IUNI Web of Science Data Enclave 102

Katy Börner and Robert Light
Cyberinfrastructure for Network Science Center
School of Informatics and Computing and IUNI
Indiana University, USA

Val Pentchev, Matt Hutchinson, and Benjamin Serrette
University Network Science Institute (IUNI)
Indiana University, USA

Thanks go to James Russell and Bill Barnett who served on the original IUNI WoS Data Advisory Board.

October 10, 2016
Data Acquisition

The IUNI Science of Science Hub acquired the complete set of Thomson Reuters’ Web of Science XML raw data (Web of Knowledge version 5) comprising

- Science Citation Index Expanded from 1900-2013
- Social Sciences Citation Index from 1900-2013
- Arts & Humanities Citation Index from 1975-2013
- Book Citation Index -- Science from 2005-2013
- Book Citation Index -- Social Sciences & Humanities from 2005-2013
- Conference Proceedings Citation Index -- Science & Technical from 1990-2013

Basic Statistics:

- Web of Science Core Collection: The number of total items from 1900 through 2013 is 56,442,146.
- There are 1,005,597,828 references to all items in the collection.
- Items By Edition (some documents span multiple editions)
  - SCIE (Science Citation Index Expanded) - 42,263,961 [828.9 M references]
  - SSCI (Social Sciences Citation Index) - 7,690,154 [131.6 M references]
  - AHCI (Arts & Humanities Citation Index) - 4,281,088 [35.3 M references]
  - BSCI (Book Citation Index – Science) - 307,091 [15.6 M references]
  - BHCI (Book Citation Index – Social Sciences & Humanities) - 452,559 [14.2 M references]
  - ISTP (Index to Scientific & Technical Proceedings) - 7,291,457 [72.7 M references]
  - ISSHP (Index to Social Sciences & Humanities Proceedings) - 564,970 [9.4 M references]
Web of Science Annual Total Documents

56,441,146 documents total
Years with <1,000 documents excluded

World War 1
World War 2
Web of Science Annual Total References

1,005,597,828 references total
Years with <10,000 references excluded
Data Cleanliness

Not all data is as pristine as we’d wish. For instance, grant agencies are very much taken as is from the Acknowledgement section of papers, making a simple question like “How many papers are derived from NIH funding?” rather challenging to address until some form of data cleaning is done.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2008</td>
<td>National Institutes of Health (NIH) of the United States</td>
</tr>
<tr>
<td>2</td>
<td>2009</td>
<td>National Institutes of Health (NIH) of the United States</td>
</tr>
<tr>
<td>1</td>
<td>2011</td>
<td>National Institutes of Health (NIH) of the United States</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) of the United States</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) of the United States</td>
</tr>
<tr>
<td>1</td>
<td>2011</td>
<td>National Institutes of Health (NIH) of the United States of America</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) of the United States of America</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) of the United States of America</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) of the USA</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) of the USA (GPI synthesis)</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) of the U.S. Department of Health and Human Services</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) of United States of America</td>
</tr>
<tr>
<td>1</td>
<td>2008</td>
<td>National Institutes of Health (NIH) of USA</td>
</tr>
<tr>
<td>1</td>
<td>2011</td>
<td>National Institutes of Health (NIH) of USA</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) of USA</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH) Pacific Southwest Regional Center of Excellence</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) part of the NIH Roadmap for Medical Research</td>
</tr>
<tr>
<td>2</td>
<td>2011</td>
<td>National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research</td>
</tr>
<tr>
<td>1</td>
<td>2013</td>
<td>National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research</td>
</tr>
<tr>
<td>1</td>
<td>2012</td>
<td>National Institutes of Health (NIH) (Pathogenesis and Diagnosis of Multiple System Atrophy)</td>
</tr>
</tbody>
</table>

How do we avoid every team doing this data cleaning repetitively?
Raw Data

Provided by Thomson Reuters in the form of 166 XML files (561 GB). The 127-page documentation of the XML data format is available on the Enclave at /WoS/Documentation.
WoS Database

- PostgreSQL database with a fully parsed set of ALL data.
- Database schema and Data Dictionary documentation for ease of use.
- Planned: Simple user interface for running common queries.

