Web of Science as a Research Dataset

Workshop Organizers:
Katy Börner, Victor H. Yngve Distinguished Professor of Information Science, IUB
Eamon Duede, Executive Director, Knowledge Lab, University of Chicago
James Pringle, Head of Industry Development, Clarivate Analytics

Place and Time:
Social Science Research Commons (SSRC), Woodburn Hall, Room 200
1100 East Seventh Street, Indiana University, Bloomington, IN
November 14-15, 2016

This effort is supported by Clarivate Analytics, the KnowledgeLab at the University of Chicago, the Indiana University Network Science Institute (IUNI), and the Cyberinfrastructure for Network Science Center at Indiana University.

Workshop Goals
This workshop brings together data scientists and data stewards from research centers that are using the Web of Science™ at scale. We will explore WoS from the perspective of a research dataset and work together on practical ways to better support our research in the future. While the main focus will be on the Web of Science, the results should be extensible to all similar metadata aggregations. This unique focus—bringing data stewards and data scientists from these centers together to work on shared needs in tandem with the Web of Science team—will enable us to redefine and fully repurpose WoS to fit our research goals. We intend to launch an ongoing community in which we will learn techniques and develop tools to improve the data that underlies our research.

Advance Preparations
- Data stewards will provide a short profile of how WoS as a dataset is being implemented in the context of their research center/university and the technical, content, and other challenges they are facing.
- Researchers will prepare a short profile of current research projects leveraging the WoS dataset, focusing on key challenges such as linking, disambiguating, mining, etc. that, if solved, would offer greater research opportunities.

Mention the Workshop on Twitter
Tweet #WoSRD

Web site: http://cns.iu.edu/workshops/event/161114.html
Scratch Space: http://bit.ly/2fz3wOJ - please reveal your GitHub username
Schedule:

Monday, November 14, 2016

Morning: Informal meetings will be coordinated around specific interests and arrival times.

12:00pm Arrival, Light Lunch
12:30pm Welcome and Introductions
1:00pm Session 1: Web of Science “Outside the Box”
   Facilitator: Katy Börner
   The Web of Science and similar metadata datasets are housed, maintained, and enhanced in local institutional enclaves. This session looks at how data scientists and managers set up the WoS dataset in an enclave, disambiguate the data (names, institutions, and geolocations), link it to other datasets, and build predictive models.
   - Matt Hutchinson, Indiana University: “Implementing Web of Science data for university-wide research”
   - Yoda Babji, University of Chicago: “Web of Science data for secure, cloud-based cross-institutional access”
   - Orion Penn, École Polytechnique Fédérale de Lausanne: “Managing and exploring Web of Science data at the research group level”
   - Nicholas Morey, University of Massachusetts Amherst: “Entity Resolution and Clustering Author Disambiguation and Related Tasks”
   - Vokie Torve, University of Illinois: “Three applications of integrating PubMed with Web of Science”
   - Lee Giles, Pennsylvania State University: “Web of Science data in the context of open tools and infrastructures”

2:30pm Break

3:00pm Session 2: Breakthrough Research
   Facilitator: Eamon Duede
   A “lightning round” look at examples of new questions asked and answered in recent research across network science, science of science, scientometrics, and science policy. We will explore how WoS and related large-scale datasets enable these new questions to be addressed and whether enhancements to the dataset could have better supported the research.
   - Approximately twelve participants to be confirmed.

2:00pm Wrap Up: Discussion will focus on identifying key data challenges in these areas.

4:30pm Day 2 Preparation:

5:00pm Adjourn
6:30pm Dinner

Tuesday, November 15, 2016

8:00am Breakfast, Initial Breakout Team Meetings

9:00am Session 3: Understanding Web of Science as Research Data
   Facilitator: Jason Rollins
   For over 50 years, the Web of Science evolved as a dataset in response to changing research contexts and priorities. Today, more researchers are using the Web of Science “at scale” to ask and answer powerful new questions about the shape, dynamics, and veracity of science and scholarship. The Web of Science now appears both an object of inquiry in its own right and a vast sensor network for discerning large-scale trends. What is changing in this dataset to support these new uses, and what could change further? Presentations and discussion led by Clarivate Analytics team.
   - Jim Pringle: “WoS Metadata as Research Data”
   - Patricia Brennan: “Web of Science and Open Identifiers”
   - Ling Li: “Data Unification and Disambiguation: Institutions and Authors”

10:30am Break

11:00am Session 4: Hackathon Breakout Sessions
   Facilitators: Eamon Duede, Jason Rollins, Ted Lawless, Others TBD.
   A mix of sessions determined by 3-4 “big questions” prioritized on Day 1, grouped as:
   A. Technical Hackathon(s): Practical Focus on applying code across research centers in such areas as data disambiguation (names, institutions, geolocations), linking WoS data to other datasets, building models to predict gender, ethnicity, etc.
   B. Topical Hackathons: Working across research centers on Authorship & Collaboration; Gender in Science; Topic Modeling and/or other topics defined by attendee interest.
   C. Community Hackathon: Focus on establishing an ongoing community (e.g. setting up an enclave, tools & mechanisms for sharing code, citing and acknowledging contributions, and/or what is appropriate for cross-enclave sharing).

