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

Dana Rotman
College of Information Studies, University of Maryland, College Park College Park, Maryland 20742. E-mail: drotman@umd.edu

Kezia Procita
College of Information Studies, University of Maryland, College Park College Park, Maryland 20742. E-mail: kprocita@umd.edu

Derek Hansen
College of Information Studies, University of Maryland, College Park College Park, Maryland 20742. E-mail: dlhansen@umd.edu

Cynthia Sims Parr
National Museum of Natural History, Smithsonian Institution. Washington, DC 20013. E-mail: parrc@si.edu

Jennifer Preece
College of Information Studies, University of Maryland, College Park College Park, Maryland 20742. E-mail: preece@umd.edu

This article 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 article, 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 and validate multiple sources and (b) integrate traditional peer reviewed sources with user-generated, nonpeer-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.

Received July 8, 2011; revised December 14, 2011; accepted December 15, 2011

 DOI: 10.1002/asi.22633

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 helped 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, validate, and annotate *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 content creation and content curation communities share some aspects with organizational repositories, mainly the need to "shape" knowledge—whether created or curated—into a more useful format that will create a "dynamic memory system" (Yates, Wagner, & Majchrzak, 2010) that will serve the community members. As such, both communities fill important niches in the information landscape and have already proven their worth. Although 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, 2007; Kriplean, Beschastnikh, & McDonald, 2008; Nov, 2007), there have not been comparable studies of content curation communities. 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 article 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 (EOL) as a case study for providing insights into the design of such communities. Our key research questions are as follows:

- 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 content curation communities?

Content Curation Communities

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 collective effort. Although not necessary, they typically include volunteers and reflect the social roles and participation patterns that mirror those of other types of online communities (Preece, 2000). Much of the curation effort is done individually, but what separates content curation communities from personal curation is the coordinated curation work that pulls content into a mass repository.
- 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 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, validating, organizing, describing, maintaining, and preserving existing artifacts. This term follows the term "content" because the artifacts being curated by the community are content, not necessarily the objects themselves (e.g., rocks or animals aren't curated, but content about them is curated). 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.

Although there are various examples for content curation communities (see Table 1), in this article, we focus on a scientific content curation community, which builds 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 are shared via advances in information technology independent of time and location (Bos et al., 2007; Finholt, 2002). Collaboratories and their more recent instantiations led to an increase in the diversity of participants and participation levels (Birnholtz & Bietz, 2003; Bos et al., 2007; Finholt, 2002). Yet issues of data reuse, upholding scientific standards, and maintaining high-quality data persist (Zimmerman, 2007).

Bos and his colleagues (2007) identified seven distinct types of collaboratories, including "community data systems," where scientific data is created, maintained, and improved by a geographically distributed community. 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 are then vetted and annotated by paid professionals. Content curation communities are not primarily about collecting original data. For example, the EOL reverses the traditional community data system approach by having its volunteer content curators annotate and vet data that have 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.

DOI: 10.1002/asi

DOI: 10.1002/asi

TABLE 1. Comparison of various genres of content curation communities.

