

**USING GUIDED DATA COLLECTION TO IMPROVE THE
QUALITY OF CITIZEN SCIENCE USER-GENERATED CONTENT**

by

© Haniyesadat Razavi

**A Dissertation submitted to the
School of Graduate Studies
in partial fulfillment of the requirements for the degree of**

Master of Science

**Faculty of Business Administration
Memorial University of Newfoundland**

**May 2018
St. John's, Newfoundland and Labrador**

Abstract

With the advent of Web 2.0, there has been tremendous growth in User-Generated Content (UGC), wherein members of the general public participate in contributing information online. Citizen science is a popular form of UGC in which participants support scientific data collection or analysis. However, in projects that rely on citizens to contribute data, obtaining data that is of sufficient quality to be useful for research is challenging. Among the challenges in obtaining data are: lack of control over the content of data supplied; lack of incentive to contribute; and lack of system flexibility to capture unanticipated data. Any of these challenges may lead to low-quality data that might not be useful in scientific research. Improving the data collection phase in online citizen science may facilitate capturing higher quality data. The primary purpose of this research is to propose and evaluate guidance features to support data entry to increase the quality of data collected. An experiment under three different conditions was conducted based on a citizen science project in the biology domain. Three types of guidance were tested to determine which is more effective in assisting the contributors in species identification (a widely used level of classification that is useful in biology research). The results demonstrate that using a guidance feature assists contributors in identifying species. Moreover, the guidance enables contributors to provide data of better quality in terms of relevance and objectivity. This thesis concludes by summarizing implications and provides suggestions for future study.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor, Dr. Jeffrey Parsons, for the continuous support of my M.Sc. study and research, his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor for my study.

My sincere thanks also go to all of those with whom I have had the pleasure to work during this research project.

I wish to thank my loving and supportive spouse, Mojtaba, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without him.

TABLE OF CONTENTS

| | |
|--|-----------|
| ABSTRACT..... | I |
| ACKNOWLEDGEMENTS | II |
| LIST OF TABLES | V |
| LIST OF FIGURES | VI |
| 1 INTRODUCTION..... | 1 |
| 1.1 BACKGROUND AND MOTIVATION | 1 |
| 1.1.1 User-generated Content | 1 |
| 1.2 THE PROBLEM OF DATA QUALITY IN UGC..... | 4 |
| 1.3 OBJECTIVE | 6 |
| 1.4 THESIS ORGANIZATION..... | 7 |
| 2 LITERATURE REVIEW | 8 |
| 2.1 DATA & DATA QUALITY | 8 |
| 2.2 DATA QUALITY IN UGC | 11 |
| 2.2.1 Approaches to Improving Data Quality in UGC..... | 12 |
| 2.3 RECOMMENDATION SYSTEMS | 17 |
| 2.3.1 Content-Based Recommendation systems | 18 |
| 2.3.2 Collaborative Recommendation Systems | 18 |
| 2.3.3 Hybrid Recommendation Systems | 19 |
| 3 IMPACT OF USING RECOMMENDATION SYSTEM ON DATA QUALITY | |
| 22 | |
| 3.1 IMPACT OF USING RECOMMENDATION SYSTEM ON SPECIES IDENTIFICATION..... | 24 |

| | | |
|----------|---|-----------|
| 3.2 | IMPACT OF USING RECOMMENDATION SYSTEM ON DATA RELEVANCE | 26 |
| 3.3 | IMPACT OF USING RECOMMENDATION SYSTEM ON DATA OBJECTIVITY & NUMBER OF CAPTURED ATTRIBUTES..... | 27 |
| 4 | EXPERIMENTAL DESIGN..... | 29 |
| 4.1 | METHOD | 30 |
| 4.2 | MATERIALS & PROCEDURES..... | 30 |
| 4.2.1 | <i>Condition 1 (No Guidance)</i> | 32 |
| 4.2.2 | <i>Condition 2 (Traditional Guidance)</i> | 33 |
| 4.2.3 | <i>Condition 3 (Cognitive Guidance)</i> | 37 |
| 5 | DATA ANALYSIS & DISCUSSION | 42 |
| 5.1 | DATA ANALYSIS | 42 |
| 5.1.1 | <i>Bird Identification (H-1.1)</i> | 46 |
| 5.1.2 | <i>Data Relevance (H-1.2)</i> | 51 |
| 5.1.3 | <i>Data Objectivity & Data Completeness (H-1.3, H-1.4)</i> | 54 |
| 5.2 | DISCUSSION | 56 |
| 6 | CONCLUSION, IMPLICATIONS, AND FUTURE STUDY..... | 58 |
| 6.1 | CONCLUSION | 58 |
| 6.2 | IMPLICATIONS..... | 59 |
| 6.3 | FUTURE STUDY | 60 |
| | REFERENCES..... | 63 |
| | APPENDIX 1: IMAGES USED IN THE LABORATORY EXPERIMENTS..... | 68 |

List of Tables

| | |
|---|----|
| TABLE 1.1 WELL-KNOWN CITIZEN SCIENCE PROJECTS FROM DIFFERENT DOMAINS | 2 |
| TABLE 2.1 TYPES OF DATA | 8 |
| TABLE 2.2 NOTABLE DATA QUALITY DIMENSIONS..... | 10 |
| TABLE 5.1 THE RESULT OF NO GUIDANCE CONDITION FOR EACH IMAGE | 43 |
| TABLE 5.2 THE RESULT OF TRADITIONAL GUIDANCE CONDITION FOR EACH IMAGE..... | 44 |
| TABLE 5.3 THE RESULT OF THE COGNITIVE GUIDANCE CONDITION FOR EACH IMAGE | 45 |
| TABLE 5.4 BASIC-LEVEL IDENTIFICATION IN NO GUIDANCE (NG), TRADITIONAL GUIDANCE (TG), AND COGNITIVE GUIDANCE (CG) CONDITIONS..... | 47 |
| TABLE 5.5 EXPERT-LEVEL IDENTIFICATION IN NO GUIDANCE (NG), TRADITIONAL GUIDANCE (TG), AND COGNITIVE GUIDANCE (CG) CONDITIONS..... | 48 |
| TABLE 5.6 A COMPARISON OF THE NUMBER OF IDENTIFICATIONS AT BASIC-LEVEL & EXPERT-LEVEL | 49 |
| TABLE 5.7 COMPARISON OF THE NUMBER OF TOTAL ATTRIBUTES CAPTURED IN EACH CONDITION | 51 |
| TABLE 5.8 T-TEST FOR THE NUMBER OF TOTAL CAPTURED ATTRIBUTES..... | 52 |
| TABLE 5.9 T-TEST FOR PERCENTAGE OF RELEVANT ATTRIBUTES EACH CONDITION CAPTURED..... | 52 |
| TABLE 5.10 THE COMPARISON OF DATA RELEVANCY IN NO GUIDANCE (NG), TRADITIONAL GUIDANCE (TG), AND COGNITIVE GUIDANCE (CG) CONDITIONS..... | 53 |
| TABLE 5.11 OBJECTIVITY OF RESPONSES IN NO GUIDANCE (NG), TRADITIONAL GUIDANCE (TG), AND COGNITIVE GUIDANCE (CG) CONDITIONS..... | 54 |
| TABLE 5.12 T-TEST FOR PERCENTAGE OF OBJECTIVE ATTRIBUTES EACH CONDITION CAPTURED | 55 |

List of Figures

| | |
|--|----|
| FIGURE 4.1 THE EXPERIMENT WELCOME PAGE | 31 |
| FIGURE 4.2 THE INTERFACE OF THE “NO GUIDANCE” CONDITION | 33 |
| FIGURE 4.3 THE INTERFACE OF THE “TRADITIONAL GUIDANCE” CONDITION | 35 |
| FIGURE 4.4 HOW THE INTERFACE CHANGES IN THE “TRADITIONAL GUIDANCE” | 35 |
| FIGURE 4.5 THE PROCEDURE FOR THE “TRADITIONAL GUIDANCE” CONDITION | 36 |
| FIGURE 4.6 TRADITIONAL GUIDANCE WITH A TEXT BOX FOR ADDITIONAL INFORMATION | 37 |
| FIGURE 4.7 THE INTERFACE OF THE COGNITIVE GUIDANCE CONDITION | 39 |
| FIGURE 4.8 THE ALGORITHM DEVELOPED FOR THE “COGNITIVE GUIDANCE” | 41 |

1 Introduction

1.1 Background and Motivation

1.1.1 User-generated Content

The advent of Web 2.0, along with the development of mobile technology, allows people to produce digital content, resulting in an enormous volume of data being created online. The term user-generated content (UGC) is used to describe *any form of content- such as tweets, blogs, wikis, discussion forum posts, video, and audio - created by members of general public who voluntarily contribute, rather than by employees or others closely associated with an organization* (Krumm et al., 2008) .

UGC has received significant interest in recent years. Factors contributing to this growing interest include: (1) the increasing number of users on social networks like Facebook and Twitter, (2) the proliferation of crowdsourcing projects wherein members of the general public are asked to perform certain tasks (Lukyanenko, 2014), and (3) gratification from being recognized and being able to articulate views, thoughts, and experiences through online data contribution (Krumm et al., 2008; Leung, 2009).

UGC supports decision making and analysis in different domains. Businesses can use UGC to monitor what their customers are saying (Di Gangi et al., 2010). On the other hand, customers also have access to opinions from other customers which can be very helpful in purchasing (Dhar & Chang, 2009). In healthcare, generating UGC by developing a digital platform to capture patients' reviews can be considered a beneficial tool to improve the quality of services provided (Gao et al., 2010). Governments and social

applications can also benefit by using UGC. For example, OpenStreetMap (Haklay & Weber, 2008) and CitySourced (www.citysourced.com) are instances of UGC projects launched in recent years to support governmental and civic activities.

Scientists are also increasingly using UGC as a tool for expanding scientific knowledge and literacy. Citizen science projects have been developed to involve member of the general public in scientific research. According to Bhattacharjee (2005), citizen science can be defined as *a research technique that involves the public in obtaining scientific data across a broad geographic region (like a country) over a large period of time*. For this purpose, a specific type of information system has been built to capture the knowledge of the general public. With the aim of advancing scientific knowledge, citizen science has been employed in many projects, including classifying galaxies, deciphering ancient scripts, identifying species, and mapping the planet (Lukyanenko, 2014). Table 1.1 lists some well-known citizen science projects in different contexts.

Table 1.1 Well-known citizen science projects from different domains

| Project Name | Sponsor(s) | Start Date | Target | Website |
|---------------------------|---|-------------------|------------------------|-----------------|
| eBird | Cornell Lab of Ornithology and National Audubon Society | 2002 | Orthinology | ebird.org |
| Galaxy Zoo | Zooniverse collaboration | 2007 | Galaxies | Galaxyzoo.org |
| Atlas of living Australia | Birdlife Australia | 2010 | Plants, Animals, Fungi | ala.org.au |
| iSpot | The Open University | 2008 | Nature | ispotnature.org |
| Naturewatch | Nature Canada | 2000 | Nature | naturewatch.ca |

As shown in Table 1.1, citizen science has been applied in different context, mostly ecology. Interest in the conservation of species has led scientists to gather citizen-generated data on the distribution, abundance, habitat preferences, and movements of organisms across broad geographic areas and over long period of time (Hochachka et al., 2012). However, the cost and availability of experts to collect data has always been a challenge. Citizen science promises to reduce information acquisition costs and facilitate discoveries (Hochachka et al., 2012). Another example of citizen science projects in the ecology domain is Volunteered Geographical Information (VGI). Sites such as Wikimapia and OpenStreetMap are enabling citizens to create a global collection of geographic information and learn a great deal about remote places (Goodchild, 2007).

Whereas data quality in biology has established standards (e.g. biological terminology), this thesis analyzes the quality of content generated in a citizen science context and proposes that the quality of supplied data can be positively influenced by developing a guidance feature to better match contributor capabilities.

Citizen science projects now yield both scientific and educational outcomes. They help participants learn about the objects they are observing and to experience the process of a scientific investigation, rather than just gathering a vast amount of data (Bonney et al., 2009). Although most citizen science projects benefit human participations to collect data, “Gravity Spy”¹ relies on machine learning methods as well as contributors’ data. In that project, machine learning algorithms learn from the contributors’ data and produce new

¹ www.arxiv.org/abs/1611.04596

data (Zevin et al., 2017). Another example of coupling of machine learning and human contribution is “Zooniverse,”² which focuses on data mining activities and machine learning algorithms that are being applied to the contributed data (Simpson et al., 2014).

A local example of citizen science project in the biology domain, NLNature³ was initiated in 2007 by Dr. Yolanda Wiersma, a biologist at Memorial University, with the aim of engaging the general public with issues of environmental change. In 2013, a new approach for modeling the captured data was proposed and the website was redesigned accordingly (Lukyanenko, 2014). NLNature now features an instance-based model for data collection. Although the instance-based model is promising in terms of data accuracy and dataset completeness, it can result in some limitations (Lukyanenko et al., 2014). This thesis proposes a new method for data collection that can mitigate the negative impact of using an instance-based model in a citizen science project.

