Facilitating scientific collaboration through content curation communities

Dana Rotman
College of Information Studies
University of Maryland, College Park, MD 20742 USA
drotman@umd.edu

Jennifer Preece
College of Information Studies
University of Maryland, College Park, MD 20742 USA
preece@umd.edu

Derek Hansen
Brigham Young University
Provo, UT 84602 USA
dlhansen@byu.edu

Kezia Procita
College of Information Studies
University of Maryland, College Park MD 20742 USA
kprocita@gmail.com

CSCW'12, February 11–15, 2012, Seattle, Washington, USA.
ACM 978-1-4503-1051-2/12/02.

Abstract
Scientific information resources, which traditionally have been created and maintained by a handful of paid scientists or information professionals, are increasingly being crowdsourced by professional and nonprofessional volunteers in what we define as "content curation communities". Content curation communities offer scientists and volunteers a novel format for collaboration, yet they also raise serious challenges of information and social integration. We see the DICOSE workshop as a opportunity to discuss these challenges and find ways in which the CSCW community can address them.

Author Keywords
Citizen science, collaboration, content curation, communities, crowdsourcing

ACM Classification Keywords
H.5.3. Computer-supported collaborative work

General Terms
Human factors

Content curation communities
Scientific progress to a large part depends 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 by professional and nonprofessional volunteers in what we define as "content curation communities." Content curation communities such as the Encyclopedia of Life (EOL) and ChemSpider, 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 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.

The scientific enterprise has always been collaborative, and the advent of the Internet has enabled large distributed scientific communities to collaborate at a scale and pace never before realized [1-3]. Content curation communities can become important tools in facing the exponential growth of scientific data that existing technological advances offer us. large scale collaborative scientific projects, including content curation communities, have already helped in developing resources on drug discovery [4], biodiversity [5], astronomic shifts [6], and language [7], to name a few. However, content curation communities face serious challenges stemming both from the need to consolidate disparate data resources and from the inherent social tensions that surface when users with different levels of expertise, coming from various domains, come together [5].

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. Specifically, data curated in many 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 some of this data is becoming increasingly complex.

An altogether different information challenge is the origin of the curated data. Some content curation communities focus solely on data collected and vetted through traditional scientific work (e.g. chemical structures, such as curated on ChemSpider), others – like the Encyclopedia of Life combine traditional scientific data with user generated content (i.e. photos contributed through a Flickr group). Inclusion of user generated content in scientific repositories raises questions of ensuring data quality and scientific integrity, as well as making the distinction between scientific content curation and general crowdsourcing projects, like Wikipedia, too vague.

Ideally, content curation communities will create a one-stop shop, in which this underlying complexity is transparent to users. This requires content curators and system designers to address these information integration challenges.

Social integration challenges

Scientific content curation communities, bring together various populations of users – from professional scientists to citizen scientists with varying degrees of expertise. Each population brings to the table its own motivations, practices, accreditation procedures and norms [8-10]. Amalgamating the differences between the two populations into an efficient work process is a major challenge. Two distinct issues shape social integration: legitimization of the work of content curation by the larger scientific community, and addressing potential conflicts between the different population of curators and users.

While content curation has proven its worth in different scientific domains, the actual curation work is still

marginalized by the larger scientific communities. As each scientific domain socially constructs what constitutes a legitimate contribution to the specific community (e.g. publication in specific journals, reviewing papers, receiving grants), open-access curation work is not yet viewed as a contribution of novel scientific knowledge worthy of the community appreciation. The result is that curation activities are not considered a priority among many scientists [5] , and do not translate to funding, promotion or time investment.

A different social integration challenge is the need to bridge between different contributor populations. While some scientists view content curation done by volunteers as a net gain that will allow them to focus on analysis rather than data collection [10], others who are reluctant to collaborate with citizen scientists fear that the latter do not uphold the same rigorous scientific standards that professional scientists are committed to. This issue is highly intertwined with issues of information integration, as the lack of trust translates into control over the type of curated content and the forms in which it is vetted and curated.