Domain/community	Curation purpose	Structure	Information integration	Social integration	Coordination mechanisms
Ecology—Encyclopedia of Life (EOL)	Aggregating information about every living species on earth. Focus on a broad community of users, ranging from the general public to academics.	Hierarchical structure–Council and executive committee in charge of setting general goals and policies, working groups in charge of day-to-day work, and volunteer curators who are in charge of the data.	Pulling data from various resources (open access and scientific resources). Different data formats. Data needs validation and quality control. Use of different classification schemes.	Interaction with various groups, organizations and individuals coming from different domains–a need to establish working processes and trust between them. Recognition of the value of content curation within the scientific community.	Detailed FAQ, mailing list, meetings in person with some curators.
Language and information research—Open language archives (OLAC); (www.languagearchives.org)	Harvesting linguistic resources, indexing, archiving and standardizing for use by linguists and others. Focus on a broad community of users.	Hierarchical structure–Governing advisory board of respected peers. Council–approved by the advisory board, making decisions for the community; Working groups–ad hoc associations that draft standards, notes, and documentation (that are approved by the council). Individuals volunteer to participate in the community efforts–can advance to governing roles in time.	Pulling data from numerous (sometimes conflicting) resources. Standardization process of data, resources, creating an accepted controlled vocabulary. Metadata standardization has to be approved by the governing advisory and council–long process, sometimes dependent on individuals coming from different domains with different viewpoints.	Interaction with various groups, organizations and individuals coming from different domains. Very centralized organizational structure. Task regulation and division of labor are not always clear.	Web portal for the community, extensive documentation, Q&A, directory of community members.
Sciences ChemSpider (www.chemspider.com/)	Database for curating chemical structures, physical properties, images, spectra, and other related information. Connected to many other databases. Focus on a broad community of users–students, academics, industry.	Semihierarchical structure–Governed by the Royal Chemistry Society (UK). Users can comment on others' actions; registered users can curate names, links, spectra, and other resources. Master curators validate and approve the data. Organizations (university departments, research groups) can deposit data. Prerequisites–registration and basic knowledge of chemistry.	Pulling data from various resources (open access resources–Wikipedia; open and private repositories). Not all data is vetted and it requires validation according to scientific standards. Links to chemical vendors' catalogues raise issues of commercialism. No controlled vocabulary or semantic markup.	Users need to formally apply to attain certain curation roles–privileges and access to curation are granted based on credentials. Some users do not have formal credentials and thus are barred from playing a more active role in the community. Emphasis on building intra-community reputation through microcontributions.	Discussion forum and pages for data-integration issues. A formal blog published by the RCS.
Literature Project Gutenberg (www.gutenberg.org/)	Digitizing copyright-free and public-domain books. Focus on public access to cultural works.	Purposefully decentralized–community members decide independently what to curate. "Posting team" cleans and approves works for publication.	No comprehensive list of cultural works in need of publication. Decentralized structure and no "selection policy" lead to cases where numerous volunteers are unknowingly working on the same book/cultural piece. Proofreading standards are upheld, yet the project but metadata is not always available.	Little attribution. In most cases there is no attribution to members who cleaned and validated the data, just to volunteers who digitized (i.e., produced) it.	Mailing list, private emails, extensive FAQ documentation and newsletters.

Genealogy GenWeb (usgenweb.org)	Providing resources for genealogy research for every county in the United States. Focus on individual genealogy researchers.	Completely decentralized–volunteers "adopt" one (or more) county and produce all the data/links for that area. Coordinating team and project coordinators are in charge of each state, as well as linking individual resources to the general website and resolving conflicts, but have minimal control over content. No clear process for becoming a member of the coordinating team.	Varying levels of completeness and accuracy, dependent on the individual volunteer. Federated repositories do not allow for comprehensive search of databases. Different data format, no controlled vocabulary.	Mostly volunteers act individually with little interaction with others or with the coordinating team.	Mailing lists and private e-mails. Grievance committee resolves conflicts between members.
Crafts Ravelry (www.ravelry.com)	Curating and sharing yarn-based crafts, projects, ideas and tools. Focus on a broad community of crafters.	Emphasizes individual curation and social networking. Each member creates their content and can deposit data into a collective database. Other members can comment. Annotation is done only by members who have joined an editing group and for data that is open to editing. Removal of data is only done by select editors.	Data validation. No controlled vocabulary, standardization, or regulation regarding annotation or contribution of data makes search difficult.	Social integration is done mostly on social networking tools and less on databased components of the community.	Wiki, forums, subgroups, social networking (comments, friends lists). Specific discussion forums for data editing and cleaning purposes.
Web Indexing Open Web Directory (www.dmoz.org)	Bookmarking, cataloguing, and curation of web pages. Focus on curation for the greater good.	Hierarchical structure: editors have to apply for a position. Senior editors approve junior editors applications, curated content and editorial suggestions. Curation is done on all levels but has to be vetted. Admins control all editing privileges and content. Long serving senior editors can become admins through recommendation from other admins.	Data are structured ontologically, based on existing formal cataloguing principles (usnet inspired). Changes to the structure made through discussion among the senior editors until reaching consensus. No annotation tools available. Different vocabularies for general content and adult content. Difficulty maintaining quality content and removing spam.	The community promoted self-regulation; therefore no strict policies for social integration were publicly published. Allegations of blacklisting editors and content, personal conflicts and lack of control over removal of editing privileges cause fragmentation in the community.	Discussion forums open only to editors. Although the site has been one of the first examples for content curation (established in 1998), it lost some of its appeal due to the rise of wikis and folksonomies, and limited its work beginning in 2007.

