Column Editor: **Donald T. Hawkins** (Freelance Conference Blogger and Editor) <dthawkins@verizon.net>

Data Sparks Discovery: The 2016 NFAIS Annual Conference

Column Editor's Note: *Because of space limitations, this is an abridged version of my report on this conference. You can read the full article which includes descriptions of additional sessions at* ***http://www.against-the-grain.com/2016/04/v28-2-dons-conference-notes/****. — DTH*

About 150 information professionals assembled in Philadelphia on February 21-23 for the **58th NFAIS Annual Conference**, which had the theme "Data Sparks Discovery of Tomorrow's Global Knowledge." The meeting featured the usual mix of plenary addresses and panel discussions, the always popular **Miles Conrad Memorial Lecture** (see sidebar), and a well-received "Shark Tank Shootout," in which representatives of four startup companies were asked a series of probing questions by a panel of judges.

Opening Keynote: Preparing the Next Generation for the Cognitive Era

Steven Miller, Data Maestro in the IBM Analytics Group, opened his keynote address by noting that data is transforming industries and professions, and the demand for data engineers is skyrocketing. New data-based professions such as "data scientist," "data engineer," "data policy professional," and even "chief data officer" are emerging. The Internet of Things and software analytics have been significant drivers in the emergence of data-based professions and services, for example:

Steven Miller

- The Uber ride-sharing service uses GPS data to determine where a car is and how long it will take to arrive at the customer's location.
- Trimet, the rapid transit system in Portland, OR, has teamed with Google to integrate real-time transit data with Google Maps, allowing smartphone users to easily plan their journeys. (See *bit.ly/trimetandgoogle*)
- HackOregon is transforming public data into knowledge; for example, large sets of geology data are used in a system to help Oregon residents plan for possible future earthquakes (*bit.ly/after-shock*).
- London has become the leading city working with open data The London Datastore (*http://data.london.gov.uk/*) provides access to wide range of over 500 data sets about London.
- Kabbage (*https://www.kabbage.com*) has helped over 100,000 small business owners qualify for loans or lines of credit and has redefined the process of obtaining business loans.
- NewRelic (*http://newrelic.com/*) uses predictive analytics to monitor and analyze software applications and warn developers about overloads and potential outages.
- Many cities are now embedding smart sensors in roadways, buildings, waterways, etc. to send and receive data, thus improving their services.
- IBM's Watson for Oncology (*http://www.ibm.com/smarter-planet/us/en/ibmwatson/health/*) system combines patient data with large volumes of medical literature to deliver evidence-based suggestions to oncologists.

Obviously, in all these systems, confidence and trust in the data are critical. Many companies have failed in their responsibility to protect data, resulting in significant breaches.

Data is a core business asset, but few colleges have courses in data policy skills. In today's cognitive era, everyone must become data literate and be able to manage and analyze data. **Miller** said that "T-shaped" skills (boundary-crossing competencies in many areas coupled with in-depth knowledge of at least one discipline or system) are required.

A look at job listings at *www.linkedin.com/jobs/data-engineer-jobs* (or data-scientist-jobs) and *indeed.com* provides a picture of how strong the demand for such professionals is. According to **Miller**, there are two types of data scientist: *human data scientists* who advise businesses, and *machine data scientists* who write advanced algorithms.

Data Usage Practices

Courtney Soderberg from the Center for Open Science said that a reproducibility crisis is occurring in science, and the literature is not as reproducible as we would like to believe. Journals and funders have therefore implemented data sharing policies and are mandating that authors publish their raw data. The Center for Open Science is working to increase openness, reproducibility, and transparency in science. It has built an open source space (see *http://osf.io*) where scientists can manage their projects, store files, and do research. It is important to make it easy to share data, and not require researchers to invest time learning how to do it.