Overcoming these challenges

In order to realize the ambitious goals of content curation communities and make them a valuable addition to the broader mission of novel scientific cyber-infrastructures, we must understand how to effectively design the technologies, information standards, processes, and social practices that support them. Creative design solutions may ameliorate some of the challenges outlined above. Some examples may include:

- Standardization: creating and maintaining uniform (yet flexible) data standards within an individual community or across domain-specific communities. This requires not only establishing agreed upon standards to facilitate different formalisms and nomenclatures within a discipline (a process that should be based on domain-specific discourse), but also systematically providing the appropriate technological support for various standards.
- Using visualizations to document contribution and activity: visualizations can be a powerful tool on both the macro and the micro levels. On the macro level visualizations can be used to uncover the implicit activity behind the curation scene, as well as power relations between curators, and areas that require intense community efforts. On the micro level, visualization tools can offer curators and users insights about the history of each contribution, whether it replaced previous content, what are its quality and its value as assessed by the community.
- Emphasizing the role of information professionals in content curation communities: information professionals' extensive experience working with multiple standards, open access resources, content databases, and interfaces could be leveraged, as they bridge between the different sub-communities that form the larger content curation community. They can address standardization challenges, and create new mechanisms for data curation.
- Maintaining quality control through personal task routing: task-routing emulates the traditional scientific apprenticeship process, in which a novice is accepted into the scientific community through legitimate peripheral participation which grows into more prominent roles in time. 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 or "contamination" of resources, and scientists' hesitation of association with non-professionals.
- Recognizing participation: various techniques for motivating contributions, such as feedback notes, recognizing contributors' participation through

reputation systems, featuring valued contributors within the community, or highlighting data reuse, can all contribute to social cohesion within the community as well as emphasize the broader impact of the specific community on its domain.

Boarder questions and topics for discussion

Content curation communities present both opportunities and challenges. While we outlined and address some of the challenges, other questions beg a broader discussion that will cross disciplinary borders and engage the CSCW community as well as the scientific community. Some topics that would benefit from such a discussion are:

- Should the abundance of data and the growing role volunteers and citizen-scientists play in the scientific world cause a paradigm shift in the ways traditional forms of science and engineering are done?
- How can we legitimize and reward content curation, crowdsourcing, and open-access resources within the broader scientific community?
- What are the best technological tools to facilitate effective content curation? What level of elasticity is needed in creating tools that will fit the needs of different scientific domains and various scientific standards?
- Should scientific content curation communities learn from more generic social media initiatives and tools, such as recommender systems, aggregates, and personal curation tools?
- How can we efficiently motivate different curating populations while overcoming the social conflicts inherent to multi-level multi-disciplinary communities?

We are excited about the opportunity to discuss these – and other related topics – in the Data Intensive Collaboration in Science and Engineering Workshop.

References

[1] Borgman, C. L. *Scholarship in the digital age: Information, infrastructure, and the Internet*. The MIT Press, 2007.

[2] Birnholtz, J. P. and Bietz, M. J. Data at work: supporting sharing in science and engineering. In *Proc. GROUP* 2003. ACM Press (2003).

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

[4] Li, Q., Cheng, T., Wang, Y. and Bryant, S. H. PubChem as a public resource for drug discovery. *Drug discovery today*, 15, 23/24 (2010).

[5] Rotman, D., Procita, K., Hansen, D., Parr, C. and Preece, J. Supporting Content Curation Communities: The Case of the Encyclopedia of Life. HCIL tech report-2011-19

[6] 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. and Vandenberg, J. Galaxy zoo: an experiment in public science participation. *Bulletin of the American Astronomical Society*, 39 (2007), 892.

[7] Hughes, B. Metadata quality evaluation: Experience from the open language archives community. *Digital Libraries: International Collaboration and Cross-Fertilization* (2005), 135-148.

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

[9] Hara, N., Solomon, P., Kim, S. L. and Sonnenwald, D. H. 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 (2003), 952-965.

[10] Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D. and Jacobs, D. Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. In *Proc. CSCW 2012* ACM Press (2012).