In addition to being an influence on collaboratories, scientific content curation communities are also a subset of the growing trend of crowdsourcing scientific work to interested enthusiasts and volunteers. Crowdsourced projects can be found across academic domains, from astronomy (Galaxy Zoo, Cerberus) to biochemistry (FoldIt, ChemSpider), geography and climatology (Old Weather Tracking), history (Civil War Diaries and Letters transcription project, FamilySearch, The Great War archive), archeology (Ancient Lives), and even peer review (IMPROVER). The premise of crowdsourced projects is that engaging a large number of users in solving complex problems that require many resources (time, effort, expertise, or computing powers) will facilitate quick and potentially novel findings. The website scistarter.com lists more than 400 crowdsourced scientific projects, across different domains. From this list we can learn that crowdsourced projects are more prevalent in domains that are labor intensive and where large-scale questions can be broken down to smaller, manageable tasks. Most crowdsourced scientific projects focus on generating new data rather than curating existing data. Content curation communities are less popular than scientific crowdsourcing projects for various reasons that will be outlined in the article.

Content curation communities have also been influenced by more generic social media initiatives and tools. As web content has proliferated, the need to curate general content, and 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 to filter out the low-quality discussions about news (Lampe & Resnick, 2004). Social sharing sites like Diigo, Flickr, and YouTube also allow popular content to rise to the top (e.g., Rotman & Preece, 2010). Although 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 article is to learn from strategies employed by existing social media tools to effectively aid in negotiating the challenges faced by content curation communities.

EOL

> 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, p. 77)

The EOL 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). EOL is an open-access aggregation portal (www.eol.org). One of its bold goals is to engage individuals, including scientists and nonscientists, on contributing to scientific information curation, so that they will "study, protect and prolong the existence of Earth's precious species" (EOL, 2011). EOL is structured such that each species is presented on a dedicated page, where all available and relevant information, vetted and nonvetted alike, is presented. Currently, there are 3 million pages, representing 1.9 million known species on Earth. Most of these are stubs, but 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 to the site. 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, 200 of whom are currently active. 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's curation work process brings together automatic tools and human input: EOL automatically aggregates objects such as images or text onto publicly available pages based on metadata from various content sources. Curators are granted access to controls, allowing them to "trust," "untrust," and "hide" objects such as images or text. EOL curators cannot change the objects but they can leave comments and make a trust/untrust/hide decision. The system shares these comments and curator actions with the original content providers who may choose to make a correction and resubmit their data. Curators can rate objects on EOL, based on their comprehensiveness and quality. EOL uses this rating to sort the objects on the page so the highest rated objects show first. To a lesser extent, curators are also encouraged to be contributors by adding their own text to pages or improving third-party content (e.g., editing Wikipedia pages) before importing them into EOL. Although there is no formal training and accreditation process for curators, they are expected to learn about EOL's policies and tools by reading documentation on EOL's website or by following discussions in an opt-in mailing list.

Curator activity is only loosely monitored by EOL staff and other curators, and social conflicts, as discussed later, are usually resolved through backchannels such as personal e-mails. To retain established curators, EOL emphasizes personal feedback and recognition through e-mails but also public recognition through highlighting curators' work on their personal EOL pages. Curators are also invited to test

DOI: 10.1002/asi

new features and provide input into the design of future tools. EOL predecessors include sites that provide freely available, biodiversity-focused, and comprehensive, species coverage, such as 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 preapproved users from the professional science community directly add information (Maddison, Schulz, & Maddison, 2007).