1.1.2 The Problem of Data Quality in UGC

In spite of numerous advantages, UGC has its risks and challenges. First, in projects that rely on citizens to contribute data, it is difficult to control the content of data supplied in term of accuracy. Participants may have little knowledge about the domain, so they will not care too much about the project success. For this problem, having some level of domain knowledge is desirable. Second, there is a lack of incentives among contributors, specifically if the data contribution process is difficult. For instance, if a computer interface requires data with a high scientific precision level, which contributors are unable to

² www.zooniverse.org

³ www.nlnature.com

provide, this can lead to low levels of contribution. Finally, a lack of system flexibility to capture unanticipated data may result in information loss. When participants are asked to report a particular observation, they may wish to report data which is not anticipated by the project sponsor or data consumer. The system should be flexible in order to let contributors to provide whatever they perceive (Lukyanenko et al., 2014).

Any of these challenges may lead to low quality data. To make effective use of UGC, the quality of data supplied by the general public should be maintained (Alabri & Hunter, 2010). If an information system is considered as a manufacturing system in which information is produced, three different roles, dealing with data quality, can be identified: (1) Data suppliers who are responsible for data collection, (2) Data manufacturers who are responsible for maintaining data, and (3) Data consumers who use data for analysis and decision making process (Wang, 1998). Given that data entry is the first step to produce data in an information system, finding an effective way to collect data of better quality can be considered as one of the first steps in a Data Quality Management (DQM) process. Developing a proper way of data collection can mitigate the negative impact of misspelling, missing information, or invalid data which impairs the quality of data (Barchard & Pace, 2011).

Traditional solutions to DQM, like training the operators who are responsible for data entry (Redman, 1996), developing a proper application to validate or control data supplied, and designing a user-friendly interface, can decrease the negative impact of data errors. Since UGC is provided by online users, it is often infeasible to train online contributors to provide high quality data (Lukyanenko & Parsons, 2015). Moreover, in

UGC data is often accepted by the system with little or no control/validation. Given the limitations of traditional approaches to data quality in UGC, novel approaches are needed.

This thesis will examine the impact of using “guidance” based on a recommendation system to assist contributors to supply data of better quality.

1.2 Objective

Considering the importance of data quality in UGC, this thesis examines the impact of using an extant method in information system (IS), but new in UGC settings - *recommendation systems* - on the quality of data supplied online. Recommendation systems (RS) have been used extensively in E-commerce to help online users find desirable information (Fernández-Tobías et al., 2012). They aim to filter irrelevant data, and present those that better suit the users’ interest according to user’s personal profile or behavior.

This thesis claims that in some UGC settings (specifically a citizen science project in the biology domain), an RS can guide contributors in providing data at a level of value to data consumers (i.e. species identification) when reporting an observation. The recommendations are offered based on what a contributor has submitted to the system and an existing database. A verified database, by an expert, from a citizen science project can be used as the resource for the RS. For this thesis, the database contains all key features of species by which they can be identified, such as overall shape, pattern, and coloring. This thesis claims that using a guidance feature based on an RS may result in a higher number of correct species identification. Besides, this thesis claims that using an RS-based guidance results in better data objectivity and higher level of relevance – captured data is applicable and helpful for the intended purpose (Wang & Strong, 1996). Therefore, the

research question of this thesis is:

Research question: How might a recommendation system affect the quality of data in a citizen science setting?

1.3 Thesis Organization

The next chapter takes a closer look at the definition of data quality, and the problem of data quality in the context of UGC. Moreover, it reviews current approaches to improve data quality in UGC settings. Also, recommendation systems are fully explained in this chapter.

Chapter 3 provides the rationale for using an RS-based guidance to improve the quality of data. It also shows that which dimensions of quality will be affected by an RS.

Chapter 4 presents a laboratory experiment under three conditions to test hypotheses. The hypotheses, which are based on hypotheses from chapter3, examine the effect of using an RS-based feature on *species identification*, *data relevancy*, *data objectivity*.

Chapter 5 includes the data analysis and general discussion. The limitation of the proposed method is also explained.

The thesis concludes by summarizing the research contribution and its implication to practice, and suggests several areas for future research.

2 Literature Review

2.1 Data & Data quality

Data is defined as *things known or assumed as facts, making the basis of reasoning or calculation* in philosophy. In computing science, data represent *real world objects, with ability of storing, retrieving and elaborating through a software process and can communicate via a network* (Batini & Scannapieco, 2010). Researchers have suggested different classifications for data in different domains. In the field of data quality, three types of data can be identified. Table 2.1 presents this classification.

Table 2.1 Types of Data

| Type of Data | Definition | Example |
|----------------------|--|--|
| Structured Data | Data with a pre-defined data model | Relational Databases |
| Unstructured Data | Data with no pre-defined data model | Body of survey Questionnaire with free form data entry |
| Semi-structured Data | Data that have a structure with some degree of flexibility | XML |

There is also another classification for data types which considers data as a product. Based on this model, data have been classified into three types: (1) raw data- which are unprocessed (2) component data- which are constructed from the raw data and stored temporarily until the final information is derived, and (3) information products- which are the final outcome of data processing (Batini & Scannapieco, 2010).

Data quality also has different meanings in different contexts. According to DQM, data quality can be defined as *data that is appropriate for use or to meet user needs* (Sidi et al., 2012) or *data fitn for use* (Wang, 1998), which implies relativity in the sense that

data with good quality for one use may be considered poor or insufficient for another use (Wand & Wang, 1996).

The consequences of poor quality data can be experienced in every day's life. For example, a letter which was delivered wrongly is usually seen as the result of a malfunctioned postal service, but a closer look often reveals that data-related errors are the main cause. Moreover, poor data quality can have a severe impact on the overall effectiveness of an organization (Wand & Wang, 1996), as well as impose costs and risks on businesses (Forbes Insight Report, May 2017)⁴.

Given its importance, the concept of data quality has been extensively studied in terms of many dimensions. A data quality dimension is a *“characteristic or part of information for classifying information and data requirements”* (Sidi et al., 2012, p. 302). Accuracy, completeness, consistency, and timeliness are among the most important dimensions that have been studied so far (Scannapieco et al., 2005; Sidi et al., 2012; Wand & Wang, 1996). Sidi et al. (2012) have performed a comprehensive review of data quality dimensions. Table 2.2 reviews those dimensions of data quality which have been mentioned more frequently in the literature.

⁴ https://www.forbes.com/forbesinsights/pitney_bowes_data_quality/index.html

Table 2.2 Notable data quality dimensions

| Dimension | | Definition |
|-------------------------|------------|---|
| Accuracy | | The extent to which data represent a real-world phenomenon (Batini & Scannapieco, 2010) |
| Completeness | | The extent to which data is of sufficient breadth, depth, and scope for the intended task (Batini & Scannapieco, 2010) |
| Consistency | | The extent to which data is presented in the same format and compatible with previous data (Wang & Strong, 1996) |
| Time-related Dimensions | Currency | The degree to which data is up to date. Data value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value (Batini & Scannapieco, 2010) |
| | Timeliness | The extent to which age of the data is appropriate for the intended task (Wang & Strong, 1996) |
| Accessibility | | The extent to which data is available, or easily and quickly retrievable (Wang & Strong, 1996) |
| Reliability | | The extent to which data can be counted on to convey the right information (Batini & Scannapieco, 2010) |
| Relevancy | | The extent to which data is applicable and helpful for the intended task (Wang & Strong, 1996) |

In information systems associated with organizations, data and data quality are vital for managers and operating processes in order to detect related performance issues, and to ensure the improvement of business processes, making smart decisions, and the creation of strategic advantages (Madnick et al., 2009). Data of poor quality may lead to high and unnecessary costs, lower customer satisfaction, lower job satisfaction, and organizational mistrust (Redman, 1996). Within organizations, there are usually three main segments in which all DQM activities are concentrated; (1) data collection, which is generally implemented by operators, (2) data protection, which is executed by IT experts who are responsible to maintain data and make it ready for use, and finally (3) data consumption, which includes managers who use data for analysis and decision making process (Lee & Strong, 2003). To achieve high quality data, all three parties must collaborate properly

(Lukyanenko, 2014). All members in all three processes must share a common knowledge of which data is good, how to capture it, and why it is important.

Generally, approaches to improve data quality in the information systems field can be characterized as theory-based or design-oriented (Wand & Wang, 1996). Theoretical approaches are particularly relevant to data characteristics themselves, however design-oriented approaches intend to study characteristics of data within information systems in terms of actual design and implementation (Wang & Strong, 1996). Design-oriented approaches focus on the structure and values of the data in a system and provide actual guidance to system designers.

2.2 Data Quality in UGC

The quality of UGC can be defined as “the extent to which stored information represents the phenomena of interest to data consumers (and project sponsors), as perceived by information contributors” (Lukyanenko et al., 2014, p. 15). This definition is different from the common “fitness for use” definition, as it emphasizes the important role of contributors in describing a real-world phenomenon.

To understand the concept of quality in UGC settings, it is useful to distinguish between traditional settings and UGC settings. In organizational settings, the data consumer’s perspective is usually highlighted in the information system design process. In such consumer-oriented systems, the extent of matching the captured data with the consumer's needs determines the level of design success, while the perspective of data contributors may be ignored. On the other hand, UGC projects are financially and technically managed by sponsors, however the key creators of data are ordinary people,

thus the abilities, motivation, and domain knowledge of contributors may have a strong impact on engagement level and the quality of supplied data (Lukyanenko et al., 2014). Contributors usually provide what they are able to provide, which may be totally different from what is needed in some cases. Such data may be collected with one use in mind, but used for many different tasks (anticipated or unanticipated future uses) (Lukyanenko, 2014).

In organizational settings, contributors often know about the main goal of the system, while the contributors to UGC projects are not usually aware of how the supplied data is going to be used, and usually do not have the motivation to meet the expectations of data consumers (Nov et al., 2011). To design an effective information system for UGC projects, contributors' capabilities as well as data consumers' needs should be considered. Moreover, in UGC projects, when the data entry is voluntarily, contributors may simply abandon data entry if the consumers' needs are not aligned with what contributors can supply.

2.2.1 Approaches to Improving Data Quality in UGC

Traditional DQM principles cannot be straightforwardly applied because of the nature of user-generated data. Approaches to improve the quality of UGC, can be classified in two main streams. The dominant stream is “consumer-oriented” approach which mainly focuses on traditional definition of quality, fitness for use, and data consumers. The main aim of this approach is to better align captured data with data consumers' needs. However, this approach may ignore the characteristics of user-generated data, which reflect the contributors' perspective. On the other hand, the “contributor-oriented” approach, which

is the scope of this thesis, examines ways to develop an effective information system in which contributors' perspective is highlighted, to better capture data of real-world phenomenon.

In citizen science projects, the quality of contributed data has also been criticized (Newman et al., 2012). New technologies, such as mobile applications, wireless sensor networks, and online computer/video gaming have emerged to empower contributors to explore, collect, and share data of better quality (Kim et al., 2013). Another example of consumer-oriented approach is “Creek watch”; an iPhone application designed and implemented allowing contributors to report data about waterways (Kim et al., 2011). To aid water management program, this project was sponsored by state and local officials. The main use of creek watch is to ensure that the captured data is useful for consumers or not.

Quality improvement can be implemented during data entry. Using collaboration among users in social media to enhance the quality of data generated online has been examined as a consumer-oriented approach. A UGC project called iSpot (www.ispot.org.uk), uses social network collaboration for species identification (Silvertown, 2009). Social collaboration refers to processes during which people interact and share data to achieve common goals. The iSpot platform was designed as a social network to support learning by providing feedback – from experts and other contributors – on supplied data. According to Silvertown (2010), in the first year of the iSpot website operation, 25,000 observations of 2500 species were identified by 6000 users. Based on Lukyanenko (2014), “social networking is suggested to increase data quality through the increased scale of data” (p. 18). A study conducted by Prestopnik and Crowston (2011)

demonstrated that a computational system working based on social networking and gamification methods can motivate participants and increase the number of participations in a citizen science project. They argued that using game along with social networking improve the quality of supplied data.

While social network collaboration seems to be promising, it has a number of limitations. UGC projects with a small number of users will not have sufficient user activity per unit of data to ensure adequate improvement (Lukyanenko et al., 2014). In addition, data can be verified and corrected in the collaborative process, thus there is no way to specify whose data is being stored and represented – the original or the modified version.

Another technique for quality improvement via data entry is defining different roles for online users in the contribution process. For example, Wikipedia defines different roles for online users (e.g., moderator, editor, and beginner) based on users background and expertise (Liu & Ram, 2009). Based on the study by Liu and Ram (2009), it is assumed that users in different roles will provide data of different quality.

Quality improvement via online training or providing online feedback is another strategy. For example, in Galaxy Zoo (www.galaxyzoo.com), users are required to pass a tutorial before they are allowed to classify galaxies, however, this strategy has its own limitations. Online training can sometimes cause issues like biased contributions, tendency to exaggerate certain observations and to under-report others, and a general reluctance of observers to enter data when they see only common phenomena (Bonney et al., 2009). In addition to those disadvantages, short-term training is not always promising. For example, in a study conducted by Crall et al. (2011), participants attended a one-day training

workshop before contributing in a citizen science project. The results demonstrate that some skills, like taxonomic identification, will not be acquired by short-term learning, and need a longer period of training.

