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

IUNI Web of Science Data

Data Acquisition

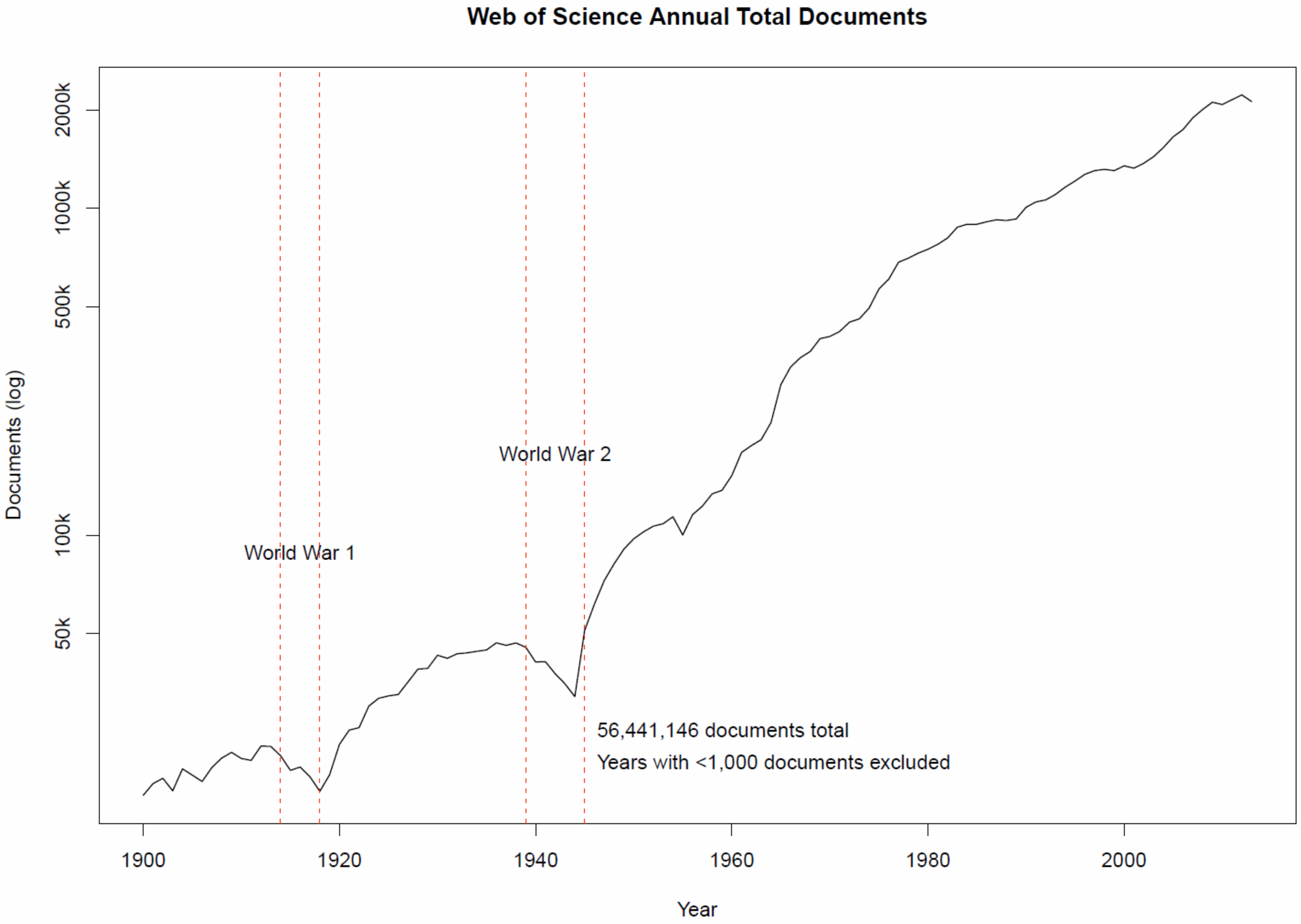
The IUNI Science of Science Hub acquired the complete set of Thomson Reuters' Web of Science XML raw data (Web of Knowledge version 5) comprising

- 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]