Table Schema is at /WoS/Documentation on the Enclave and will be added to the public site soon.

Data Dictionary work is in progress (expected mid-to-end of Jan, 2017).

Reduced need for XML parsing will free up Enclave compute cycles for running network analysis and visualization algorithms.
Data Enclave

Setup
• Created by IU as a secure environment for working with the data.
• Powered by the Karst cluster and maintained by UITS administrators.
• Overseen by the Data Steward and the Data Advisory Board.
• Opened for testing in September 2015.
• Currently houses the raw XML files, along with statistical and analysis software.
• In August 2016, a ‘simple database’ and a simple query interface became available.

Usage
• Login from campus (or via VPN) using any computer.
• Use virtual desktop setup to run queries, analyzes, generate visualizations.
• To extract data/results, save files in dedicated /mbx directory for your research team and contact Data Steward.
• So far, over thirty data extraction requests have been issued and the average response time is less than 30 minutes (median response is less than ten minutes)
• Consider writing papers in the enclave to reduce data extraction requests.
Data Access

This data can be used by any employee of Indiana University for academic research and without any sharing of data. Data details and information on how to access the data are detailed at http://www.indiana.edu/~iuni/resources/wos.html

To request data access, please complete the IUNI Web of Science Data (WoS) Access Request Form

So far, eleven teams have requested and gained access to the data.
Open Firefox. Goto URL localhost:11001
Step 1: Initial search

Author and Journal searches use simple text comparisons against the database (LIKE). Search terms can be separated by a pipe character (|). A minus sign (-) can be used to exclude a search term from the results (i.e. if a search term begins with -, the results will not include any records containing that search term).

Keywords will search the database using a fuzzy text search, so it will pull up records that include similar words to the search term (e.g. query of cats may include results with cat or kitten). Search terms can be separated using spaces.

Enter search terms. Hit ‘Submit Query’
Step 2: Review results and refine search

Browser-Based Query Interface (beta)

This interface allows users who are unfamiliar with SQL to access Web of Science data by using a simple browser-based form to query the database. During Step 1, users can begin their search by entering authors, journals or keywords. Step 2 allows users to filter down their result set before starting the export process. Step 3 lets users customize the output of the file export to suit their specific needs.

Step 2: Review Results

Summary

13 Articles
4 Journals

First 50 Rows

<table>
<thead>
<tr>
<th>Article Title</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Security</td>
<td>Public Health</td>
<td>BEARING, SOLIDARITY AND</td>
<td></td>
</tr>
</tbody>
</table>

Review basic statistics regarding result set. Revise search terms as needed.
Step 2: Review results and refine search

Refine Query

- Author Name:
  - Boerner/Boemer

- Journal Name:
  - 

- Title or Abstract Keywords:
  - 

- Publication Date:
  - From: 1900  
  - To: 2013

Submit Query

Refine query and see query that was used to generate the preview dataset.
Select relevant meta-data. Save in dir within data enclave.
Step 4: Save Resulting Data

Select relevant meta-data. Save in dir within data enclave.
### Step 5: Explore Data

Open in LibreOffice Calc  
Save results within data enclave or ask Data Steward to extract results.

<table>
<thead>
<tr>
<th></th>
<th>wos_id</th>
<th>Title</th>
<th>Year</th>
<th>Authors Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>WOS:0003040179000010 Science Policy and the Challenges for Modeling</td>
<td>2012</td>
<td>Boener, Katly</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Science Policy and the Challenges for Modeling Science</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>WOS:0003269688000006 Debating the Expansion of Solidarity</td>
<td>2012</td>
<td>Boener, KatlyBoerner, S</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Towards a New Historical Sociology?</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>An Introduction to Modeling Science: Basic Model Types, Key Definiti</td>
<td>2012</td>
<td>Boener, KatlySchannelhorst, A</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>WOS:0003269688000009 Belonging, Solidarity and Expansion in Social</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Policy Conclusion</td>
<td></td>
<td>Boener, SBoerner, Stefanie</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Social Security Beneath the Nation State</td>
<td>2013</td>
<td>Boener, SBoerner, Stefanie</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>WOS:0003269688000005 Collectively Against the State</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>WOS:000293625800001 Introduction</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>WOS:0003269688000008 Beyond the National: The Construction of Solida</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Beyond the National: The Construction of Solidarity in the Europ</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>WOS:0003269688000003 Solidarity and Social Policy</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>WOS:0002930338000006 Designing Three-Dimensional Materials at the Int</td>
<td>2011</td>
<td>Boener, H.Gentsch, R</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Interface to Biology</td>
<td>2013</td>
<td>Boener, StefanieBoerner, S</td>
</tr>
</tbody>
</table>
Sci2 Tool is in /usr/local/sci2. Copy it into your dir—it needs permission to write work log. Save results within data enclave or ask Data Steward to extract results.
Step 6: Data Analysis and Visualization
Design and Deployment of WoS Data Interface