Improving the quality of UGC can also be obtained via content filtering, which is classified as a technology-based method. In this method, contributors can enter data without any modification, but only data of certain quality will be stored and retrieved. Content filtering usually is a set of verification mechanisms, developed by experts, and implemented through the process of data storage. For example, with the aim of species identification, the eBird project (www.ebird.org) uses a combination of smart filters to evaluate submissions and identify the species (Sullivan et al., 2009). Its filtering system contains two stages of verification: (1) an automated verification filter evaluates the submitted data instantly based on species count limits for a given location and time, and flags unusual observations in terms of location, exceptional counts, and extreme rarity, (2) a network of local experts carefully examines stored data flagged by an automated filter. Since the task of verification requires human cognitive abilities, content filtering will be infeasible in terms of cost and time if the size of data set increases.

Another example of content filtering is developing a set of validation and verification tools to improve the quality of contributed data. Alabri and Hunter (2010) demonstrated that using filtering feature along with social networking in a “CoralWatch” project as a case study can significantly improve the quality of captured data. They also argued that the reliability or trustworthiness of citizen science data can be measured by using a weighted aggregation of both direct and inferred attributes supplied by contributors.

On the other hand, a contributor-oriented approach tries to improve data quality by considering contributors as the main key in data quality management. Lukyanenko et al. (2014) introduced an approach in which contributors' perspectives were highlighted to improve data quality. They examined the impact of using an instance-based model for data collection on information quality in a citizen science project in the biology domain. In the instance-based model, individual entities (instances) are stored only with their attributes, rather than classes (Parsons & Wand, 2000). This model suggests a two-layered structure in which data of instances is stored separately from any specific classification. The first layer consists of information about instances and their attributes, and the second layer consists of information about classes in terms of attributes (Parsons & Wand, 2000). Using an instance-based model in a UGC setting improves the quality of data in terms of accuracy and dataset completeness (Lukyanenko et al., 2014). In addition to those advantages, they argued that collecting data in a flexible way can facilitate researchers in capturing unanticipated data which can be appropriated for additional uses, such as monitoring environmental change. A real citizen science information system (www.nlnature.com) was redesigned based on those principles, featuring an instance-based model for data entry.

In spite of advantages in data quality management, the instance-based approach has a number of challenges. The first challenge is that the method will result in a large number of attributes reported by contributors. Managing this volume of attributes to make it ready for analysis can be challenging. Second, to make the instance-based data ready for querying, a novel query tool should be developed. Third, it can increase data irrelevancy. Irrelevant data refers to those which are not helpful or applicable for the intended purpose

(Sidi et al., 2012). Irrelevancy will arise when contributors are allowed to supply data without any restriction (Lukyanenko et al., 2014). Consequently, the database is vulnerable to misidentifications. Proper species identification is crucial in biology citizen science projects (Sullivan et al., 2009). A standardization mechanism will be needed for such data to make data ready for further analysis.

Given the limitations of the instance-based model, and the importance of data quality - specifically species identification - on citizen science projects success, new approaches are needed. This thesis examines the effect of using a “recommendation system” in UGC setting. Recommendation systems have traditionally been used in E-commerce to guide users making a better decision in online-purchasing. Although recommendation system and UGC seem quite different approaches, a recommendation system may guide contributors in a UGC setting to provide data of better quality in terms of higher level of relevancy, objectivity and species identification.

2.3 Recommendation Systems

Recommendation systems have emerged to provide mechanisms to help online users find desirable information (Fernández-Tobías et al., 2012). They aim to filter irrelevant data, and present those data that better suit the users’ interests according to their personal profile or online behaviour. In other words, an RS assists users in the decision making process; it captures the user information as input and develops personalized recommendations as output. Recommendation systems are being successfully used in various fields, mainly in e-commerce (Fernández-Tobías et al., 2012).

Recommendation systems use different types of filtering methods to create

personalized recommendations. These methods can be classified into three main techniques: content-based, collaborative, and hybrid approaches. Following is a brief summary of three main techniques that recommendation systems are using.

2.3.1 Content-Based Recommendation systems

A content-based RS relies on two main components: user profile and item profile. The item profile refers to all information about items' attributes and features, while the users profile contains users' interests and preferences (Pazzani & Billsus, 2007). The user profile can be explicitly provided by the user, or implicitly extracted by analyzing the user's search history, online-purchasing history, or items rated by the user. To create a personalized recommendation by a content-based algorithm, the first step is determining the best match between user profile and item profile. Then the system recommends items similar to others that already match the user's interest. This procedure is at the core of most RS applications such as Amazon and TripAdvisor.

2.3.2 Collaborative Recommendation Systems

The main components of this approach are user profile, item profile, and other users' profiles. In the first step, Collaborative RS (or collaborative filtering systems), unlike content-based RS, compares users' profiles in order to identify users with similar preferences. Next step, they recommend items to a particular user, based on the items previously rated by other users with similar interests (Adomavicius & Tuzhilin, 2005). The process of comparing users' profiles will be done based on the similarity of ratings that users already have provided for the same items.

A good example of collaborative RS is the book recommendation system of

Amazon. Users will be recommended those books that received highest rating from users with similar tastes. Facebook, LinkedIn, and other social networks use collaborative RS to recommend new friends, groups, and other social connections.

2.3.3 Hybrid Recommendation Systems

Hybrid RS use a combination of content-based and collaborative filtering approaches. This combination is intended to mitigate the weaknesses and highlight the advantages of content-based and collaborative filtering methods. A well-known example of the Hybrid approach is Netflix. The application makes recommendations based on a content filtering method as well as a collaborative RS. Netflix compares the watching habits and search history of similar users (collaborative filtering) to make suggestions. It also offers movies based on an individual user's profile, search history, and movies characteristics (Content-based filtering).

The combination of content-filtering and collaborative RS can be implemented in several ways. Based on a study by Adomavicius and Tuzhilin (2005), the combination methods can be classified into four main categories:

- Content-based and collaborative approaches implemented separately and make their own recommendations, then the recommendations are combined.
- A content-filtering approach is the base of the hybrid system, but some collaborative approach features will be added.
- A collaborative approach is the base of the hybrid system, but some content-based approach features will be added.
- Construct a novel model that feature both content-based and collaborative approach

characteristics.

Recommendation systems have been applied successfully in many different domains (Fernández-Tobías et al., 2012): E-commerce (eBay and Amazon), Entertainment (Netflix, YouTube), Services (LinkedIn, and TripAdvisor), News (Twitter). The use of recommender systems in other domains is also promising and should be explored and researched (Adomavicius & Tuzhilin, 2005). One area in which RSs has almost been ignored is UGC projects. Although an RS has been applied in YouTube (a UGC website) to recommend videos/audios, to our knowledge RS have rarely been used as a data entry tool for UGC projects. One exception is the study conducted by (Vandecasteele & Devillers, 2015), in which a recommender system approach was proposed to improve the semantic quality and reducing dataset heterogeneity.

As mentioned earlier, data quality is a critical issue for any citizen science project. Errors resulting from misidentified species can be a major issue for citizen science because similar species can be confused (Bonney et al., 2009). To enable the general public to provide data of better quality, three main steps should be conducted (Bonney et al., 2009): developing a clear and user-friendly data collection method: designing an appropriate interface that properly reflects the collection method, and providing support for participants to better understand how to submit their information.

The proposed RS-based guidance for this thesis is a content-based recommendation system. The RS-based guidance feature will utilize a database which contains the key characteristics of species that are going to be tested in this experiment. It will assist contributors to better identify the common name of species. Also, it is expected to result

an acceptable level of relevancy and objectivity in data while the number of captured data is expected to be the same as that of instance-based model.

3 Impact of using Recommendation System on Data Quality

Data quality is a critical issue for any citizen science project (Bonney et al., 2009). For citizen science projects to be successful, the captured data must have an acceptable level of quality. Based on the study conducted by Lukyanenko (2014), the instance-based model for capturing data enhances the accuracy and completeness of a dataset, and decreases the information loss in a citizen science project. Although their proposed model was promising in terms of data accuracy and dataset completeness, it resulted in a very large quantity of attributes. While contributors are allowed to submit data without any restriction, supplied data may have low/no relevancy to what data consumers need. Managing a large amount of data to find relevant data can be challenging. Another problem is lack of data objectivity, meaning that supplied data may reflect contributors' feelings and opinions, rather than focusing on objective description of the real world phenomenon. To make such data ready for analysis, data standardization is required.

Recommendation systems have emerged as methods helping filter irrelevant data to make data more compatible with what the user really wants (Fernández-Tobías et al., 2012). Considering the advantages of the RS in other domains, I argue that using a guidance feature utilizing a content-based recommendation system can assist contributors to provide data of better quality. As mentioned before, data quality has different dimensions in organizational settings. This thesis analyzes the impact of using a guidance on two dimensions of quality – *relevance and objectivity*. Also, I claim that using guidance in a specific citizen science domain (biology) may assist contributors in species identification.

Whereas using the guidance feature will not interfere with the instance-based

approach principles, contributors are free to use guidance feature or report data in a free form. Thus, it is expected that the number of accurate attributes (attributes with no typo or misspelling) captured by this feature will be the same as the number of accurate attributes captured by an instance-based approach.

The guidance feature works based on a database consisting of all key characteristics of species, and an initial attribute as a trigger, provided by contributors. The RS-based guidance feature enables contributors to choose attribute from the recommendation list in which each attribute is relevant to the intended purpose of project – species identification, thus relevancy will be guaranteed. In addition, choosing attributes from the list of recommended attributes may prevent data influenced by contributors’ personal feelings, or opinions when reporting a species, and increase data objectivity.

This method may assist contributors in better identifying the species at the *expert-level*, which is more fine-grained than *basic-level* identification. Contributors generally are not biology experts, thus it is expected those with low biology background are only able to identify very few species. As an alternative, they are able to identify species at the *basic level* which is an intermediate identification level in biology (e.g. “duck” is a level higher than “American Black Duck”, and a level lower than “bird”). Basic level categorization is the more preferred classification level for non-experts and widely used in cognitive psychology (Lukyanenko, 2014). Providing the basic-level identification and species attributes, the RS-based guidance can find the common name of species at the expert level.

This chapter investigates the impact of having an RS in an instance-based model citizen science project on two dimensions of data quality as well as species identification.

3.1 Impact of using Recommendation System on Species Identification

Proper species identification is crucial in field observation studies in citizen science research in the biology context (Sullivan et al., 2009). Methods such as smart filtering or online training have been used in citizen science projects to help contributors in species identification. However, in the citizen science project which utilizes an instance-based data collection, there is no distinct method for helping contributors to identify species.

Identification in biology is defined as *the process of assigning a pre-existing taxon name to an individual organism*⁵. When confronting an unknown organism, biologists usually use a tool called an *identification key*, containing the written description of species which are discovered and classified up to date. The key provides a series of choices about the characteristics of the unknown organisms; by making the correct choice at each step of the key, the user is ultimately led to the identity of a specimen. Different techniques were established to aid biologist in identifying species accurately such as: photographic identification which is based on appearance and visible body features (e.g. overall shape, color, and camouflage pattern) (Katona & Kraus, 1979), genetic identification which is based on genetic attributes (Hebert et al., 2004), chemical-based identification in which chemical compounds of species are tested (Lavine & Carlson, 1987), and microscopic identification in which types of cells and cell structure are tested (Cooper et al., 2007).

Whereas in many citizen science projects the general public reports real-world phenomena by describing their appearance, the appearance and visible trait method is

⁵ [www.wikipedia.org/wiki/Identification_\(biology\)](http://www.wikipedia.org/wiki/Identification_(biology))

frequently the only practical option to identify instances. In a citizen science project in the biology context, species often can be identified only by appearance and visible trait. This thesis focuses on birds as the species that contributors intend to identify.

According to the Forests Ontario Bird identification Guide⁶ and Cornell Lab of Ornithology⁷, the appearance-based method of bird identification relies on four keys; (1) shape and size, (2) coloration and pattern, (3) behavior, and (4) habitat. If an individual with species identification skill is provided with accurate data for the main identification keys, the species will be identified successfully. However, species identification skill is rare among members of the general public who are the main contributors in citizen science projects. Since an instance-based model data collection enables contributors to submit data about all four identification keys, it has a good potential for species identification. A guidance feature which is based on a content-based filtering method can compensate the lack of ability to identify species in an instance-based model.

By filtering the irrelevant data, an RS can find the name of species which the contributor is reporting via its attributes. Using the data supplied to the system by the contributors, and matching it to the database of species characteristics, RS will recommend the name of species which match with the supplied data.

Hypothesis 1.1 *In an instance-based data collection task, an RS-based guidance feature in data entry will enable contributors to identify species better compared to using no guidance.*

⁶ www.forestsontario.ca/wp-content/uploads/2016/04/Ontario-Bird-Identification-Guide.pdf

⁷ www.allaboutbirds.org/four-keys-to-bird-identification/

3.2 Impact of using Recommendation System on Data Relevance

According to Wang and Strong (1996), relevance is defined as “*the extent to which data is applicable and helpful for the task at hand* (p.31).” Relevance issues are not unique to citizen science projects, and a number of approaches have been proposed to improve them (Wang & Strong, 1996). However proposed approaches are typically based on organizational settings where data contributors’ perspectives are largely ignored.