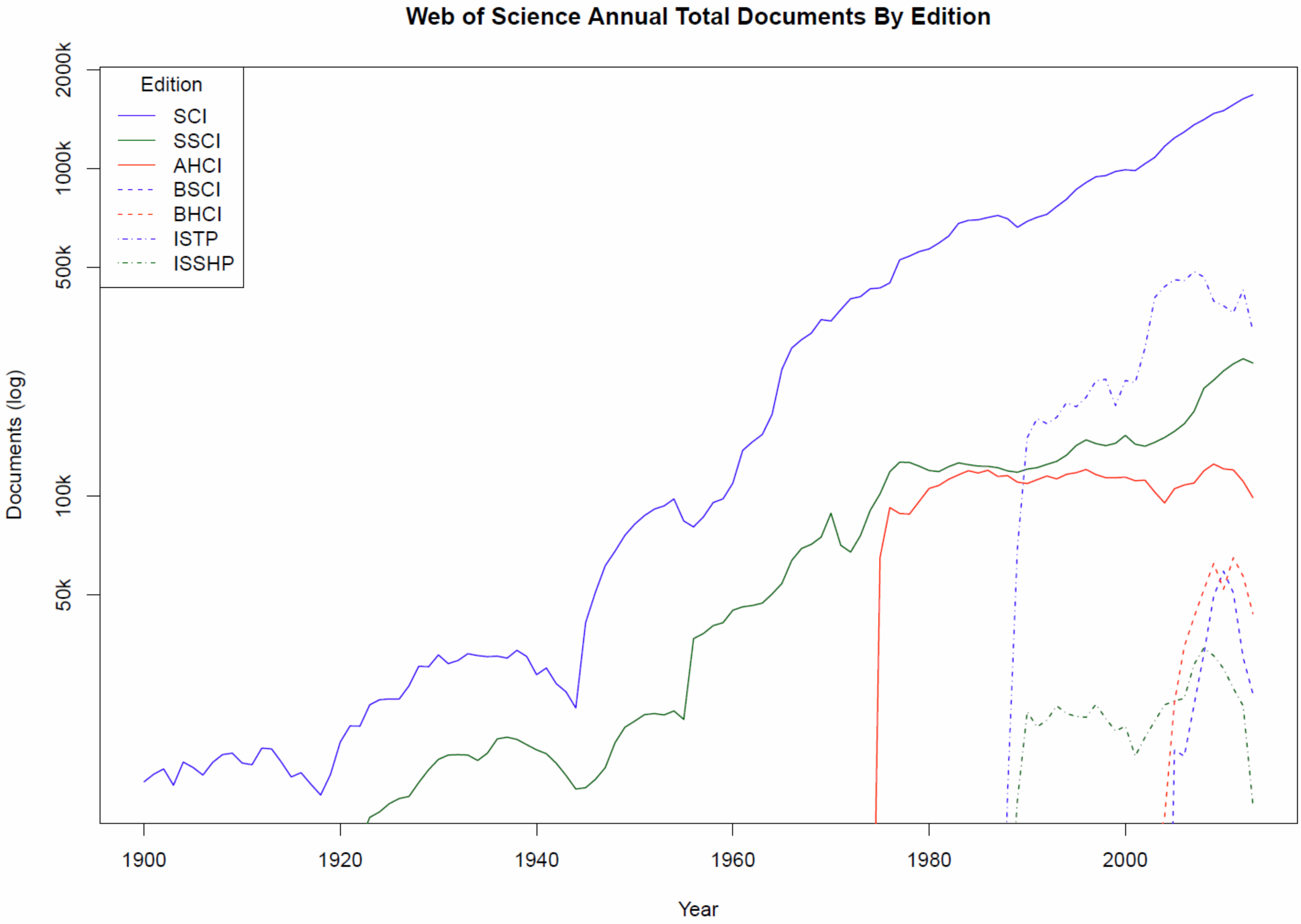
IUNI Web of Science Data



56,441,146 documents total

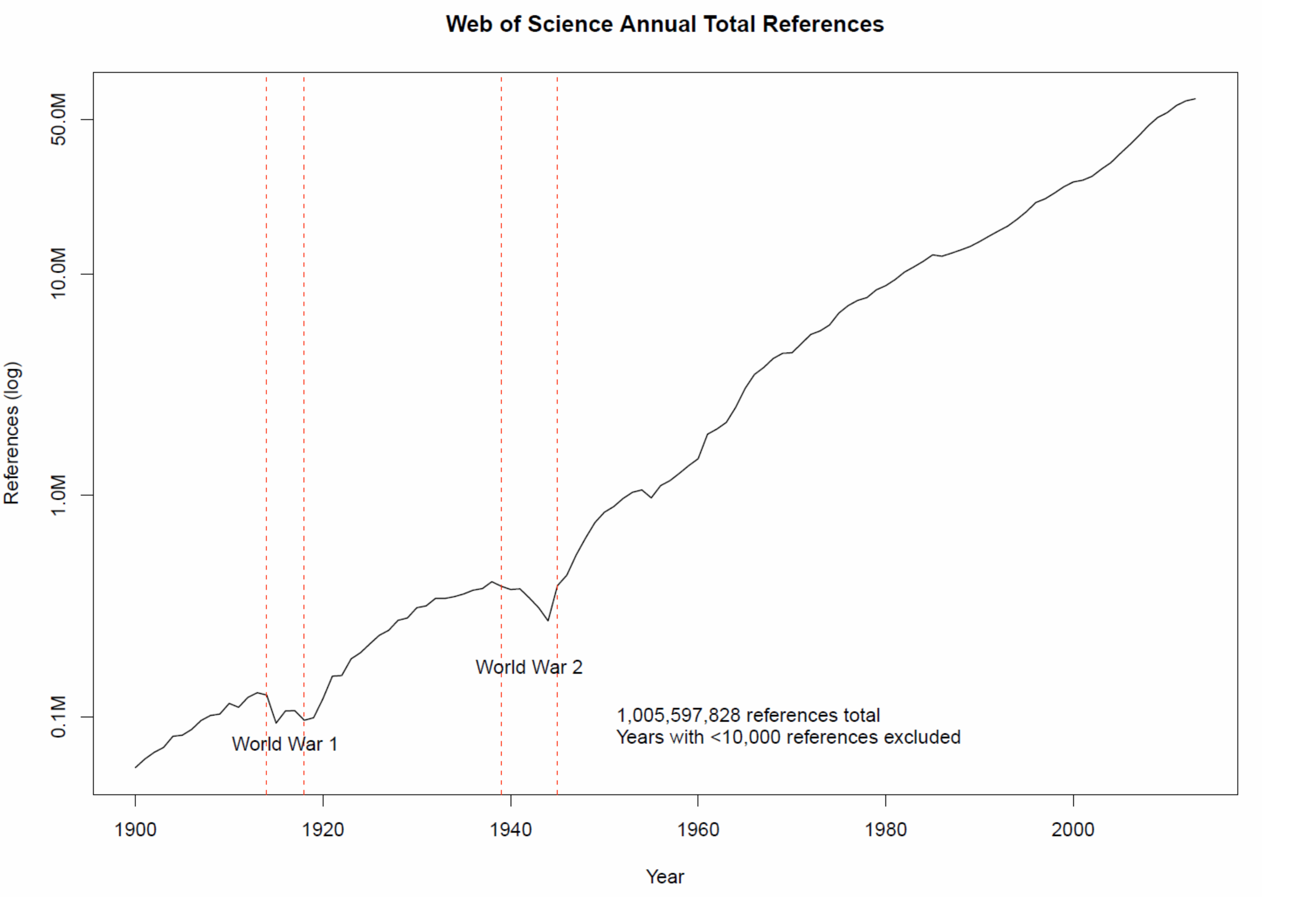
Years with <1,000 documents excluded

IUNI Web of Science Data



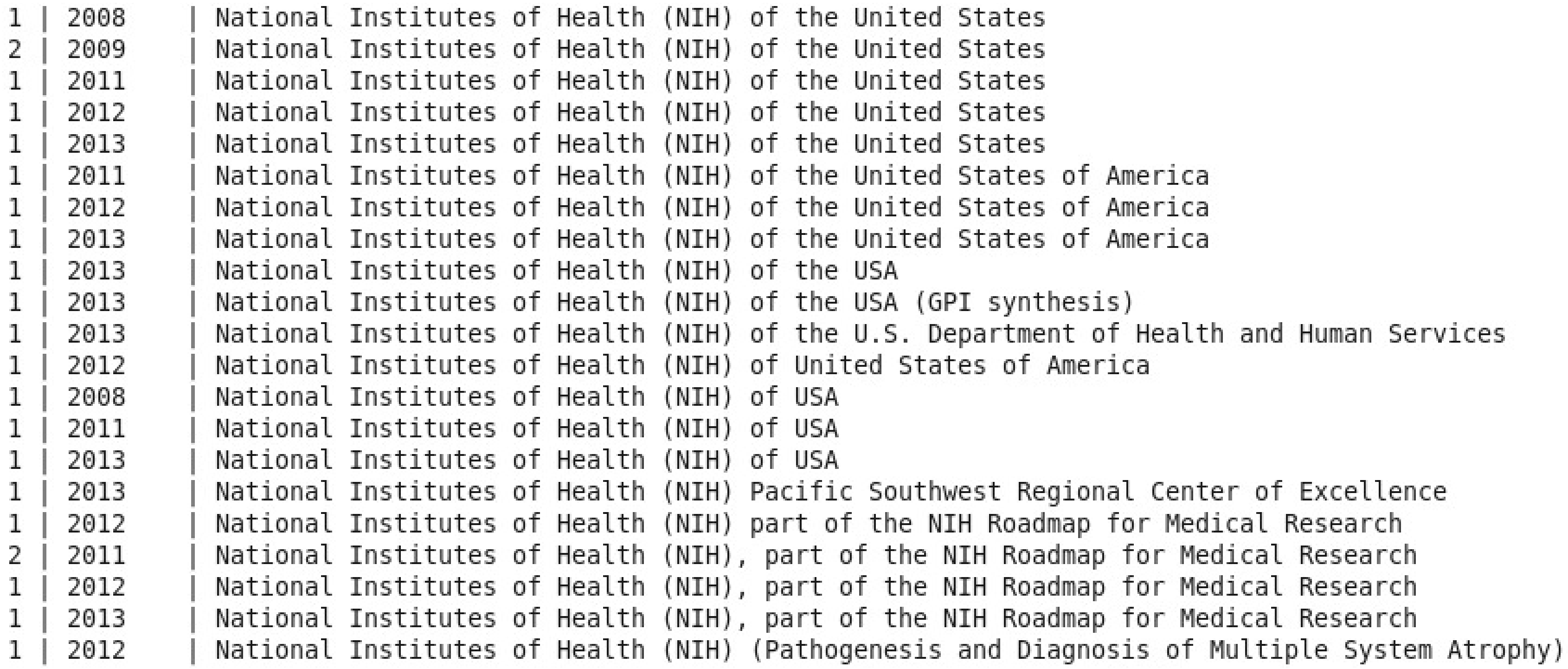
CNS Cyberinfrastructure for Network Science Center

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.

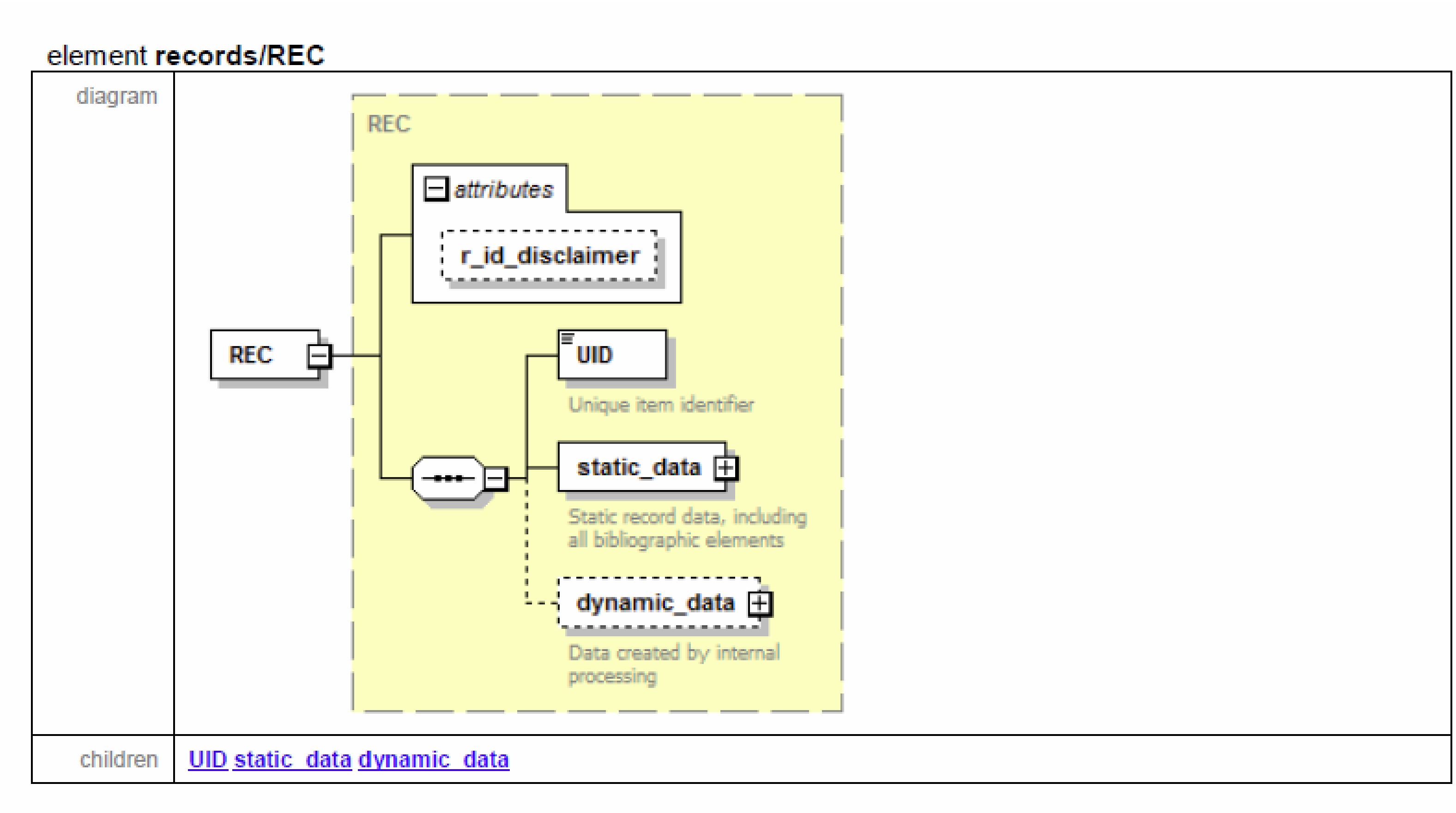


```
 1 | 2008    | National Institutes of Health (NIH) of the United States
 2 | 2009    | National Institutes of Health (NIH) of the United States
 1 | 2011    | National Institutes of Health (NIH) of the United States
 1 | 2012    | National Institutes of Health (NIH) of the United States
 1 | 2013    | 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 Atrophy)
```

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.



element **records/REC**

diagram	
children	UID static_data dynamic_data

REC
attributes
r_id_disclaimer
REC
UID
Unique item identifier
static_data
Static record data, including all bibliographic elements
dynamic_data
Data created by internal processing

IUNI 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.

