Don's Conference Notes

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

Information Transformation: Open. Global. Collaborative: NFAIS's 60th Anniversary Meeting

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* ***https://against-the-grain.com/2018/04/nfaiss-60th-anniversary-meeting/****. — DTH*

In 1958, **G. Miles Conrad**, director of **Biological Abstracts**, convened a meeting of representatives from 14 information services to collaborate and cooperate in sharing technology and discussing issues of mutual interest. The **National Federation of Abstracting and Indexing (now Advanced Information) Services (NFAIS)** was formed as a result of that meeting. Today, **NFAIS** is a diverse community that networks and develops forward-looking information services and products.

This year's **NFAIS** meeting in historic Alexandria, VA on February 28-March 2 celebrated the diamond anniversary of the organization. It drew about 140 attendees and featured not only the usual plenary presentations and the traditional lecture in honor of **G. Miles Conrad** (see sidebar), but also a "startup shootout" in which representatives of four industry startup companies pitched their products and business models to a panel of judges, and a series of six 6-minute "lightning talks" on a variety of current technologies and issues.

Opening Keynote

Will We Still Recognize Ourselves? Identity and Community in a Transforming Information Environment

Cameron Neylon, Professor of Research Communications, **Curtin University**, wondered if we will still recognize ourselves in an information environment where everything is rapidly changing. Are we sure who we are any more — publishers, providers, scholars, researchers, aggregators, etc? (**Neylon** wondered if he might be the last of a generation that remembers physically going to a library.)

Cameron Neylon

By 1964, we were in the midst of an information flood, and we had no way to deal with it. The answer was found in technology; if only everyone would use a standard format for the information, we would all be able to process it, and if we could build an open network of standardized information flows, it would be a simple task to get everyone to adopt it. Unfortunately, that did not happen because people like to do things the way they have always done them.

Neylon drew attention to these two important assertions:

1. Knowledge grows and matters.[1] We have an interest in ensuring that this growth continues.
2. Knowledge is made by groups. Individuals have ideas but until they are shared, they will not spread.

Robert Maxwell saved scholarly publishing because he brought a scale of distribution to it, significant growth resulted, and scholarly publication was finally able to make money. Many discussions about the future end with the question "How will we ever keep track of such a large project?"

The next big problem we face is one of *trust*. How do we know what we should use in a body of information? What is "scholarly" culture? Even in the earliest days of scholarly publishing, the idea of reproducibility was appearing. We must continue to open up our institutions or communities to include people that are under-represented. The more boundaries we can cross and the more general or powerful the knowledge under development is, the better will be the community identity. We need to build institutions that support productive conflict and a culture of openness and result in a network that we can all trust. Who might we be well placed to do that? **Neylon** suggested that **NFAIS** members are.

Plenary Sessions

Regina Joseph, founder of **Sibylink** (*http://www.sibylink.com/*) and co-founder of **pytho** (*http://www.pytho.io/*), consultancies that specialize in decision science and information design, said that we are gatekeepers of knowledge and information. Information has never been more accessible, in demand, but simultaneously under attack. There is both a challenge and an opportunity in information system availability and diversity. News outlets have become organs of influence, and social networks are changing our consumption of information (for example, 26% of news retrieval is through social media). We are willingly allowing ourselves to be controlled. How will we be able to harness the advantages of open access to information when the ability to access it might be compromised? We need people with multiple areas of specialist knowledge but who are also connected with broad and general knowledge.

Driving How We Do Business

Jason Priem, Co-founder of **Impactstory** (*https://impactstory.org/*), spoke on thriving in an era of ubiquitous open access (OA), as illustrated by lessons learned in the development of the **Unpaywall** system. He noted that OA is the new default model for information access, and the number of OA articles has grown significantly in the last 18 years. (In the mid-1990s, 75% of articles published were behind paywalls, but by 2015, nearly half of them were OA.) Based on current growth, by 2040, all articles published will be OA. Even now, the most widely used articles are likely to be published as OA. The value is in moving from articles to groups of articles.