In citizen science projects, the absence of a defined protocol for data collection often leads to irrelevant data which is not useful for the intended purpose of research (Paulos, 2009). Developing validation protocols, screening methods to prevent typing errors, and smart filtering are examples of several methods proposed for making the contributed data more compatible with consumers’ need (Bonter & Cooper, 2012). In the instance-based model, contributors are able to report attributes without any constraint or a standard data entry procedure. Thus a large number of attributes with no standard format will be captured. In such a large database, it is also likely that the amount of irrelevant data (for species identification) is increased. An RS can reduce the negative impact of having irrelevant data.

By filtering the irrelevant data, an RS can assist contributors in providing the system with more relevant data. For instance, when a contributor submits an attribute to the system, the RS will take the initial attribute as a clue and search the database for other attributes which are relevant to the initial data. After finding all possible matches, the relevant attributes will be recommended to the contributor as a list to choose from. Therefore, the relevance of data will be guaranteed.

Hypothesis 1.2 *In an instance-based data collection task, using an RS-based guidance feature in data entry will enable contributors to provide data of a higher level of relevance compared to using no guidance.*

3.3 Impact of using Recommendation System on Data Objectivity & Number of Captured Attributes

Data objectivity is defined as “*the extent to which data is unbiased, unprejudiced and impartial*”(Wang & Strong, 1996, p. 32). Citizen science projects have been criticized as sacrificing objectivity in exchange for lower cost and higher data quantity (Jasanoff, 2003). Training in the data entry phase is a proposed method to reduce subjectivity of data in a citizen science research (Kremen et al., 2011). Developing a class-based data collection is another way to reduce the subjectivity, however the class-based method results in information loss (Lukyanenko, 2014). The instance-based approach was developed to mitigate the negative impact of information loss in citizen science research. However, data subjectivity may increase since contributors are free to report whatever data they are willing to contribute. Contributors may include their personal feelings and opinions unintentionally while reporting an observation (Connor-Greene, 2007).

An RS-based guidance feature enables contributors to choose from a list of recommended attributes, thus contributors have less chance to contribute subjective data. Whereas using the guidance feature will not interfere with the instance-based approach principles, contributors are also able to report data in a free form rather than the recommended list to prevent information loss.

Hypothesis 1.3 *In an instance-based data collection task, an RS-based guidance*

feature in data entry will enable contributors to provide data of higher level of objectivity compared to using no guidance.

Data completeness is defined as “*the extent to which data is of sufficient breadth, depth, and scope for the intended task* (Wang & Strong, 1996, p. 32).” According to Wand and Wang (1996), if an IS is not capable of capturing every relevant state of the world, data completeness may be threatened. The instance-based model developed by Lukyanenko et al. (2014) has the ability to capture any data - anticipated or unanticipated – of an observation. That work showed that the instance-based model of data collection leads to more accurate data compared to class-based models.

Similarly, it is expected that using an RS-based guidance feature with the instance-based model principles will also result in same amount of captured data. The RS will recommend relevant attributes to contributors to choose from while the system is also capable of capturing any data other than recommended ones by contributors.

Hypothesis 1.4 *In an instance-based data collection task, an RS-based guidance feature in data entry will enable contributors to provide as much data as when there is no guidance.*

To test the proposed hypotheses, an experiment under three conditions was conducted. The experiment design and method are explained in the next chapter.

4 Experimental Design

The previous chapters introduced the notion of data quality in UGC settings, and current approaches to improve the quality of data in UGC settings were reviewed. Also, it was suggested that using an RS might mitigate the negative impact of using instance-based data collection in UGC. Particularly, it was argued that using an RS in the data collection phase of a citizen science website can enhance the quality of data provided by improving data relevancy and objectivity without sacrificing data accuracy and completeness. Moreover, it was claimed that using an RS in data entry task of a citizen science website will assist contributors to identify species properly. To evaluate the propositions regarding species identification, data relevancy, data objectivity and data completeness in a UGC setting, a laboratory experiment was conducted under three conditions in the context of a citizen science project in the biology domain. The proposed hypotheses are as follows:

H 1.1 *In an instance-based data collection task, an RS-based guidance feature in data entry will enable contributors to identify species better compared to using no guidance.*

H 1.2 *In an instance-based data collection task, using an RS-based guidance feature in data entry will enable contributors to provide data of a higher level of relevance compared to using no guidance.*

H 1.3 *In an instance-based data collection task, an RS-based guidance feature in data entry will enable contributors to provide data of higher level of objectivity compared to using no guidance.*

H 1.4 *In an instance-based data collection task, an RS-based guidance feature in*

data entry will enable contributors to provide as much data as when there is no guidance.

4.1 Method

To test the hypotheses, I conducted a study involving 60 undergraduate/graduate business students (28 female, 32 male) from the Memorial University of Newfoundland. The participants were randomly assigned to 3 conditions of the experiment. According to (Cleary et al., 2014), recruiting 20 participants for each condition yields result of acceptable margin of error (95% confidence). To align with the definition of citizen science, in which contributors are non-experts with respect to the intended use of data, business students were chosen to make sure that the participants have low/no biology background knowledge. The participants were asked orally about their biology background. Participants had no idea about the purpose of the study until the beginning of the experiment so they could not prepare in advance. Each participant was asked to work individually with a computer, and shown the same set of stimuli on the computer screen. While viewing the stimuli, they were asked to describe what they were watching.

4.2 Materials & Procedures

The stimuli consisted of 14 images of birds common in the region of Newfoundland and Labrador (see Appendix 1). Since the experiment involves human participants, the research proposal was reviewed and approved by the Interdisciplinary Committee on Ethics in Human Research (ICEHR).

Participants were randomly assigned to one of three conditions. For the purpose of this study, I developed and implemented a web-enabled information system using the

RStudio software package (version 3.3.2). Although the experiment could be carried out anywhere with an internet connection, a computer laboratory was selected to make sure that the participants did not access to other sources of knowledge. For all conditions, the system interface showed the birds' images, provided a data entry area, and used a database to maintain the captured data. Each condition had a different interface, and a different data entry format. At the beginning of all conditions, participants were shown a welcome page, consisting of an introduction and purpose of the study and the procedure in brief. Figure 4.1 shows the experiment welcome page.

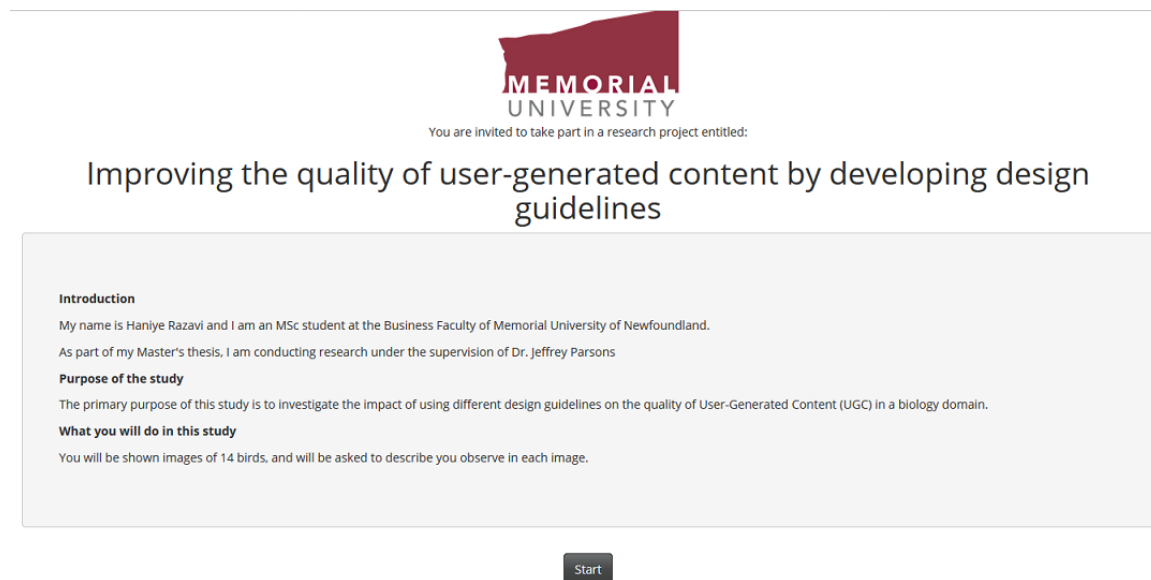


Figure 4.1 The experiment Welcome Page

After clicking the start button, the first bird image was displayed on the screen. The image was shown while participants entered data about the bird. In a pre-test, it was determined that 25 minutes was sufficient to elicit several attributes for all bird's images, however the duration of experiment session for each participant was not constrained.

For the purpose of analyzing captured data, SPSS Statistics software (version 24)

was chosen. To test the hypotheses, the independent t-test, also called the two sample t-test, was used to compare conditions in terms of the number of identifications, level of data relevance, and objectivity. A t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known⁸. Since the sample size for each condition is small (20 participants for each condition), and the distribution variance for each group is unknown, t-test was chosen.

4.2.1 Condition 1 (No Guidance)

The first condition is referred to as “No Guidance”. This condition featured an instance-based model for data collection. A real world example of an instance-based model of data collection in UGC setting is NLNature⁹, in which contributors have no restriction in reporting observations. This condition was used as the control group to make a better judgment on how an RS-based guidance feature can assist contributors in bird identification and providing data of higher level of relevance and objectivity compared to that when there is no guidance. Participants were asked to describe the bird on the computer screen while watching it. In this condition, there was no restriction on data entry, so participants were free in terms of the format and number of attributes. The system was capable of capturing as many attributes as the participants decided to submit. Participants might want to identify birds, however it was expected that they could identify only a few of the birds at the species level. Identifying the birds will help to compare the conditions

⁸ <https://statistics.laerd.com/statistical-guides/independent-t-test-statistical-guide.php>

⁹ www.nlnature.com/

in terms of number of bird identifications occurred in each of them. By clicking on the “submit” button, participants could see all attributes as a list on screen, and submitted attributes were stored in the data base simultaneously. As illustrated in Figure 4.2, all submitted data was shown as a list on the screen.

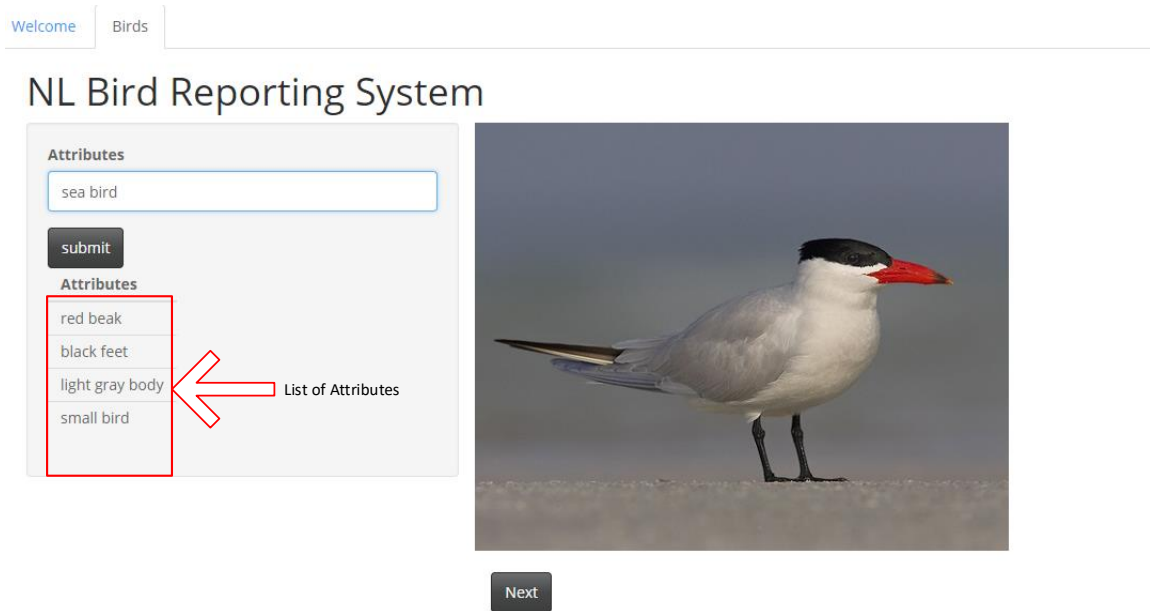


Figure 4.2 The interface of the “No Guidance” condition

4.2.2 Condition 2 (Traditional Guidance)

In this condition, guidance was provided to the participants to assist them in species identification, a popular approach in practice that uses common “keys” to guide in species identification (e.g., www.whatbird.com). This visual feature worked based on a database consisting of key visible characteristics of birds and a filtering technique. This guidance, which is referred to “Traditional Guidance”, was a step by step approach. In each step, participants were questioned about a visible key character about birds. Participants had a

number of possible answers for the question in each step. The possible answers to each question were derived from the database. Based on which answer was chosen by the participants, the possible answers to the next question were updated, which was performed by the filtering technique.