EOL derives its content from scientific content partners, including many of its predecessors, but also from user-generated content: In December of 2009, EOL began a content partnership with Wikipedia, importing Wikipedia species pages into EOL. Although the addition of these articles increased EOL's species coverage considerably, it was controversial because of the nontraditional authoring process of Wikipedia that allows anyone to contribute content. Other examples of nonauthoritative 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 not yet been empirically substantiated (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 in an important scientific domain; (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.

Data Collection

EOL conducted a survey in July and August of 2010. The survey was sent to all of EOL's scientist participants (contributors, database partners, content curators) to obtain information to guide the next phase of EOL's development. Participants weren't offered any incentive for their participation in the survey. Considering the lack of reward, participants' time-demands and heterogeneity, the survey was designed such that the questions were mostly open-ended and each question was independent of other questions, so that each participant could choose to answer only the questions that were relevant to him, and skip questions that he found irrelevant. 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. A link to the survey was sent by e-mail to 1,225 participants; 281 (23%) responded and completed the survey. Typical time to completion was 15–20 minutes. Because of the survey structure, response rates varied by question, 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, and the challenges and opportunities they and others at EOL face. These interviewees were contacted based on their active involvement in EOL. A semistructured interview approach was used to allow for follow-up questions that encouraged further elaboration of important findings. Additionally, seven 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 e-mail 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, and included a roundtable 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.

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

DOI: 10.1002/asi

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, p. 536). 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 information 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 types of data (e.g., photographs, text, video, audio), 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. This requires content curators to address several information integration challenges, two of which were identified as particularly important by EOL participants: (a) dealing with multiple and at times competing standards and (b) ensuring information quality. In this section, we discuss these challenges and their effect on the content curation community.

Standards and 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 standards of classification 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, nearly half (49%) of respondents used the term taxonomist or systematics, both of which refer to 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.

Standards and 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 the research focused on specific types of organisms, while others have been largely ignored (Tautz et al. 2003). One result of these longstanding 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 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 standard 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." Such 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. 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,

[1]See Society of Systematic Biologists: http://systbiol.org/

DOI: 10.1002/asi

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 standard as the basis for which pages exist; a surveyed 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."

Obstacles for creating master taxonomic classifications. 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. 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. According to survey participants, 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" and "I'm not really familiar with the features"—and effort—"[I don't have the] time."

Quality Control

Quality control challenges faced by EOL arise because it aggregates content from a variety of sources, including both traditional scientific content and nontraditional user-generated content. The decision to incorporate data from nontraditional sources including Wikipedia and Flickr was meant to help increase species' coverage, while taking steps to maintain the validity of the data. Wikipedia articles imported into EOL are initially marked as "nonreviewed" material. Curators can 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 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 2,500 members and 90,000 images. These are automatically harvested using user-generated tags 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 nontraditional sources (e.g., Wikipedia, Flickr), but would like this content vetted and/or clearly identified. Though an arguably controversial aspect of EOL, many survey respondents and interviewees shared positive feelings regarding EOL's aggregation of nontraditional sources. When surveyed about integration of such sources, curators 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 nontraditional source on EOL. One interviewee captured this sentiment by stating that integrating nontraditional 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: "I do not see any problem with that in [the] case that the source is clearly shown" and "As long as it is clearly marked up as unvetted." Several other curators stressed the need to allow them to closely monitor nontraditional 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 nontraditional content from sources such as Wikipedia. A small but vocal minority of survey participants expressed strong opposition to the inclusion of nontraditional 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 information that has not been vetted and may be incorrect." Some survey participants expressed an even stronger disapproval when asked their opinion of integrating nontraditional content from sources like Wikipedia and Flickr using

