Big Data and MOOCs Herald Change for Academic Publishers

By Paula Gantz
Paula Gantz Publishing Consultancy

Big Data, MOOCs and author/researcher tools were the main themes of the International Association of Science Technical and Medical Publishers (STM) Innovations Seminar 2013 held last week in Washington, DC. Speakers from both within and outside the academic publishing industry converged on these three topics as this segment of the industry moves rapidly to a digital environment.

Keynote speaker, Dennis Gannon, Director of Cloud Research Strategy at Microsoft Research, emphasized that a data revolution is transforming science. "We are now in the fourth paradigm of science, evolving from prior scientific methodologies: experimental, theoretical, and computational. "This paradigm is not hypothesis driven. Now data can be used for exploration and data mining to reach scientific findings," he said.

But, Gannon cautioned, the data need to be scalable, sustainable and curatable. Algorithms to process massive amounts of data must be derived. Semantics are necessary to search and concept clusters are essential. All these developments are bringing a revolution in machine learning. Gannon envisions a time when a machine could generate an article abstract from a scientific paper, improving on the abstracts which are currently author-generated.

Gannon pointed to the work of Eric Horvitz, co-director of Microsoft Research, who created a predictive model for readmissions to hospitals by processing 25,000 variables from 30,000 emergency room visits.

Deep learning is another trend to watch. Deep learning can be both supervised and unsupervised. "Unsupervised machine learning is what is really interesting. You can find the hidden structure in data without labels and present the data with no hypothesis. You just start grouping data," Gannon explained. As an example, Geoffrey Hinton at University of Toronto processed more than 10 million unlabeled YouTube images to produce classifications.

A soon-to-be released Microsoft product for the scientific community is Microsoft Academic Search which explores over 48 million articles to show connections between scientist networks, creating citation graphs of the research.

Gannon also looks forward to extending the long tail of data analytics to more researchers, not just in the life and biomedical science communities, but in areas like economics and history. But he cautions that scientists want to focus on science not on the programming, and emphasizes that cloud resources are available to facilitate this explorative process.

Heather Ruland Staines of SIPX explained why publishers should care about MOOCs (massive open online courses). Staines defined MOOCs in a very broad sense to include what she termed "free-range" MOOCs that don't necessarily provide any actual course structure, but just focus on a desire to learn. She explained that there are a lot of players in the MOOC space including Coursera, edX and Udacity. Khan Academy provides practical learning. Futurelearn is a UK universities' initiative and Open2Study is an Australian effort. There are also online course enablers that help you to create courses.

Business models for MOOCs are still in a state of flux, according to Staines. In most cases the MOOCs are still venture-funded, but they will need to become sustainable soon. Some of the models now being explored include payment for academic credit, certificates, proctored exams, leads for job recruiters, leads for admissions officers and competency-based awards through federal funding.

MOOCs are growing rapidly because they represent a new form of brand extension, especially for prestigious universities. They are also useful for remedial education, oversubscribed courses, introductory courses, satellite campus connections, test preparation, recruiting and admissions, CME, executive programs and corporate training. Celebrity MOOCs have also become popular.

Although the completion rates for MOOCs is atrocious, according to Staines, academic publishers can benefit from providing content to MOOCs. "Why should publishers care?" Staines queried. "Because courses will need content and it is an opportunity to reach new markets."

Since most MOOCs are library-based and students are university-based, students could be taking course at a different university. SIPX clears content for MOOCs. MOOCs often charge for content even if the course is free. Staines recommended that publishers experiment with pricing for MOOC usage. She pointed out that lots of user analytics will be available from this source.

Paul Uhlir**.** director of the Board on Research Data and Information (BRDI) at the U.S. National Academies, spoke about the need for establishing data citation standards and practices. This is particularly true as existing literature is being supplemented by data sets and other supplemental information resident in repositories and other non-publisher-based environments. BRDI will be issuing a white paper in the near future that will address some of these issues. The implementation strategy will include all major stakeholders: data centers, universities, research funders, researchers, publishers and editors.

According to Uhlir, some of the major questions are: Why institute a data citation requirement? Would the benefits outweigh the costs? How would the process be implemented? Who would do it? At what point in the publishing process? Are there other issues to consider?

Eefke Smit of STM pointed out that research shows that article citations go up 40 to 70 percent when datasets are included with a published article. "Supplemental data dumping is very difficult so the industry needs a proper data citation structure to alleviate this problem," she said.

Expanding on the topic of proper citations, Micah Altman of MIT Libraries spoke about the need to register researchers through a unique number system. There are many initiatives, including ORCID which now has 100,000 registered researchers, and ISNI, which has 3.9 million assigned identifiers. Altman

identified the stakeholders in this space as researchers themselves, publishers, funders and research organizations. He stressed that new governmental reporting requirements will strengthen this initiative.

A regular feature of the Innovations Seminar is the presentation of the STM Future Lab trends. David Martinsen of the American Chemical Society pointed to the three overarching trends: a move to servicing individuals from institutions, a shift from closed to open content, and the primacy of the article as the transmitted object, often accompanied by supplemental data.

Martinsen emphasized that as mobile devices are becoming ubiquitous, the end-user experience is the ultimate goal of publishers. This experience should be enriched with additional content and tools, especially as users generally arrive at content through search engines and not through publishers' site.

As regards open content, Martinsen encouraged publishers to find new ways to collect fees, possibly at an article level and when supplying content to MOOCs. He also recommended that publishers make use of the user analytics now available to help direct future activities.

Additional presentations by industry innovators highlighted the many projects now in development to address the issues of author tools and data curation and analysis. Among those presenting were Kaitlin Thaney, Digital Science, who spoke about figshare and other researcher tools; Kent Anderson, *Journal of Bone and Joint Surgery*, who explained Case Connector; Jeff Lang, American Chemical Society, demonstrating the ACS's new interactive and composed full-text article in an active view PDF format; Andrea Powell of CABI who discussed Direct2Farm; Mike Takats, Thomson Reuters, who discussed a new initiative to create a data citation index; Marty Picco, Atypon, who demonstrated the newly retooled Atypon; Anita de Waard, Elsevier, discussing three pilot projects geared to data preservation; and Tom Carpenter of NISO, speaking about the need for article-level open access metadata.