In our experiment, there were four steps in which participants were asked about (1) body shape, (2) body color, (3) head color, and (4) bill color, respectively. For this thesis, these four main visible characteristics of birds are sufficient to identify birds. For example, the system initially asks about the overall body shape. For this question, the system provides six possible options using silhouettes. When participants choose the body shape, the system search the database to find matching data with the selected silhouette. After finding a match/matches, the possible answers for the second question are updated. This procedure is repeated for the remaining steps. After answering all questions, the bird will be identified. If the system is provided with accurate data, (which means the answers to all questions are accurate), the bird will be identified successfully. As shown in Figure 4.3, the interface has three main parts: (1) bird's image, (2) silhouettes, and (3) questions and answers box.

NL Bird Reporting System

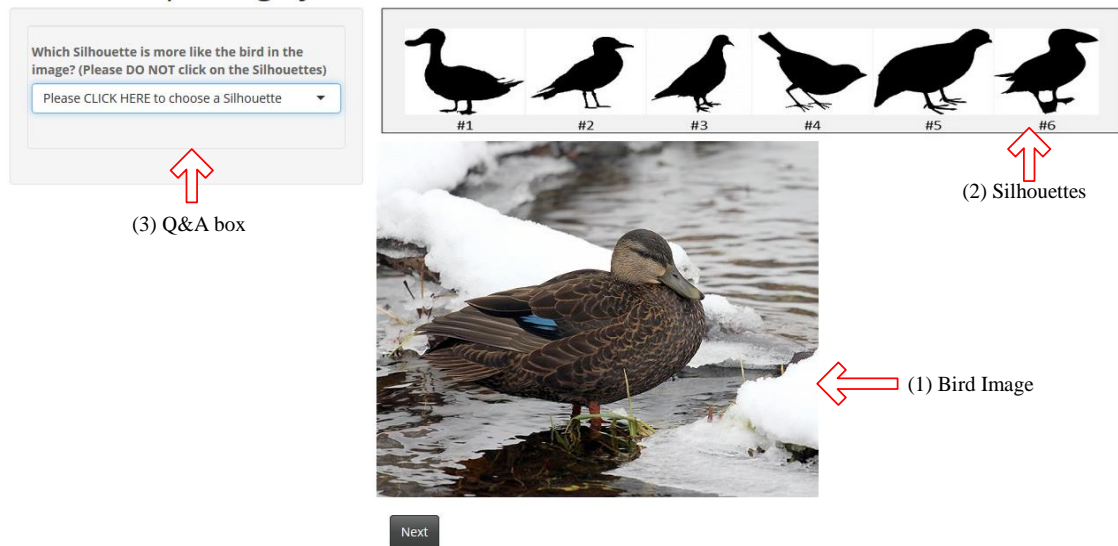


Figure 4.3 The interface of the “Traditional Guidance” condition

To make the initial interface similar for all conditions, participants initially see the first question and answer box. As shown in Figure 4.4, the other three questions appear on the screen after participants answer the first question. The procedure for this condition is shown as a flowchart in Figure 4.5.

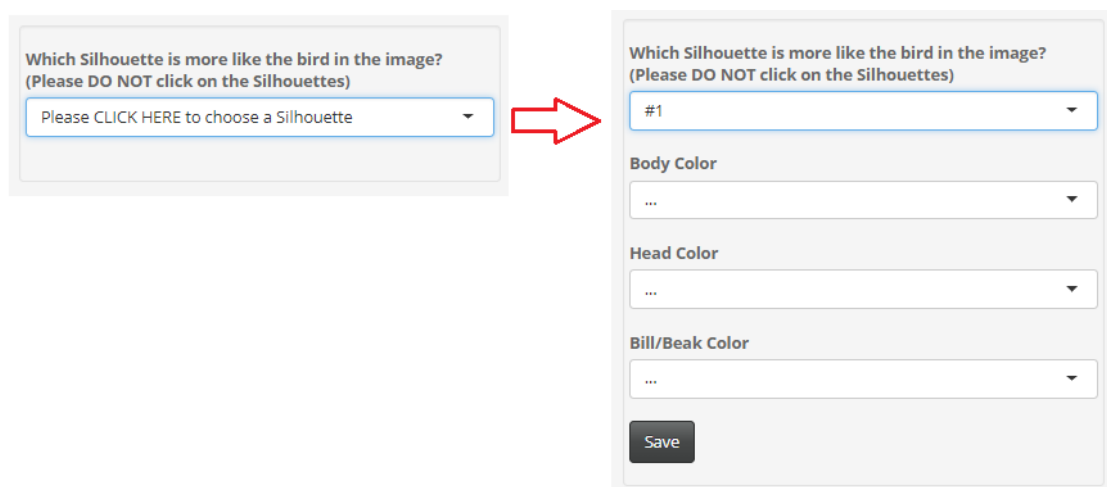


Figure 4.4 How the interface changes in the “Traditional Guidance”

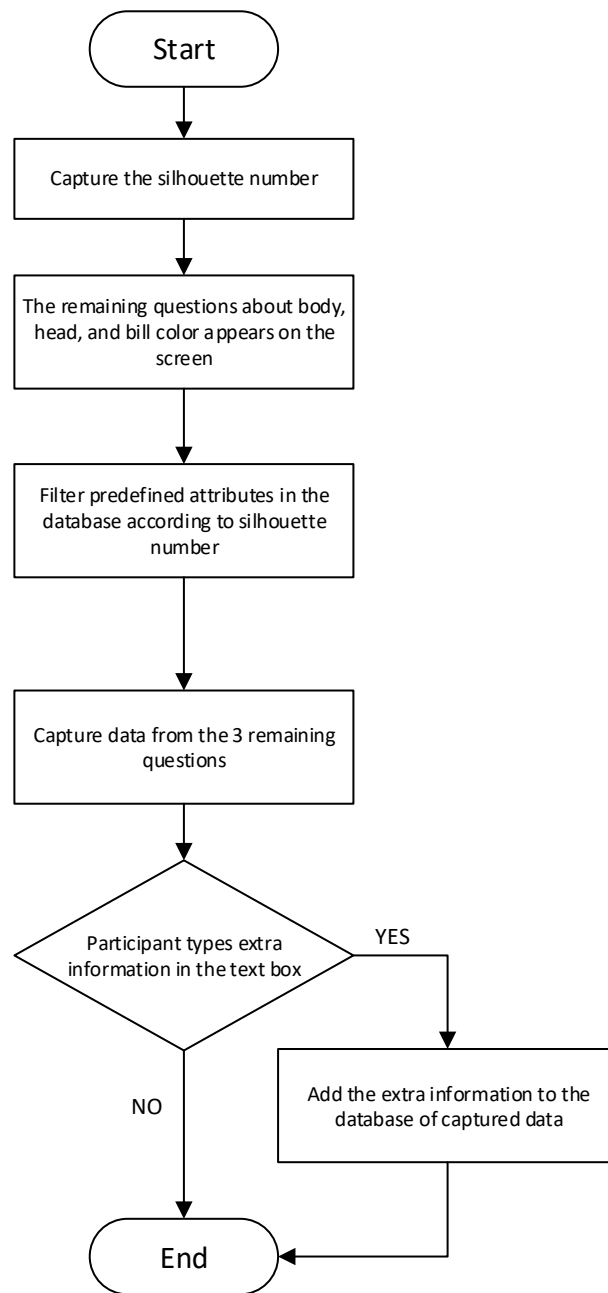


Figure 4.5 The procedure for the "Traditional Guidance" condition

To mitigate information loss, another box for free-form data entry was provided for this condition to ensure the system is capable of capturing all the information about birds

which are on the screen. This box appears once the contributor is finished answering the guidance questions. All additional data was shown as a list on the screen in Figure 4.6.

Figure 4.6 Traditional guidance with a text box for additional information

4.2.3 Condition 3 (Cognitive Guidance)

Cognition is defined as *the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses*¹⁰. The process of cognition uses previous knowledge to generate new knowledge. Since this condition enable participants to identify birds by using the sense of sight, thought, and experience, this condition is referred to as “Cognitive Guidance”.

The cognitive guidance condition uses a content-based RS to filter irrelevant data and to assist contributors in species identification. The system works based on a database consisting of birds’ attributes, and the attribute supplied by contributors.

¹⁰ en.oxforddictionaries.com/definition/cognition

The attribute database, which was created for the purpose of this thesis, consists of all visible attributes of birds such as overall body shape, body pattern, bill color, head shape, head color, and etc. These attributes were derived from the handbook of bird identification guide and the bird guide of the eBird project¹¹.


Initially the system prompts for an attribute, then the system searches the database of birds attributes to find matches. This process is repeated until the system can identify a bird corresponding to the captured attributes. If the system cannot find a bird that matches the captured data or if the participant is unwilling to continue the iteration, a failure mode occurs. If failure happens, all submitted attributes are stored in the database of captured data, and the system proceeds to the next bird image. The user interface designed for the cognitive guidance is shown in Figure 4.7.

¹¹ <https://www.allaboutbirds.org/guide/search/>

NL Bird Reporting System

Attribute

Please Enter an Attribute



Next

Figure 4.7 The interface of the Cognitive Guidance Condition

While the bird's image is shown, participant submit the initial attribute. Following popular practice on social media websites (e.g., Facebook, Twitter) and search engines (e.g., Google), this condition has an “autocomplete” feature to help participants in providing attributes. The feature will suggest matches by using the database. Participants are totally free to supply the first data. They can either use the autocomplete feature or type something else. Then, the RS-based guidance feature will search the database of birds attributes and find a match/matches based on what is captured from the participants. If the system finds only one bird matching the initial attribute, a message of “Bird is Successfully identified” along with the bird's name will pop up on the screen. Otherwise, the system will recommend a list of relevant attributes from which contributors can choose. Participants can either select one by one from the recommended list or select all attributes

that apply at once. This loop will repeat till the system find only one match for captured attributes. If no match is found, system will store all captured data from participants without any bird identification. The algorithm for the above procedure is shown in Figure 4.8.

To prevent information loss, the system is also capable of capturing free-form data. Participants can either use the RS-based guidance feature for data entry or submit any data they are willing to submit without using the guidance at all.

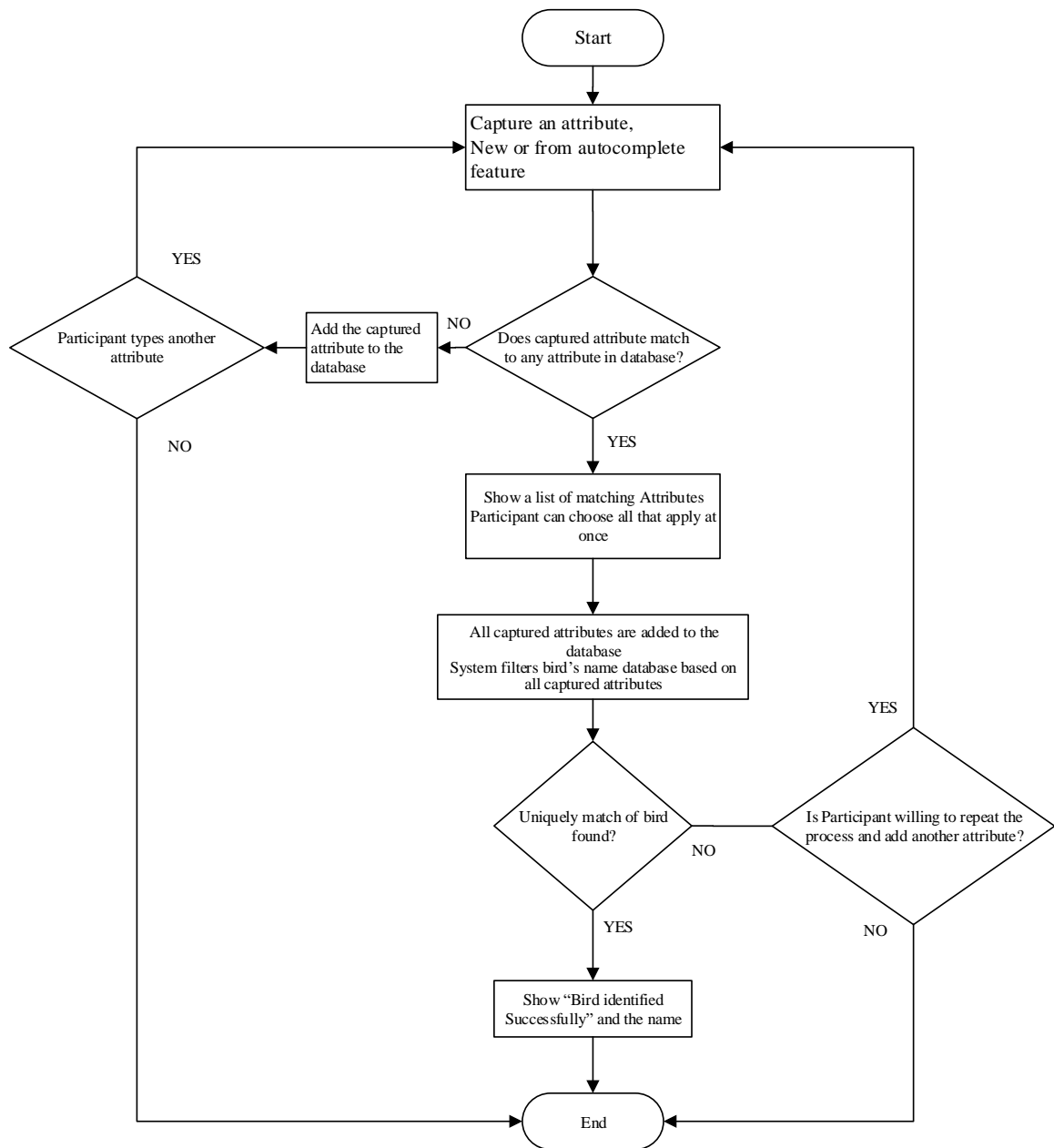


Figure 4.8 The algorithm developed for the “Cognitive Guidance”

