# Supporting Content Curation Communities: The Case of the Encyclopedia of Life

**Dana Rotman[1]+, Kezia Procita[1], Derek Hansen[1], Cynthia Sims Parr[2], Jennifer Preece[1]**

[1] College of Information Studies
University of Maryland, College Park
College Park, Maryland 20742
{drotman, kprocita, dlhansen, preece} @umd.edu
Phone: 301.405.7185 Fax: 301.314.9145
+ - corresponding author

[2] National Museum of Natural History
Smithsonian Institution
Washington, DC 20013
parrc@si.edu
Phone: 202.633.9513 Fax: 202.633.8742

DRAFT

## Abstract

This paper explores the opportunities and challenges of creating and sustaining large-scale, "content curation communities" through an in-depth case study of the Encyclopedia of Life (EOL). Content curation communities are large scale crowdsourcing endeavors that aim to curate existing content into a single repository, making these communities different from content creation communities such as Wikipedia. In this paper we define content curation communities and provide examples of this increasingly important genre. We then follow by presenting EOL, a compelling example of a content curation community, and describe a case study of EOL based on analysis of interviews, online discussions, and survey data. Our findings are characterized into two broad categories – *information integration* and *social integration*. Information integration challenges at EOL include the need to (a) accommodate multiple taxonomic classification sources and (b) integrate traditional peer reviewed sources with user-generated, non-peer reviewed content. Social integration challenges at EOL include the need to (a) establish the credibility of open-access resources within the scientific community, and (b) facilitate collaboration between experts and novices. After identifying the challenges, we discuss the potential strategies EOL and other content curation communities can use to address them, and provide technical, content, and social design recommendations for overcoming them.

## Introduction

Scientific progress is in large part dependent on the development of high-quality, shared information resources tailored to meet the needs of various scientific communities. Traditionally created and maintained by a handful of paid scientists or information professionals, scientific information resources such as repositories, databases and archives, are increasingly being crowdsourced to professional and nonprofessional volunteers in what we define as "content curation communities". Content curation communities are distributed communities of volunteers who work together to curate data from disparate resources into coherent, validated, and oftentimes freely-available repositories. Content curation communities help to develop resources on drug discovery (Li, Cheng, Wang, & Bryant, 2010), worm habitats or bird migration (Sullivan , et al., 2009), astronomic shifts (Raddick, et al., 2007), and language (Hughes, 2005), to name a few.

Content *curation* communities are related to content *creation* communities like Wikipedia, but face different challenges, as curation and creation are fundamentally different activities. They both require a community of contributors to help maintain free content, but the tasks performed by the contributors differ, as does the requisite skill set. Content curation communities aggregate *existing* content with its associated intellectual property claims into a comprehensive repository, while content creation communities avoid intellectual property concerns by *creating* their own content from scratch. Both types of communities fill important niches in the information landscape and have already proven their worth. While content creation communities, particularly Wikipedia, have been examined extensively to uncover the principles that have led to their successes and failures (c.f. Forte & Bruckman, 2005; Kittur, Chi, Pendleton, Suh, & Mytkowicz, 2006; Kriplean, Beschastnikh, & McDonald, 2008; Nov, 2007), there have not been comparable studies of content curation communities. In order to realize the ambitious goals of such communities, we must understand how to effectively design the technologies, information standards, processes, and social practices that support them.

The goal of this paper is to characterize content curation communities and their increasing prominence in the information landscape, and then to use the formative years of the Encyclopedia of Life

DRAFT

(EOL) as a case study for providing insights into the design of such communities. Our key research questions are:

- What are the unique informational and social challenges experienced by EOL – as an example of a content curation community?
- What are the key information and social design choices available to designers of EOL?

## Content curation communities

In this section we discuss the novel concept of content curation communities, their evolution, and the differences between them and other types of community crowdsourcing.

Content curation communities can be defined by describing their core elements:

- *Community*: The word "community" emphasizes that a large number of people are involved in some type of coordinated effort. Although not necessary, they typically include volunteers and reflect the social roles and participation patterns that mirror those of other types of volunteer-based online communities (Preece, 2000).
- Content: The word "content" conveys the primary emphasis of the community. Curation communities exist for the purpose of providing content, which can take on many forms such as text, multimedia (e.g., images, videos), and structured data (e.g. metadata). The content is typically free and web-based, but doesn't necessarily have to be. The nature and scope of the content is based on topics and resources that that the specific community finds valuable enough to curate.
- Curation: The word "curation" describes the type of work that is performed by these communities; it is their *raison d'être*. Curation connotes the activities of identifying, selecting, verifying, organizing, describing, maintaining, and preserving existing artifacts. This term follows the term "content" since the artifacts being curated by the community are not the objects themselves but the content they represent. Curators of museum objects or physical samples are typically experts in their domain, suggesting that this role requires some level of content expertise, which can range from lay interest in the specific domain to professional-level expertise, although this is not a requirement for the definition.

DRAFT

We focus on scientific content curation communities, which build on a long tradition of sharing and curating scientific data and literature. The scientific enterprise has always been collaborative. However, the advent of the Internet has enabled large distributed scientific communities to collaborate at a scale and pace never before realized (Borgman, 2007). Distributed communities of scientists were coined “collaboratories”, a term first used in the early 1980s, to indicate a laboratory without walls where data is shared via advances in information technology and independent of time and location (Finholt & Olson, 1997). Collaboratories and their more recent instantiations facilitate data sharing by helping bridge time and space, as well as reduce status barriers, thus leading to an increase in the diversity of participants and participation levels (Birnholtz & Bietz, 2003; Bos, et al., 2007; Finholt, 2002). A core function of many collaboratories is data sharing. Yet, despite successful examples of data sharing in collaboratories and related efforts, barriers continue to exist, and include finding and re-using data appropriately, meeting the current standards of scientific practice, and providing increased access to data while decreasing concerns about data quality (Zimmerman, 2007).

Bos and his colleagues identified seven distinct types of collaboratories, including “community data systems”, where scientific data is created, maintained, and improved by a geographically distributed community (Bos, et al., 2007). Unlike content curation communities that strive to aggregate various types of content, whether traditional peer-reviewed scientific data or user generated data, community data systems and related efforts such as “curated databases” (Buneman, Cheney, Tan, & Vansummeren, 2008) focus solely on curating scientific data. In many of these community data systems volunteers play the role of content contributor by submitting new data that is then vetted and annotated by paid professionals. Content curation communities don’t follow the role of original data collectors. For example, the Encyclopedia of Life reverses the traditional community data system approach by having its volunteer content curators annotate and vet data that has already been collected in community data systems or other publications. In many of these cases the EOL curators’ primary contribution is to identify and combine data from reputable sources into a meaningful whole.

