhibitor or a facilitator of system evolution [8, 20, 22]. Traditionally, database design results in a representation of the application domain as a set of related classes (translated to tables in a relational database). In addition, once the database structure is established it is assumed to be relatively static, allowing other application elements, such as program code, to be created based on the static structure. Altering the database structure once a system is built is costly. Thus, traditional database design is subject to the inherit limitations of a rigid classification [8].

The collection of user-supplied information based on attributes of observations suggests the need for a database structure that supports variability in the data collected from observers, including failure to classify an observation. Support for flexible attribute collection can be implemented using a traditional relational database, as illustrated in Fig. 2. We propose storing attributes in a generic table "Attributes" that contains attribute name and a unique identifier. A separate "Attributes-Relationships" table links one attribute to another and creates relationships between attributes. The table contains the primary key from the parent attribute and a primary key from the child attribute, thus making many-to-many relationships possible. For example, if the user selects the attribute "was flying" then "lives in water" will be automatically removed from the interface, and the system will respond by presenting a new set of potential attributes that can be inferred from "was flying" (e.g., "has feathers", "seen at night", "six legs"). The choice of the first attribute narrows the observation to a bird (subsequent attributes could focus on feather color, beak size, habitat, etc.), the second to a bat, and the third to a flying insect.

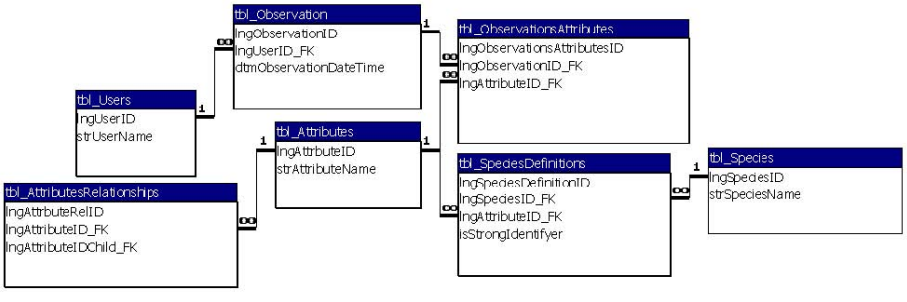

Fig. 2. ER diagram showing instance-based data structure for a typical citizen scientist project

In order to match selected attributes against a class-defining set, class blueprints for each species need to be maintained. This is achieved by a table of "Species Definitions" that links species with their attributes via a one-to-many relationship. For example, boreal felt lichen will link to the following attributes: fuzzy white fringe around the edges, grayish-brown when dry, has red dots, leafy, slate-blue when moist. User-observed properties are then matched against the class definitions to infer class membership. If necessary, new class definitions can be added or existing ones altered during the operational phase of the enterprise system without having to change the database schema. Finally, we provide tables that join objects and attributes to store the details of the observations. These tables store events in the system. Each table includes primary keys from the attributes and objects tables, attribute values, and date/time of the attribute creation/change. By recording the date of attribute creation/change, the system can document events that happen to the same phenomenon. This approach addresses a persistent issue of database design – adaptation to organizational change – that a traditional approach with its reliance on rigid classification struggles to resolve [8, 22].

5 Implications for Data Gathering

The attribute based system proposed for this citizen scientist project has the potential to increase participation rates (and, hence, data quantity). Unlike natural history websites that only present taxonomic checklists and assume a basic level of expertise from citizen scientists, the system proposed here allows for the full spectrum of volunteer contributors [7] to participate. We believe that this will provide a means of validating user-supplied data within the user community, particularly if users supply additional information with their observations (e.g., photographs) that can be reviewed by experts when necessary.

Many citizen science projects provide inventory data across space and time. Although there will be biases within the data (for example, to areas where there is high human population density and to more charismatic or easily observable species), the data do have the benefit of indicating long-term trends. For the scientific community, the biggest value is that such data sets are generated by many "eyes on the ground;" thus, there is a higher likelihood of rare or unusual species being detected or for early detection of new trends. Hence, it is important to have a usable system that promotes a broad and consistent level of participation. Some potential uses of data collected this way might be unanticipated. For example, long term data can be useful to identify benchmark conditions in the event of a natural or anthropogenic disaster (e.g., the Gulf oil spill) and can guide restoration strategies.

This research explores general ways of facilitating information transfer between users with different level of domain expertise within the context of a citizen science project. Information systems are increasingly being used to collect data from ordinary people (e.g. personal health records [23]). While a number of factors are considered to influence information quality (e.g., [24]), little attention is given to the role of data structures in ensuring quality of collected information.

6 Limitations and Future Research

Internet technologies open new opportunities for citizen science. Yet the knowledge requirements implied by rigid data structures constrain effective participation of novices and thereby limit the potential outreach of citizen science projects. A successful implementation of the approach proposed in this paper can facilitate development of citizen-scientist initiatives. We believe it also has broader applications based on user-generated content, and promises to be a practical solution to an important design problem in citizen science.

The foundation of our proposed approach to improving the quantity and quality of citizen science projects is the IBDM [8]. The primary theoretical assumption of the IBDM – that existence of *things* and *properties* (attributes) precedes classification - has generally [cf. 25] been supported in ontological [26,27] and cognitive research [28]. However, while attributes are building blocks of classification [29], not all classes can be *efficiently* expressed as sets of common attributes (e.g., radial categories [30,31,32]). Moreover, many superordinate categories, such as *furniture, animal, vehicles* tend to be abstract and reflect some rules or functions rather than observable attributes [33-34]. While this appears to limit the scope of our model,

we believe that for practical reasons little information in citizen science projects will be expressed in terms of higher-level categories. Indeed, humans prefer to avoid superordinate categories when they think of individual objects [35].

