A cooperative for big data in scholarly publishing

Kevin S. Hawkins
@KevinSHawkins

Assistant Dean for Scholarly Communication
University of North Texas Libraries

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?

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/20170804.pdf