In addition to being influenced by collaboratories, content curation communities have been influenced by more generic social media initiatives and tools. As web content has proliferated, the need to

DRAFT

curate general content, in order to separate the wheat from the chaff, has increased. For example, a number of news sites curate stories of interest. Some, like Google News, use automatic processing to group similar articles from reputable sources. Others, such as Slashdot, rely on user-generated ratings of posts to filter out the low quality discussions about news (Lampe & Resnick, 2004). Social tagging sites like Digg, Flickr, and YouTube also allow popular content to rise to the top (Rotman & Preece, 2010). While many of these approaches work well at identifying "popular" content, they are not designed for situations where specialized expertise is needed, making it unclear how to apply similar design strategies to scientific domains. Many online communities use Frequently Asked Questions (FAQs) and wiki repositories (Hansen et al. 2007) as a way of collecting high-quality content, but unlike content curation communities they are focused on their own community content, not content created by other sources. One of the goals of this paper is to learn from strategies employed by existing social media tools to effectively aid in negotiating the challenges faced by content curation communities.

## Encyclopedia of Life

*"Imagine an electronic page for each species of organism on Earth, available everywhere by single access on command. The page… comprises a summary of everything known about the species' genome, proteome, geographical distribution, phylogenetic position, habitat, ecological relationships and, not least, its practical importance for humanity."* (Wilson, 2003)

The Encyclopedia of Life (EOL) (Figure 1) is "a sweeping global effort to gather and share the vast wealth of information on every creature—animals, plants and microorganisms—and make it available as a web-based resource" (EOL, 2011). The hope is that making this information freely available will help "energize the science of biology," as well as encourage people to "study, protect and prolong the existence of Earth's precious species". EOL is an open access-based, aggregation portal (www.eol.org). One of its bold goals is to engage individuals, including scientists and non-scientists in contributing to scientific information curation. EOL is structured such that that each species is presented on a dedicated page, where all available and relevant information, vetted and non-vetted alike, is presented. Currently there are 3 million pages, representing 1.9 million known species on Earth. Most of these are stubs, but

DRAFT

trusted content is available on over 600,000 pages that collectively are viewed by nearly 3 million visitors a year. Over 40,000 people have registered on the site.

[Insert Figure 1 here]

EOL was initially supported by major foundations as well as by matching grants raised by each of the initial cornerstone institutions that host its paid staff (Smithsonian Institution, Harvard University, Field Museum, the Marine Biological Laboratory, and Missouri Botanical Gardens). Paid staff members include information, education and technical professionals. Contributors include interested volunteers, who can freely contribute information about organisms, as well as credentialed scientists and "citizen scientists" (Cohn, 2008), who agree to curate information on EOL. Seven hundred and fifty content curators are associated with EOL, two hundred of whom are currently active. Curators can trust valid information and rate it for quality and completeness, or untrust and even hide inaccurate data objects. A small number of scientific contributors and content curators receive temporary part-time financial support from a competitive Fellows program, but most content on EOL arises through volunteer partnerships with database and library digitization projects and citizen contributions. All the vetted and unvetted content is freely available online.

EOL predecessors include sites that provide freely available, biodiversity-focused, and comprehensive species coverage. A well-known project that is among the most similar to EOL is the Tree of Life Web project (TOLWeb), which was created with a focus on biodiversity and evolution to share with readers the "interrelationships of earth's wealth of species". Unlike EOL, TOLWeb is a content *creation* community where pre-approved users from the professional science community directly add information. TOLWeb also seeks to provide information on both living and extinct species, unlike EOL, which focuses on contemporary species (Maddison, Schulz, & Maddison, 2007). This slight variation of scope is common among EOL's predecessors; nothing exactly like the EOL has been created to date. Prior to TOLWeb's 1996 launch, two notable digital systems had been created. The University of California at Berkeley's Phylogeny of Life (now the History of Life Through Time) and the California Academy of Sciences LIFEMap CD-ROM project. Each of these projects sought to provide digital access to species information (Maddison, et al., 2007). Other related projects include: the Catalogue of Life, Wikispecies,

DRAFT

the Animal Diversity Web, FishBase, AmphibiaWeb, LepTree, uBio, iSpecies, and TreeBASE. Each of these sites also differs in scope from the EOL; some focus more on a specific subset of species, focus only on taxonomic data, or have different operating models. Many of EOL's predecessors are now content partners to EOL.

EOL derives its content from scientific content partners but also from user-generated content: in December of 2009, EOL began a content partnership with Wikipedia, importing Wikipedia species pages into EOL . While the addition of these articles increased EOL's species coverage considerably, it was controversial due to the non-traditional authoring process of Wikipedia that allows anyone to contribute content. Other examples of non-authoritative source material that EOL aggregates include Flickr and Wikimedia Commons photos. The inclusion of both types of sources provided an excellent window into the challenges that content curation communities face when blending content from sources that have different levels of credibility, as discussed later.

## Methods

We chose to perform a case study of EOL for several reasons. Case studies are ideal for understanding an entire system of action (Feagin, Orum, & Sjoberg, 1991), in this case, a content curation community. One of the goals of this research is to provide actionable insights to content curation community designers. A case study approach draws the boundaries of inquiry precisely around the matter (i.e., the system of action) that community designers can influence. Case studies are also ideal for describing emerging phenomena, such as content curation communities, where current perspectives have yet to have little empirical substantiation (Yin, 2008). Finally, case studies offer a way of investigating a phenomenon in depth, and within its real-life context, lending high external validity to the findings. EOL has several characteristics that support the choice of case study research: (a) it is a prototypical example of a content curation community; (b) it is a high-impact project with an international reputation; (c) it has sufficient history and is large enough to have encountered many of the potential challenges typical of these types of communities; and (d) it is still evolving so that some of the insights from this study can be incorporated into EOL's design.

DRAFT

***Data Collection***

EOL conducted a survey in July and August of 2010. The survey was sent to EOL's scientist participants (contributors, database partners, content curators) in order to obtain information to guide the next phase of EOL's development. Survey questions asked about participants' field of expertise, opinions about EOL and related efforts, involvement in various activities and initiatives, technological aspects of the site, and more. Both open-ended and closed-ended questions were used. A link to the survey was sent by email to 1225 participants; 281 responded and completed the survey. Typical time to completion was 15-20 minutes. Response rates varied by question, the overall completion rate for participants who started the survey, was 85% and for questions that were used in this study response rate varied from 18% to 86%.

Five interviews were conducted with EOL curators to better understand the role of content curators, the nature of the work they perform, their perceptions of EOL, their motivations for participating, and the challenges and opportunities they and others at EOL face. A semi-structured interview approach was used to allow for follow up questions that encouraged further elaboration of important findings. Additionally, 7 interviews were based on a convenience sample (Patton, 1990) of volunteers who contribute to Wikipedia species pages and are familiar with EOL. These participants were contacted based on their levels of participation on Wikipedia's Species Pages, which are displayed on EOL. Interviews were conducted face-to-face, via Skype, and via email and detailed notes and transcriptions were captured for analysis.