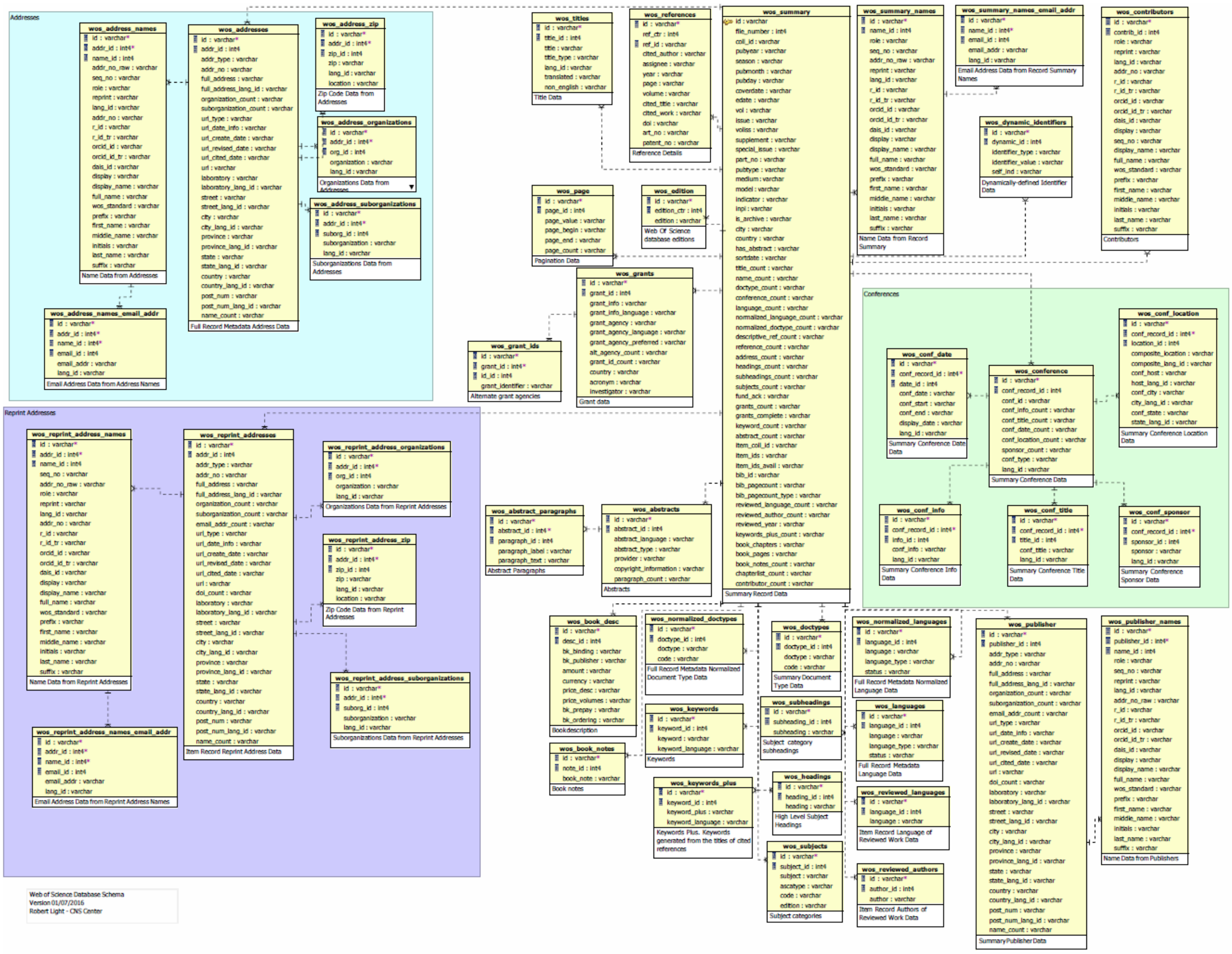
Table Schema is at /WoS/Documentation on the Enclave and will be added to the public site soon.

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

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

CNS Cyberinfrastructure for Network Science Center

IUNI Web of Science Data



Addresses

wos_address_names
- id : varchar*
- addr_id : int4*
- name_id : int4
- addr_no_raw : varchar
- seq_no : varchar
- role : varchar
- reprint : varchar
- lang_id : varchar
- addr_no : varchar
- r_id : varchar
- r_id_tr : varchar
- orcid_id : varchar
- orcid_id_tr : varchar
- dais_id : varchar
- display : varchar
- display_name : varchar
- full_name : varchar
- wos_standard : varchar
- prefix : varchar
- first_name : varchar
- middle_name : varchar
- initials : varchar
- last_name : varchar
- suffix : varchar

Name Data from Addresses

wos_addresses
- id : varchar*
- addr_id : int4
- addr_type : varchar
- addr_no : varchar
- full_address : varchar
- full_address_lang_id : varchar
- organization_count : varchar
- suborganization_count : varchar
- url_type : varchar
- url_date_info : varchar
- url_create_date : varchar
- url_revised_date : varchar
- url_cited_date : varchar
- url : varchar
- laboratory : varchar
- laboratory_lang_id : varchar
- street : varchar
- street_lang_id : varchar
- city : varchar
- city_lang_id : varchar
- province : varchar
- province_lang_id : varchar
- state : varchar
- state_lang_id : varchar
- country : varchar
- country_lang_id : varchar
- post_num : varchar
- post_num_lang_id : varchar
- name_count : varchar

Full Record Metadata Address Data

wos_address_zip
- id : varchar*
- addr_id : int4*
- zip_id : int4
- zip : varchar
- lang_id : varchar
- location : varchar

