IUNI Web of Science Data: An Introduction for Beginners

Katy Börner and Robert Light

Cyberinfrastructure for Network Science Center School of Informatics and Computing and Indiana University Network Science Institute Indiana University, USA

January 11, 2016

INDIANA UNIVERSITY NETWORK SCIENCE INSTITUTE

CNS Cuberinfrastructure for Network Science Center IUNI Web of Science Data

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]

IUNI Web of Science Data

Cyberinfrastructure for Network Science Center

Web of Science Annual Total Documents





Cuberinfrastructure for Network Science Center IUNI Web of Science Data

Web of Science Annual Total References



Cyberinfrastructure for IUNI Web of Science Data

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.

1 2008 National Institutes of Health (NIH) of the United States 2 2009 National Institutes of Health (NIH) of the United States 1 2011 National Institutes of Health (NIH) of the United States 1 2012 National Institutes of Health (NIH) of the United States	
1 2011 National Institutes of Health (NIH) of the United States 1 2012 National Institutes of Health (NIH) of the United States	
1 2012 National Institutes of Health (NIH) of the United States	
1 2013 National Institutes of Health (NIH) of the United States	
1 2011 National Institutes of Health (NIH) of the United States of America	
1 2012 National Institutes of Health (NIH) of the United States of America	
1 2013 National Institutes of Health (NIH) of the United States of America	
1 2013 National Institutes of Health (NIH) of the USA	
1 2013 National Institutes of Health (NIH) of the USA (GPI synthesis)	
1 2013 National Institutes of Health (NIH) of the U.S. Department of Health and Human Services	
1 2012 National Institutes of Health (NIH) of United States of America	
1 2008 National Institutes of Health (NIH) of USA	
1 2011 National Institutes of Health (NIH) of USA	
1 2013 National Institutes of Health (NIH) of USA	
1 2013 National Institutes of Health (NIH) Pacific Southwest Regional Center of Excellence	
1 2012 National Institutes of Health (NIH) part of the NIH Roadmap for Medical Research	
2 2011 National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research	
1 2012 National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research	
1 2013 National Institutes of Health (NIH), part of the NIH Roadmap for Medical Research	
1 2012 National Institutes of Health (NIH) (Pathogenesis and Diagnosis of Multiple System Atro	phy)

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.



Web of Science Data

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).

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

IUNI Web of Science Data



83 **IUNI Web of Science Data** Cyberinfrastructure for

Data Enclave

Setup

- Created by IU as a secure environment for working with the data.
- ٠ 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. •

Usage

- Login from campus (or via VPN) using any computer. ٠
- Use virtual desktop setup to run queries, analyzes, generate visualizations. ٠
- To extract data/results, save files in dedicated /mbx directory for your research team and contact Data Steward.
- So far, over twenty data extraction requests have been issued and the average ٠ 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. ٠





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 <u>IUNI Web of Science Data (WoS) Access</u> <u>Request Form</u>

So far, five teams have requested and gained access to the data.