Unpaywall is a system that gathers data on the articles that people actually read and cite. OA articles lurk in thousands of journals; **Unpaywall** aggregates them and makes them accessible. It contains data for every **Crossref** DOI (95 million articles), features 98% accuracy (compared with 75% for Google Scholar), and is being used by many systems and link resolvers.

Sari Francis, Manager, Digital Licensing Compliance, **IEEE**, reviewed the impact of digital piracy on publishers and libraries. Piracy is a threat to the entire publishing community, and Sci-Hub is the biggest global threat. **IEEE** has taken a leadership position in detecting and combatting this piracy.

Here are several harmful results of using Sci-Hub:

- Declines in the usage of publishers' products lead to cancellations, so publishers will not be able to support their mission.
- Illegal sites may acquire users' personal information.
- Users of Sci-Hub may not be getting accurate or complete information.
- Usage statistics become skewed.

IEEE has taken these actions in response to Sci-Hub:

- Tracking and alerting users.
- Requesting compromised users to provide data on the sites they accessed. Over 100,000 IP addresses have been collected and put into an "IP intrusion" database.
- Collaborating with other publishers to stop misuse of their data.
- Educating libraries, users, and other publishers about their efforts.
- Partnering with the academic community to protect electronic resources and personal data.

Jan Reichelt, Co-founder, described **Kopernio.com** (*https://kopernio.com/*), another way to provide rapid and convenient access to publishers' websites and databases. Most users simply want a PDF of an article (the version of record) and resist logging into different systems, each with different password, which may take several clicks and two to three minutes, resulting in frustration and turning to sites like Sci-Hub.

Reichelt estimated that by making access convenient for users, we have an opportunity to delight them 2.5 billion times a year. The concept

continued on page 69

of **Kopernio** is to provide access with a single click and give them a consistent user experience, thus increasing the reach of the publisher, journal, and article. **Kopernio** is a win for the user and also for the publisher because content gets out into the community and is used. *[Editor's note: **Kipernio** has just been acquired by **Clarivate Analytics**.]*

Shirley Dexter-Locke, Publishing Director, **Social Sciences Research Network (SSRN)**, **Elsevier**, discussed the role of preprints. She said that the scholarly world is under intense pressure to produce research that is open, accessible, collaborative, measurable, useful, and quickly shared. Preprints (articles written but not yet published in a peer-reviewed journal) have existed for a long time, but they have recently become more recognized as acceptable proofs of research in scholarly publishing. They also have the advantage of exposing early stage research. Published journal articles are like gourmet meals that have been perfectly prepared; preprints are convenient like food trucks. Users want both options. Even though preprints are not peer reviewed, there is a place for them in disseminating information. Sharing early stage research globally helps authors demonstrate their productivity, establish priority of their discoveries, and obtain feedback from other authors.

Thursday Plenary Session

Transforming a Backward Business Model in a Fast Forward World

Ralf Schimmer, Head of Information, **Max Planck Digital Library**, presented a strong denunciation of today's paywall system. He likened today's whirlwind of change, velocity, and turbulence in the information industry to a storm and noted that like many storms, there is an eye where there is no change. For us, that is a stagnant paywall system. Even after 15 years of OA, the paywall system has been largely unaffected. It is the antithesis to a world of openness, comes with excessive costs and outrageous price increases, restrictive copyright, and budgets still based on print legacies. It is a barrier to digital publishing, an avenue for piracy, and hinders use, reuse and interoperability.

Ralf Schimmer

Workaround systems like **Unpaywall** and **Kopernio** are relieving the symptoms and helping to make our lives easier, but they are not curing the cause. Real innovation will come only in an open environment. **Schimmer** said that Sci-Hub and RA21 are the evil twins of the system. They are symptomatic of a dysfunctional and decaying system. RA21 is unneeded and unwanted. OA is the only legitimate resource access at the present time. The principle of openness must be adopted as the default in the scholarly communication system. For that, *the paywall must come down*! It is the primary roadblock to openness, innovation, and sustainability. Do we really think that we can do business as usual? The underlying business model of today's information industry must be changed and reorganized, and journals flipped to a truly open model. It is time to unplug the paywall system, leave the subscription model behind, and find new ways to finance the system.