Lisa Federer, an "Informationist" at the National Institutes of Health (NIH) Library, said that many definitions of "Big Data" revolve around the four "Vs":

- Velocity (the speed of gathering data),
- Variety (many types of data),
- Volume (a lot of it), and
- Veracity (good data).

Faster and cheaper technology and an increase in "born digital" data are also providing new roles for librarians to assist researchers who have never before been required to share their data.[1] **NIH** has created a publicly available planning tool to help researchers create data management plans that meet funders' requirements, (see *http://dmptool.org*) and the **NIH** library has developed a comprehensive guide to data services resources (*http://nihlibrary.campusguides.com/dataservices*).

Managing Data and Establishing Appropriate Policies

Heather Joseph, Executive Director of the Scholarly Publishing and Academic Resources Coalition (SPARC) said that data policy development is an evolutionary and iterative process involving the entire research community. It is focused on four major areas:

1. *Policy drivers*. U.S. funders invest up to $60 billion a year in research to achieve specific outcomes, which require free access to the research results and the data.
2. *Policy precedents and developments*. The Open Data Executive Order, issued in 2013 by **President Obama**, mandated open and machine-readable data as the default for all government information, and a subsequent Public Access Directive begins to lay out the rules for accessing data.
3. *Emergence of research data policies*. Today, three years after the Executive Order, draft or final policy plans have become available for 14 federal agencies.
4. *Policies supporting a robust research environment*. Reiteration of evolutionary policy development, consistent policy tracking, and regular input are vital to promoting a reasonable level of standardization.

Anita De Waard, VP, Research Data Collaborations at **Elsevier**, discussed the research data life cycle that was developed at **Jisc**.[2] Important steps in the lifecycle and **Elsevier's** involvement include:

- *Collection and capture of data and sharing of protocols at the moment of capture*. **De Waard** mentioned Hivebench (*https://www.hivebench.com/*), a unified electronic notebook allowing a researcher to collaborate with colleagues, share data, and easily export it to a publication system.

continued on page 67

- *Data rescue*. Much data is unavailable because it is hidden (such as in desk drawers). **Elsevier** has sponsored the International Data Rescue Award to draw attention to this problem and stimulate recovery of such data.
- *Publishing software*. **Elsevier's** open access *SoftwareX* journal (*http://www.journals.elsevier.com/softwarex/*) supports the publication of software developed in research projects.
- *Management and storage of data*. **Mendeley** and **GitHub** provide versioning and provenance.
- *Linking between articles and data sets* will allow a researcher to identify data sets in repositories. Data sets must therefore be given their own DOIs which will allow them to be linked to articles.

Larry Alexander, Executive Director of the Center for Visual and Decision Informatics (VDI) at **Drexel University**, said that the Center has a visualization and big data analytics focus and supports research in visualization techniques, visual interfaces, and high performance data management strategies. Some of its noteworthy results include:

- Exploration of the volume and velocity of data streams in a 3D environment,
- Analysis of crime data in Chicago to find hot spots and how they change,
- Prediction of flu hot spots using environmental conditions (temperature, sun exposure, etc.),
- A gap analysis of U.S. patents to predict where new breakthroughs would occur, and
- Mining of PLoS ONE articles to determine popularity of software use.

New Data Opportunities

Ann Michael, President of the consulting firm **Delta Think**, kicked off the second day by saying that anything to do with data is a new career opportunity. Publishing and media businesses are leveraging data today. Article impacts have become more important; companies such as **Plum Analytics** and **Altmetrics** are in the business of providing usage data for journal articles, and **Springer's** Bookmetrix does the same for books. Other systems provide different applications; for example:

- Impact Vizor from **HighWire Press** (*http://blog.highwire.org/tag/impact-vizor/*) looks at analytics of rejected articles and helps publishers decide if it would be worthwhile to start a new journal to accommodate them;
- UberResearch (*http://www.uberresearch.com*) builds decision support systems for science funding organizations;
- The *New York Times* uses predictive algorithms to increase sales and engagement (it is careful to emphasize that such data are not used to make editorial decisions);
- RedLink (*https://redlink.com*) helps marketing and sales teams at academic publishers focus on the needs of their customers; and
- Tamr (*http://www.tamr.org*) connects, cleans, and catalogs disparate data so that it can be used effectively throughout organizations to enhance productivity.