[2]http://www.flickr.com/groups/encyclopedia_of_life/

DOI: 10.1002/asi

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." Although these opinions were about nonpeer-reviewed, nontraditional 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 another: "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 nontraditional (user-generated). Traditional resources are often slow to make changes and fix errors because of their hierarchical and measured work processes, leading to frustration on the part of the content curators who were surveyed: "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. Unlike traditional resources that necessitate preapproval for changes and fixing errors, on Wikipedia curators can directly fix errors without having to wait for editorial approval. Some curators who were interviewed liked 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" and "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: (a) the role of open-source resources in the scientific community and (b) 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 three challenges associated with this issue including (a) a lack of familiarity with the project, (b) lack of legitimacy of curation work within the larger scientific community, and (c) 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. As was evident from the survey, 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 prerequirements 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

DOI: 10.1002/asi

scientific community can be thought of as a constellation of communities of practice (Wenger, 1998); it comprises individuals belonging to one or more research disciplines and subdisciplines (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, to advance (Latour & Woolgar, 1979).

Although 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 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 who were surveyed: "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." Curation activities are also almost completely unsupported by the scientific community as a whole, as several curators commented: "iI has been difficult to convince tenured scientists to contribute" and "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. As one interviewee noted, 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 associated funding, recognition, and time aspects: "Lack of time/funding . . . I have no research grants . . . volunteer work is a luxury" and "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 "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 nonprofessionals: Collaboration and conflict. A different social integration challenge, which EOL faces, 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 both egoistic and altruistic motivations, such curiosity, sense of community, and commitment to conservation and related educational efforts (Evans et al., 2005; Raddick et al., 2010; Rotman et al., 2012), while professional scientists contribute to the formal processes and structure of "the scientific method" to advance science and further their own professional career (Cohn, 2008; Latour & Woolgar, 1979). Designing a community that appeals to these distinct populations can be challenging, 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 nonexperts shifts in groups that are popular with amateurs and untrained dilettantes."

For other scientists who were interviewed, collaborations with citizen scientists or nonprofessionals were perceived as a net gain: "[Experts and nonexperts] 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 nonprofessionals (e.g., "amateurs," "untrained dilettantes," "nonexperts"). Surveyed scientists expressed nearly universal desire to "vet," "approve," and "control" content submitted by citizen scientists, or at least

DOI: 10.1002/asi

have it be clearly differentiated from expert content (e.g., "content should be marked with different levels of trustability"). They 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 andreducing some of the collaborative effect of content curation communities.

Discussion—Facing the Challenges That Content Curation Communities Present

Content curation communities flourish in various fields. From science to marketing to hobbies, they serve users and communities who find content aggregation, annotation, and archiving valuable. They may differ in the type of content they curate (images, texts, preformed data, etc.), in their use (personal or collective) and in their structure (centralized and hierarchal vs. decentralized and flat), but, surprisingly, many of them share similar challenges. Table 1 provides examples of content curation communities from different domains, and uses the "informational" and "social" challenges framework identified in EOL to comment on them. Although detailed analysis of each example was beyond the scope of this article, the table provides insights into the commonalities and differences among content curation communities.

When comparing the various examples of content curation communities with EOL (Table 1), we can clearly see that most similarities can be found between EOL and content curation communities in the academic domain, such as chemistry and linguistics. This is not surprising given the overreaching purpose of these communities: to share and exchange data and knowledge among a broad spectrum of users, emphasizing professional academically oriented users. A more important distinction is that these academic communities' need to uphold scientific principles and academic rigor, in both the work processes they endorse and the quality of the data they offer. These needs emphasize the challenges that pertain to information validation and structure as well as social integration aspects that are tied to the information. Communities addressing more "leisurely" activities, such as news consumption, reading and other hobbies, represent a different genre of content curation, where the consumption of curated materials can be individual or collaborative. More importantly, the data quality issue, while relevant, can be less controversial and plays a smaller role in establishing the structure and the legitimacy of the community.

Information Integration

In this section, we address the second research question: "What are the key information and social design choices available to designers of content curation communities?" and place it in the broader context of scientific content curation communities. We outline design recommendations that may help to address the challenges related to information integration and social integration. Some of these recommendations defy neat classification as they apply to both themes, so they are discussed jointly.