5 Data Analysis & Discussion

5.1 Data Analysis

In the free-form data entry condition (No Guidance), 20 participants provided a total of 1445 attributes (on average 5.16 per image per participant). To ensure consistency, pre-analysis data preparation was performed; obvious typing errors were corrected (e.g., “tale” was corrected to “tail”); redundant words (e.g., “It has a blue color in the body” was coded as “blue body”); symbols (e.g., parenthesis, comma) and characters that did not carry additional meaning were removed. After data cleaning, 1426 attributes (98.7%) remained which are referred as “correct attributes” in Table 5.1. Complex attributes were broken down into individual components (e.g., “long yellow beak” was coded as “long beak” and “yellow beak”). Based on the psychological research done by Tanaka and Taylor (1991), attributes with similar meanings for the same birds were grouped together (e.g., “bill”, “beak” and “mouth”). Also, colors with similar meaning were grouped together (e.g., “black”, “gray” and “dark blue” were grouped as black, and “orange”, “brown”, “gold” and “amber” were grouped as orange). Color naming is affected by color perception which can be affected by several factors such as language, learning, and cultures (Kay & Regier, 2007; Özgen, 2004). To mitigate the effects of those factors, color grouping was implemented. Attributes were also coded as either “bird-related”, or “background-related,” to see which condition is better capable of capturing data unrelated to the identification task. All attributes related to body, overall shape, size, and color pattern of the bird were coded as bird-related (e.g., green spots on neck was coded as bird-related while sitting on

a rocky beach was coded as background-related). Table 5.1 shows the total number of attributes before and after data preparation (typo and misspelling removal), along with the total number of bird-related and background-related attributes for each bird in the No Guidance condition (Free-form data entry).

Table 5.1 The result of No Guidance condition for each image

| | Attributes | Correct Attributes | % Attribute Correctness | Bird-related Attributes | Background-related Attributes |
|--------|------------|--------------------|-------------------------|-------------------------|-------------------------------|
| Bird1 | 99 | 99 | 100.00% | 77 | 22 |
| Bird2 | 100 | 100 | 100.00% | 94 | 6 |
| Bird3 | 99 | 97 | 97.98% | 91 | 6 |
| Bird4 | 101 | 101 | 100.00% | 96 | 5 |
| Bird5 | 97 | 95 | 97.94% | 90 | 5 |
| Bird6 | 93 | 91 | 97.85% | 87 | 4 |
| Bird7 | 106 | 104 | 98.11% | 96 | 8 |
| Bird8 | 98 | 96 | 97.96% | 82 | 14 |
| Bird9 | 112 | 110 | 98.21% | 102 | 8 |
| Bird10 | 103 | 102 | 99.03% | 96 | 6 |
| Bird11 | 113 | 111 | 98.23% | 107 | 4 |
| Bird12 | 106 | 106 | 100.00% | 99 | 7 |
| Bird13 | 108 | 106 | 98.15% | 97 | 9 |
| Bird14 | 110 | 108 | 98.18% | 99 | 9 |
| TOTAL | 1445 | 1426 | 98.68% | 1313 | 113 |

In the traditional condition, which includes four main questions and a free-form data entry task, 1427 attributes (5.10 per image per participant) in total were captured. After pre-analysis data preparation, which includes typo removal, redundant words elimination,

and blank answers removal, almost 92% (1309 attributes) of supplied data remained. In this condition, participants were questioned about the body shape, color pattern of body, bill color, and color of birds' head. To answer each question, participants have a number of pre-defined attributes as possible answers. For example, if a bird of duck family was showing on the screen, "Brown-Dark Gray with a Blue Patch" or "Spotted Brown with a Blue and White Patch" were two possible answers for the question about the color pattern of body.

Moreover, the system also enables contributors to supply free form data to enable additional attributes to be captured. The number of correct attributes (attributes remained after preparation) captured as the pre-defined and non-predefined data are 1002 (almost 77%) and 307 (23%) respectively. Table 5.2 shows the results for the traditional condition.

Table 5.2 The result of Traditional Guidance condition for each image

| | Attributes | Correct attributes | % Attribute correctness | Bird related attributes | Background Related attributes | Predefined attributes | Non-predefined attributes |
|--------|------------|--------------------|-------------------------|-------------------------|-------------------------------|-----------------------|---------------------------|
| Bird1 | 102 | 93 | 91.18% | 90 | 3 | 71 | 22 |
| Bird2 | 99 | 99 | 100.00% | 97 | 2 | 80 | 19 |
| Bird3 | 103 | 75 | 72.82% | 72 | 3 | 52 | 23 |
| Bird4 | 94 | 85 | 90.43% | 84 | 1 | 71 | 14 |
| Bird5 | 101 | 95 | 94.06% | 93 | 2 | 74 | 21 |
| Bird6 | 107 | 102 | 95.33% | 99 | 3 | 75 | 27 |
| Bird7 | 106 | 93 | 87.74% | 89 | 4 | 67 | 26 |
| Bird8 | 103 | 91 | 88.35% | 83 | 8 | 68 | 23 |
| Bird9 | 104 | 104 | 100.00% | 102 | 2 | 80 | 24 |
| Bird10 | 97 | 90 | 92.78% | 88 | 2 | 73 | 17 |
| Bird11 | 100 | 100 | 100.00% | 98 | 2 | 80 | 20 |

| | | | | | | | |
|--------|------|------|---------|------|----|------|-----|
| Bird12 | 105 | 105 | 100.00% | 103 | 2 | 80 | 25 |
| Bird13 | 99 | 99 | 100.00% | 96 | 3 | 80 | 19 |
| Bird14 | 107 | 78 | 72.90% | 77 | 1 | 51 | 27 |
| TOTAL | 1427 | 1309 | 91.73% | 1271 | 38 | 1002 | 307 |

In the cognitive guidance condition, a total of 1415 attributes (5.05 per image per participant) were obtained from participants. In this case, there is no typo/misspelling since all the respondents used the autocomplete feature to enter the initial attribute to trigger the RS-based guidance feature and then they chose data from the list of recommended attributes. The intention for using the autocomplete and guidance feature can be explained by the ease of use provided by these mechanisms. The initial attribute can be any kind of data which means participants can report either bird-related data or any other useful data (unanticipated data) of the image they were watching. However, all 1415 captured attributes are bird-related data. In other word, all 20 respondents preferred choosing from the recommended list of attributes rather than typing the attributes.

Table 5.3 The result of the Cognitive Guidance condition for each image

| | Attributes | Bird related attributes | Background related attributes |
|-------|------------|----------------------------|----------------------------------|
| Bird1 | 117 | 117 | 0 |
| Bird2 | 137 | 137 | 0 |
| Bird3 | 71 | 71 | 0 |
| Bird4 | 117 | 117 | 0 |
| Bird5 | 100 | 100 | 0 |
| Bird6 | 101 | 101 | 0 |
| Bird7 | 123 | 123 | 0 |
| Bird8 | 122 | 122 | 0 |

| | | | |
|--------|------|------|---|
| Bird9 | 68 | 68 | 0 |
| Bird10 | 81 | 81 | 0 |
| Bird11 | 123 | 123 | 0 |
| Bird12 | 76 | 76 | 0 |
| Bird13 | 70 | 70 | 0 |
| Bird14 | 109 | 109 | 0 |
| TOTAL | 1415 | 1415 | 0 |

5.1.1 Bird Identification (H-1.1)

To test bird identification (H-1.1), I analyzed data for the number of cases in which the bird was identified by the participants in each condition. All the identifications were coded as either “basic-level” or “expert-level” identifications. Basic-level identification means that participants reported the birds at a commonly understood high level of categorization, such as: duck, pigeon, and seabird. Expert-level identification means that participants identified the bird at the species level, such as: American Robin, Green Winged Teal, or Northern Fulmar. For each level of identification, I assigned a binary variable for each case (each bird for each participant), indicating whether it was identified correctly or not. For example, if an “American Black duck” is identified as a “duck”, the basic-level identification was coded as correct. Also, “American Black Duck” was coded as correct at the expert-level identification, while Mallard, Mottled Duck, and Gadwall were coded as incorrect.

For the “No Guidance” condition, in 60 cases out of 280 (21%), participants categorized birds at the basic level correctly which means the basic-level identification is successful, as shown in Table 5.4. This can be explained by the familiarity with these

animals among participants. However, in most cases participants were not able to have an expert-level identification. In less than 3% of cases (8 out of 280) contributors provided the bird's scientific name along with bird's attributes, as illustrated in Table 5.5. The result of identifications in both level are consistent with the results of prior research (Lukyanenko et al., 2014).

In the “Traditional Guidance” condition, each silhouette represented a distinct group of birds. For example, the first and the second silhouettes represented duck, and seabird, respectively. In 258 cases out of 280 (92%), respondents made a correct choice of the bird's silhouette, shown in Table 5.4, that means basic-level category of the bird was successfully identified. For the expert-level identification, if the participant could answer all the questions about the bird correctly, the system would identify the common name of the bird at the species-level. In 231 out of 280 cases, the bird was successfully identified at the species-level, as illustrated in Table 5.5.

Table 5.4 Basic-level identification in No guidance (NG), traditional Guidance (TG), and Cognitive Guidance (CG) conditions

| | No | | Traditional | | Cognitive | |
|-------|---------------|----------------|---------------|----------------|---------------|----------------|
| | Guidance (NG) | | Guidance (TG) | | Guidance (CG) | |
| | Correct/ | % Correct | Correct/ | % Correct | Correct/ | % Correct |
| | Incorrect | Identification | Incorrect | Identification | Incorrect | Identification |
| Bird1 | 11/9 | 55 | 18/2 | 90 | 19/1 | 95 |
| Bird2 | 0/20 | 0 | 20/0 | 100 | 20/0 | 100 |
| Bird3 | 0/20 | 0 | 13/7 | 65 | 20/0 | 100 |
| Bird4 | 9/11 | 45 | 20/0 | 100 | 19/1 | 95 |
| Bird5 | 5/15 | 25 | 19/1 | 95 | 18/2 | 90 |
| Bird6 | 7/13 | 35 | 19/1 | 95 | 16/4 | 80 |

| | | | | | | |
|---------|-------|-------|------|-------|------|-------|
| Bird7 | 6/14 | 30 | 18/2 | 90 | 19/1 | 95 |
| Bird8 | 2/18 | 10 | 19/1 | 95 | 15/5 | 75 |
| Bird9 | 2/18 | 10 | 20/0 | 100 | 19/1 | 95 |
| Bird10 | 2/18 | 10 | 19/1 | 95 | 19/1 | 95 |
| Bird11 | 1/19 | 5 | 20/0 | 100 | 19/1 | 95 |
| Bird12 | 0/20 | 0 | 20/0 | 100 | 19/1 | 95 |
| Bird13 | 10/10 | 50 | 20/0 | 100 | 18/2 | 90 |
| Bird14 | 5/15 | 25 | 13/7 | 65 | 20/0 | 100 |
| Average | | 21.43 | | 92.14 | | 92.86 |

For the “Cognitive Guidance” condition, almost 93% of cases (260 out of 280) were successfully identified at the basic-level, as shown in Table 5.4. Participants were either able to report the basic-level category of birds directly (by typing it as the initial attribute) or choose the correct basic-level category from the recommended list. In the expert-level identification, if respondents provide sufficient number of relevant attributes, the system was able to identify the birds ‘species name (expert-level identification), which happened in 245 cases (out of 280), or 87.5% of the time, as shown in Table 5.5.

Table 5.5 Expert-level identification in No guidance (NG), Traditional Guidance (TG), and Cognitive Guidance (CG) conditions

| | No | | Traditional | | Cognitive | |
|-------|---------------|----------------|---------------|----------------|---------------|----------------|
| | Guidance (NG) | | Guidance (TG) | | Guidance (CG) | |
| | Correct/ | % Correct | Correct/ | % Correct | Correct/ | % Correct |
| | Incorrect | Identification | Incorrect | Identification | Incorrect | Identification |
| Bird1 | 0/20 | 0 | 17/3 | 85 | 19/1 | 95 |
| Bird2 | 0/20 | 0 | 20/0 | 100 | 19/1 | 95 |
| Bird3 | 2/18 | 10 | 13/7 | 65 | 20/0 | 100 |
| Bird4 | 0/20 | 0 | 13/7 | 65 | 19/1 | 95 |

| | | | | | | |
|---------|------|------|------|-------|------|------|
| Bird5 | 0/20 | 0 | 17/3 | 85 | 18/2 | 90 |
| Bird6 | 0/20 | 0 | 18/2 | 90 | 15/5 | 75 |
| Bird7 | 0/20 | 0 | 13/7 | 65 | 16/4 | 80 |
| Bird8 | 0/20 | 0 | 12/8 | 60 | 15/5 | 75 |
| Bird9 | 5/15 | 25 | 20/0 | 100 | 18/2 | 90 |
| Bird10 | 1/19 | 5 | 16/4 | 80 | 19/1 | 95 |
| Bird11 | 0/20 | 0 | 20/0 | 100 | 15/5 | 75 |
| Bird12 | 0/20 | 0 | 20/0 | 100 | 16/4 | 80 |
| Bird13 | 0/20 | 0 | 20/0 | 100 | 17/3 | 85 |
| Bird14 | 0/20 | 0 | 12/8 | 60 | 19/1 | 95 |
| Average | | 2.85 | | 82.05 | | 87.5 |

As shown in Table 5.6, generally there was significantly more basic-level identification in the “Cognitive Guidance” condition than in the “No Guidance” condition. Using independent samples t-test confirmed that the mean difference of “Cognitive Guidance” and “No Guidance” condition was significant ($t=15.607$, $d.f.=26.933$, t -test p -value <0.001). In addition, for each of 14 birds, the basic-level identification was higher in the “Cognitive Guidance” than the “No Guidance” condition, as shown in Table 5.4.