Zip Code Data from Addresses

wos_address_organizations
- id : varchar*
- addr_id : int4*
- org_id : int4
- organization : varchar
- lang_id : varchar

Organizations Data from Addresses

wos_address_suborganizations
- id : varchar*
- addr_id : int4*
- suborg_id : int4
- suborganization : varchar
- lang_id : varchar

Suborganizations Data from Addresses

wos_address_names_email_addr
- id : varchar*
- addr_id : int4*
- name_id : int4*
- email_id : int4
- email_addr : varchar
- lang_id : varchar

Email Address Data from Address Names

wos_titles
- id : varchar*
- title_id : int4
- title : varchar
- title_type : varchar
- lang_id : varchar
- translated : varchar
- non_english : varchar

Title Data

wos_references
- id : varchar*
- ref_ctr : int4
- ref_id : varchar
- cited_author : varchar
- assignee : varchar
- year : varchar
- page : varchar
- volume : varchar
- cited_title : varchar
- cited_work : varchar
- doi : varchar
- art_no : varchar
- patent_no : varchar

Reference Details

wos_page
- id : varchar*
- page_id : int4
- page_value : varchar
- page_begin : varchar
- page_end : varchar
- page_count : varchar

Pagination Data

wos_edition
- id : varchar*
- edition_ctr : int4
- edition : varchar

Web Of Science database editions

wos_grants
- id : varchar*
- grant_id : int4
- grant_info : varchar
- grant_info_language : varchar
- grant_agency : varchar
- grant_agency_language : varchar
- grant_agency_preferred : varchar
- alt_agency_count : varchar
- grant_id_count : varchar
- country : varchar
- acronym : varchar
- investigator : varchar

Grant data

wos_grant_ids
- id : varchar*
- grant_id : int4*
- id_id : int4
- grant_identifier : varchar

Alternate grant agencies

wos_summary
- id : varchar
- file_number : int4
- coll_id : varchar
- pubyear : varchar
- season : varchar
- pubmonth : varchar
- pubday : varchar
- coverdate : varchar
- edate : varchar
- vol : varchar
- issue : varchar
- voliss : varchar
- supplement : varchar
- special_issue : varchar
- part_no : varchar
- pubtype : varchar
- medium : varchar
- model : varchar
- indicator : varchar
- inpi : varchar
- is_archive : varchar
- city : varchar
- country : varchar
- has_abstract : varchar
- sortdate : varchar
- title_count : varchar
- name_count : varchar
- doctype_count : varchar
- conference_count : varchar
- language_count : varchar
- normalized_language_count : varchar
- normalized_doctype_count : varchar
- descriptive_ref_count : varchar
- reference_count : varchar
- address_count : varchar
- headings_count : varchar
- subheadings_count : varchar
- subjects_count : varchar
- fund_ack : varchar
- grants_count : varchar
- grants_complete : varchar
- keyword_count : varchar
- abstract_count : varchar
- item_coll_id : varchar
- item_ids : varchar
- item_ids_avail : varchar
- bib_id : varchar
- bib_pagecount : varchar
- bib_pagecount_type : varchar
- reviewed_language_count : varchar
- reviewed_author_count : varchar
- reviewed_year : varchar
- keywords_plus_count : varchar
- book_chapters : varchar
- book_pages : varchar
- book_notes_count : varchar
- chapterlist_count : varchar
- contributor_count : varchar

Summary Record Data

wos_summary_names
- id : varchar*
- name_id : int4
- role : varchar
- seq_no : varchar
- addr_no_raw : varchar
- reprint : varchar
- lang_id : varchar
- r_id : varchar
- r_id_tr : varchar
- orcid_id : varchar
- orcid_id_tr : varchar
- dais_id : varchar
- display : varchar
- display_name : varchar
- full_name : varchar
- wos_standard : varchar
- prefix : varchar
- first_name : varchar
- middle_name : varchar
- initials : varchar
- last_name : varchar
- suffix : varchar

Name Data from Record Summary

wos_summary_names_email_addr
- id : varchar*
- name_id : int4*
- email_id : int4
- email_addr : varchar
- lang_id : varchar

Email Address Data from Record Summary Names

wos_dynamic_identifiers
- id : varchar*
- dynamic_id : int4
- identifier_type : varchar
- identifier_value : varchar
- self_ind : varchar

Dynamically-defined Identifier Data

wos_contributors
- id : varchar*
- contrib_id : int4
- role : varchar
- reprint : varchar
- lang_id : varchar
- addr_no : varchar
- r_id : varchar
- r_id_tr : varchar
- orcid_id : varchar
- orcid_id_tr : varchar
- dais_id : varchar
- display : varchar
- seq_no : varchar
- display_name : varchar
- full_name : varchar
- wos_standard : varchar
- prefix : varchar
- first_name : varchar
- middle_name : varchar
- initials : varchar
- last_name : varchar
- suffix : varchar

Contributors

Conferences

wos_conf_date
- id : varchar*
- conf_record_id : int4*
- date_id : int4
- conf_date : varchar
- conf_start : varchar
- conf_end : varchar
- display_date : varchar
- lang_id : varchar

Summary Conference Date Data

wos_conference
- id : varchar*
- conf_record_id : int4
- conf_id : varchar
- conf_info_count : varchar
- conf_title_count : varchar
- conf_date_count : varchar
- conf_location_count : varchar
- sponsor_count : varchar
- conf_type : varchar
- lang_id : varchar

Summary Conference Data

wos_conf_location
- id : varchar*
- conf_record_id : int4*
- location_id : int4
- composite_location : varchar
- composite_lang_id : varchar
- conf_host : varchar
- host_lang_id : varchar
- conf_city : varchar
- city_lang_id : varchar
- conf_state : varchar
- state_lang_id : varchar

Summary Conference Location Data

wos_conf_info
- id : varchar*
- conf_record_id : int4*
- info_id : int4
- conf_info : varchar
- lang_id : varchar

Summary Conference Info Data

wos_conf_title
- id : varchar*
- conf_record_id : int4*
- title_id : int4
- conf_title : varchar
- lang_id : varchar

Summary Conference Title Data

wos_conf_sponsor
- id : varchar*
- conf_record_id : int4*
- sponsor_id : int4
- sponsor : varchar
- lang_id : varchar