Building Value Through a Portfolio of Software and Systems

Three entrepreneurs followed **Michael** and described their products for using data. Overleaf (*http://www.overleaf.com*) provides a set of writing, reviewing, and publishing tools for collaborators and removes many of the frustrations that authors experience, especially when articles have many of them. Many articles today have more than one international author, and the traditional way of collaboration was to email versions of documents to them, which leads to long email chains, multiple versions of the same document, reference maintenance problems, and lengthy revision times. Now, documents can be stored in the cloud and managed by Overleaf, so most of these problems are removed. Some journals are now receiving up to 15% of their submissions from Overleaf.

Etsimo (*http://www.etsimo.com*) is a cloud-based visual content discovery platform combining an intelligent search engine and an interactive visual interface to a document collection. In traditional ("lookup") searching, the user's intent is captured only in the initial query, so the query must be reformulated if revisions are needed. Etsimo works with keywords in a full-text index and uses artificial intelligence (AI) or machine learning to create connections in the content. A demonstration is available at *http://wikipedia.etsimo.com*.

Meta (*http://meta.com/*) is a scientific knowledge network powered by machine intelligence that seeks to solve some of the common problems caused by the current flood of scholarly articles. It is powered by the world's largest knowledge graph and is coupled with ontologies to unlock the information in scientific articles. Researchers can receive recommendations and discover unknown articles or historical landmark articles based on the concepts and people they follow. Bibliometric intelligence can be integrated into author workflows, which provides significant benefits to all parties in the publication process.

Creating Value for External Institutions and Systems

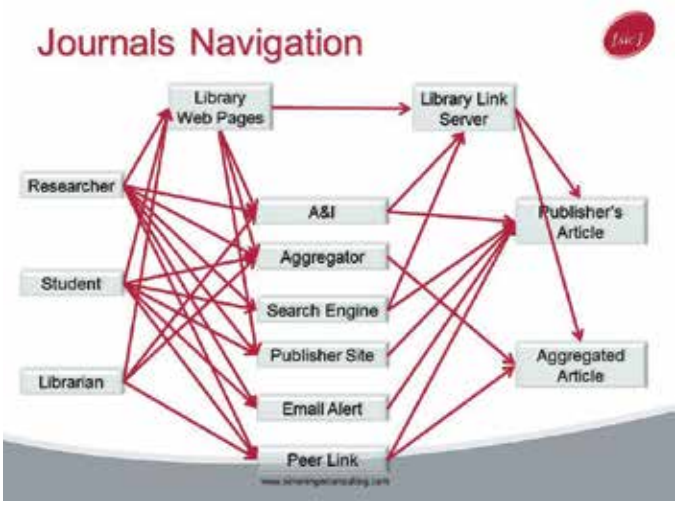
James King, from the **NIH** library, described its vision to be the premier provider of information solutions by enabling discovery through its "Informationist" program (see *http://nihlibrary.nih.gov/Services/Pages/Informationists.aspx*), in which information professionals are embedded into NIH workflows to focus on delivering knowledge-based solutions.

A unique feature of the Informationist program is its custom information services: a "Geek Squad" that provides support to informationists by offering digitization of government publications, database access via APIs, and consulting services.