Table 5.6 A comparison of the number of identifications at basic-level & expert-level

| | Number of Basic-level Identification (out of 280) | Mean (Identifications per Participant) | Number of Expert-level Identification (out of 280) | Mean (Identifications per Participant) |
|---------------------------------|--|---|---|---|
| No Guidance (NG) | 60 | 3 | 8 | 0.4 |
| Traditional Guidance (TG) | 258 | 12.9 | 231 | 11.55 |
| Cognitive Guidance (CG) | 260 | 13 | 245 | 12.25 |

In terms of basic-level identification, the result showed that the means of “Cognitive Guidance” and “Traditional Guidance” conditions was not significantly different (using independent samples t-test, $t=0.276$, $d.f.=37.420$, t-test p-value= 0.784). However, in 2 of 14 birds, participants in the “Cognitive Guidance” condition provided a higher percentage of basic-level identification compared to those in the “Traditional Guidance” condition.

For expert-level identification, as Table 5.6 shows, there were significantly more expert-level identifications in the “Cognitive Guidance” condition than in the “No Guidance” condition. The difference of means between “Cognitive Guidance” and “No Guidance” condition was 11.85 (12.25-0.4) per respondent. Using independent samples t-test confirmed that the means of “Cognitive Guidance” and “No Guidance” condition were significantly different ($t=39.299$, $d.f.=35.610$, t-test p-value <0.001). In addition, for each of 14 birds, the difference of expert-level identification between the “Cognitive Guidance” condition and the “No Guidance” condition was significant as shown in Table 5.5.

In terms of expert-level identification, the result showed that the means of “Cognitive Guidance” and “Traditional Guidance” conditions were not significantly different (using independent samples t-test, $t=1.321$, $d.f.=28.131$, t-test p-value= 0.197). As shown in Table 5.5, in 3 cases participants in the “Cognitive Guidance” condition provided a significantly higher percentage of expert-level identification compared to those in the “Traditional Guidance” condition.

Considering the result of basic and expert level identification, having a guidance in

data entry enables contributors to identify species better compared to free-form data entry task, thus the result provides good support for H1.1.

5.1.2 Data Relevance (H-1.2)

To test data relevance, I measured data in terms of (1) total number of attributes, (2) the number of attributes related to birds, and (3) the number of attributes related to surroundings reported by participants for each bird.

As shown in Table 5.7, all three conditions had same performance in terms of the number of attributes each condition captured. However, the “Cognitive guidance” only captured bird-related attributes. The “No Guidance” condition was more successful in capturing attributes related to background, comparing to the “Traditional Guidance” condition and “Cognitive Guidance” conditions.

Table 5.7 Comparison of the number of total attributes captured in each condition

| | Total number of attributes | Mean (attributes per participants) | Number of bird-related attributes | Number of background-related attributes |
|---------------------------|----------------------------|------------------------------------|-----------------------------------|---|
| No Guidance (NG) | 1426 | 71.3 | 1313 (92%) | 113 |
| Traditional Guidance (TG) | 1309 | 65.45 | 1271 (97%) | 38 |
| Cognitive Guidance (CG) | 1415 | 70.75 | 1415 (100%) | 0 |

Using an independent t-test showed the difference of mean in total number of attributes between each two conditions was not significant, as shown in Table 5.8. All p-values from the test are greater than 0.05 (significant value).

Table 5.8 T-test for the number of total captured attributes

| | Mean difference | t value | d.f. | p-value |
|---|-----------------|---------|--------|---------|
| Cognitive guidance vs. No guidance | -0.55 | -0.055 | 35.978 | 0.957 |
| Cognitive guidance vs. Traditional guidance | 5.30 | 0.776 | 26.709 | 0.445 |
| Traditional guidance vs. No guidance | -5.85 | -0.696 | 23.885 | 0.493 |

To determine which conditions had a better performance in data relevance, I calculated the percentage of bird-related data in each condition.

Table 5.9 T-test for percentage of relevant attributes each condition captured

| | Mean difference | t value | d.f. | p-value |
|---|-----------------|---------|--------|---------|
| Cognitive guidance vs. No guidance | 7.9614 | 6.107 | 13 | <0.001 |
| Cognitive guidance vs. Traditional guidance | 2.9228 | 5.680 | 13 | <0.001 |
| Traditional guidance vs. No guidance | 5.038 | 3.595 | 26.955 | 0.002 |

As shown in Table 5.10, the overall percentage of the “Cognitive Guidance” was almost 8% higher than the “No Guidance” condition, and 3% higher than the “Traditional Guidance” condition. Using independent samples t-test confirmed that the “Cognitive Guidance” had the best performance among three conditions in terms of capturing relevant data for species identification purpose. All p-values were <0.05 (significant value), confirming that “Cognitive Guidance” condition performed significantly better in

capturing data of better level of relevancy than the other two conditions. The test of relevance was also performed for each bird in all conditions, and the result was consistent.

Table 5.10 illustrates the percentage of bird-related attributes for each bird in all conditions.

Table 5.10 The comparison of data relevancy in No guidance (NG), Traditional Guidance (TG), and Cognitive Guidance (CG) conditions

| | No | | | Traditional | | | Cognitive | | |
|---------|--------------------|-------------------|--------------------------------|--------------------|-------------------|--------------------------------|--------------------|-------------------|--------------------------------|
| | Guidance (NG) | | | Guidance (TG) | | | Guidance (CG) | | |
| | Total attribute | % Bird related | % backgro und related | Total attribute | % Bird related | % backgro und related | Total attribute | % Bird related | % backgro und related |
| Bird1 | 99 | 77.78 | 22.22 | 93 | 96.77 | 3.23 | 117 | 100 | 0 |
| Bird2 | 100 | 94.00 | 6.00 | 99 | 97.98 | 2.02 | 137 | 100 | 0 |
| Bird3 | 97 | 93.81 | 6.19 | 75 | 96.00 | 4.00 | 71 | 100 | 0 |
| Bird4 | 101 | 95.05 | 4.95 | 85 | 98.82 | 1.18 | 117 | 100 | 0 |
| Bird5 | 95 | 94.74 | 5.26 | 95 | 97.89 | 2.11 | 100 | 100 | 0 |
| Bird6 | 91 | 95.60 | 4.40 | 102 | 97.06 | 2.94 | 101 | 100 | 0 |
| Bird7 | 104 | 92.31 | 7.69 | 93 | 95.70 | 4.30 | 123 | 100 | 0 |
| Bird8 | 96 | 85.42 | 14.58 | 91 | 91.21 | 8.79 | 122 | 100 | 0 |
| Bird9 | 110 | 92.73 | 7.27 | 104 | 98.08 | 1.92 | 68 | 100 | 0 |
| Bird10 | 102 | 94.12 | 5.88 | 90 | 97.78 | 2.22 | 81 | 100 | 0 |
| Bird11 | 111 | 96.40 | 3.60 | 100 | 98.00 | 2.00 | 123 | 100 | 0 |
| Bird12 | 106 | 93.40 | 6.60 | 105 | 98.10 | 1.90 | 76 | 100 | 0 |
| Bird13 | 106 | 91.51 | 8.49 | 99 | 96.97 | 3.03 | 70 | 100 | 0 |
| Bird14 | 108 | 91.67 | 8.33 | 78 | 98.72 | 1.28 | 109 | 100 | 0 |
| Average | | 92.07 | 7.93 | | 97.1 | 2.9 | | 100 | 0 |

Overall, the results provide good support for H-1.2, meaning that having a guidance feature in a data entry task results in data of better level of relevancy compared to using no guidance.

5.1.3 Data Objectivity & Data Completeness (H-1.3, H-1.4)

To test objectivity, I assigned a binary variable to each response indicating whether it was objective/subjective for the bird it described. Aligned with the definition of data objectivity, *the extent to which data is unbiased, unprejudiced and impartial*, Objective data in this research is referred to that data which has no personal opinion or feeling. On the other hand, if data is influenced by a contributor's feelings or opinions, it is called subjective. For example, in description of Mallard duck, "colorful bird" was coded as objective, while "lonely" was coded as subjective. I measured total number of objective responses for each conditions as well as the percentage of objective to subjective responses.

Table 5.11 Objectivity of responses in No Guidance (NG), Traditional Guidance (TG), and Cognitive Guidance (CG) conditions

| | No | | Traditional | | Cognitive | |
|--------|--------------------------------------|-----------------------|--------------------------------------|-----------------------|--------------------------------------|-----------------------|
| | Guidance (NG) | | Guidance (TG) | | Guidance (CG) | |
| | % | | % | | % | |
| | Objective/ Subjective Response | Objective Response | Objective/ Subjective Response | Objective Response | Objective/ Subjective Response | Objective Response |
| Bird1 | 99/0 | 100 | 93/9 | 91.18 | 117/0 | 100 |
| Bird2 | 100/0 | 100 | 99/0 | 100 | 137/0 | 100 |
| Bird3 | 97/2 | 97.98 | 75/28 | 72.82 | 71/0 | 100 |
| Bird4 | 101/0 | 100 | 85/9 | 90.43 | 117/0 | 100 |
| Bird5 | 95/2 | 97.94 | 95/6 | 94.06 | 100/0 | 100 |
| Bird6 | 91/2 | 97.85 | 102/5 | 95.33 | 101/0 | 100 |
| Bird7 | 104/2 | 98.11 | 93/13 | 87.74 | 123/0 | 100 |
| Bird8 | 96/2 | 97.96 | 91/12 | 88.35 | 122/0 | 100 |
| Bird9 | 110/2 | 98.21 | 104/0 | 100 | 68/0 | 100 |
| Bird10 | 102/1 | 99.03 | 90/7 | 92.78 | 81/0 | 100 |
| Bird11 | 111/2 | 98.23 | 100/0 | 100 | 123/0 | 100 |

| | | | | | | |
|---------|-------|-------|--------|-------|-------|-----|
| Bird12 | 106/0 | 100 | 105//0 | 100 | 76/0 | 100 |
| Bird13 | 106/2 | 98.15 | 99/0 | 100 | 70/0 | 100 |
| Bird14 | 108/2 | 98.18 | 78/19 | 72.9 | 109/0 | 100 |
| Average | | 98.68 | | 91.73 | | 100 |

To determine which condition has better performance in terms of data objectivity, I measured the percentage of objective response for each condition. Using two independent samples t-test on the means revealed that the “Cognitive Guidance” has the best performance in capturing objective data. In this condition, all data is objective which means an excellent level of objectivity was achieved. As shown in Table 5.12, the overall percentage of objectivity in the “Cognitive Guidance” condition was 1.32% higher than the “No Guidance” condition, and 8% higher than the “Traditional Guidance”.

Table 5.12 T-test for percentage of objective attributes each condition captured

| | Mean difference | t value | d.f. | t-test p-value |
|---|-----------------|---------|-------|----------------|
| Cognitive guidance vs. No guidance | 1.311 | 5.430 | 13 | <0.001 |
| Cognitive guidance vs. Traditional guidance | 8.1721 | 3.320 | 13 | 0.006 |
| Traditional guidance vs. No guidance | -6.86 | -2.774 | 13.25 | 0.016 |

The difference of data objectivity between each two conditions are statistically significant, since all p-values for t-test are < 0.05 (significant value).¹² In other words, data

¹² The objectivity was tested for each bird in all three conditions, and the result was consistent with the result of t-test performed for the percentage of objective attributes in all conditions.

of “Cognitive Guidance” are more objective than data in other two conditions. Overall, the result of data objectivity provides good support for H-1.3.

For ensuring data completeness, all the conditions are capable of capturing anticipated and unanticipated data. In all conditions, contributors were enabled to report unanticipated data through the free-form text box along with data related to birds. After pre-analysis preparation, the total number of attributes in each conditions shows that all three conditions had a similar performance, as shown in Table 5.7. Using t-test for the means shows that the difference between the total numbers of attributes was not significant (Table 5.8). Although “No Guidance” condition had a better performance in capturing background-related data compared to the other conditions, the “Cognitive Guidance” was resulted in more bird identifications. The weak capability of “Cognitive Guidance” condition in capturing unanticipated data can be explained by the high intention of contributors to choose attributes from the recommended list rather than typing new information. Overall, the result of comparing the number of captured attributes provides good support for H-1.4.