Additional data sources were used. One author took detailed notes during the EOL 2010 Rubenstein Fellows and Mentors Orientation Workshop. The workshop brought together nearly all 2010 fellowship recipients and their mentors. The workshop included a round-table discussion on expanding the utility of EOL for scientists, which elicited information from participants about their practical needs and their views of the project. The research team also reviewed the nearly 200 posts sent to the EOL curators' Google Group since its inception in May 2009 until March 17, 2010. These messages helped us to develop our interview protocol and identify salient issues to the community that are used as evidence throughout the study.

DRAFT

### *Data Analysis*

Notes and transcripts from interviews, workshop observations, open-ended survey questions, and online forum posts were analyzed using a Grounded Theory approach (Corbin & Strauss 2007). Grounded Theory calls for the systematic analysis of data so that common concepts and ideas are extracted and then axially referenced to produce higher-level themes and concepts that frame the theoretical understanding of the researched phenomenon. In the spirit of Grounded Theory we have tried to begin "as close as possible to the ideal of no theory under consideration and no hypothesis to test" (Eisenhardt, 1989). This was particularly important for this project, which is a first look at a content curation community. Once the major themes were identified and analyzed, we were able to relate them to existing theories and literature in the findings and discussion. All data were reviewed by at least two different authors to reduce personal biases.

## Findings

Our first research question asked: "What are the unique informatio and social challenges experienced by EOL as an example of a content curation community?" We found that the primary challenges faced by EOL relate to integration: both integration of information from disparate sources and integration of social practices and norms from different communities. Although information integration and social integration are highly intertwined, we discuss them separately below.

### *Information integration challenges*

Content curation communities deal with large corpora of content from diverse sources, which must be selected, organized, managed, and integrated into a holistic resource. This is no small feat, given the complexity of dealing with different taxonomies, meta-data standards, licensing, and competing incentives for data sharing. Ideally, content curation communities will create a one-stop shop, in which this underlying complexity is transparent to users, while allowing them to easily access the information they seek. This requires content curators to address several information integration challenges, two of which were identified as particularly important by EOL participants: (1) dealing with multiple and at times

DRAFT

competing classifications, and (2) ensuring information quality. In this section we discuss these challenges and their effect on the content curation community.

*Classification challenges*

A dominant theme that emerged from all of our data sources was the importance of biological classification in the work that the EOL content curation community performs. Many challenges inherent in the curation work of EOL relate to dealing with different classifications and their highly contested nature. Biological classification – the process of describing, naming, and classifying species into stable hierarchies – provides frameworks for all biological research. Indeed, the current body of taxonomic information and practice represents 250 years of research, starting with Linnaeus in the 18th century (Tautz, Arctander, Minelli, Thomas, & Vogler, 2003) . When asked to describe their professional role in an open-ended survey question, 6% of respondents used the term taxonomist and 43% used the related term "systematics", which is a well-defined field of study that refines and applies taxonomic theory and practice as a means to further the study of biological diversity[1]. Likewise, although only one of the EOL curator interviewees identified himself as a "senior taxonomist", all of them work closely with biological classifications as part of their everyday jobs. The frequency and passion associated with comments about how EOL deals with biological classification (as detailed below) suggest that not only is biological classification important to the field of biology, but it is also seen as fundamental to the work occurring at EOL.

Classification challenges at EOL emanate from debates within the greater biology community, as well as usability decisions made by EOL. Although we discuss these issues separately, they are inter-related since the broader classification debates in biology are reified in the ways that EOL presents its aggregated content. Disputes on how to classify species are woven through the history of taxonomy and evolutionary biology. Classification faces persistent challenges in the form of deep knowledge gaps in how much is unknown about biodiversity, arguments over nomenclature and species' names, as well as the extent to which species should be grouped or separated (Wheeler, 2008). In addition, every year new species are discovered as well as go extinct, constantly changing the scope and hierarchy of classification (Gonzalez-Oreja, 2008). Furthermore, existing work on taxonomy is sporadic, with much of

DRAFT

the research focused on specific types of organisms, while others have been largely ignored (Tautz et al. 2003).. One result of these long-standing issues is the existence of competing classifications within the biological sciences, each of which is recognized as imperfect. Technology for supporting a consensus has not yet solved the biological and sociological challenges of this work (Hine, 2008).

The debates from the greater biological community come to a head at EOL because the work occurring at EOL requires that there be a uniform type of classification chosen to organize the effort and the information in the community. One curator described how she encountered "conflicts in higher taxonomies" including one about a particular species that prompted someone to be "really mad about the way it was done and who was doing it." Other interviewees and discussions in the forums agreed with the sentiment that "in taxonomy there are huge disputes." These strong feelings emerge because taxonomists and species specialists grow accustomed to using a preferred schema. EOL, as an aggregator of content based on different classifications, is unavoidably thrown into the middle of such debates, causing tensions within the content curators' community.

**Having only one biological classification option as a point of contention**. The tensions in the general biology community permeate EOL. Early in its history, all EOL species pages were organized using the Catalogue of Life (COL) annual checklist and classification, the product of global partnerships. Many content curators preferred to use familiar classifications and did not like being forced into using COL, even though it is considered one of the most comprehensive species catalogues available (Orrell, 2011). When asked which classification they preferred, survey respondents named multiple systems (e.g. CAS, COF, COL, EOL, Global Checklist, ITIS, TOLWeb, WoRMS); no single classification system was preferred by the majority of respondents. Early Google Group discussions clarified the problems associated with having a rigid classification scheme as the basis for which pages exist; one curator explained how EOL "species pages [are] based on nomenclatorial lists derived second- or third-hand from various other databases… [and] … species names that are nomenclatorially available are not necessarily equivalent to species taxa as currently conceptualized by biologists… this nomenclatorial chaos has made it extremely difficult to curate." One survey respondent described how "the [choice of] taxonomy on the main part of EOL is a big turn off for recruiting the group that I am targeting."

DRAFT

In October 2009 the EOL informatics team responded to initial curator complaints about COL and added a function that allows content curators to choose the default taxonomic classification by which they could navigate the site. Figure 1 shows the current taxonomic drop- down menu that allows curators as well as EOL site users to view a given species page in the context of their choice of classification.

[Insert Figure 2 here]

**Obstacles for creating master taxonomic classifications.** Participants expressed widely varying opinions about being able to manipulate specific classifications and integrate new self-designated classifications on EOL. One survey question asked users of a particular tool (LifeDesk), that is used to create a modified or new classification on EOL; of all the participants who answered the question only a third have used this tool for creating new classifications, but the vast majority of them (80%) indicated that they would like to have their classification integrated into EOL, and a smaller majority (56%) would have liked to submit their classifications to other large-scale catalogues as well. However, in practice, most users do not modify or create new classifications. The main reasons for this have to do with usability:"[it is] difficult to determine how to add or change taxa previously listed in classification"; "I'm not really familiar with the features"; and effort:"[I don't have the] time".