OA2020 (*https://oa2020.org*) is a global alliance committed to accelerating the transition to OA. **Schimmer** estimates that there is more than $5,000 per research paper being spent currently in the worldwide publishing model, which could be made available to support a move to **OA2020**. Over 100 institutions, including all German research organizations, have already signed an Expression of Interest in **OA2020** (the list is available on the website).

Examining Models in Support of Research

Impact of OA on Access and Subscriptions

Michael Levine-Clark, Dean of University Libraries, **University of Denver**, asked what OA means to library budgets and workflows and how open is the scholarly literature. Journal subscription models are generally based on print spending (often over a period of time) and are often negotiated at a consortium level. As a result, it has become difficult to understand costs at a single journal level and even more difficult at an article level. Cost per use pricing assumes that all use is good, which may not be true. It is hard to move beyond subscriptions, so it is time to move to a new model linked to outcomes.

Even though there is an uneven distribution of publishing and subscriptions, can we flip journals to OA? **Levine** presented several considerations:

- If someone pays for an article to be open, it should benefit all subscribers, but everyone's subscription costs are different.
- If one library in a consortium subscribes to a journal and another one does not, how can the discount for the subscribing library be calculated?
- As more journals become OA, the institution may reduce the library's subscription budget? Is this a bad thing?
- What is the resulting role for library discovery? Will we pay more for it?

One possible solution to the problems of transitioning to OA is to renew big deals, and devote a portion of the costs to opening up all articles to authors at subscribing institutions.

Startup Shootout

Three startup company representatives made 10-minute presentations of their products and business plans to a panel of three judges. The contestants were:

- **David Celana**, **Science Prose on-Demand** (POD, *https://sciencepod.net/*), a cloud-based digital content creation system that translates articles into abstracts that tell stores.
- **Mads Holmen**, **Bibblio** (*http://www.bibblio.org/*). **Bibblio** helps solve the discovery problem by recommending the right content to the right person at the right time. The system generates metadata, models topic, and clusters content. **Bibblio's** content recommendation system increases engagement on web pages by suggesting other relevant content from across the user's site.
- **Craig Tashman**, **LiquidText** (*http://liquidtext.net/*). Research is the heart of knowledge work. It has been estimated that 40% of a researcher's time is spent reading and analyzing documents, and 80% of knowledge workers want to print their documents and read them. **LiquidText** is a platform allowing knowledge workers to make reading more efficient and enjoyable, helping them to understand what they are reading. It collects information from diverse documents, creates a personal semantic web, and then creates an aggregated document for the user.

The winner of the shootout was **LiquidText**.

Members-Only Lunch

Creating Connections and Tying Preprints to NIH-Funded Research

One of the benefits of **NFAIS** membership is admission to a special lunch presentation during the annual conference. This year's speaker was **Neil Thakur**, Special Assistant to the **National Institutes of Health (NIH)** Deputy Director for Extramural Research. He began by noting that we have much more research that can be funded, and there is strong competition among researchers for grants, which is distracting them from doing their research. Fortunately, recognition is growing that preprints may be an answer to this problem because they speed the dissemination of science, even though they may report only interim results. In some disciplines, particularly the sciences, preprints have been in use for years. **NIH** has therefore changed its policies and, because of its interest in fostering a stable infrastructure to advance science, is now permitting citation of preprints in grant applications, providing that the preprint is permanent (i.e., has a DOI), contains a statement that it has not been peer reviewed, and acknowledges its funding source.[2] This policy change has been welcomed by scientists.

Neil Thakur

The current publication tracking structure could be applied to grants and contracts to help solve some of these problems, and **NIH** is therefore considering assigning DOIs to grants.

continued on page 70

Innovative Library Projects Impacting Scholarly Communications: We Are the Change We Want to See

Carl Grant, Associate Dean and Chief Technology Officer, **University of Oklahoma Libraries**, said that we need to move up the value chain. Libraries are being squeezed in ways that hurt their abilities to serve. We must do something about this.