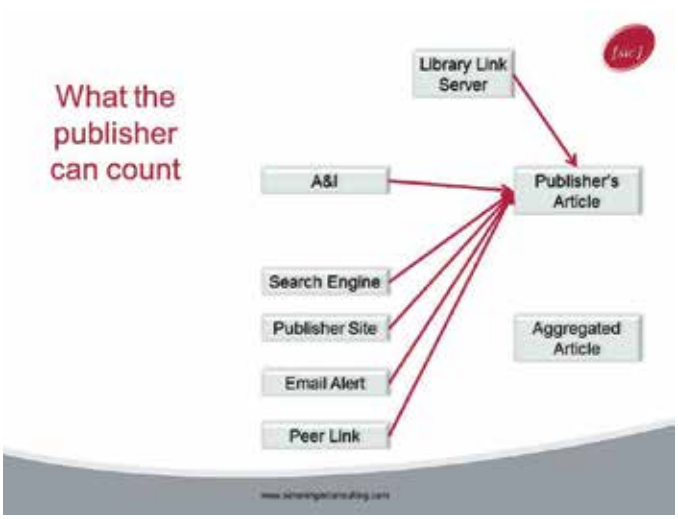
How Readers Discover Content in Scholarly Publications

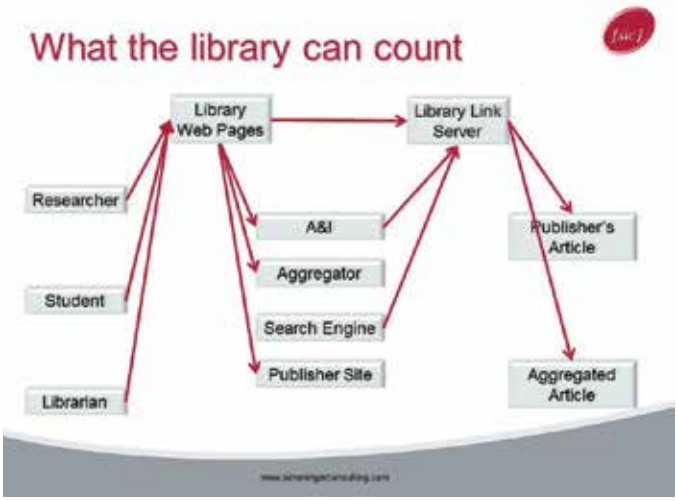
At the members-only lunch, **Simon Inger** presented the results of a large survey that stemmed from the recognition that search and discovery do not happen in the same silo. For journals, there are many ways of discovering articles. For example, this figure shows how some academic readers might navigate to various incarnations of the publisher's article.

Simon Inger



continued on page 68





However, publishers cannot get a complete picture of access to their content because they can see only the last referring Website; in contrast, libraries can view the complete access paths of the users, as these diagrams show.

Inger's consulting company conducted a large survey of reader navigation in 2015 to gain a view of the importance of access channels to information publishers and buyers. About two million invitations were sent to potential respondents, and 40,000 responses were received from all over the world (there was even one from the French Southern and Antarctic Lands!). Literally millions of hypotheses can be tested using an analytical tool. Sponsors will receive access to the full set of data and the analytical tool; a summary of the main conclusions is freely available at *http://sic.pub/discover*. Here are some of the findings:

Headlines

- A&Is show decline in search importance, but still #1 in aggregate in STEM across all sectors
- Academic researchers rate library discovery as high as A&Is (in high-income countries)
- Academic researchers rate Google Scholar #1
- Over half of article downloads are free versions — PubMed Central a major factor
- ToC alerts in decline
- Increased role for social media in discovery

Further observations

- Many free discovery resources, like PubMed and Google Scholar, are used less in poorer countries.
- Use of mobile devices is increasing, but smartphone use remains marginal in most territories: greatest use is in low-income countries.
- Publisher Websites are becoming a more popular place to do a search.

It is important to note that publishers always report that they receive more traffic from Google than from Google Scholar, but traffic from Google Scholar typically comes from link resolvers which are not the original source of the research. Understanding the origins of reader navigation helps publishers, libraries, indexing organizations, and technology companies optimize their products for different sectors across the world.