*Quality control*

Quality control related challenges faced by EOL arise because it aggregates content from a variety of sources, including both traditional scientific content and non-traditional user-generated content. The decision to incorporate data from non-traditional sources including Wikipedia and Flickr was meant to help increase species' coverage, while taking steps to maintain the validity of the data. As EOL imports Wikipedia articles on species the content is initially marked as "non-reviewed" material. Curators can then approve the content, which removes the markings. The specific version of the Wikipedia article that was imported and approved remains unchanged even if the original Wikipedia article is modified. Content curators are encouraged to fix any problems with the page on Wikipedia itself before approving the text, but cannot make changes to the EOL version of the content. This is similar to the recommendation that EOL curators should submit problems they identify to scientific publishers and information providers

DRAFT

rather than changing them on EOL. The difference is that Wikipedia allows them to directly make their own edits. Images of species are offered to EOL by members of the EOL Images Flickr group[2] which, as of this writing, includes over 2,000 members and 70,000 images. These are automatically harvested and placed on EOL pages, initially with the same marking, suggesting that they have not yet been reviewed. As they are reviewed, they are then trusted or not trusted by content curators on EOL.

**Most participants support the integration of content from non-traditional sources (e.g., Wikipedia, Flickr), but would like this content vetted and/or clearly identified.** Though an arguably controversial aspect of EOL, many participants in both the interviews and survey shared positive feelings regarding EOL's aggregation of non-traditional sources. Survey participants asked about integration of such sources responded: "It's a good start", "Many of them are [a] good source of information", "Fine by me", and, "Many hands make light work – can't expect all the pages to be populated by experts." However, even the most enthusiastic respondents expressed a desire to assure that the content was vetted and/or clearly identified as a non-traditional source on EOL. One interviewee captured this sentiment by stating that integrating non-traditional content was a "good idea as long as you keep things apart. The most important thing is that the end user needs to know where stuff comes from and as long as that is given, it is a really good thing to mix specialist content with amateur content." A large group of survey respondents posted similar thoughts including, "I do not see any problem with that in [the] case that the source is clearly shown", "as long as it is clearly marked up as unvetted". Several content curators stressed the need to allow them to closely monitor non-traditional resources: "It is often wrong – curators should be able to 'exclude' content or mark it as 'erroneous'", and 'Fine as long as it is clearly marked as unvetted'".

**Some participants strongly oppose the use of non-traditional content from sources such as Wikipedia.** A small but vocal minority expressed strong opposition to the inclusion of non-traditional data. They used harsh words such as "contamination", "muddling", and "vandalism" to express their concerns. This opposition primarily stems from integrating what participants view as lower quality content. One interviewee stated, "It may be better to have no information about a taxon available on the web than

DRAFT

information that has not been vetted and may be incorrect." Some survey participants expressed an even stronger disapproval when asked their opinion of integrating non-traditional content from sources like Wikipedia and Flickr using the following language: "not acceptable", "many sources are full of errors", "should not be allowed. Too many reams of spurious data", "unvetted content continues to perpetuate errors that slip through the scientific peer review process that have been passed into the public domain", and - "a lot of what I have seen is either old or of poor quality". While these opinions were about non-peer reviewed, non-traditional content in general, some survey participants went so far as to single out Wikipedia and Flickr as undesirable content providers. For example, "in several cases, typos in Flickr contributions have resulted in 'phantom' species being added to EOL", "Wikipedia, I'm not so sure about… Wikipedia has lots of mistakes", and "not a fan of Flickr material."

A different set of respondents was concerned that using Wikipedia content in EOL makes the distinction between the two, and the role of EOL as a scientifically trusted resource, less clear. One interviewee stated, "If EOL = Wikipedia, then why have EOL? A worthy question." Another said, "I suspect that many professionals will get frustrated with the proliferation of unvetted material and lose interest in contributing, and that as a result EOL will become so much like Wikipedia that it will lose its relevance, and the point of EOL was to have vetted content that was checked by specialists. If this is the route to be taken then why would anyone come to EOL instead of just going to Wikipedia?", and "don't think this is a good idea… what would then make EOL different from Wikipedia?",

**There is no consensus about where the work of fixing errors should occur.** The EOL content curation community, by its nature, faces numerous quality control challenges that stem from the variance in accuracy, terminology and quality of the curated resources, whether these resources are traditional (scientific) or non-traditional (user-generated). Content curators may find classification mistakes embedded in both types of resources that are pulled into EOL, but traditional resources are often slow to make changes and fix errors due to their hierarchical and measured work processes, leading to frustration on part of the content curators: ("EOL suggests I communicate with the original authors of the source database… where nobody is currently funded to update or repair it"), yet the alternative of fixing errors directly on EOL is problematic as well, as the incorrect information persists in the original source.

DRAFT

Unlike traditional resources that necessitate pre-approval for changes and fixing errors, on Wikipedia curators can directly fix errors without having to wait for editorial approval. Some curators like being able to directly edit Wikipedia pages before importing them to EOL: "[EOL aggregating Wikipedia content] is in some ways easier because you can go in and edit on Wikipedia much easier than on EOL"; "when I heard [EOL] was going to [aggregate Wikipedia] I thought, 'Oh, great, I'll make a lot of changes to Wikipedia… and have them load into EOL instead of contributing to EOL, this is easier". A potential problem is having multiple versions of a Wikipedia page, one inside EOL and one outside EOL, or in Wikipedia itself, which could "divide contributors and decrease the quality of content."

Despite the positive reaction of some participants to making edits on Wikipedia, those with a low opinion of the quality of Wikipedia content did not like being forced to make edits there. For example, one interviewee noted that, "Wikipedia is full of errors… I never have time to go clean them all up, there's always errors… it's like a full time job and nobody has full time to do it." Others were less condemning of Wikipedia, but still recognized that it is "Hard to keep up with the volume" of data from there and it is a "big job to vet it all" let alone fix it.

### ***Social integration challenges***

EOL, as a content curation community, brings together two distinct populations – professional scientists and citizen scientists. Each population brings to the table its own motivations, practices, accreditation procedures and norms (Hara, Solomon, Kim, & Sonnenwald, 2003; Ling, et al., 2005; O'Hara, 2008; Van House, 2002a). Amalgamating the differences between the two populations into an efficient work process is a major challenge. This section details the various obstacles to social integration: (1) the role of open-source resources in the scientific community, and (2) conflict among scientists and citizen scientists.

#### *The Role of content curation communities within the larger scientific community*

Despite endorsements from prominent scientists such as Wilson (2003), and EOL providing the esteemed and competitive Rubenstein Fellowships, the larger scientific community has not yet explicitly legitimized curation work that occurs at EOL. This happens for a number of reasons and has multiple implications for EOL and other content curation communities in the scientific realm. Below we address

DRAFT

three challenges associated with this issue including (1) a lack of familiarity with the project, (2) lack of legitimacy of curation work within the larger scientific community, and (3) lack of resources to support the effort.

