Identifying Crossover Documents in an Interdisciplinary Research Environment

Rich Gazan
University of Hawaii
gazan@hawaii.edu

Abstract

AIRFrame is a NASA project to analyze and integrate astrobiology documents from diverse disciplines to catalyze new knowledge. This paper outlines the technical infrastructure of the current system and reports on an ongoing iterative evaluation, to address the question of how scientists perceive and integrate crossover documents in their research. Some of the obstacles preventing AIRFrame from gaining traction with its target audience of astrobiology researchers include representing their research output accurately, effectively translating and relating diverse metadata, and understanding disciplinary norms and the broader knowledge production infrastructure. The skills required to address these needs suggest a role for both researchers and information professionals to work in tandem with technical tools to catalyze interdisciplinary knowledge. A graduate seminar in interdisciplinary knowledge production, targeted at both researchers and graduate students at the University of Hawaii, has been designed to elicit and impart needed information as input to ongoing AIRFrame development.

Keywords: astrobiology, interdisciplinarity, scientific communication

Introduction

Traditional inputs to scientific research have been largely limited to discipline-specific scholarly publications and datasets, but integrative science requires that information be translated and transported across fields, and applied to common problems. This requires not just technology-based support systems, but nuanced understanding of the knowledge production infrastructure in which scientific research takes place. This paper discusses the iterative evaluation of the Astrobiology Integrative Research Framework (AIRFrame), a NASA Astrobiology Institute (NAI) project designed to identify particular documents from an astrobiology corpus that might be used to connect and mutually inform researchers from different disciplines, and how scientists perceive and integrate crossover documents in their research. The development of a seminar designed to instill interdisciplinary practices and perspectives in both researchers and graduate students is proposed in response to the findings reported here.

Background

Interdisciplinary research is notoriously difficult to quantify (Morillo et al., 2001). Heuristic indicators such as co-authorship or joint project participation are often used to assess interdisciplinary science (Rafols & Meyer, 2010), but the extent to which new knowledge is actually catalyzed as a result of these interactions is open to question. Scientometric analysis of research documents has developed in response to the need to quantify and evaluate the actual and potential impact of scholarly works (Huutoniemi et al., 2010).

Acknowledgements: This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA08DA77A issued through the Office of Space Science. I would also like to thank Michael G. Gowanlock and Lisa J. Miller for their contributions to this work.

Gazan, R. (2013). Identifying crossover documents in an interdisciplinary research environment. *iConference 2013 Proceedings* (pp. 457-460). doi:10.9776/13254

Cyberinfrastructure research and education has devoted significant attention to the “last mile”— the last section of connectivity from the network to the user (Gabridge, 2009). Equally important is the “first mile”: the range of inputs available to a scientist conducting research. Scientists are traditionally trained to draw from data sources within their discipline to ground their contributions; to do otherwise risks the perception that the contributions are less relevant or valid. However, this insular approach can work against boundary-crossing research. To do integrative work, researchers must have the means to develop a working understanding of the knowledge production infrastructure within which their research takes place, in its technical, logical and social dimensions (Friedlander, 2008).

The NASA Astrobiology Institute (NAI) consists of 18 teams and over 800 researchers studying the possibility of life beyond earth. This is an inherently interdisciplinary endeavor, with astronomers, biologists, chemists, engineers, hydrologists, meteorologists, oceanographers and researchers from many other fields addressing different aspects of the same question. One of the primary goals of NAI is to catalyze interdisciplinary knowledge, and it provides a hospitable environment for exploration beyond one’s home discipline.

Our approach has been to develop a set of tools and methods to identify potential “crossover” documents within astrobiology, the boundary objects (Star & Griesemer, 1989) across which diverse scientists might communicate and inform one another. To catalyze interdisciplinary knowledge in astrobiology, we have developed the Astrobiology Integrative Research Framework (AIRFrame), a conceptual framework and set of tools and methods for collecting, analyzing and integrating astrobiology documents (Gazan, 2010). Our approach has focused on textual analysis of documents from diverse disciplines to reveal implicit relationships.

Our previous work (Gowanlock & Gazan, 2012) represented the research tracks of scientists at the University of Hawaii NASA Astrobiology Institute (UHNAI) over a ten-year period. Documents were represented by their abstracts and the abstracts of the publications they cite, and their source discipline by the Journal Subject Category from the Thomson Reuters Web of Knowledge database suite. Using an unsupervised machine learning clustering technique, the sequential Information Bottleneck (Slonim et al., 2002), each paper was assigned a cluster, where publications in the same cluster share mutual information. The publications of most UHNAI researchers clustered with those in similar disciplines across multiple clustering runs. However, some publications were found in multiple clusters, and some authors’ work consistently clustered with that of non-obvious colleagues. The clustering and classification processes yielded the first step in a data-driven analysis designed to identify actual and potential crossover points for the integration of diverse knowledge.

Method

While much of our initial work has been accomplished via automated document harvesting from databases such as the NASA Astrophysics Data System (ADS), and clustering using the Weka machine learning and data mining toolkit (http://www.cs.waikato.ac.nz/ml/weka/), integrating documents from other databases presented challenges that a purely algorithmic approach could not solve, motivating the present study. Critical tasks such as searching both Web and professional databases for relevant astrobiology literature, and interpreting, creating and equating diverse metadata, required the participation of individuals with integrative information skills, including researchers, graduate students and others supporting their research, warranting a focus-group data collection approach to iterative evaluation including stakeholders from all these groups.