Globalization and Internationalization of Content

James Testa, VP Emeritus, **Thomson Reuters**, noted that the ten countries with significant growth in journal coverage in the Web of Science (WoS) database each added at least 40 journals to their coverage in the last ten years. China, Spain, and Brazil have all increased their coverage; coverage of Turkish journals has grown from 4 to 67 journals; and China has quadrupled its annual output to about 275,000 articles in 2015. Here are his interpretations of this data:

- Many obscure journals were revealed with the introduction of the Internet, and the WoS user base became more internationally diverse, so **Thomson Reuters** began to add more journals to its coverage.
- Australia is ranked first of those journals by citation impact, probably because its journals all publish in English.
- By citation impact, Chinese journals rank in 7th place, which is an indication of their lower quality. Chinese authors receive rewards for publishing in journals covered by the WoS, but their articles tend to be shallow in inventiveness and originality.
- Serious side effects of globalization include the practice of rewarding scholars disproportionately for publishing in high-impact journals, a lower regard for peer review, and excessive self-citations.
- Major progress in communication of scholarly results has been achieved, but efforts to gain higher rankings by unethical behavior are suspect.
- Secondary publishers must demonstrate that their procedures to remove questionable entries from their publications are effective.

Stacy Olkowski, Senior Product Manager at **Thomson Reuters** said that patents contain extremely valuable technical information; the claims are like recipes in a cookbook. Some 70% of the information in patents cannot be found anywhere else in the research literature. Patents are more than just technical documents and can provide answers to marketing and business information questions. Focusing on China, **Olkowski** said that there has been an incredible growth in the numbers of papers being published in Chinese journals, and the same trend has occurred with Chinese patents. The Chinese government initially questioned whether they should establish a patent office, but they did so in 1984, and now it is first in the world in numbers of patent applications with an annual growth rate of about 12.5%. About 85% of the applications to the Chinese patent office are from Chinese nationals. Since 2003, the Chinese government has been paying people to apply for patents.

Donald Samulack, President of U.S. Operations for **Editage** (*http://www.editage.com*), presented a concerning picture of the globalization of the Chinese published literature. He said that Western publication practices have typically been built on trust and rigorous peer review, but there has been a tsunami of articles from China, and there is an entrepreneurial element of commerce in every part of Asian society, which has led to an erosion of this trust and honesty. There are irresponsible and in some cases predatory commercial elements in Asia that prey on facets of the Chinese publication process, such as authorship for sale (see *http://scipaper.net*), plagiarism, writing and data fraud, paper mills,

continued on page 69

hijacked and look-alike journals, and organizations that sell fake impact factors and misleading article metrics. Without appropriate guidance regarding publication ethics and good publication practices, Chinese researchers fall prey to these scams.

In response to unethical practices that led to the retraction of many articles by Chinese authors, the China government has recently issued a policy of standard conduct in international publication. According to **Samulack**, some scientists have been removed from their academic positions and forced to repay grant money to the government. He and others have proposed the formation of a Coalition for Responsible Publication Resources (CRPR, *http://www.rprcoalition.org/*) to recognize publishers and vendors that are "vetted as conducting themselves and providing services in alignment with current publishing guidelines and ethical practices, as certified through an audit process..." so that authors can readily identify responsible publication resources. Articles describing efforts to combat unethical publication activities been published in *Science, Nature*, and on the **Editage** Website (*http://www.editage.com/insights/china-takes-stern-steps-against-those-involved-in-author-misconduct*).

Shark Tank Shoot Out

The final day of the **NFAIS** meeting began with a very informative session in which four entrepreneurs briefly described their companies and products and then were subjected to questioning by three judges: **James Phimister**, VP, **ProQuest Information Solutions**; **Kent Anderson**, Founder, **Caldera Publishing Solutions**; and **Christopher Wink**, Co-Founder, **Technical.ly**. The questions were very intense and probing and mainly centered around the companies' business models (they reminded me of the process one goes through when taking oral exams for a PhD degree!). Here are summaries of the four companies.