The ongoing debates regarding standardization choices are a major challenge for information integration on EOL and other taxonomic content curation sites: Standardization, e.g., creating taxonomies and controlled vocabularies, provides a universal framework for scientists to communicate and it presents a shared representation of a field of knowledge that provides the structure for understanding and collaboration. Information standards can be viewed as boundary objects (Bowker & Star, 2000; Star, 2010) 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, p. 602). EOL focuses on species classification systems that allow for communication between scientists with different subspecialties, languages, and resources. A careful approach is needed for choosing taxonomies that will serve as boundary infrastructures and hence support shared understanding and communication. Some possible options could include the following.

Option 1: Choose or develop a master standard. We use the example of classification to address the broader challenge of creating and maintaining uniform data standards. The initial approach was for EOL to select one master taxonomy to serve as a unifying boundary object; EOL decided on the COL as the default consensus classification. Although 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, although this option may be relevant to other content curation communities that face standardization 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).

Option 2: Simultaneously support multiple standards. 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 helps support content curators' preferences and allows for additional taxonomic information to be pulled into EOL. However, it does not push the larger community toward the creation of a unifying boundary standard,

DOI: 10.1002/asi

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 and help and discussion pages, or by automating and streamlining the selection process.

Option 3: Flexibility in determining the default standard. Alternative options could also be explored. A 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 which standard 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) to supplement taxonomic classifications as meaningful ways to browse and explore EOL.

Noticeably absent from the EOL participants' dialogues around standards 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; Ehrlich & Cash, 1999). Information professionals' extensive experience working with multiple standards, open-access resources, content databases, and interfaces could be leveraged, as they bridge between the different subcommunities that form the larger content curation community, address standardization challenges, and create new mechanisms for data curation. EOL currently employs information professionals on the project, but may benefit from active solicitation of microcontributions from academic science librarians and other information professionals who would join the ranks of EOL curators. These microcontributions may include work related to standardization, 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).

Other factors that complicate the role of curators arise from the rapidly changing digital information world. Data curated in many content curation communities, and specifically scientific content curation communities, are submitted in different formats (text, tables, graphical, visualizations) using multiple media (e.g., static and dynamic visualizations, audio, video, and multimedia) with different technical standards. Knowing the source of the data is becoming increasingly complex. Content curation communities like EOL start to draw from not only scientific (e.g., GBIS), vetted, and known sources (e.g., Wikipedia) but also the growing number of volunteer citizen science projects that are springing up across the world (iNaturalist, iSpot, Project Noah). Keeping track of the sources of this data, checking the validity of the data and ensuring appropriate credit is given to contributors will become increasingly complex as more people become involved.

To assist both curators and contributors, more visualization tools that clearly articulate changes over time could be developed, akin to "Historyflow," which was used for the analysis of wiki contributions (Viégas, Wattenberg, & Dave, 2004) or "Tempoviz," which is used for tracking networks of contribution over time (Ahn, Taieb-Maimon, Sopan, Plaisant, & Shneiderman, 2011; Bongwon, Chi, Kittur, & Pendleton, 2008). Visualizations can be a powerful tool at both the macrolevel and the microlevel. On the macro level, visualizations can be used to uncover the implicit activity behind the curation scene, as well as power relations between curators, data gaps, and areas that require intense community efforts. On the microlevel, visualization tools can offer curators and users insights about the history of each contribution, whether it replaced previous content, what is its quality and values as assessed by the community, and who curated or modified it.

Social Integration

Large-scale content curation communities such as EOL can succeed only 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: (a) establishing active subgroups within large content curation communities, (b) establishing the role of open-access resources in the scientific community, (c) facilitating appropriate task-routing, (d) recognizing other contributors' efforts, and (e) building on and developing new techniques to motivate and recognize participation.

Establishing active subgroups within large content curation communities. The size and scope of crowdsourced content curation communities 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., for EOL, 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, content curation communities like EOL focus on creating their own community of practice related to curation of content. When they grow sufficiently, we propose that they actively develop interconnected subgroups that mirror other existing communities of practice (e.g., for EOL, invasive species, birds, phylogenetics). Such

DOI: 10.1002/asi

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 their walls, content curation communities will not only be places from which to send and receive content, they will also 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 rolling out tools that will facilitate a federated network of subcommunities.

