"Data" Is Not Enough
Case Studies in Managing and Preserving Digital Research Outputs

Elizabeth Rolando
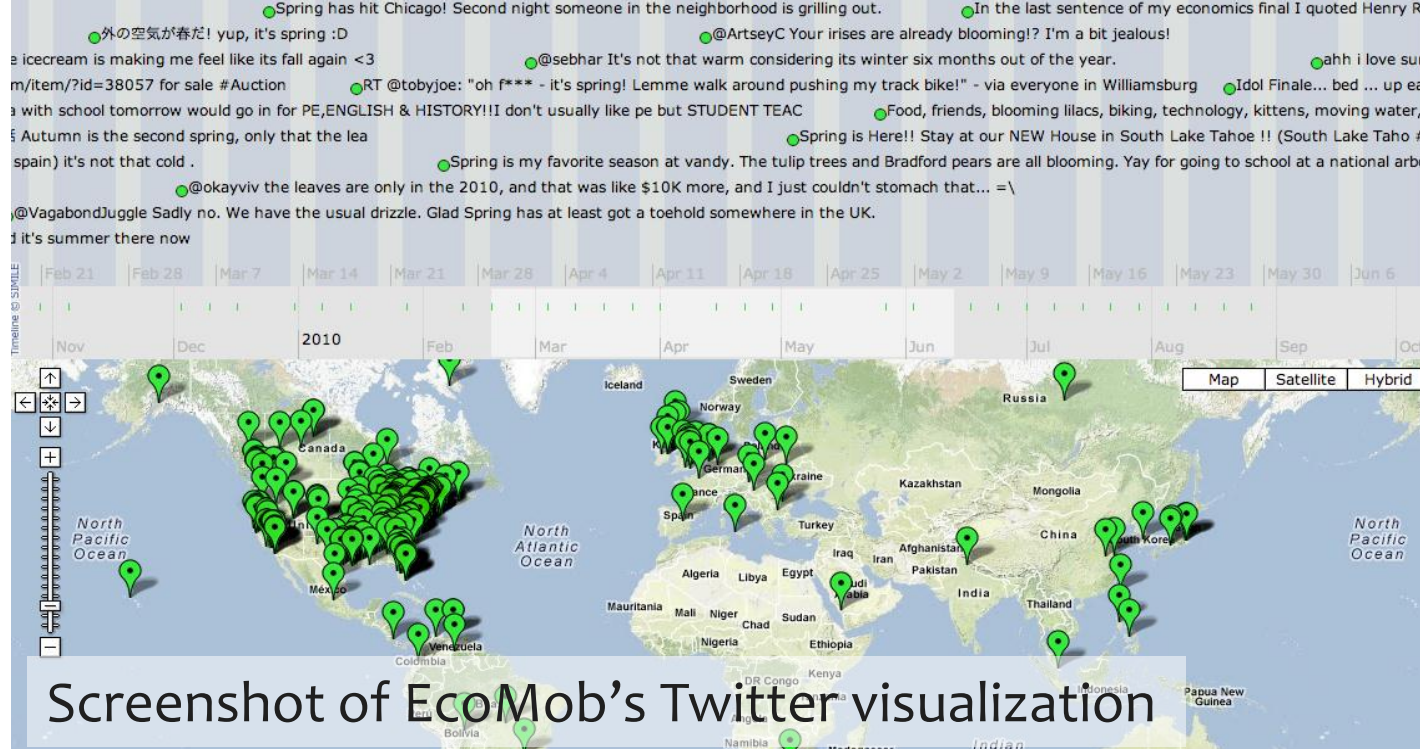Georgia Institute of Technology

Rachel Mandell
UCLA Information Studies

Introduction

"Data is the new special collection." – Sayeed Choudhury
(as quoted in Palmer et al., 2010)

- To an ever greater degree, librarians are helping researchers manage and curate their data, especially with increased emphasis on data management and sharing plans.
- Data are often a form of scholarship in and of themselves, an "end product of research" (Borgman, 2007, p. 115).
  - Recognizing that data are a product of research complements the more traditional library focus on "digital output[s] of scientific research, namely, the results of research published by researchers as the articles in the scientific journals" (Arzberger et al., 2004, p.135).
- "Data" is a nebulous concept. Even funding agencies recognize that data means different things to different communities.
  - From the National Science Foundation (2010):
    "What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models."
- Are libraries and information professionals meaningfully capturing the intellectual output of researchers? What about materials that are neither data nor articles?

Case Study 1: EcoMob

- EcoMob tracks Tweets about the weather – can we learn about climate change by looking at what people say on Twitter?
- 2012 – Consulted with EcoMob to develop a strategy for the preservation of their data.

Screenshot of EcoMob's Twitter visualization

- Through the lens of a data management plan, we determined that the data for the project were the Tweets, stored in a MySQL database, along with the text files containing the scripts. This was what could be deposited in a repository.
- But what about the dynamic visualizations? Aren't these important products of the research?
- Re-creation of the visualization from the scripts requires that a future user has the necessary software and know-how.
- Our library-centric preservation plan was unable to capture the true intellectual contribution of the project.

Case Study 2: UCLA Data Registry

- The UCLA Registry is a new online tool designed to assist researchers in managing and discovering surrogate records of datasets.
- To determine how to actually build the tool, we conducted 20 in-depth interviews with researchers from disparate fields—from neurology to archaeology, and everything in between.
- In some disciplines, such as the Digital Humanities, we found that research often involves a multitude of material—such as archival documents, computer code, and visualization software.
- What counts as data and furthermore what counts as a dataset is often very difficult to define.
- For the purposes of the Data Registry, we left these definitions up to the researcher to decide, but it left us convinced that "data" was not enough.

Conclusions

- While interest in managing data has broadened the role of many academic libraries, information professionals must recognize the need to care for an even wider variety of materials – a variety often not covered by "data."
- A focus on "data curation" risks the development of services that are too narrow – many projects create work that require more support than most libraries are currently prepared to offer.
- If these materials are to be shared meaningfully, we should focus on how to care for "research outputs."
- A preservation program that seeks to manage and curate the intellectual contributions of the community beyond the narrow conceptions of data and publications will be best able to provide for meaningful reuse of these materials.

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 3(29), 135-152.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

National Science Foundation. (2010, November 30). Data management & sharing frequently asked questions.

Palmer, C., Cragin, M., MacMullen, J., Chao, T., Renear, A., Dubin, D., Sacchi, S., Michael Welge, M., & Auvil, L. (2010). *The Data Conservancy: Research on data curation and repositories*.