- **Expernova** (*http://en.expernova.com/*) finds an expert to solve a problem by accessing a database of global expertise that contains profiles of 10 million experts and 55 million collaborations.
- **Penelope** (*http://www.peneloperesearch.com/*) reviews article manuscripts, detects errors such as missing figures or incorrect references, and checks for logical soundness, statistics, etc. to eliminate errors and shorten review times.
- **Authorea** (*https://www.authorea.com/*) is collaborative writing platform for research. It contains an editor for mathematics equations, makes it easy to add citations, comments, etc., and provides 1-click formatting to create and export a PDF of the completed article. A selection of published articles written on Authorea is available at *https://www.authorea.com/browse*.
- **ResearchConnection** (*https://researchconnection.com/*) is a centralized database of university research information that allows students to search for prospective mentors and is searchable by location, university, and subject. Its target market is the top 200 U.S. universities and 3 million students seeking advisors and applying to graduate schools.

At the end of the session, the judges declared **Authorea** the winner of the shoot out because of its network potential, freemium business model, and likelihood of attracting investors.

Leveraging Data to Build Tomorrow's Information Business

Marjorie Hlava, President, **Access Innovations**, said there are three levels of artificial intelligence (AI):

1. Artificial narrow intelligence is limited to a single task like playing chess;
2. Artificial general intelligence can perform any intellectual task that a human can; and
3. Artificial superintelligence, in which the machine is smarter than a human, is the realm of science fiction and implies that the computer has some social skills.

The major AI technologies include computational linguistics, automated language processing (natural language processing, co-occurrence, and

continued on page 70

The Miles Conrad Memorial Lecture

Long time attendees at **NFAIS** annual meetings will know that the **Miles Conrad Memorial Lecture**, given in honor of one of the founders of **NFAIS**, is the highest honor bestowed by the Federation. This year's lecturer was **Deanna Marcum**, Managing Director of Ithaka S+R (*http://www.sr.ithaka.org/*), who was previously Associate Librarian for Library Services at the **Library of Congress**. She presented an outstanding and challenging lecture on the need for leadership changes in academic libraries in today's digital age. The complete transcript of **Marcum's** lecture is available on the **NFAIS** Website at *https://nfais.memberclicks.net/assets/docs/MilesConradLectures/2016_marcum.pdf*.

Deanna Marcum

Marcum said we have moved beyond simply providing support for searches and are educating students on Web technologies; all libraries are digital now, so they have become leaders in the digital revolution. But there is more to do; academic libraries must make dramatic changes, and a different kind of leadership is necessary, especially at the executive level.

According to **Marcum**, most academic library executives have at least one foot in the print world and have been trained to focus on local collections. However, a national and global mindset is essential, which requires a different kind of leadership. **Marcum** applied 10 practices of digital leaders that were found in a study of successful digital organizations[3] to the library profession:

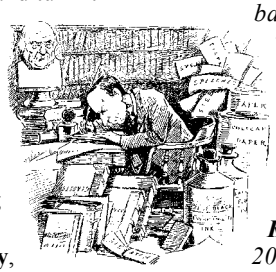
1. Build a comprehensive digital strategy that can be shared repeatedly. Users need immediate access to electronic information.
2. Embed digital literacy across the organization. Librarians must know as much about digital resources as they do about print ones.
3. Renew a focus on business fundamentals. We must integrate digital and legacy resources to give currency to our mission.
4. Embrace new rules of customer engagement. Users are now in control and can decide what is most important and how much it is worth.
5. Understand global differences in how people access and use the internet. We must provide services to a widely diverse population.
6. Develop the organization's data skills. Leaders must rely on data-driven decisions instead of past practices.
7. Focus on the customer experience. Design services from the customer's perspective; there is no "one size fits all."
8. Develop leaders with skill sets that bridge digital and traditional expertise. Help staff on both sides of the digital divide see the value the other brings.
9. Pay attention to cultural fit when recruiting digital leaders. Minimize silos and focus on customers. Empower leaders who can advance digital objectives in an inspirational rather than a threatening way.
10. Understand the motivations of top talent. Make it attractive to remain with the organization by making sure that there is excitement in the library.