5.2 Discussion

This chapter evaluates the impact of using a guidance on data relevance and objectivity, as well as on bird identification. The results demonstrate that there is no significant difference among the total amount of captured data among the 3 conditions. This suggests that using a guidance along with an instance-based approach will not prevent contributors from providing as many attributes as they are willing to submit. However, cognitive guidance produced no background-related data. The results also demonstrate that

level of data relevance will be improved by using a guidance feature in data entry. By filtering irrelevant data and suggesting relevant data, an RS-based guidance can assist contributors to supply data of higher level of relevancy to the intended purpose of research – species identification.

The result also confirmed that using an RS-based guidance can result in more objective data. Having a list of recommended attributes, contributors have less chance to mix their own opinions or feelings with data while reporting a species.

Although the number of captured attributes was statistically the same in all conditions, the “Cognitive Guidance” condition better enabled participants to correctly identify species.

The results also showed that there was no significant difference in the level of data accuracy among the three ways of data collection. However, the “Cognition Guidance” produced more accurate species identification at species and basic level. Both species and basic level of identification in the “Cognitive Guidance” condition were significantly higher than the “No Guidance” condition.

Finally, the results provide empirical evidence of the advantages of using a guidance in an instance-based approach data collection. Currently, the instance-based model approach for data collection is promising in terms of data accuracy and dataset completeness. However, it can result in some limitations. This thesis tried to mitigate the limitation of an instance-based model data collection task. The results highlighted an opportunity to produce data of higher level of objectivity and relevancy from the citizen scientists, which leads us to identify species properly.

6 Conclusion, Implications, and Future Study

6.1 Conclusion

User-generated content enables organizations to utilize customers' sensory capability and insight to support analysis and decision making. Among other uses, projects that involve citizen scientists are expanding rapidly, particularly in biology. Citizen science projects recruit members of the general public to gather scientific data, and/or to participate in the design of studies, analyze the data or interpretation of results. Although citizen science is capable of producing large amounts of data over a broad geographical or temporal scale, the quality of contributed data has frequently been criticized. Among proposed approaches to enhance data quality, an instance-based data modeling approach is capable of improving the data accuracy, and dataset completeness as two important dimensions of data quality. Despite its potential to improve some dimensions of data quality, an instance-based approach has a number of challenges. This thesis focused on challenges of an instance-based approach to data collection, and proposed a method to mitigate the limitations of the instance-based approach data collection by providing contributors with guidance during data entry.

The empirical evidence demonstrate that using a guidance feature data lead contributors to supply data of higher level of objectivity and relevance. The results also show that the guidance feature assists contributors in species identification (a widely used level of classification that is useful in biology research).

6.2 Implications

6.2.1 Addressing Challenges to the Instance-based Approach

The instance-based approach is capable of capturing data of better quality in terms of data accuracy and dataset completeness. However, when contributors can supply data without any constraint or guidance, a large number of idiosyncratic attributes will be produced. To manage a large amount of data, an RS-based guidance feature can filter irrelevant data to make data ready for further analysis. The guidance feature guides contributors to attributes by recommending potential matches and allows contributors to choose from them. Selecting attributes from the recommended list increases the likelihood that captured data are related to consumer's needs. This method also assisted contributors to identify species properly, even with low expertise and domain knowledge.

6.2.2 Reducing discrepancy between contributors and data consumers

One of the main problems of citizen science projects is the discrepancy between what consumers need and what contributors are able to submit. Members of the general public often lack domain knowledge or expertise, thus data consumers data needs are often incompatible with the data a contributor can provide. Setting constraints on data entry – such as restricting data to classes of interest to data consumers – may help consumers to gain data of better quality, but such input restrictions may prevent contributors from sharing valuable insights.

A guidance feature along with the instance-based approach guides contributors to data which is compatible with consumer's perspective while not restrict contributors to freely communicate their valuable perspectives.

In the context of biology, species identification is one of the key consumer's requirements. Proper identification may reduce the discrepancy between the consumer's need and the contributor's capabilities. However citizen scientists often lack of domain knowledge, thus identifying species may be threatened by their low expertise. This thesis demonstrated that a guidance feature on input leads contributors to a proper species identification at basic and species-level. By choosing attributes from the recommended list, contributors with low or even no domain expertise are able to identify the species.

6.3 Future study

6.3.1 Developing a Framework for the UGC Quality Dimensions

This thesis demonstrated advantages of using a guidance feature on two dimensions of data quality – data relevance and objectivity. Data quality dimensions have been extensively studied before; however, prior research usually considered data quality in traditional organizational settings. In traditional settings, the definition of data quality reflects the consumer's perspective. However, in the context of UGC contributors may not be capable of fully understand what the consumer really needs.

Considering differences between traditional organizational settings and UGC, a novel definition of quality in the context of UGC is needed. The definition should highlight the data contributors' perspectives, as they are key data creators in UGC projects. Traditional dimensions should also be mapped to the context of UGC. Thus one possible area for future research is developing a framework for quality dimensions for UGC.

6.3.2 Addressing Challenges to RS-based Guidance Approach

Although the RS-based guidance feature demonstrated potential in improving data relevance and objectivity, it has a number of challenges that can be addressed in future studies.

One is developing a database which is the base of operating a content-based recommendation system. The database consists of key characteristics of phenomenon which contributors are trying to report. Creating such a database can be challenging if the scope of a citizen science project is very large.

Another issue when using a guidance in data entry is that contributors may tend to select from the recommended list of attributes rather than doing free form data entry, resulting in information loss. This limitation can be addressed by developing a hybrid approach which motivates contributors to supply their own data as well as using a guidance in data entry.

The content-based filtering methods normally base their recommendations on the contributor's information. They usually ignore what other contributors are supplying to the system. Future work can investigate the applicability of using the collaborative filtering method as an alternative guidance feature for data entry.

6.3.3 From Bird Identification to Other Purposes

Another area for future research is applying the proposed method for a different purpose. Although major science projects, such as eBird focus on species identification, an RS based guidance feature can be applied to improve other dimensions of data quality. In Any kind of data entry, an RS-based guidance can assist the contributor to provide data of better quality

- extent and depth of data.

Finally, it is suggested that the study is repeated in a field setting. The field setting may better represent the type of users and forms of participation that are typical of citizen scientists.

References

- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.
- Alabri, A., & Hunter, J. (2010). *Enhancing the quality and trust of citizen science data*. Paper presented at the e-Science (e-Science), 2010 IEEE Sixth International Conference.
- Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behavior*, 27(5), 1834-1839.
- Batini, C., & Scannapieco, M. (2010). *Data Quality: Concepts, Methodologies and Techniques*.
- Bhattacharjee, Y. (2005). Citizen Scientists Supplement Work of Cornell Researchers. *Science*, 308(5727), 1402-1403.
- Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 977-984.
- Bonter, D. N., & Cooper, C. B. (2012). Data validation in citizen science: a case study from Project FeederWatch. *Frontiers in Ecology and the Environment*, 10(6), 305-307.
- Cleary, M., Horsfall, J., & Hayter, M. (2014). Data collection and sampling in qualitative research: does size matter? *Journal of advanced nursing*, 70(3), 473-475.
- Connor-Greene, P. A. (2007). Observation or interpretation? Demonstrating unintentional subjectivity and interpretive variance. *Teaching of Psychology*, 34(3), 167-171.
- Cooper, J. K., Sykes, G., King, S., Cottrill, K., Ivanova, N. V., Hanner, R., & Ikononi, P. (2007). Species identification in cell culture: a two-pronged molecular approach. *In Vitro Cellular & Developmental Biology-Animal*, 43(10), 344-351.
- Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters*, 4(6), 433-442.

- Dhar, V., & Chang, E. A. (2009). Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4), 300-307.
- Di Gangi, P. M., Wasko, M. M., & Hooker, R. E. (2010). Getting Customers' Ideas to Work for You: Learning from Dell how to Succeed with Online User Innovation Communities. *MIS Quarterly Executive*, 9(4).
- Fernández-Tobías, I., Cantador, I., Kaminskis, M., & Ricci, F. (2012). *Cross-domain recommender systems: A survey of the state of the art*. Paper presented at the Spanish Conference on Information Retrieval.
- Gao, G., McCullough, J. S., Agarwal, R., & Jha, A. K. (2010). *Are doctors created equal? An investigation of online ratings by patients*. Paper presented at the Proceedings of the Workshop on Information Systems and Economics (WISE).
- Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221.
- Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4), 12-18.
- Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS biology*, 2(10), e312.
- Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in ecology & evolution*, 27(2), 130-137.
- Jasanoff, S. (2003). Technologies of humility: citizen participation in governing science. *Minerva*, 41(3), 223-244.
- Katona, S. K., & Kraus, S. D. (1979). *Photographic identification of individual humpback whales (Megaptera novaeangliae): evaluation and analysis of the technique*: NTIS.
- Kay, P., & Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition*, 102(2), 289-298.
- Kim, S., Mankoff, J., & Paulos, E. (2013). *Sensr: evaluating a flexible framework for authoring mobile data-collection tools for citizen science*. Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work.

- Kim, S., Robson, C., Zimmerman, T., Pierce, J., & Haber, E. M. (2011). *Creek watch: pairing usefulness and usability for successful citizen science*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
- Kremen, C., Ullman, K., & Thorp, R. (2011). Evaluating the quality of citizen-scientist data on pollinator communities. *Conservation Biology*, 25(3), 607-617.
- Krumm, J., Davies, N., & Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4), 10-11.
- Lavine, B., & Carlson, D. (1987). European bee or Africanized bee? Species identification through chemical analysis. *Analytical Chemistry*, 59(6), 468A-470A.
- Lee, Y. W., & Strong, D. M. (2003). Knowing-why about data processes and data quality. *Journal of Management Information Systems*, 20(3), 13-39.
- Leung, L. (2009). User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. *New media & society*, 11(8), 1327-1347.
- Liu, J., & Ram, S. (2009). Who does what: Collaboration patterns in the wikipedia and their impact on data quality.
- Lukyanenko, R. (2014). *An information modeling approach to improve quality of user-generated content*. (Doctoral dissertation, Memorial University of Newfoundland).
- Lukyanenko, R., & Parsons, J. (2015). Information quality research challenge: adapting information quality principles to user-generated content. *Journal of Data and Information Quality (JDIQ)*, 6(1), 3.
- Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2014). The IQ of the crowd: understanding and improving information quality in structured user-generated content. *Information Systems Research*, 25(4), 669-689.
- Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)*, 1(1), 2.
- Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012).

- The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6), 298-304.
- Nov, O., Arazy, O., & Anderson, D. (2011). *Dusting for science: motivation and participation of digital citizen science volunteers*. Paper presented at the Proceedings of the 2011 iConference.
- Özgen, E. (2004). Language, learning, and color perception. *Current Directions in Psychological Science*, 13(3), 95-98.
- Parsons, J., & Wand, Y. (2000). Emancipating instances from the tyranny of classes in information modeling. *ACM Transactions on Database Systems (TODS)*, 25(2), 228-268.
- Paulos, E. (2009). *Designing for Doubt Citizen Science and the Challenge of Change in Engaging Data*. Paper presented at the First International Forum on the Application and Management of Personal Electronic Information. .
- Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems *The adaptive web* (pp. 325-341): Springer.
- Prestopnik, N. R., & Crowston, K. (2011). *Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system*. Paper presented at the e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on.
- Redman, T. C. (1996). *Data quality for the information age*: Artech House, Inc.
- Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. *Datenbank-Spektrum*, 14(January), 6-14.
- Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). *Data quality: A survey of data quality dimensions*. Paper presented at the Information Retrieval & Knowledge Management (CAMP), 2012 International Conference.
- Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24(9), 467-471.
- Silvertown, J. (2010). Taxonomy: include social networking. *Nature*, 467(7317), 788-788.

- Simpson, R., Page, K. R., & De Roure, D. (2014). *Zooniverse: observing the world's largest citizen science platform*. Paper presented at the Proceedings of the 23rd international conference on world wide web.
- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10), 2282-2292.
- Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology*, 23(3), 457-482.
- Vandecasteele, A., & Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system: the case of OpenStreetMap *OpenStreetMap in GIScience* (pp. 59-80): Springer.
- Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.
- Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2), 58-65.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.
- Zevin, M., Coughlin, S., Bahaadini, S., Besler, E., Rohani, N., Allen, S., . . . Larson, S. L. (2017). Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6), 064003.

Appendix 1: Images Used in the Laboratory Experiments



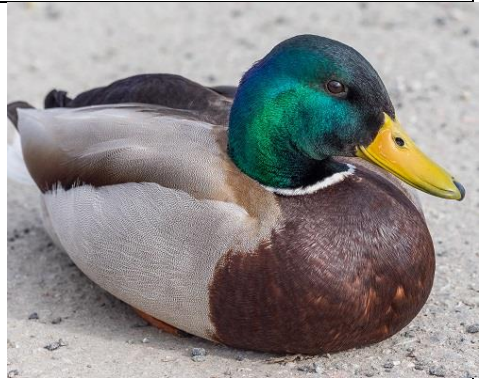
American Black Duck



Caspian Tern



Blue Jay



Mallard Duck



Willow Ptarmigan



Rock Pigeon



Blue Winged Teal



Ruffed Grouse



Atlantic Puffin



Canada Goose



Common Tern



American Robin



Green Winged Teal



Northern Fulmar