**Lack of familiarity with EOL.** Many scientists do not know about EOL despite its growing popularity among Internet users; they do not have a clear mental model of how EOL relates to larger scientific endeavors, or know how they can effectively contribute to EOL. Unlike scientific journals and conferences, which have been around for centuries, content curation communities like EOL are a new and unfamiliar scientific genre. Scientists who were not familiar with EOL's aim were hesitant to participate because they did not know what was expected of them ("[I] was not aware of requirements / needs of the program"). Others were confused about the professional pre-requirements needed for an individual to be able to participate in the curation activities ("I doubt I am qualified").

**Lack of legitimacy of content curation work.** Another problem is that the scientific community does not yet consider work on EOL as legitimate scientific activity. The scientific community can be thought of as a constellation of communities of practice (Wenger, 1998); it is comprised of individuals belonging to one or more research disciplines and sub-disciplines (i.e. communities) that share a domain, a practice, and a sense of community. As communities of practice, scientific communities socially construct what constitutes a legitimate contribution to the community. Publishing in certain journals (and not others), presenting at certain conferences, reviewing papers, and receiving grants, are all well-established academic practices that are legitimate activities among most scientific communities. Scientists, especially junior scientists, must present a strong research track, situated within the discipline's major publication outlets and other legitimized activities, in order to advance (Latour & Woolgar, 1979).

While content curation communities like EOL contribute indirectly to advancing science through improved outreach and provision of information that can support science, the work of curating existing scientific knowledge is not as valued by scientific communities as creating novel scientific knowledge. This marginalizes the contributions of content curators who are torn between using their precious time on

DRAFT

traditional, legitimate activities, and curation activities that are not yet fully legitimized, but are highly valued by the curators themselves and other interested parties (e.g., the public). The result is that curation activities are not considered a priority among many scientists: “This does not count with respect to my university accounting – not publications, not grants, not teaching – so it is very hard to find time to do this sort of thing”, and are almost completely unsupported by the scientific community as a whole, as several curators commented: “it has been difficult to convince tenured scientists to contribute”; “A number of scientists in my field have expressed reservations about sharing content with EOL. So, there is not [sic] impetus from the community, and I have not pursued possible sharing of content due to the concerns that people have expressed”.

Recognizing the importance of individual attribution in scientific domains, EOL has documented each curator’s contributions. For example, following a link to “Who can curate this page?” on a species page, takes a reader to a list of content curators. Clicking on any of their names will provide information and links to all of their contributions within the site. This strategy is rewarding to some participants, “having your content represented on EOL is more rewarding because your name is associated with it, people can know what you’ve done”. However, it does not address the larger issue of the scientific community recognizing these curation activities as legitimate contributions to science, as would be evidenced by, for example, tenure committees counting the number and quality of curated species pages or improvements to the EOL taxonomic classification scheme.

**Lack of resources for content curation work.** Issues of funding, time availability, and legitimacy of curation work are all intertwined. Almost 40% of the survey participants who were asked about their reason for not actively curating EOL content mentioned time constrains as the major reason for their lack of activity, “I try to change things, [but] it’s like a full time job and nobody has full time to do it”. Many of them associated funding, recognition and time aspects: “Lack of time/funding…I have no research grants… volunteer work is a luxury”; “Time is my biggest limitation. I've thought of using LifeDesk for the fauna of my region but need some support to do this – hint, hint.” Without formal recognition by the scientific community of open-access content curation efforts, significant sustainable funding will be lacking. This will keep curation activities in the realm of “volunteer work” rather than

DRAFT

"scientific work," which will in turn make it harder to receive certain types of funding. For now, the opportunity cost of the time spent curating at EOL is borne by the content curators themselves, which limits participation.

*Professionals and non-professionals: collaboration and conflict*

A third social integration challenge, which EOL and other volunteer curation sites face, is the need to bridge between different contributor populations. Scientists and citizen scientists participate in scientific endeavors for different reasons: citizen scientists participate in collective scientific activities out of curiosity, love of nature and commitment to conservation and related educational efforts (Evans, Abrams, Reitsma, Roux, Salmonsen, & Marra 2005; Raddick, et al., 2010), while professional scientists contribute to the formal processes and structure of "the scientific method" in order to advance science and further their own professional career (Cohn, 2008; Latour & Woolgar, 1979). Designing a community that appeals to diverse populations can be challenging, especially when the community operates within professional settings, although it also affords opportunities to leverage the differences to accomplish more than could be accomplished with only one of the groups.

EOL curators in the study held highly divergent views regarding the role citizen scientists can and should play in curation activities at EOL, and their willingness to work with them. The main reason that scientists were reluctant to collaborate with citizen scientists was their fear that citizen scientists do not uphold the same rigorous scientific standards that professional scientists are committed to: "I think that the benefit/cost ratio of participation by non-experts shifts in groups that are popular with amateurs and untrained dilettantes."

For other scientists, collaborations with citizen scientists or non-professionals were perceived as a net gain: "[experts and non-experts] contributing to the same resource pages… can be both useful and potentially damaging, though I haven't seen any of the later," and "[I] really like working with the Flickr community, I like the images, [it is] always fun working with people and seeing their different approaches." However, even those who supported collaboration with citizen scientists expressed a desire to distinguish clearly between trusted, professional scientists and untrusted non-professionals (e.g., "amateurs"; "untrained dilettantes"; "non-experts"). Scientists expressed nearly universal desire to "vet", "approve",

DRAFT

and “control” content submitted by citizen scientists, or at least have it be clearly differentiated from expert content (e.g., “content should be marked with different levels of trustability”). Scientists preferred to retain control of their expert opinions and privileged status that gives them the final word. This view positions citizen scientists at the bottom of the curation ladder and professional scientists at the top, maintaining the traditional power-balance between them, reducing some of the collaborative effect of content curation communities.

## Discussion - broader design issues for content curation communities

Our second research question asked: “What are the key information and social design choices available to designers of EOL?”

In this section we explore some design recommendations for addressing the challenges discussed above: information integration and social integration. Some of these recommendations defy neat classification as they apply to both themes, and are discussed jointly.

### *Information integration*

The ongoing debates regarding classification choices are a major challenge for information integration on EOL and other taxonomic content curation sites. Classifications and taxonomies provide a universal framework for scientists to communicate; they present a shared representation of a field of knowledge that provides the structure for understanding and collaboration. Biological classifications, such as those available in the EOL dropdown menus (Figure 1), can be viewed as boundary objects (Bowker & Star, 2000) or boundary infrastructures (Star, 2010). Boundary infrastructures, in their broadest sense, are items that create a shared space or “a sort of arrangement that allow different groups to work together without consensus” but within organic settings that require them to address “information and work requirements” (Star, 2010). In this case they are species classification systems that allow for communication between scientists with different subspecialties, languages, and resources. Content curation communities, are amenable to merging these different perspectives. But a careful approach is needed for choosing taxonomies that will serve as boundary infrastructures. We suggest the following options.