Establishing the value 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 integrate 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 by consolidating 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 content curation community (Pence and Williams, 2010) is integrated with Royal Society of Chemistry journals and databases like PubMed, which use the curated content to facilitate new types of searches.
- *Providing financial support for active content curators*. EOL's Rubenstein Fellowships is a program for budding scientists that provides financial support and mentorship for up to 1 year. Rubinstein Fellowships are competitive and 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 communities, 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 some scientific communities acknowledge 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 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 contribution is needed and unique (Ling et al., 2005). What is needed, then, 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, Frankowski, Terveen, and Riedl (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.

Many other options exist as well, such as allowing individual content curators to post lists of microcontributions 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 in regards to content and to 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 nonprofessionals. Such task routing emulates the traditional scientific apprenticeship process, in which a novice is accepted into the scientific community

DOI: 10.1002/asi

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. To enable social integration, users first have to acknowledge other users' presence and contributions to the community. In large-scale communities, such as EOL, this is an especially difficult task because of 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 reuse 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 curricula 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) or GenWeb's "grievance committee." The proposed federated subcommunities 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.

Building on and developing new techniques to motivate and recognize participation. Some techniques for motivating contributions, such as "thank you" feedback notes, and recognizing contributors participation, such as reputation systems, have already been mentioned. But these examples are just the tip of today's iceberg in promoting the social participation needed for content curation communities to flourish. iSpot, a site for teaching about biodiversity developed by the UK Open University (http://www.ispot.org.uk/), provides activities for students and the public to engage, motivate, and reward contributors. iSpot's state-of-the-art reputation system combines recognizing expertise and accuracy of suggestions with effort and social interaction: Users who correctly helped identify species receive "reputation points" based on the level of agreement and helpfulness they reached. Reputation is based on collective agreement and discussion and cannot be easily manipulated. Building similar, task-based reputation systems is an important step towards improved social integration. An even bolder step would be to implement systems that "transfer" reputation from one content curation community to another. This way, participants need not start afresh when contributing to several communities, as their expertise will already have been substantiated. This is a potent idea, but one that needs to be implemented carefully: Transferring reputation is most relevant to content curation communities that share topical similarities, as different domains require different types and levels of knowledge and expertise.

Conclusion

In this article, we define a new genre of online community, "content curation communities," using the EOL as a case study. We conducted a multimethod analysis, combining survey and 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. Although 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, 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, but this is changing. 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 article. By rethinking and redesigning some of the

DOI: 10.1002/asi

online tools available to content curators, the community is attempting to face these challenges. This is done based on not only the lessons learned from the daily operation of EOL and the findings of this study but also 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. EOL is already implementing some of these changes in its current version that was recently released. These changes 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 although we have attempted to focus on the most pressing issues faced by EOL, other types of content curation communities, within and outside the scientific realm, may face other challenges that were not addressed in this article. Comparable examples of collaborations between scientists and citizen scientists in other disciplines (e.g., astronomy or chemistry) may surface different challenges. Similarly, our analysis of scientific content curation communities follows the current structure of the academic world. The advance of crowdsourcing citizen-science projects may bring about a fundamental change to how science is done and to the balance between scientists and volunteers. This change will necessitate a new look at the informational and societal infrastructures that facilitate scientific work. Despite these limitations, we believe the concept of content curation communities is enduring and a useful frame for starting a dialogue about related communities. Furthermore, the focus on information and social challenges seems to apply well to multiple content curation communities as shown in Table 1.

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

References

Ahn, J., Taieb-Maimon, M., Sopan, A., Plaisant, C., & Shneiderman, B. (2011). Temporal visualization of social network dynamics: prototypes for nation of neighbors. Social Computing, Behavioral-Cultural Modeling and Prediction. In Salerno, J., Yang, S., Nau, D. & Chai, S.K. (Eds.), (Vol. 6589, 309–316). Berlin: Springer. DOI: 10.1007/978-3-642-19656-0_43

Bates, M.J. (1999). The invisible substrate of information science. Journal of the American Society for Information Science and Technology, 50(12), 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.