12:00pm Working Lunch

3:00pm Group Discussion
   Facilitators: Katy Börner, Eamon Duede, and Jim Pringle
   Group Discussion: Lessons, and results. What kind of community should we be?

4:00pm Wrap-up
4. Please select valuable activities at this event:

- Gain deeper understanding of Web of Science
- Discuss best practices for Web of Science
- Share data cleaning/integration strategies
- Develop plans for future Web of Science
- General networking and discussions
- Other

5. Topical research issues of high interest to you:

- Authorship
- Name disambiguation
- Scientific collaboration
- Gender in science
- Topic modeling
- Data/text mining
- Linking and data integration
- Funding outcomes
- Science policy analytics
- Innovation policy analytics
- New metrics
- Other
6. What would you most want to come away with from attending the workshop?

**Data Management and Usage**
- A deep understanding of the possibilities and qualities of the WoS dataset, its completeness / limitations, comparison to other data sets (e.g. Google Scholar, or other open databases).
- Best practices for linking and data integration.
- Getting a better sense of the data,
- Author disambiguation data sets.
- Understand what data cleaning, linking, prediction algorithms exist (as open source).
- The publications always cover the analysis and the results, but what's 'under the hood' is almost never discussed, at least in terms of specifics.
- Additional insights into using WoS for author name disambiguation; both for training and evaluation of machine learning approaches for disambiguation.

**Community Building**
- Get to know new people in the field.
- Seeing group consensus on prioritizing data cleaning/integration/disambiguation tasks for the WoS community
- Developing an action plan for divvying up these shared tasks across user groups
- A collective pipeline for cleaning our WOS data
- Community adoption of infrastructure we've developed to enable simple, secure and cost-effective compute on WoS datasets.
- Better understanding of the challenges faced by the community from a compute/data perspective and some discussions on reasonable solutions.
- Create a community of practice to perform collaborative R&D on mining, modelling, and mapping science, technology, innovation.
Research
• New ideas about how to model innovation through WoS data.
• A knowledge of some interesting research done with the WoS dataset.
• Enhanced collaboration opportunities.
• What do the researchers see as strength and places to improve in WOS data.

Commercial Partnerships
• Learning how WoS works and how we might utilize this in our work related to IUB and translational research and commercial partnerships
• More knowledge of how people are using WoS data for research and what tools can be developed to support that.

7. Other thoughts, comments, or suggestions:
• Google docs list of people that are attending the event with GitHub account or link to personal page.
• I am delighted that so many data managers will be able to attend, meet and hack together!
• … discuss the extent of support and cooperation with IU WoS staff they are willing to provide for particular research projects.
• Difficult to select issues for the focused sessions and hackathon. …
• Provide suggested readings/tools/software/data for focused sessions / hackathon a few days in advance so we may do some preparation.
• Regarding questions 4 and 5, I have gone to great lengths to restrict my answers … I could easily have selected all options for both question as our challenges and research questions touch upon every option to some extent or another.
• This will be awesome.
Schedule:
Monday, November 14, 2016
Morning: Informal meetings will be coordinated around specific interests and arrival times.
12:00pm Arrival, Light Lunch
12:30pm Welcome and Introductions
1:00pm Session 1: Web of Science “Outside the Box”
Facilitator: Kaye Börner
The Web of Science and similar metadata datasets are housed, maintained, and enhanced in local institutional enclaves. This session looks at how data scientists and managers set up the WoS dataset in an enclave, disambiguate the data (names, institutions, and geolocations), link it to other datasets, and build predictive models.
- Matt Huchinson, Indiana University: "Implementing Web of Science data for university-wide research”.
- Yada Babuji, University of Chicago: “Web of Science data for secure, cloud-based cross-institutional access”.
- Orion Penner, École Polytechnique Fédérale de Lausanne: “Managing and exploring Web of Science data at the research group level”.
- Nicholas Monash, University of Massachusetts Amherst: “Entity Resolution and Clustering Author Disambiguation and Related Tasks”.
- Vele Torvik, University of Illinois: “Three applications of Integrating PubMed with Web of Science”.
- Lee Giles, Pennsylvania State University: “Web of Science data in the context of open tools and infrastructures”.
2:00pm Wrap Up: Discussion will focus on identifying key data challenges in these areas.
2:30pm Break
3:00pm Session 2: Breakthrough Research
Facilitator: Eamon Dueke
A “lightning round” look at examples of new questions asked and answered in recent research across network science, science of science, scientometrics, and science policy. We will explore how WoS and related large-scale datasets enabled these new questions to be addressed and whether enhancements to the dataset could have better supported the research.
- Approximately twelve participants to be confirmed.