Summary Conference Sponsor Data

Reprint Addresses

wos_reprint_address_names
- id : varchar*
- addr_id : int4*
- name_id : int4
- seq_no : varchar
- addr_no_raw : varchar
- role : varchar
- reprint : varchar
- lang_id : varchar
- addr_no : varchar
- r_id : varchar
- r_id_tr : varchar
- orcid_id : varchar
- orcid_id_tr : varchar
- dais_id : varchar
- display : varchar
- display_name : varchar
- full_name : varchar
- wos_standard : varchar
- prefix : varchar
- first_name : varchar
- middle_name : varchar
- initials : varchar
- last_name : varchar
- suffix : varchar

Name Data from Reprint Addresses

wos_reprint_addresses
- id : varchar*
- addr_id : int4
- addr_type : varchar
- addr_no : varchar
- full_address : varchar
- full_address_lang_id : varchar
- organization_count : varchar
- suborganization_count : varchar
- email_addr_count : varchar
- url_type : varchar
- url_date_info : varchar
- url_create_date : varchar
- url_revised_date : varchar
- url_cited_date : varchar
- url : varchar
- doi_count : varchar
- laboratory : varchar
- laboratory_lang_id : varchar
- street : varchar
- street_lang_id : varchar
- city : varchar
- city_lang_id : varchar
- province : varchar
- province_lang_id : varchar
- state : varchar
- state_lang_id : varchar
- country : varchar
- country_lang_id : varchar
- post_num : varchar
- post_num_lang_id : varchar
- name_count : varchar

Item Record Reprint Address Data

wos_reprint_address_organizations
- id : varchar*
- addr_id : int4*
- org_id : int4
- organization : varchar
- lang_id : varchar

Organizations Data from Reprint Addresses

wos_reprint_address_zip
- id : varchar*
- addr_id : int4*
- zip_id : int4
- zip : varchar
- lang_id : varchar
- location : varchar

Zip Code Data from Reprint Addresses

wos_reprint_address_suborganizations
- id : varchar*
- addr_id : int4*
- suborg_id : int4
- suborganization : varchar
- lang_id : varchar

Suborganizations Data from Reprint Addresses

wos_reprint_address_names_email_addr
- id : varchar*
- addr_id : int4*
- name_id : int4*
- email_id : int4
- email_addr : varchar
- lang_id : varchar

Email Address Data from Reprint Address Names

wos_abstract_paragraphs
- id : varchar*
- abstract_id : int4*
- paragraph_id : int4
- paragraph_label : varchar
- paragraph_text : varchar

Abstract Paragraphs

wos_abstracts
- id : varchar*
- abstract_id : int4
- abstract_language : varchar
- abstract_type : varchar
- provider : varchar
- copyright_information : varchar
- paragraph_count : varchar

Abstracts

wos_book_desc
- id : varchar*
- desc_id : int4
- bk_binding : varchar
- bk_publisher : varchar
- amount : varchar
- currency : varchar
- price_desc : varchar
- price_volumes : varchar
- bk_prepay : varchar
- bk_ordering : varchar

Bookdescription

wos_book_notes
- id : varchar*
- note_id : int4
- book_note : varchar

Book notes

wos_normalized_doctypes
- id : varchar*
- doctype_id : int4
- doctype : varchar
- code : varchar

Full Record Metadata Normalized Document Type Data

wos_keywords
- id : varchar*
- keyword_id : int4
- keyword : varchar
- keyword_language : varchar

Keywords

wos_keywords_plus
- id : varchar*
- keyword_id : int4
- keyword_plus : varchar
- keyword_language : varchar

Keywords Plus. Keywords generated from the titles of cited references

wos_doctypes
- id : varchar*
- doctype_id : int4
- doctype : varchar
- code : varchar

Summary Document Type Data

wos_subheadings
- id : varchar*
- subheading_id : int4
- subheading : varchar

Subject category subheadings

wos_headings
- id : varchar*
- heading_id : int4
- heading : varchar

High Level Subject Headings

wos_subjects
- id : varchar*
- subject_id : int4
- subject : varchar
- ascatype : varchar
- code : varchar
- edition : varchar

Subject categories

wos_normalized_languages
- id : varchar*
- language_id : int4
- language : varchar
- language_type : varchar
- status : varchar

Full Record Metadata Normalized Language Data

wos_languages
- id : varchar*
- language_id : int4
- language : varchar
- language_type : varchar
- status : varchar

Full Record Metadata Language Data

wos_reviewed_languages
- id : varchar*
- language_id : int4
- language : varchar

Item Record Language of Reviewed Work Data

wos_reviewed_authors
- id : varchar*
- author_id : int4
- author : varchar

Item Record Authors of Reviewed Work Data

wos_publisher
- id : varchar*
- publisher_id : int4
- addr_type : varchar
- addr_no : varchar
- full_address : varchar
- full_address_lang_id : varchar
- organization_count : varchar
- suborganization_count : varchar
- email_addr_count : varchar
- url_type : varchar
- url_date_info : varchar
- url_create_date : varchar
- url_revised_date : varchar
- url_cited_date : varchar
- url : varchar
- doi_count : varchar
- laboratory : varchar
- laboratory_lang_id : varchar
- street : varchar
- street_lang_id : varchar
- city : varchar
- city_lang_id : varchar
- province : varchar
- province_lang_id : varchar
- state : varchar
- state_lang_id : varchar
- country : varchar
- country_lang_id : varchar
- post_num : varchar
- post_num_lang_id : varchar
- name_count : varchar

SummaryPublisherData

wos_publisher_names
- id : varchar*
- publisher_id : int4*
- name_id : int4
- role : varchar
- seq_no : varchar
- reprint : varchar
- lang_id : varchar
- addr_no_raw : varchar
- r_id : varchar
- r_id_tr : varchar
- orcid_id : varchar
- orcid_id_tr : varchar
- dais_id : varchar
- display : varchar
- display_name : varchar
- full_name : varchar
- wos_standard : varchar
- prefix : varchar
- first_name : varchar
- middle_name : varchar
- initials : varchar
- last_name : varchar
- suffix : varchar

Name Data from Publishers

Web of Science Database Schema
Version 01/07/2016
Robert Light - CNS Center

IUNI Web of Science Data

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.