DRAFT

*Option 1: Choose or develop a master biological classification*

The initial approach was for EOL to select one master taxonomy to serve as a unifying boundary object; EOL decided on the Catalogue of Life (COL) as the default consensus classification. While this approach is the most parsimonious, it is not without problems. As indicated by the findings above, many participants were not pleased with this option. In addition, using EOL as a platform to foster a new master or universal classification schema would drastically shift the focus of the project, in terms of purpose and resource allocation. It also may not be stable and entirely consistent, as agreement on breakpoints may not be universal or enduring. Therefore, while this option may be relevant to other content curation communities that face classification challenges, it is not a plausible solution for EOL, unless it can be achieved in a fluid way that does not promise consistency or authority. This is reminiscent of the notion that these databases "do not sweep away the past but engage with it in dynamic fashion (Hine, 2008, p. 150), providing the basis on which later classifications will be constructed.

*Option 2: Support multiple taxonomic classifications*

An alternative option is to do what EOL currently does, and allow content curators and users to choose which taxonomic classification to use when viewing species pages. It is clear that this approach has helped to support content curators' preferences and allowed for additional taxonomic information to be pulled into EOL. However, it does not push the larger community toward the creation of a unifying boundary infrastructure, which could benefit the scientific community and enable more meaningful interdisciplinary exchanges. This option also necessitates extensive support for users, whether by providing detailed documentation, help and discussion pages, or by automating and streamlining the selection process.

*Option 3: Flexibility in determining the default classification*

Alternative options should also be explored. A healthy compromise between the two options would be to support multiple taxonomic classifications, but let the content curators determine the default classification for any given set of species. This has the benefit of helping novices who may not know

DRAFT

which classification is most appropriate to choose, as well as assisting in the development of a unified taxonomic classification. In addition, EOL could allow users and content curators to create and share lists of species (e.g. species in a particular ecosystem, species that are significant to users) in order supplement taxonomic classifications as meaningful ways to browse and explore EOL.

Noticeably absent from the EOL participants' dialogues around taxonomic classification was a recognition of the potential role of information professionals. This is not surprising given the "invisible" nature of information work more generally (Bates, 1999; Kate Ehrlich & Cash, 1999). Information professionals' extensive experience working with multiple classification systems, open access resources, content databases, and interfaces could be leveraged. EOL currently employs information professionals on the project, but may benefit from soliciting micro-contributions from academic science librarians and other information professionals who would join the ranks of EOL curators. These micro-contributions may include work related to classification, as well as other traditional information work such as verifying citations and writing species summaries (a familiar activity to those who have done abstracting work).

***Social integration***

Large-scale content curation communities such as EOL can only succeed if they find ways of facilitating effective collaboration among different groups of contributors, who bring significant domain-specific knowledge with them, whether they are professional credentialed scientists, volunteer ecology citizen scientists, or information professionals. We envision a successful model for a content curation community as one that is based on a hybrid structure, where professional scientists collaborate with citizen scientists, with the assistance of professional information specialists, all committed to the same idea of open-access content curation. To achieve this and overcome the social integration issues discussed above and issues pertaining to quality control, which are intertwined with social integration, we suggest the following design recommendations: (1) establishing active subgroups within large content curation communities; (2) establishing the role of open-access resources in the scientific community; (3) facilitating appropriate task-routing, and (4) recognizing other contributors' efforts.

*Establishing active subgroups within large content curation communities*

DRAFT

The size and scope of content curation communities, such as EOL, make community-building efforts difficult. One challenge is that members have diverse interests and are associated with multiple communities of practice, each of which has its own shared domain of interest (e.g., insects, plant biology), mutual engagement (e.g., conferences, journals, relationships), and shared repertoire (e.g., terminology, values, ways of doing things). In its early stages of development, EOL has focused on creating its own community of practice related to curation of content. Now that it has grown sufficiently, we propose that it actively develops interconnected subgroups that mirror existing communities of practice outside of EOL (e.g. invasive species, birds, phylogenetics). Such subgroups could create master lists ("collections") of pages of shared interest and can discuss relevant topics. This can promote users' sense of community, emphasizing collaboration over individual data curation. By welcoming subgroups within its walls, EOL will not only be a place to send and receive content from, it will become a place to engage directly in the core practices of its members and subgroups.

Understanding the need to support the various practices and norms of its sub-communities, EOL is currently developing tools that will facilitate a federated network of sub-communities, which are planned to be rolled out in the coming months.

*Establishing the role of community-curated content in the scientific community*

Scientists are committed to the traditional standards that are set by their individual scientific communities. Currently, work on open-access data repositories is not highly valued in the scientific promotion process, or in acquiring funding for research. Several routes of action can help in changing this predicament:

- Active efforts to consolidate community curated data into scientific publications. This can be achieved by directing efforts to promote content curation communities and their products within the scientific community (in scientific journals and conferences). It can also be done structurally, by creating easy to use import tools that will enable scientists to pull curated data into their databases for continuous work (e.g. via the EOL API), or consolidate vetted content into scientific publications (e.g. citing the curated source in background material, text mining the community for new insights, or aggregating and repurposing analyzable data). For example, the ChemSpider

DRAFT

content curation community (Williams, 2011) is integrated with RSC journals and databases like PubMed, which use the curated content to facilitate new types of searches.

- Providing financial support for active content curators – through funding and recognition. EOL, for example, provides a fellowship program for budding scientists that provides financial support and mentorship for up to one year. EOL's Rubenstein Fellowships are prestigious positions, and can contribute to the scientists' credentials and future funding. Similar efforts can help in establishing the scientists' status within their respective scientific community, and simultaneously enhance the role of the content curation communities as reputable entities within the scientific establishment. Interestingly there is an emergence of a novel "biocurator" profession in molecular biology, (Sanderson, 2011), suggesting that the scientific world acknowledges the merits of such activities. However, developing and continuously supporting positions that uphold these activities requires a substantial financial investment that not all content curation communities have. This is especially true for budding content curation communities that arise out of community members' interests and are not backed up by well-established institutions;
- Endorsing curation work at academic institutions. Many activities performed by scholars such as reviewing articles and grants, serving on editorial boards, and engaging with local communities, are endorsed by academic institutions that expect their faculty members to engage in them even though they are not explicitly paid to do so. Recognizing contributions to content curation communities as a form of highly valued professional service, like journal reviewing, could go a long way towards motivating participation in curating tasks.

*Facilitating the process – getting the right task to the right people*

In all content curation communities, but especially in hybrid ones that are built upon collaboration between different user populations, a special emphasis should be put on facilitating a streamlined curation process. Part of what makes commons-based peer-production communities, like those who develop open source software and Wikipedia, successful is the ability of individuals to assign themselves to the tasks that they can complete most readily (Benkler, 2002). Nobody knows a person's interests and skills better than himself; furthermore, people are highly motivated to contribute when they feel that their

DRAFT

contribution is needed and unique (Ling, et al. 2005). What is needed than, are tools that facilitate the matching of individuals skills and interests with curation needs. The use of Watchlists (as on Wikipedia) that notify a user when changes are made to a page they are "watching" can help people stay abreast of changes, but doesn't point users to gaps in the existing content. Cosley et al. (2006) suggested bridging this gap in the Wikipedia community with their automatic task routing tool that uses past editing behavior to recommend new pages a user may want to edit. Similar approaches could be modified and applied to content curation communities. However, many other options exist, such as allowing individual content curators to post lists of micro-contributions or "wanted ads" that other contributors can monitor (i.e. subscribe to) or have sent to them based on a profile they have filled out describing their interests and level of expertise. Posting EOL needs to the EOL Flickr group has worked well for soliciting needed photos, but this has not yet been scaled up or automated for other photos or other types of content. Such routing processes can also recommend relatively simple tasks at first and, after an "initiation" period, advance to more complex and high-level tasks, such as vetting resources. This would also allow users to grow and progress, and work towards attaining a higher status within the content curation community. No matter the mechanism, helping potential contributors know what is needed and empowering them to accomplish those tasks is paramount to the success of content curation communities.