Classification is a ubiquitous activity and an attribute-centered approach to knowledge management needs to be tested to determine its technological, economic, scientific and business utility. We are currently designing empirical studies to measure the practical impact of the above approach on data collection and storage, user participation and satisfaction, data quality, and usefulness to scientists. The experiment will also test the overall effectiveness and feasibility of applying the IBDM to empower citizen scientists.

References

1. Silvertown, J.: A new dawn for citizen science. Trends in Ecology and Evolution 24, 467–471 (2009)
2. Goodchild, M.: Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221 (2007)
3. Hand, E.: People power. Nature 466, 685–687 (2010)
4. Foster-Smith, J., Evans, S.M.: The value of marine ecological data collected by volunteers. Biological Conservation 113, 199–213 (2003)
5. Flanagin, A., Metzger, M.: The credibility of volunteered geographic information. GeoJournal 72, 137–148 (2008)
6. Wiersma, Y.F.: Birding 2.0: citizen science and effective monitoring in the Web 2.0 world. Avian Conservation and Ecology 5, 13 (2010)
7. Coleman, D.J., Georgiadou, Y., Labonte, J.: Volunteered geographic information: The nature and motivation of producers. International Journal of Spatial Data Infrastructures Research 4, 332–358 (2009)
8. Parsons, J., Wand, Y.: Emancipating instances from the tyranny of classes in information modeling. ACM Transactions on Database Systems 25, 228 (2000)
9. Parsons, J., Lukyanenko, R., Wiersma, Y.: Easier citizen science is better. Nature 471, 37 (2011)
10. Dickinson, J.L., Zuckerberg, B., Bonter, D.N.: Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41, 112–149 (2010)
11. Aaron, W.E.G., Tudor, M.T., Haegen, W.M.V.: The Reliability of Citizen Science: A Case Study of Oregon White Oak Stand Surveys. Wildlife Society Bulletin 34, 1425–1429 (2006)
12. Hamel, N.J., Burger, A.E., Charleton, K., Davidson, P., Lee, S., Bertram, D.F., Parrish, J.K.: Bycatch and beached birds: Assessing mortality impacts in coastal net fisheries using marine bird strandings. Marine Ornithology (2009)
13. Bishr, M., Mantelas, L.: A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal 72, 229–237 (2008)
14. Silvertown, J.: Taxonomy: include social networking. Nature 467, 788–788 (2010)
15. Komiak, S.Y.X., Benbasat, I.: The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS Quarterly 30, 941–960 (2006)
16. Gefen, D., Karahanna, E., Straub, D.W.: Trust and TAM in online shopping: An integrated model. MIS Quarterly 27, 51–90 (2003)

17. Palvia, P.: The role of trust in e-commerce relational exchange: A unified model. Information & Management 46, 213–220 (2009)
18. Alabri, A., Hunter, J.: Enhancing the quality and trust of citizen science data. In: IEEE eScience 2010, Brisbane, Australia (2010)
19. Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S.: eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142, 2282–2292 (2009)
20. Parsons, J., Su, J.: Analysis of data structures to support the instance-based data model. In: International Conference on Design Science Research in Information Systems and Technology (DESRIST), pp. 107–130 (2006)
21. Parsons, J., Wand, Y.: A question of class. Nature 455, 1040–1041 (2008)
22. Allen, B.R., Boynton, A.C.: Information architecture: In search of efficient flexibility. MIS Quarterly 15, 435–445 (1991)
23. Agarwal, R., Angst, C.M.: Technology-Enabled Transformations in U.S. Health Care: Early Findings on Personal Health Records and Individual Use. In: Galletta, D., Zhang, P. (eds.) Human-Computer Interaction and Management Information Systems: Applications. M.E. Sharpe, Inc., Armonk (2006)
24. Nicolaou, A.I., McKnight, D.H.: Perceived information quality in data exchanges: Effects on risk, trust, and intention to use. Information Systems Research 17, 332–351 (2006)
25. Grill-Spector, K., Kanwisher, N.: Visual recognition. Psychological Science 16, 152–160 (2005)
26. Bunge, M.A.: Treatise on Basic Philosophy: The furniture of the world. Reidel, Dordrecht (1977)
27. Wand, Y., Weber, R.: An ontological model of an information system. IEEE Transactions on Software Engineering 16, 1282–1292 (1990)
28. Bowers, J.S., Jones, K.W.: Detecting objects is easier than categorizing them. Quarterly Journal of Experimental Psychology 61, 552–557 (2008)
29. Wand, Y., Monarchi, D.E., Parsons, J., Woo, C.C.: Theoretical foundations for conceptual modeling in information systems development. Decision Support Systems 15, 285–304 (1995)
30. Raccoon, L.S.B., Puppydog, P.O.P.: A middle-out concept of hierarchy (or the problem of feeding the animals). SIGSOFT Softw. Eng. Notes 23, 111–119 (1998)
31. Young, J.J., Williams, P.F.: Sorting and comparing: Standard-setting and "ethical" categories. Critical Perspectives on Accounting 21, 509–521 (2010)
32. Lakoff, G.: Women, fire, and dangerous things: what categories reveal about the mind. University of Chicago Press, Chicago (1987)
33. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyesbraem, P.: Basic objects in natural categories. Cognitive Psychology 8, 382–439 (1976)
34. Murphy, G.L., Wisniewski, E.J.: Categorizing objects in isolation and in scenes - What a superordinate is good for. Journal of Experimental Psychology: Learning 15, 572–586 (1989)
35. Rorissa, A.: User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. Information Processing & Management 44, 1741–1753 (2008)