Major changes happened about 11 years ago. **Grant** quoted *Thank You For Being Late*, by **Thomas Friedman**[3] and showed an amazing list of many of those changes, which include such familiar systems as the iPhone, Kindle, Android operating system, and Twitter. He said that it was one of the greatest leaps of technology in history.

The average laptop can store about 500gB of information, but the average person has only 1gB! We are therefore totally dependent on tools to filter information and get what we need. Information is at commodity status; additional value is locked away in database silos, document containers, behind paywalls, in legal contracts and restrictions, and difficult access protocols.

Many people do not know how to unleash the value of eBooks. Unless we can see our future in a broader context, we may not have a future. The cost of information is now far exceeding its value, because of OA initiatives and OER initiatives. To create new value, we need to move from seeing information as the source of value to unleashing its full potential via virtual tools and physical spaces. Academic libraries are the best environments for this.

Ken Parker, Co-founder and CEO of **NextThought** (*https://nextthought.com/*) said that connections are everywhere, and we must use them. Connections in nature range from the minute (nerve cells, 10^{-7} meters apart) to the vast (the cosmic web, 10^{25} meters). There are similarities in these widely varying dimensions. Our brains are made to connect; how do we harness that to let people connect and add value to information?

- Sharing: Everyone creates and it is all shared: code, knowledge, judgement, photos. Facebook is the top consumer for human attention and now has over two billion users, growing at 17% year over year.
- Education: Education is less hierarchical because knowledge comes from everywhere. Technology increases access by removing much of the friction. Now everyone can get content and people are connected with resources. Education leverages connections and increases information value

David King, Founder and CEO, **Exaptive** (*https://www.exaptive.com/*) discussed how to increase researchers' productivity and thus increase value. The most successful innovation engines of today are the Web, crowdsourcing, and incubation labs. Cities are where innovation develops, and the internet creates a virtual city. Innovation does not happen unless people find common ground to connect.

Friday Plenary Sessions

The Next Big Paradigm Shift in IT: Chatbots and Conversational Artificial Intelligence (CAPs)

Matthew Devapiryam, Director of Technology, **ProQuest**, said that **ProQuest** curates content, simplifies workflows, and connects communities, so that finding answers and deriving insights is straightforward and leads to extraordinary outcomes. Everyone is now talking about digital reinvention, creating new experiences, and disrupting business models. Although many processes are expected to occur without any direct human-to-human interaction (for example, **Devapiryam** suggested that by 2020, 80% of the buying process will occur this way), it is important to recognize that some things must be the province of humans, such as common sense decisions, morals, immigration issues, compassion, or abstractions. Other operations such as locating knowledge, pattern identification, natural language, and machine learning can be done by cognitive systems. Uses for conversational agents include customer service, mobile apps, messaging channels, the internet of things, and robots.

A chatbot is conversational software that can be interacted with using text. **ProQuest** has developed a beta version of a chatbot, Aristotle, to interact with its data conversationally, asking questions such as "What was our revenue last year?" Aristotle is now being used with internal **ProQuest** sales data, and it will eventually be made available to customers.

Big Data and AI Technologies Coming of Age

Ruth Pickering, Co-founder of **Yewno**, said that AI stands in a long line of human innovation to help people find what they are looking for, but every time progress is made, we encounter huge challenges. The first one is knowing how to look; many times it is necessary to know what you are looking for. In an AI environment, one needs less advance knowledge; AI can help find connections to information, making researchers more productive, starting by applying filters and eliminating irrelevant information, but as the information is narrowed down, the context may be lost. We cannot talk about the future of information without introducing the tools powered by AI. For example, a book must be looked at on a chapter level because frequently, there is information in chapters which is outside the domain of the book.

Transforming 50 Years of Data: A Collaborative Approach to Creating a New Revenue Stream

Jonathan Griffin, Head of Product Development, **IFIS Publishing**, and **Jignesh Bhate**, Founder and CEO, **Molecular Connections (MC)**, collaborated in this presentation to describe how **IFIS** and **MC** worked together to add new value to data and create new market segments. **Griffin** noted that about 50 years ago, two trade associations founded the **International Food Information System (IFIS)** to produce an A&I database on food and nutrition that is largely used by academic and corporate organizations. Recently, usage has not increased because younger researchers prefer to use Google Scholar. **IFIS** staff are specialists in content, not technology, so they formed a very successful partnership with **MC**. The relationship started with indexing the database and then grew to workflow solution development and new products.