IUNI Web of Science Data

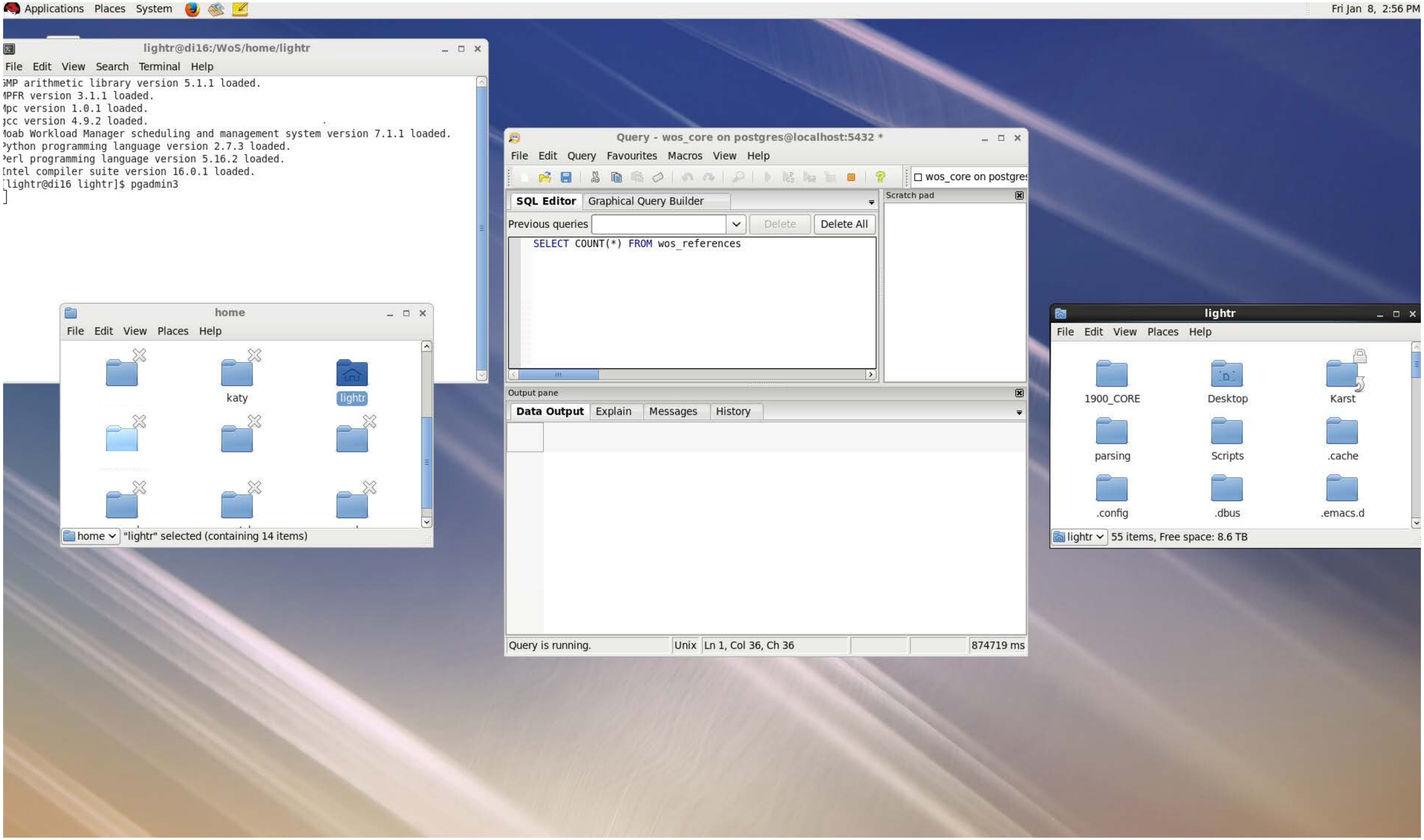
Data Access

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.

Data Enclave Screenshot



```
GMP arithmetic library version 5.1.1 loaded.
MPFR version 3.1.1 loaded.
Mpc version 1.0.1 loaded.
gcc version 4.9.2 loaded.
Moab Workload Manager scheduling and management system version 7.1.1 loaded.
Python programming language version 2.7.3 loaded.
Perl programming language version 5.16.2 loaded.
Intel compiler suite version 16.0.1 loaded.
[lightr@di16 lightr]$ pgadmin3
]
```

```
SELECT COUNT(*) FROM wos_references
```

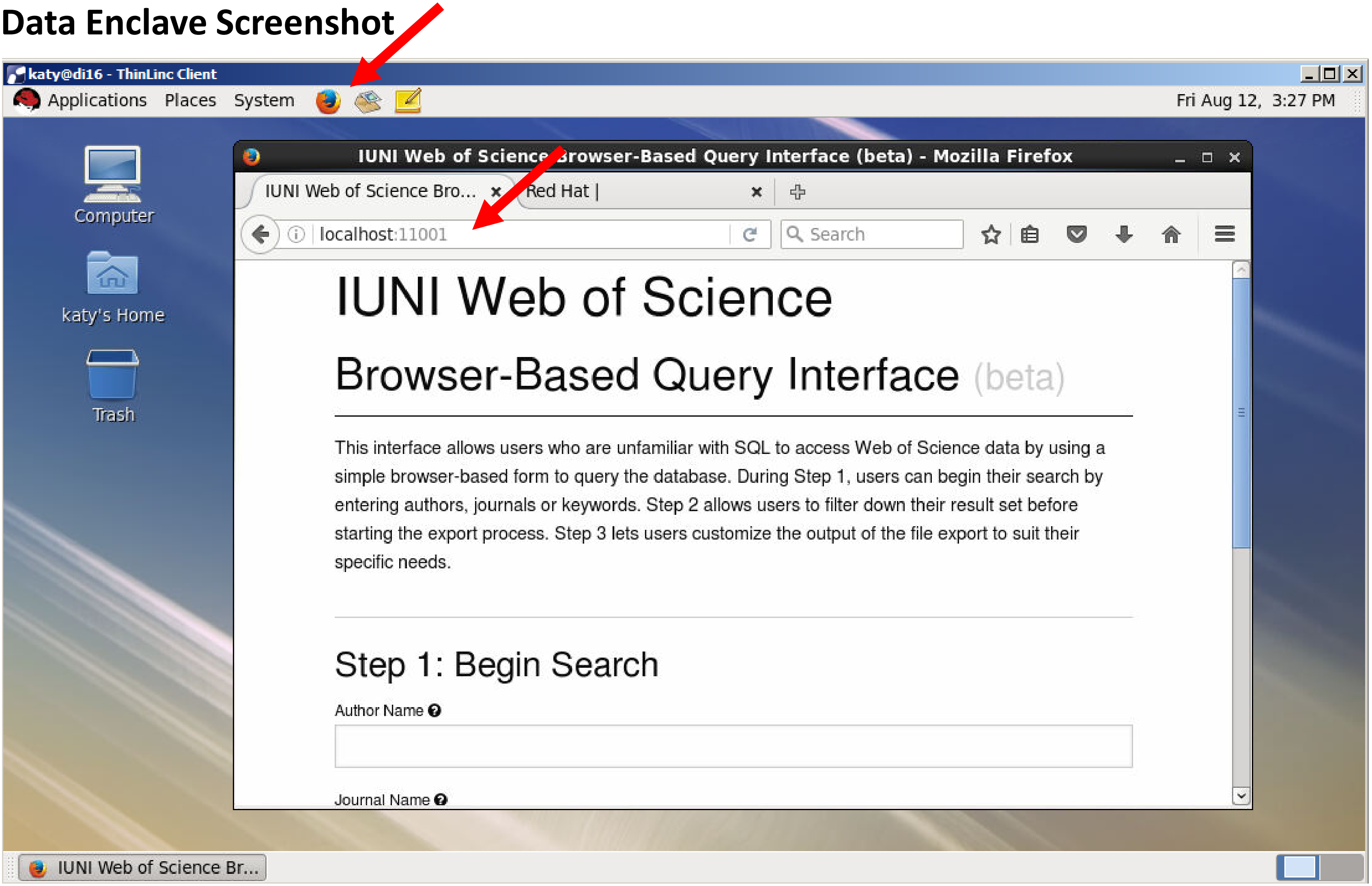
Query is running. | Unix | Ln 1, Col 36, Ch 36 | 874719 ms

