A cooperative for big data in scholarly publishing

Kevin S. Hawkins
@KevinSHawkins

Assistant Dean for Scholarly Communication
University of North Texas Libraries

A vision conceived in collaboration with

- Joyce Chapman (Duke University)
- Sarah V. Melton (Boston College)
- Lucy Montgomery (Curtin University)
- Katherine Skinner (Educopia Institute)

at the 2015 Scholarly Communication Institute (trianglesci.org). Others are now involved as well:

- Peter Berkery (Association of American University Presses)
- Martin Paul Eve (Birbeck, University of London / Open Library of Humanities)
- Christina Drummond (Educopia Institute)
- James MacGregor (Public Knowledge Project)
- Cameron Neylon (Curtin University)
- Lisa Schiff (California Digital Library)

What kind of big data?

Not

- datasets created by researchers
- other types of research outputs sometimes grouped together under “research data”

but rather *big data about published research.*

What kind of big data about published research? (1)

Data generated by publishers and aggregators of content

- **purchasing data**: customer type, number of copies, how much they paid, when they purchased
- **licensing data**: who licenses, how much they pay
- **online usage data / web analytics**: number of hits or visits, demographics of users, types of use (search vs. browse vs. download, part vs. whole)
- **subject classification** of products

What kind of big data about published research? (2)

Data from research institutions

- **Library data**: holdings, circulation, link resolver stats, subject classification
- **Structured productivity data** captured in an online faculty CV system, which may be referred to by any of the following names:
  - current research information system (CRIS)
  - faculty profile system
  - research profiling tool
  - research networking tool
  - research information system
  - research information management system (RIMS)

What kind of big data about published research? (3)

Data from third parties

- from **bibliometrics services**: journal-level metrics, article-level metrics, author-level metrics (including altmetrics)
- from **social networking sites**: Academia.edu, ResearchGate

All of these, like other forms of big data, can be used for various types of assessment and also for *predictive analytics:*

Which publications are most likely to be purchased, used, and cited?