Bhate described market research which revealed that **IFIS** users were having difficulty finding important food regulations from different geographical areas. The research also indicated that people wanted to receive their information in different ways. A new product, Escalex, was created by combining **IFIS** legacy content with information freely available on the web. Domain expertise is critically important to ensure that the quality of the content is not compromised.

Lightning Talks

Each participant in these six Lightning Talks had six minutes to present a focused presentation on a critical issue of interest to them.

Libraries are Really AI Services: Improving Discovery Access to Library Special Collections — Presented by **Marjorie Hlava**, President, **Access Innovations, Inc.**

Libraries of today have an emphasis on storage, not discovery and retrieval. The challenge of discovery was exemplified in a project done with the **Smathers Libraries** at the **University of Florida** to create the Portal of Florida history. Over 14 million pages on microfilm were digitized, increasing access to the digital collection. Improvements included:

- Use of XML instead of MARC headings,
- 23 fields instead of 900 in the MARC system,
- Implementation of the JSTOR Thesaurus instead of Library of Congress Subject Headings,
- Metadata records created in an XML Intranet System which could then be exported as needed. (Catalogers became metadata librarians.)
- Installation of new tools: Data Harmony XIS and MAIstro.

Through application of automated processes such as digitization and OCR that supported indexing, the increase in search accuracy was persuasive and impressive.

Rapid Digital Product Innovation — Presented by **Michael Cairns**, CEO, **Digital Transformation**

Digital Transformation helps companies make transitions from legacy environments and execute on their strategies. Problems have arisen when executions do not deliver on the strategy because of lack of clarity, silos, lack of transparency, shifting priorities, and accountability. The dPrism system offers a solution and provides a team to guide people, processes, and technology.

continued on page 71

Open Infrastructure: Come On In, The Water's Fine — Presented by **Jennifer Kemp**, Head, Business Development, **Crossref**

When an infrastructure is open, everyone can benefit from it; the process is no longer "DOI and Done." Event data like blog posts, feeds, etc. and other new content types can added to a system as useful enhancements. Metadata is powerful for libraries, authors, and editors.

CHORUS Institution Dashboard Services: A Collaborative Solution to Article Access — Presented by **Susan Pastore**, Director of Business Development, **CHORUS**

CHORUS is creating a future where the output from funded research is easily and permanently discoverable, accessible, and verifiable by anyone in the world. It maximizes interoperability by employing widely used standards and infrastructure, and supports funder policies, OA business models, and diverse publishing systems. The CHORUS dashboard is used by authors to help institutions comply with funding requirements. Lessons learned from pilot projects:

- Accurate article metadata can be hard to obtain.
- Linking authors to a university can be a complex problem.
- Faculty research is being deposited, but it is not necessarily compliant with funding agency requirements. Researchers need help in this.
- Preservation in perpetuity has value.
- Researchers are confused by their usage rights and funder obligations.

CHORUS works with Scopus and the Web of Science — see *https://www.chorusaccess.org/*.

The New Dimension in Scholarly Communication: How a Global Scholarly Community Collaboration Created the World's Linked Research Knowledge System — Presented by **Ashlea Higgs**, Founder, **ÜberResearch**

Higgs described recent work with the Dimensions system in which over 100 global research institutions have been collaborating to solve challenges in the existing research environment. Dimensions is no longer just a database of awarded grants; it now includes patents, articles, and over 4 billion links between them. Further information on this work is available on Figshare[4] and at *https://app.dimensions.ai/discover/publication*.

Flipping the Script — Presented by **Joseph Lerro**, OA Sales Executive, Routledge, **Taylor & Francis (T&F) Group**

T&F has embarked on a project to convert, or "flip" scholarly hybrid titles to full OA. Twenty subscription titles and over 70 society-owned journals were flipped in 2017. The goals of the project are to accelerate OA conversion and make it as easy as possible for researchers, societies, institutions, funders, and governments to achieve their open scholarship aims.