One of the outcomes of AIRFrame will be a system that presents researchers who seek to do interdisciplinary science with a suggested set of related papers outside their areas of expertise— essentially a recommender system for crossover documents. But like any recommender system, user feedback is critical. In the initial phase of iterative evaluation, members of the UHNAI team were presented with a summary of the results published in Gowanlock and Gazan (2012) in a weekly astrobiology seminar, and participated in open-ended questioning and discussion about their reaction to the design and functions of the system, and recommendations for its productive evolution.

Results and Discussion

Researchers' responses included concerns about how accurately the document abstracts and abstracts of cited articles represented their research areas and points of potential crossover. Some were pleased at the potential interdisciplinary breadth of their work suggested by AIRFrame, while others questioned why the system had not represented tangible points of crossover they knew to be present in their work. Researchers were made aware of the probabilistic and relativistic nature of clustering, and that documents might cluster differently given the content of other documents in the analyzed corpus.
The extent to which users understand and trust the system inputs and representations is understandably related to their likelihood of using the system.

Another weakness of relying on published works is that the research tracks represented are necessarily backward-looking, when the goal is to generate useful crossover documents to catalyze interdisciplinary knowledge for present and future research. Indeed, several researchers found that papers they had published in areas entirely unrelated to astrobiology had been harvested and used to represent their interdisciplinary potential in the field.

The ability to curate data from diverse sources, distinguish relevance, and understand the nuances of different disciplinary literatures and publication norms are some of the critical skills necessary to represent documents and authors accurately, and to normalize representations of their work for comparative analysis. The University of Illinois recently introduced a Data Curation Education Program within the Graduate School of Library & Information Science that integrates many of these skills: "Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time." (http://cirss.lis.illinois.edu/CollMeta/dcep.html)

To this end, a seminar in interdisciplinary knowledge production is in development at the University of Hawaii, which will engage graduate students and researchers across campus, and cover the following topics:

- The global information infrastructure
- Scientific knowledge production
- Scientometrics
- Disciplines and interdisciplinary collaboration
- Intercultural communication
- Data mining
- Searching across disciplinary literatures
- Ownership of scientific information
- Data access and visualization
- Science policy
- Economics of science
- Citizen science

Where a generic seminar offering might be easily ignored by overcommitted researchers, especially one outside their domain of interest, the institutional infrastructure of the NASA Astrobiology Institute requires that funded participants submit regular reports on their progress, and that they specifically address how their work crosses disciplinary and institutional boundaries. Thus, NAI researchers are highly motivated to learn how they can maximize and articulate the interdisciplinary aspects of their work. Existing astrobiology seminars are one form of boundary object where diverse researchers share information with students and one another, and adding components of interdisciplinary knowledge production provides a more explicit mechanism to promote meaningful and actionable interdisciplinary collaboration.

The seminar addresses researchers' roles as generators, seekers, consumers, integrators and stewards of scientific information (ARL 2006). Information that is inaccessible, untranslatable, informal or simply lost cannot be used to represent and suggest areas of potentially productive crossover. The source of these hidden representations is the researcher, but eliciting them often requires the contributions of information professionals. Unpacking the affordances and constraints of the research infrastructure necessitates a sociotechnical approach, and will allow the next generation of researchers and information professionals to confront and address issues of data management, intellectual property rights, data representation and integration, interdisciplinary and international collaboration, and the publication, dissemination and preservation of scientific information.

Modern researchers, and those who help provide access to the products of their research, must incorporate and critically evaluate relevant data from diverse sources. Works published in non-core databases as well as unpublished and informal documents also represent science practice, and should form part of any system designed to measure and promote interdisciplinarity (Sugimoto 2011). Individuals with skills of data curation and integration will be well positioned to develop not only a broad, systematic view of scientific practice, but be in a position to identify, create or act as boundary objects in future interdisciplinary efforts.

Conclusion

The crossover documents suggested by AIRFrame are designed to increase the interdisciplinary aspects of current and future research, but might also be used to inform the development of astrobiology syllabi and curricula, to train the next generation of astrobiology researchers. AIRFrame can also be used to measure the breadth of disciplines from which papers and researchers draw, allowing fine- grained analysis of interdisciplinary work. Through iterative evaluation, the results are shared with researchers, providing them a working understanding of some of the scientometric indicators by which their work will be evaluated. Understanding how research is represented and transported in formal and informal documents, and some of the barriers researchers face to interdisciplinary work, will allow more accurate representations of scientists' work, and the identification of more potential points of crossover to catalyze cross-disciplinary research.

References

ARL (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering. *Proceedings of the ARL/NSF Workshop on Long-Term Stewardship of Digital Data Collections*. Retrieved from http://www.arl.org/pp/access/nsfworkshop.shtml

Friedlander, A. (2008). The triple helix: Cyberinfrastructure, scholarly communication, and trust. *Journal of Electronic Publishing 11*(1). doi:10.3998/3336451.0011.109

Gabridge, T. (2009). The last mile: Liaison roles in curating science and engineering research data. *Research Library Issues: A Bimonthly Report* from ARL, CNI, and SPARC, No. 265: 15–21. Retrieved from http://www.arl.org/resources/pubs/rli/archive/rli265.shtml

Gazan, R. (2010). AIRFrame: Integrating diverse digital collections in astrobiology. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 375-376.

Gowanlock, M. G., & Gazan, R. (2012): Assessing researcher interdisciplinarity: A case study of the University of Hawaii NASA Astrobiology Institute. *Scientometrics*. doi:10.1007/s11192-012-0765-y

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy 39*(1), 79-88.

Morillo, F., Bordons, M., & Gomez, I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics 51*, 203–222.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometric*s *82*, 263–287.

Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 129–136.

Star, S., & Griesemer, J. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science 19*(3): 387–420.

Sugimoto, C. (2011). Looking across communicative genres: a call for inclusive indicators of interdisciplinarity. *Scientometrics 86*, 449–461.