Libraries are at a pivotal point now, and survival depends on becoming a node in a national and international ecosystem. Information needs are enormous and vast; digital technology has opened the doors for us.

inference engines, text analytics, and automatic indexing), and automatic translation. Semantics underlie all these systems which work more accurately with a dictionary or taxonomy.

Access Innovations is pushing the edges of AI and is developing practical applications for publishers. Support for Level 1 AI includes concepts, automatic indexing, and discovery. Semantic normalization tells us what the content is about, so we can now issue verbal commands, retrieve relevance results, filter for relevance to the requester, and sometimes give answers.

Expert System (*http://www.expertsystem.com/*) develops software that understands the meaning of written language. Its CEO, **Daniel Mayer**, said that publishers have enormous archives of unstructured content and are looking for ways to exploit it and turn it into products. They want to help users find information faster and easier, focus on the most relevant content, find insights, and make better decisions. Faceted search, a recurring feature of online information products is supported by taxonomies and offers users an efficient way to access information. Content recommendation engines let users discover things unknown to them using AI technologies. The end goal is to provide a faster way of getting to an answer, not just to the content.



C. Lee Giles, Professor at **Pennsylvania State University**, defined scholarly big data as all academic or research documents, such as journal and conference papers, books, theses, reports, and their related data. The CiteSeer^{X} system (*http://csxstatic.ist.psu.edu/about*) has a digital library and search engine for computer and information science literature and provides resources to create digital libraries in other subjects. It can extract data from tables, figures, and formulas in articles.

Closing Keynote: AI and the Future of Trust

Stephane Bura, Co-Founder, **Weave** (*http://www.weave.ai/*) said that trust is a guiding principle and will have the most impact on our information systems. He presented illustrations in the context of video games, which are designed to cater to players' emotions by using their motivations. *Extrinsic* motivations come from outside of us; we experience them when we choose to use a service. But the real motivations that drive us are *intrinsic*:

Stephane Bura

- *Mastery*: the desire to be good, or competence,
- *Autonomy*: the desire to be the agent in your life, set your goals, and reach them, and
- *Relatedness*: the desire to connect and find one's place in the community.

Photos of some of the attendees at the meeting are available on the **NFAIS** Facebook page. The **2017 NFAIS** meeting will be in Alexandria, VA on February 26-28, 2017.

Donald T. Hawkins *is an information industry freelance writer based in Pennsylvania. In addition to blogging and writing about conferences for* ***Against the Grain****, he blogs the* ***Computers in Libraries*** *and* ***Internet Librarian*** *conferences for* ***Information Today, Inc. (ITI)*** *and maintains the* ***Conference Calendar*** *on the* ***ITI*** *Website (****http://www.infotoday.com/calendar.asp****). He is the Editor of* ***Personal Archiving*** *(Information Today, 2013) and Co-Editor of* ***Public Knowledge: Access and Benefits*** *(Information Today, 2016). He holds a Ph.D. degree from the* ***University of California, Berkeley*** *and has worked in the online information industry for over 40 years.*

Endnotes

1. See **Federer's** article, "Data literacy training needs of biomedical researchers," *J Med. Libr. Assoc*., 104(1): 52-7 (January 2016), available at *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722643/*. Also see *http://data.library.virginia.edu/data-management/lifecycle/*, which describes the data management lifecycle and roles librarians can play.
2. "How and why you should manage your research data: a guide for researchers," *https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data*.
3. *http://www.slideshare.net/oscarmirandalahoz/pagetalent-30-solving-the-digital-leadership-challenge-a-global-perspectives*