We have identified quality control, both of content and curation, as a crucial factor for the success of a content curation community. Creating a mechanism for task routing can also allow for better quality control: improved control and feedback of the work that is being done within the content curation community will rid the community of some of the inherent weariness of "dalliances" and "contamination" of resources, and hesitation of association with non-professionals. Such task-routing emulates the traditional scientific apprenticeship process, in which a novice is accepted into the scientific community through legitimate peripheral participation (Lave & Wenger, 1991). Through this process of initiation, citizen scientists and novice curators can become trustworthy and established partners in the content curation community. EOL expects to launch tools to support this process in the near future.

*Recognizing other contributors' efforts*

DRAFT

To enable social integration, users first have to acknowledge other users' presence and contributions to the community. On large-scale communities such as EOL this is an especially difficult task, due to the federated nature of the activity. Therefore explicit mechanisms are required to create such awareness. Portraying users and their contribution, by way of attributing the data to the users who curated it, on the front page of the community on a rotating basis, as well as on other designated pages, can increase other users' awareness of them, and heighten their status within the community (Kriplean, et al., 2008; Preece & Shneiderman, 2009). EOL currently uses "featured pages", as does Wikipedia, which could be augmented by "featured curators" or "featured contributor" pages to recognize important contributors, make clear to first-time visitors that it is a community-created site, and point to model content curators for others to emulate. Another form of attribution could be facilitated by creating a feedback mechanism that will notify users whether the data they contributed or curated was vetted and approved. Similarly, a feedback mechanism could also alert a user to the use and re-use of the data he contributed both within the community and by third parties such as when it is downloaded, viewed, or is cited in publications. By providing periodic personal reports of the number of data downloads, citations, and reuse, and sharing these statistics with the entire community, the contributor is openly appreciated and his social presence within the community is heightened. Scientists could include data from these reports on their curriculum vitae to help legitimize the work being done.

Other mechanisms that are needed for social integration are internal conflict resolution tools, such as Wikipedia "talk" pages (cf. Kittur & Kraut, 2008). The proposed federated sub-communities can provide a place to discuss disagreements and issues that stem from different disciplinary backgrounds and practices. In this context, dedicated and long-time contributors (both scientists and citizen scientists) may also be conferred the role of "mediators" in case topical and procedural conflicts arise. EOL is currently instituting a role of "master curators", who will take upon themselves some of the aforementioned social facilitation roles.

## Conclusion

In this paper we define a new genre of online community, "content curation communities", using the Encyclopedia of Life as a case study. We conducted a multi-method analysis, combining survey and

DRAFT

interview data with observations and analysis of online interaction. Our aim was to explore and understand the challenges faced by content curation communities, and to recommend potential design solutions. While our focus has been on EOL, we believe many of the challenges and design recommendations will apply to other content curation communities as well.

In the past, the contribution of citizen scientists, naturalists and enthusiasts to the development of science, and specifically astronomy, biology and geology, has been well documented (Ellis & Waterton, 2004; Schiff, Van House, & Butler, 1997; Van House, 2002b), but their participation was usually considered peripheral, and limited to one-sided data transfer, with little or no mutual knowledge exchange between scientists and enthusiasts. Content curation communities offer the opportunity to shift the balance towards more collaborative practices, where both scientists and citizen scientists work together to create a vetted and reputable repository that can be used for traditional scientific endeavors and for education beyond the scientific community.

Bringing together heterogeneous populations of contributors is challenging from both practical and social standpoints. Creative design solutions, such as those discussed above, may ameliorate these challenges.

EOL is larger and more ambitious than most content curation communities, yet it can be viewed as a typical and representative instance of the new genre of collaborative information resources. It is still evolving and growing, and in its growth it is currently facing the challenges outlined in this paper. By rethinking and redesigning some of the online tools available to content curators the community is attempting to face these challenges. This is done based not only on the lessons learned from the daily operation of EOL and from the findings of this study, but also based on feedback received from EOL curators. An approach that speaks to contributors' needs and allows their voice to be heard, may not only support the design changes of the content curation community, but also enhance contributors' sense of community and their willingness to further contribute to the community in the future. Some of these changes EOL is already implementing, and they will be rolled out in the foreseeable future and will be the basis for future studies of content curation communities.

It should be noted that the study is not devoid of limitations. The concept of content curation communities is a relatively new one, and while we have attempted to focus on the most pressing issues

DRAFT

that stem from the proliferation of such communities, we could not have scoped them all. Other types of content curation communities, within and outside the scientific realm, may face other challenges that were not addressed in this paper. Comparable examples of collaborations between scientists and citizen scientists in other disciplines (e.g. astronomy or chemistry) may be quite different from EOL. However, we believe the concept of content curation communities is enduring and a useful frame for starting a dialogue about related communities.

Future research on content curation communities should encompass other important issues, including: (1) motivational factors affecting professionals' and volunteers' participation in content curation communities; (2) the role of information professionals within content curation communities; (3) strategies for developing a strong sense of community within content curation communities; and (4) different types of content curation communities and design strategies that support them. Our future work will address these topics.

DRAFT

## Footnotes

[1] See: Society of Systematic Biologists - http://systbiol.org/
[2] http://www.flickr.com/groups/encyclopedia_of_life/

## References

Bates, M. J. (1999). The invisible substrate of information science. Journal of the American Society for Information Science and Technology, 50, 1043-1050.

Benkler, Y. (2002). Coase's Penguin, or, Linux and the nature of the firm. The Yale Law Journal, 112(3), 369-446.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: supporting sharing in science and engineering. Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work (GROUP '03) (pp. 339-348) New York, NY: ACM Press.

Borgman, C. L. (2007). Scholarship in the digital age: Information, infrastructure, and the Internet. Cambridge, MA: The MIT Press.

Bos, N., Zimmerman, A., Olson, J., Yew, J., Yerkie, J., Dahl, E., et al. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. Journal of Computer Mediated Communication, 12(2), 652-672.

Bowker, G.C., & Star, S.L. (2000). Sorting things out: classification and its consequences. Cambridge, MA: The MIT Press.

Buneman, P., Cheney, J., Tan, W.C., & Vansummeren, S. (2008). Curated databases. Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '08). (pp. 1-12). New York, NY: ACM Press.