Closing Keynote

Academic Publishing, Blockchain and Shifting Roles in a Rapidly Changing World

Joris van Rossum, Director, Special Projects, **Digital Science**, said that is amazing to realize how fast things are changing; for example, **Elsevier** used to be no more than a printing press. Our purpose is to do what scientists cannot do or do not wish to do themselves. Publishers are managing reputations by

Joris van Rossum

- Registration: Enhancing the author's precedence and ownership of an idea,
- Certification: Ensuring quality control by peer review,
- Dissemination: Communicating the findings to the relevant audience, and
- Preservation: Preserving a fixed version for a future audience and citation.

We are successful not because of our technology but in spite of technology. But we are not technology companies! Now, alternatives to these roles are emerging, and our roles are getting smaller.

Certification is increasingly relevant because of:

- Reproducibility (only one-third of published articles can be reproduced).
- Peer review crisis: transparency and recognition. People are not recognized for peer reviewing; they are recognized for what they published. Peer review is a legacy of the print era. Fraud and manipulation, and lack of transparency and trust in the process are issues.
- Limited and outdated metrics. Everything that went into the research is not known which causes many problems.

Blockchain and scholarly communication

Blockchain is the technology behind cryptocurrencies and the underlying technology of bitcoin.[5] There is a connection between payment and money flows to scientists. Processes are not rewarded; can we introduce a cryptocurrency for science, a token which one can use to buy services? Some initiatives are emerging to create tokens to reward researchers.

Blockchain lets us move from the Internet of information to an Internet of value. It establishes ownership, prevents double spending, and is therefore the ideal technology for DRM. Micropayments, which are presently very expensive, open the way for a new business model. On today's Internet, links point to the content but not back to the owner. Blockchain is a very special kind of data storage and a new way to think about databases. They can be decentralized (everybody can have a copy), shared and immutable (they cannot be changed), and transparent but pseudonymous (real identities do not have to be revealed). We can know everything that is done to the databases by users, and can track the data that the scientist has used to build the database. There is no need for a middleman to do this.

Digital Science is starting an initiative around peer review to develop a complete, authoritative, and decentralized store of reviewer data using the blockchain technology. The data storage will find, validate, and recognize reviewers; respect confidentiality and privacy requirements, and will be shared by everyone. Thus, the process will become more robust and transparent, foster developments of applications to recognize reviewers, and comply with requirements of confidentiality and privacy.

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: Preserving Our Digital Heritage****, (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 45 years.*

Endnotes

1. See the work of **Derek deSolla Price** in the 1960s in *Big Science, Little Science.*
2. See *https://grants.nih.gov/grants/guide/notice-files/not-od-17-050.html* for further information.
3. *https://www.amazon.com/Thank-You-Being-Late-Accelerations/dp/1250141222/ref=sr_1_2*
4. *https://figshare.com/articles/A_Guide_to_the_Dimensions_Data_Approach/5783094*
5. **van Rossum** is the author of a report on blockchain: "Blockchain for Research," *Digital Science*, November 2017.

continued on page 72

AI, Scholarly Big Data, and Advances in Information Services: The Miles Conrad Memorial Lecture

This year's **Miles Conrad Memorial Lecture** was presented by **Dr. C. Lee Giles**, David Reese Professor, College of Information Sciences and Technology, **Pennsylvania State University**. He began by defining artificial intelligence (AI) as machines that think, which means understanding and reasoning rationally, making plans and decisions, and following through. AI is also the science and engineering of intelligence which helps us in many ways. In the information service area, AI assists in understanding and communicating our knowledge using automated methods that operate on large numbers of documents and in creating new knowledge in formal data structures. It also includes machine learning and writing and can extract knowledge from scholarly documents.

Outgoing **NFAIS** President **Peter Simon** (R) Presents the **Miles Conrad Award** Plaque to **C. Lee Giles**.