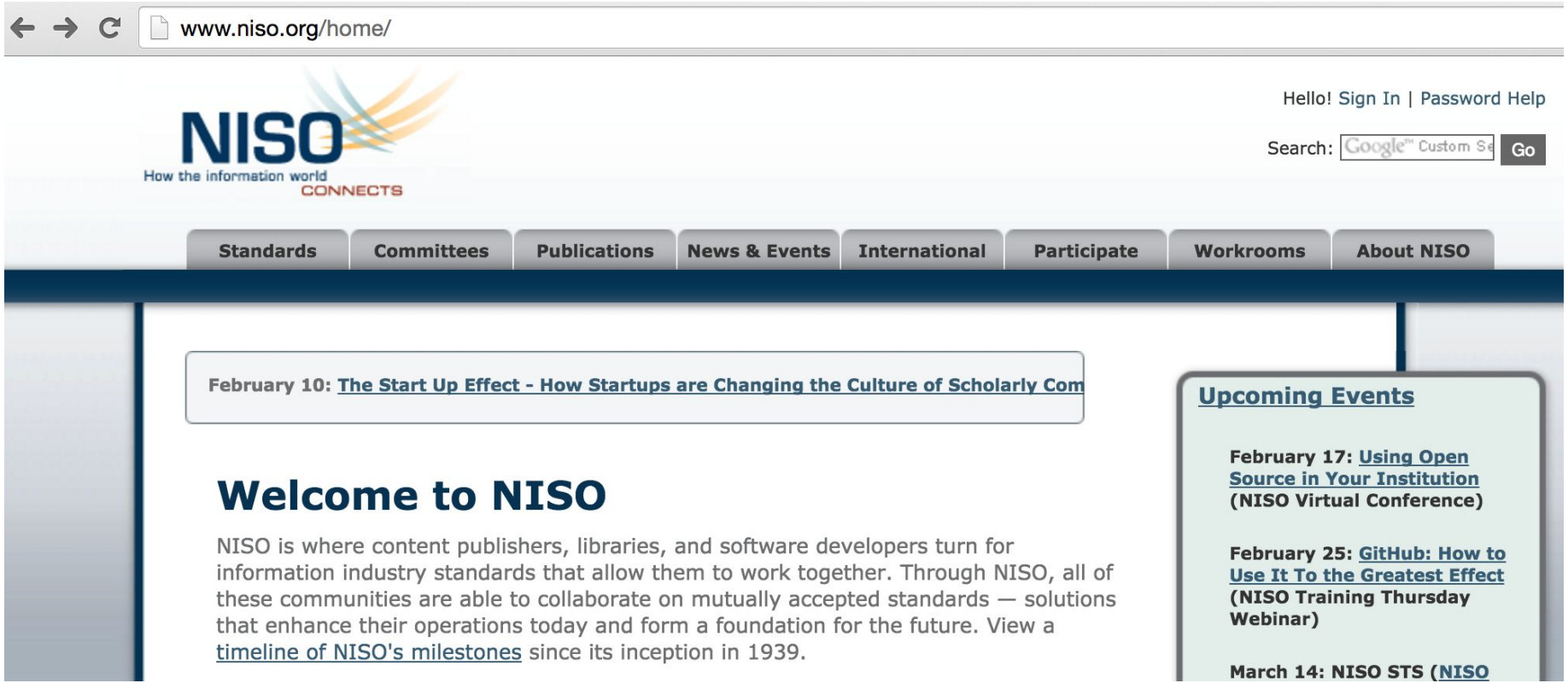
Analytics using big data

Domain	Why some want it	Why others are concerned
Retail: *Which products are currently being purchased and are most likely to be in the future?*	Companies want to gain a competitive advantage in providing and selling products (such as targeted ads).	Consumers are concerned about their privacy.
Publishing: *Which types of publications are purchased, used, and cited, and which are most likely to be in the future?*	● Publishers and aggregators want to develop products that will do better in the market. ● Libraries want to acquire resources that will be used most … or even predict what to acquire.	Researchers are concerned: ● that data is incomplete, allowing the wrong conclusions to be drawn ● that data will be used to justify decisions that don't fairly represent their work

We want the scholarly community to retain control over data related to publishing and be able to exert influence on how it’s used.

We’re going to need

- input and cooperation from all stakeholders in the system
- a neutral group taking on this work


www.niso.org/home/
NISO
How the information world CONNECTS
Hello! Sign In | Password Help
Search:
Go
Standards
Committees
Publications
News & Events
International
Participate
Workrooms
About NISO
February 10: The Start Up Effect - How Startups are Changing the Culture of Scholarly Com
Welcome to NISO
NISO is where content publishers, libraries, and software developers turn for information industry standards that allow them to work together. Through NISO, all of these communities are able to collaborate on mutually accepted standards — solutions that enhance their operations today and form a foundation for the future. View a timeline of NISO's milestones since its inception in 1939.
Upcoming Events
February 17: Using Open Source in Your Institution (NISO Virtual Conference)
February 25: GitHub: How to Use It To the Greatest Effect (NISO Training Thursday Webinar)
March 14: NISO STS (NISO

Recommended Practices

NISO Recommended Practices are "best practices" or "guidelines" for methods, materials, or practices in order to give guidance to the user. These documents usually represent a leading edge, exceptional model, or a proven industry practice. Use of any or all elements of a Recommended Practice is discretionary; it may be used as stated or modified by the user to meet specific needs.

NISO RP-24-2015, Transfer Code of Practice, version 3.0

Published March 2014 by UKSG; NISO RP edition published February 2015

Abstract: This recommended practice contains best practice guidelines for ensuring that journal content remains easily accessible by librarians and readers when there is a transfer between parties. The Code provides guidance to both the Transferring Publisher and the Receiving Publisher to ensure that the transfer process occurs with minimum disruption for libraries, intermediaries (such as serials subscription agents, link resolver administrators, and vendors of large-scale discovery systems), and readers. The Code was originally developed by the UKSG and version 3.0 was published in March 2014. Following agreement between NISO and UKSG for NISO to take over maintenance of the Code, it was republished as this NISO Recommended Practice. Except for the Foreword, the content is unchanged from the UKSG version 3.0.