Cohn, J. P. (2008). Citizen science: Can volunteers do real research? BioScience, 58(3), 192-197.

Corbin, J., & Strauss , A. C. (2007). Basics of Qualitative Research:Techniques and Procedures for Developing Grounded Theory. (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2006). Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In Grinter, R., T. Rodden, P. Aoki, E,

Cutrell, R. Jeffries & G. Olson (Eds.), Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI '06) (pp. 1037-1046). New York, NY: ACM Press.
Eisenhardt, K. M. (1989). Building theories from case study research. The Academy of Management Review, 14(4), 532-550.
Ellis, R., & Waterton, C. (2004). Environmental citizenship in the making: the participation of volunteer naturalists in UK biological recording and biodiversity policy. Science and public policy, 31(2), 95-105.
EOL. (2011). Retrieved June 12, 2011, from http://www.eol.org
Evans, C., Abrams, E., Reitsma, R., Roux, K., Salmonsen, L., & Marra, P. P. (2005). The Neighborhood Nestwatch Program: Participant Outcomes of a Citizen Science Ecological Research Project. Conservation Biology, 19(3), 589-594.
Feagin, J. R., Orum, A. M., & Sjoberg, G. (Eds.). (1991). A case for the case study. Chapel Hill, NC: The University of North Carolina
Finholt, T. A. (2002). Collaboratories. Annual review of information science and technology, 36(1), 73-107.
Finholt T. A. & Olson, G. (1997). From Labratories to Collaboratories: A New Organizational Form for Scientific Collabration. Psychological Science 8(1), 28-36.
Forte, A., & Bruckman, A. (2005). Why do people write for wikipedia? Incentives to contribute to open. content publishing. Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work (GROUP '05) (pp. 6-9). New York, NY: ACM Press.
Gonzalez-Oreja, J. A. (2008). The Encyclopedia of life vs. the brochure of life: exploring the relationships between the extinction of species and the inventory of life on Earth. Zootaxa, 1965, 61–68.
Hara, N., Solomon, P., Kim, S. L., & Sonnenwald, D. H. (2003). An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. Journal of the American Society for Information Science and Technology, 54(10), 952-965.
Hine, C. (2008). Systematics as Cyberscience: Computers, Change and Continuity in Science. Cambridge, MA: MIT Press.
Hughes, B. (2005). Metadata quality evaluation: Experience from the open language archives community. Digital Libraries: International Collaboration and Cross-Fertilization, Lecture Notes in Computer Science, 3334/2005, 135-148.
Kate Ehrlich, & Cash, D. (1999). The Invisible World of Intermediaries: A Cautionary Tale. Computer Supported Collaborative Work, 8(1-2), 147-167.
Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2006). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Alt.CHI at CHI 2007. Retrieved June 23, 2011 from http://www.viktoria.se/altchi/submissions/submission_edchi_1.pdf
Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08) (pp. 37-46). New York, NY: ACM Press.
Kriplean, T., Beschastnikh, I., & McDonald, D. W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08) (pp. 47-56). New York, NY: ACM Press.
Lampe, C., & Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. Proceedings of the 2004 SIGCHI conference on Human factors in computing systems (CHI '04) (pp. 543-550). New York, NY: ACM Press.
Latour, B., & Woolgar, S. (1979). Laboratory life: The construction of scientific facts. Princeton, NJ: Princeton Univ Press.
Lave, J., & Wenger, E. (1991). Situated Learning: legitimate peripheral participation. Cambridge, UK: Cambridge University Press.
Li, Q., Cheng, T., Wang, Y., & Bryant, S. H. (2010). PubChem as a public resource for drug discovery. Drug discovery today, 15, 23-24.
Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., et al. (2005). Using Social Psychology to Motivate Contributions to Online Communities. Journal of Computer-Mediated Communication, 10(4).
Maddison, D. R., Schulz, K. S., & Maddison, W. P. (2007). The tree of life web project. Zootaxa, 1668, 19-40.
Nov, O. (2007). What motivates wikipedians? Communications of the ACM, 50(11), 60-64.

DRAFT

O'Hara, K. (2008). Understanding geocaching practices and motivations. Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08) (pp. 1177-1186). New York, NY: ACM Press.

Orrell, T. (Producer). (2011) ITIS: The Integrated Taxonomic Information System (version Sep 2009). Species 2000 & ITIS Catalogue of Life, 3rd January 2011. retrieved from http://www.catalogueoflife.org/col, June 2011

Patton, M. Q. (1990). Qualitative evaluation and research methods. London, UK: Sage Publications, Inc.

Preece, J. (2000). Online Communities: Designing Usability, Supporting Sociability. Chichester, UK: John Wiley & Sons.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. AIS Transactions on Human-Computer Interaction, 1(1), 5.

Raddick, J., Lintott, C. J., Schawinski, K., Thomas, D., Nichol, R. C., Andreescu, D., et al. (2007). Galaxy zoo: an experiment in public science participation. Bulletin of the American Astronomical Society, *39*, 892.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., et al. (2010). Galaxy zoo: Exploring the motivations of citizen science volunteers. Astronomy Education Review*, 9*, 010103.

Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM*,* 40, 56-58.

Rotman, D., & Preece, J. (2010). The "WeTube" in YouTube – Creating an Online Community Through Video Sharing. International Journal of Web-based Communities, 6(2), 317-333.

Sanderson, K. (2011). Bioinformatics: Curation generation. Nature, 470(7333), 295-296.

Schiff, L. R., Van House, N. A., & Butler, M. H. (1997). Understanding complex information environments: a social analysis of watershed planning. Proceedings of the second ACM international conference on Digital libraries (DL '97) (pp. 161-168). New York, NY: ACM Press.

Star, S.L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. Science, Technology & Human Values, 35(5), 601-617.

Sullivan , B., Wooda, C. L., Iliffa, M. J., Bonneya, R. E., Finka, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10), 2282-2292.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. Trends in Ecology & Evolution, 18(2), 70-74.

Van House, N. A. (2002a). Digital libraries and practices of trust: networked biodiversity information. Social Epistemology, 16(1), 99-114.

Van House, N. A. (2002b). Trust and epistemic communities in biodiversity data sharing. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries (JCDL '02) (pp. 231-239). New York, NY: ACM Press.

Wenger, E. (1998). Communities of practice - Learning, meaning, and identity. Cambridge, MA: Cambridge University Press.

Wheeler, Q. (2008). The new taxonomy. New York, NY: CRC Press.

Williams, A. J. (2011). Chemspider: A Platform for Crowdsourced Collaboration to Curate Data Derived From Public Compound Databases. In S. Ekins, M. A. Z. Hupcey & A. J. Williams (Eds.), Collaborative Computational Technologies for Biomedical Research. Hoboken, NJ: John Wiley & Sons, Inc.

Wilson, E. O. (2003). The encyclopedia of life. Trends in Ecology & Evolution, 18(2), 77-80.

Yin, R. K. (2008). Case study research: Design and methods (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries, 7(1), 5-16.

DRAFT