Scholarly big data encompasses all academic and research documents, such as journal and conference papers, books, theses, and technical reports. Much of it resides in large sophisticated networks. It originates from many sources, such as the web, repositories, publishers, patents, and data aggregators, and many people are interested in it, including scholars, economists, educators, and policy makers. (**Giles** said that Science of Science conferences are very interesting and suggested attending one of them.) Although there are many applications of scholarly big data in narrow disciplines, it is limited. Knowledge representation is not universal across disciplines.

Giles and his colleagues did a study to determine how much scholarly big data is available. They estimated that there are at least 114 million articles in English on the web, 24% of which are publicly available. Google Scholar has at least 100 million articles. This research will be extended to distinguish the various types of articles and include languages other than English.

AI and machine learning are used with scholarly big data to extract and link metadata, build knowledge structures, and process natural language queries. **Giles** et al. have developed the CiteSeer^{X} system (*http://citeseerx.ist.psu.edu*) to perform some of these operations on the literature of computer science, such as author searching and name disambiguation, identification of tables in documents and extraction of the data from them, citation indexing, and full text indexing. Its impact has been far-reaching; **Giles** thinks that it is probably the first digital library search engine and has changed the methods of access to scientific research. The system will soon be available through PubMed.

The open source SeerSuite tool kit (*http://citeseerx.sourceforge.net/*) is used to build search engines for digital libraries and includes not only CiteSeer^{X} but also a search engine, Chem$_{X}$Seer, for chemical formulas (which are very different from text); a citation recommendation system (RefSeer); and several other modules. Numerical data in scientific publications are often in figures, charts, or text and must be differentiated from chemical formulas; the SeerSuite system has this capability. Shown here is a typical example:

Challenges in Formula Search

How to identify a formula in scientific documents?

Non-Formula

"... This work was funded under **NIH** grants ..."
"... YSI 5301, Yellow Springs, **OH**, USA ..."
"... action disease. **He** has published over ..."

Formula

"... such as hydroxyl radical **OH**, superoxide O2- ..."
"... and the other **He** emissions scarcely changed ..."

Machine learning algorithms (SVM + CRF) yield high accuracies for correct formula identification.

Giles concluded his lecture by referring to the knowledge graph used by Google to enhance search results (see *https://en.wikipedia.org/wiki/Knowledge_Graph*) and present them to users in the now familiar box on the right side of the results page. He said that the concept of AI as a disruption in information services is largely false; AI will change services and make them easier to use as more resources become available, thus enhancing the productivity of researchers. The challenges are to develop scalable methods for information extraction, search, and knowledge structures for scholarly data. These applications are more common than many people realize and are embodied in these goals of CiteSeer^{X}:

- Create an index of all open scholarly documents,
- Use machine learning and AI to extract all data and obtain semantics,
- Create ontologies and knowledge structures, and
- Integrate all related information in a searchable format.

The overriding goal is to automate everything!

A Few Notes From ER&L

The 13th Electronic Resources & Libraries Conference

Don't miss ***Don's*** *column coming in our June issue. Here's a very brief summary of the speaker presentations at the March meeting in Austin, Texas.*

Distinctions between media companies and platforms have begun to blur, leading to ambiguity in our perceptions of how people are consuming information.

Fake news can be detected by examining the author's intent, the type of information being conveyed, and its features.

How do students do research? One study found that asking them to draw diagrams was very useful.

Metadata is the interface between the user interface to search systems and information literacy, but it is not something that students are familiar with.

The use of online videos has become widespread in libraries. Preservation and licensing aspects are important. The process of acquiring videos involves many steps, which can be time consuming; it is important to recognize this in planning.

Discovery, the ability of users to find content they are looking for, is a shared goal that users, publishers, aggregators, and librarians must work together to achieve.

What has happened to our algorithmic culture in which AI and Big Data have assumed prominence? We need to be able to make sense of the data using these considerations:

- How we got the data,
- Seeing patterns,
- Detecting explicit corruption in the data,
- Addressing values and cultural norms.

How does Facebook impact its users' lives? It promotes nationalism and authoritarianism and scrambles social, economic, and political contexts. News feeds are constantly interrupting us and capturing our attention. It affects our relationships in commercial, personal, and political areas. Facebook's mission went wrong because its leaders believed that faith in technology could generate a better world.