For release: 07 Jan 2016

NISO Receives Two Grants To Undertake the Creation of a Framework on Data and Privacy

Baltimore, MD - January 7, 2016 - The National Information Standards Organization (NISO) has received two grants to develop a consensus framework for mitigating and managing the privacy risks related to the collection, preservation, sharing, use, and re-use of research data sets. The grants from both the Andrew W. Mellon Foundation and the Alfred P. Sloan Foundation will further advance NISO's existing privacy initiatives on user privacy in library, publisher, and software supplier systems, details of which were published in December 2015.

The 22-month project funded by the Mellon Foundation is called "Development of a Consensus Framework for Mitigating the Privacy Risks Related to the Collection, Sharing, and Use of Research Data Sets." A proposed joint NISO-Research Data Alliance (RDA) working group will develop the framework and associated metadata, use cases, and implementation support materials. Todd Carpenter, NISO Executive Director, will serve as the overall project director and co-chair of the working group. The working group will also be co-chaired by Bonnie Tijerina, a Researcher at the Data & Society Research Institute and founder of the Electronic Resources & Libraries conference. The Working Group Case Statement is currently open and available for comment (log in is required) on the RDA website.

Work on privacy issues surrounding data sets is related to and will build upon ongoing initiatives at NISO and RDA. The framework will fill a need in the international science, social science, and humanities communities, which often work with human subject research data but lack concrete guidelines as to how to safeguard those data as they are shared, used, and processed. Institutional repositories have too often dealt with the question of privacy in data sets by exclusion rather than management. Not only do repository managers risk incurring significant financial penalties in this process, they also create avoidable barriers to data sharing and reuse.

For release: 14 Dec 2015

NISO Releases a Set of Principles to Address Privacy of User Data in Library, Content-Provider, and Software-Supplier Systems

Baltimore, MD - December 14, 2015 - The National Information Standards Organization (NISO) has published a set of consensus principles for the library, content-provider and software-provider communities to address privacy issues related to the use of library and library-related systems. This set of principles developed over the past 8 months focus on balancing the expectations library users have regarding their intellectual freedoms and their privacy with the operational needs of systems providers.

The NISO Privacy Principles, available at http://www.niso.org/topics/tl/patron_privacy/, set forth a core set of guidelines by which libraries, systems providers and publishers can foster respect for patron privacy throughout their operations. The Principles outline at a high level basic concepts and areas which need to be addressed to support a greater understanding for and respect of privacy-related concerns in systems development, deployment, and user interactions. The twelve principles covered in the document address the following topics: Shared Privacy Responsibilities; Transparency and Facilitating Privacy Awareness; Security; Data Collection and Use; Anonymization; Options and Informed Consent; Sharing Data with Others; Notification of Privacy Policies and Practices; Supporting Anonymous Use; Access to One's Own User Data; Continuous Improvement and Accountability.

Independent confirmation

Such as

- ISO 9001 (quality management)
- COUNTER (online usage statistics)

The twin pines symbol, used in North America to represent cooperatives (“co-ops”)

What if

we formed a cooperative of libraries, scholarly societies, publishers, aggregators, and other stakeholders, who would each contribute to the governance of this member organization.

Members contributed data they create about scholarly communication (their small view of the world).

The cooperative, thanks to member fees, had staff and tools to aggregate, normalize, and contextualize this data for its members, showing them how their data relates to that of all members but in a way that adheres to a code of conduct.

Members would have to agree to adhere to the code of conduct in how they use the data that they get back from the cooperative.

educopia.org/research/meerkat

These slides are at
www.ultraslavonic.info/talks/20170416.pdf