Bongwon S., Chi, E.H., Kittur, A., & Pendleton, B.A. (2008). Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '08) (pp. 1037–1040). New York: 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., & Olson, G. (2007). From shared databases to communities of practice: A taxonomy of collaboratories. Journal of Computer Mediated Communication, 12(2), 652–672.

Bowker, G.C., & Leigh Star, S. (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. In Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '08) (pp. 1–12). New York: 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.

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 Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI '06), R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson (Eds.). New York, NY: ACM Press 1037–1046. DOI: 10.1145/1124772.1124928

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.

Encylcopedia of Life. (2011). Retrieved from www.eol.org/about

Ehrlich, K., & Cash, D. (1999). The invisible world of intermediaries: A cautionary tale. Computer Supported Collaborative Work, 8(1–2), 147–167.

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.

Forte, A., & Bruckman, A. (2005). Why do people write for wikipedia? Incentives to contribute to open. content publishing. Proc. of GROUP 05 Workshop: Sustaining Community: The Role and Design of Incentive Mechanisms in Online Systems. Sanibel Island, FL. 6–9.

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.

Hansen, D.L., Ackerman, M., Resnick, P., & Munson, S. (2007). Virtual Community Maintenance with a Repository. In Proceedings of the American Society for Information Science and Technology 44(1), 1–20.

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, 135–148.

Kittur, A., Chi, E., Pendleton, B.A., Suh, B., & Mytkowicz, T. (2007). Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. Alt.CHI, 2007. San Jose, CA.

Kittur, A., & Kraut, R.E. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08) (pp. 37–46). New York: ACM Press.

Kriplean, T., Beschastnikh, I., & McDonald, D.W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08) (pp. 47–56). New York: ACM Press.

Lampe, C., & Resnick, P. (2004). Slash (dot) and burn: Distributed moderation in a large online conversation space. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04) (pp. 543–550). New York: ACM Press.

Latour, B., & Woolgar, S. (1979). Laboratory life: The construction of scientific facts: 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), 1052–1057. doi: 10.1016/j.drudis.2010.10.003

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Al Mamunur, R., Resnick, P., & Kraut, R. (2005). Using social psychology to motivate contributions to online communities. Journal of Computer-Mediated Communication, 10(4), Article 10.

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.

O'Hara, K. (2008). Understanding geocaching practices and motivations. In Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems. (CHI '08) (pp. 1177–1186). New York: ACM Press.

Orrell, T. (2011). ITIS: The integrated taxonomic information system. Retrieved from http://www.catalogueoflife.org/col

Patton, M.Q. (1990). Qualitative evaluation and research methods. Newbury Park, London: Sage.

Pence, H.E., & Williams, A. (2010). ChemSpider: An Online Chemical Information Resource Journal of Chemical Education 2010 87(11), 1123–1124.

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., Bamford, S., Land, K.R., Murray, P., Slosar, A., Szalay, A.S., & Vandenberg, J. (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., Szalay, A., & Vandenberg, J. (2010). Galaxy zoo: Exploring the motivations of citizen science volunteers. Astronomy Education Review, 9(1), 010103. doi:http://dx.doi.org/10.3847/AER2009036

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.

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C.S., Lewis, D., & Jacobs, D. (2012). Dynamic changes in motivation in collaborative ecological citizen science projects. In Proceedings of the ACM Conference on Computer Supported Collaborative Work (CSCW '12). New York: ACM press.

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. In Proceedings of the Second ACM International Conference on Digital libraries (DL '97) (pp. 161–168). New York: ACM Press.

Star, L.S. (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. In Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital libraries (JCDL '02) (pp. 231–239). New York: ACM Press.

Viégas, F.B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04) (pp. 575–582). New York: 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: CRC Press.

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

Yates, D., Wagner, C., & Majchrzak, A. (2010). Factors affecting shapers of organizational wikis. Journal of the American Society for Information Science and Technology, 61(3), 543–554.

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

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.

DOI: 10.1002/asi