Cyberinfrastructure for Network Science Center IUNI Web of Science Data

Data Enclave Screenshot

	Applications Places System 🛃 🏽 📶		Fri Jan II, 2:56 PM
The first View Sect: Nerval Help Protected Links 1, 1 laddet. We restant Links 1, 1 laddet. We restant Links 1, 1 laddet. We restant Links 1, 1 laddet. If programming language retrins 5, 15, 1 laddet. Links 1, 1 laddet.			
Ø #Itherite Library version 5.1.1 landet. Wersion 1.3.1 landet. Wersion 5.1.1 landet. Wersion 6.1.1 landet. W			
Fift version 3.1.3 loaded. wersion 3.1.3 loaded. wersion 3.1.3 loaded. with in the loaded. with loaded. with in the l	File Edit View Search Terminal Help		
k vrisie 13.2 laded: ub ou Sladed: ub ou Sladed:	MP arithmetic library version 5.1.1 loaded.		
to version 4.3.2 loaded. to wersion 4.3.2 loaded. to programming language version 3.5.2 loaded. to come version 3.5.2 loa	bc version 1.0.1 loaded.		
If programming language versions 5.8.2 (about). Intervention (about) structures (about). Intervention (about). <	truersion 4.5.7 Toxind		
If programming language versions 5, 6, 2, 4 doed. Liphtretistis Liphtrij's position.	Dab Workload Manager scheduling and management system version 7.1.1 Loaded.		
Lightrights / 1 spirit / 5 spainits	erl programming language version 5.16.2 loaded.	File Edit Query Favourites Macros View Help	
SQL Editor Cospical Guery Bulder File fist Veer Place. Help Ray Ray Ray Ray Ray Ray Ray Ray	ntel compiler suite version 16.0.1 loaded.	💏 🔜 🚨 🕼 🖉 🖉 👘 🖉 🕹 🖄 👘 🖉 🗧 🔮 😭 usos_core on postgree	
SELECT COURT(-) FROM Way, references File Edit View Places Help Image: Select Court(-) Any Image: Select Court(-) Any Image: Select Court(-) Select Court(-) File Edit View Places Help Image: Select Court(-) Image: Select Court(-) Image: Select Court(-) File Edit View Places Help Image: Select Court(-) Image: Select Court(-) Image: Select Court(-) Image: Select Court(-) </td <td></td> <td>SQL Editor Graphical Query Builder . Scelich pad B</td> <td></td>		SQL Editor Graphical Query Builder . Scelich pad B	
SELECT CONT(1) FROM way, references File fait View Races Help Image: Select Control (1) Way Image: Select Control (1) Select Solution Image: Select Control (1) Image: Select Control (1) FROM way, references Image: Select Solution Image: Select Control (1) Image: Select Control (1) FROM way, references Image: Select Solution Image: Select Solution Image: Solution Image: S		Previous gueries V Division Delete All	
Pie Edit Vere Places Help kay Can kay Can base datu Catput Explain Messages History base datu Catput Explain Messages History config 0.008 Destrop Kant parsing 5.01pts			
Pie Edit Vere Places Help kay Can kay Can base datu Catput Explain Messages History base datu Catput Explain Messages History config 0.008 Destrop Kant parsing 5.01pts			
kay Conte same biologic same biologic containing 14 items) biologic containing 14 items) biologic containing 14 items) biologic containing 14 items)			- 0 4
Image: Second	File East view Places Help	Hie Edit View Places Help	
Image: Second			-
Image: Strategy of the selected (containing 14 items)			
pusing Solpts cache pusing Solpts cache corFig dous emact.d lightr vightr salected (cortaining 14 items)	katy sight?		Karst
sorting de la termina de la termi	12 12 12	Data Output Explain Messages History •	-
sorting de la termina de la termi			and an
sorting de la termina de la termi		persing Scripts	cache
Inome v / Trighter selected (containing 14 items)	A STREET AND A S		100 C
Inome v / Trighter selected (containing 14 items)			
Inome v / Trighter selected (containing 14 items)		control days	emars d
			Anada a
Overy is numing. Unix (in 1, Col 36, Ch 36 (074719 ms)	home 'light' selected (containing 14 items)	a lightr → 35 items, Free space: 8.6 TB	
Overy is numering. Unix [Jn 1, Col 36, Ch 36 (814719 ms)			
Overy is numing. Unix (in 1, Col 36, Ch 36 (1974719 ms			
Overy is numing. Unix (Jn 1, Col 36, Ch 36 (1977)39 ms			
Overy is numing. Unix (In 1, Col 36, Ch 36 (024729 ms)			
		Overy is running. Unix in 1, col 36, ch 36	

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:

- NSF Mid-West Data HUB "Impact Assessment" and "Community Engagement" IU lead is Beth Plale, SOIC
- NSF EAGER XDMoD Value Analytics (with UITS)
- NWB on Jetstream/XSEDE (with UITS)

Cyberinfrastructure for

- "Science of Science" NSF Expeditions in Computing lead by Alex Vespignani, NEU
- "Science Observatory" NSF Science and Technology Center lead by Katy Borner, SOIC
- NSF NRT 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.

CNS Cuberinfrastructure for IUNI Web of Science Data

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 <u>https://www.youtube.com/watch?v=t63c12A6bds</u>.
- Learning Analytics, IU Grand Challenges proposal
- Organizing a NSF SciSIP agenda setting conference on "Modelling Science, Technology, and Innovation" that will take place at NAS, DC on May 17-18, 2016, see <u>http://modsti.cns.iu.edu</u>
- Prototyping moderated "Science and Technology Forecasts" in collaboration with Journalism, see next slide.