**GitHub repository** is hosted on Karst. This allows read-only access to the repo from the Enclave, where the scripts will be run, but also allows developers to make changes to the interface without having to go through the many layers of Enclave security.

**Simple database** with a materialized view called ‘interface_matview’ is used to speed up queries to a reasonable length of time.

**Microserver.** The interface uses a python based microserver called Flask that runs under localhost on port 11001. It can only be accessed from within the Enclave, i.e., it is not actually connected to the internet.

**Database access:**
For information on how to use pgadmin3 or psql to access the wos_core database please see recording of demo by Matt Hutchinson at [http://cns.iu.edu/cnstalks](http://cns.iu.edu/cnstalks)
WoS Data Usage

The WoS data is a core asset of the IUNI Science of Science Research Hub. It is essential for R&D related to the Science Observatory, see http://iuni.iu.edu/about. Ultimately, the Science Observatory will provide an extensible, secure infrastructure and expertise to study the science, technology, and innovation system in near real-time and to communicate results to a broad audience of researchers, practitioners, educators, patients, and interested policymakers.

The IUNI Science of Science Research Hub is collaborating in/leading the following grant development efforts:

- EAGER XDMoD Value Analytics, NSF funded (with UITS)
- NWB on Jetstream/XSEDE, NSF funded (with UITS)
- Science Observatory NSF Science and Technology Center lead by Katy Borner, SOIC
- NRT NSF proposal lead by Luis Rocha, SOIC

Please let us know how we can maximize the value of the IUNI WoS data for your R&D.
Planned WoS Data Usage

Special agreements exist for using WoS publications by IU faculty. Data can be used to
• Bulk-fill and pre-populate FAR-like systems on an annual basis.
• Visualize IU's impact—since 1900—as part of the bicentennial celebrations.
• Serve Researcher Networking systems like http://nrn.cns.iu.edu that empower all faculty, staff, students, alumni, funders to identify expertise based on publications but also funding, teaching data provided by IU production systems.

The IUNI Science of Science Research Hub is currently
• Working on the deployment of the Network Workbench that is used by 100k experts around the globe on Jetstream/XSEDE, see IU to launch Jetstream video at https://www.youtube.com/watch?v=t63c12A6bds.
• Educational Data Science, IU Emerging Areas of Research proposal
• NSF SciSIP agenda setting conference on “Modelling Science, Technology, and Innovation” took place at NAS, DC on May 17-18, 2016, and report is now available at http://modsti.cns.iu.edu
• Prototyping moderated “Science and Technology Forecasts” in collaboration with Journalism, see next slide.
Workshop Goals

This practical 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 data scientists 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.

Meeting Place:

Social Science Research Commons (SSRC), Woodburn Hall, Room 200, 1100 East Seventh Street, Bloomington, IN 47405

Web Indiana University Campus Map »

Organizers:

Katy Börner
Victor H. Yngve Distinguished Professor of Information Science, Department of Information and Library Science, School of Informatics and Computing, Indiana University, Bloomington; Director, Cyberinfrastructure for Network Science Center & Curator of Mapping Science exhibit, Bloomington, IN
katy@indiana.edu

Eamon Duede
Executive Director, Knowledge Lab. Administrator, Metaknowledge Research Network, University of Chicago
eduede@uchicago.edu

James Pringle
Head of Industry Development & Innovation at Thomson Reuters IP & Science