"lightr" selected (containing 14 items)

55 items, Free space: 8.6 TB

IUNI Web of Science Data

Data Enclave Screenshot


katy@di16 - ThinLinc Client
Applications
Places
System
Fri Aug 12, 3:27 PM
Computer
katy's Home
Trash
IUNI Web of Science Browser-Based Query Interface (beta) - Mozilla Firefox
IUNI Web of Science Bro...
Red Hat |
localhost:11001
Search
IUNI Web of Science Br...

IUNI Web of Science

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.

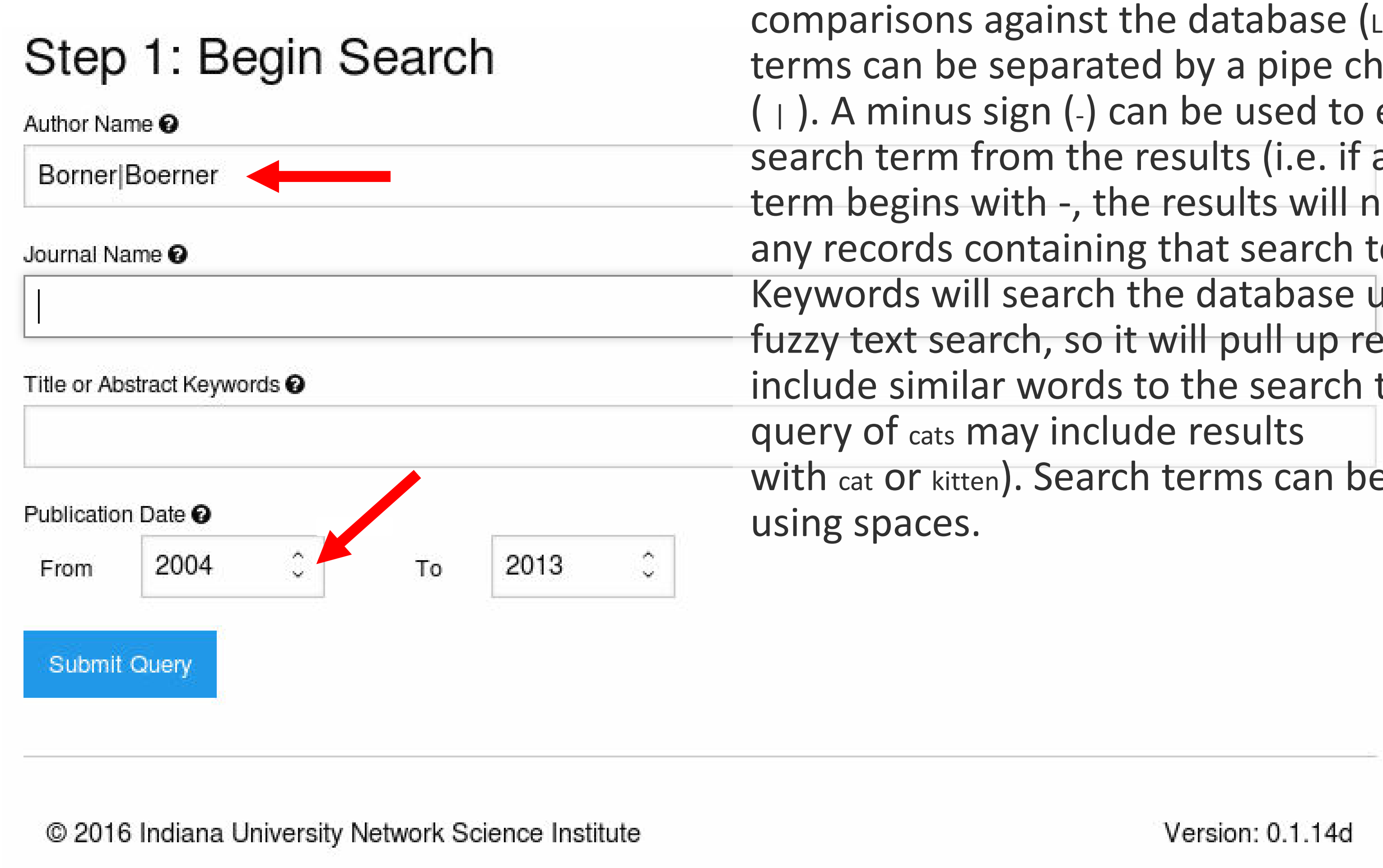
Step 1: Begin Search

Author Name ❓

Journal Name ❓

Open Firefox. Goto URL localhost:11001

Step 1: Initial search



Step 1: Begin Search

Author Name: Borner|Boerner

Journal Name:

Title or Abstract Keywords:

Publication Date From: 2004 To: 2013

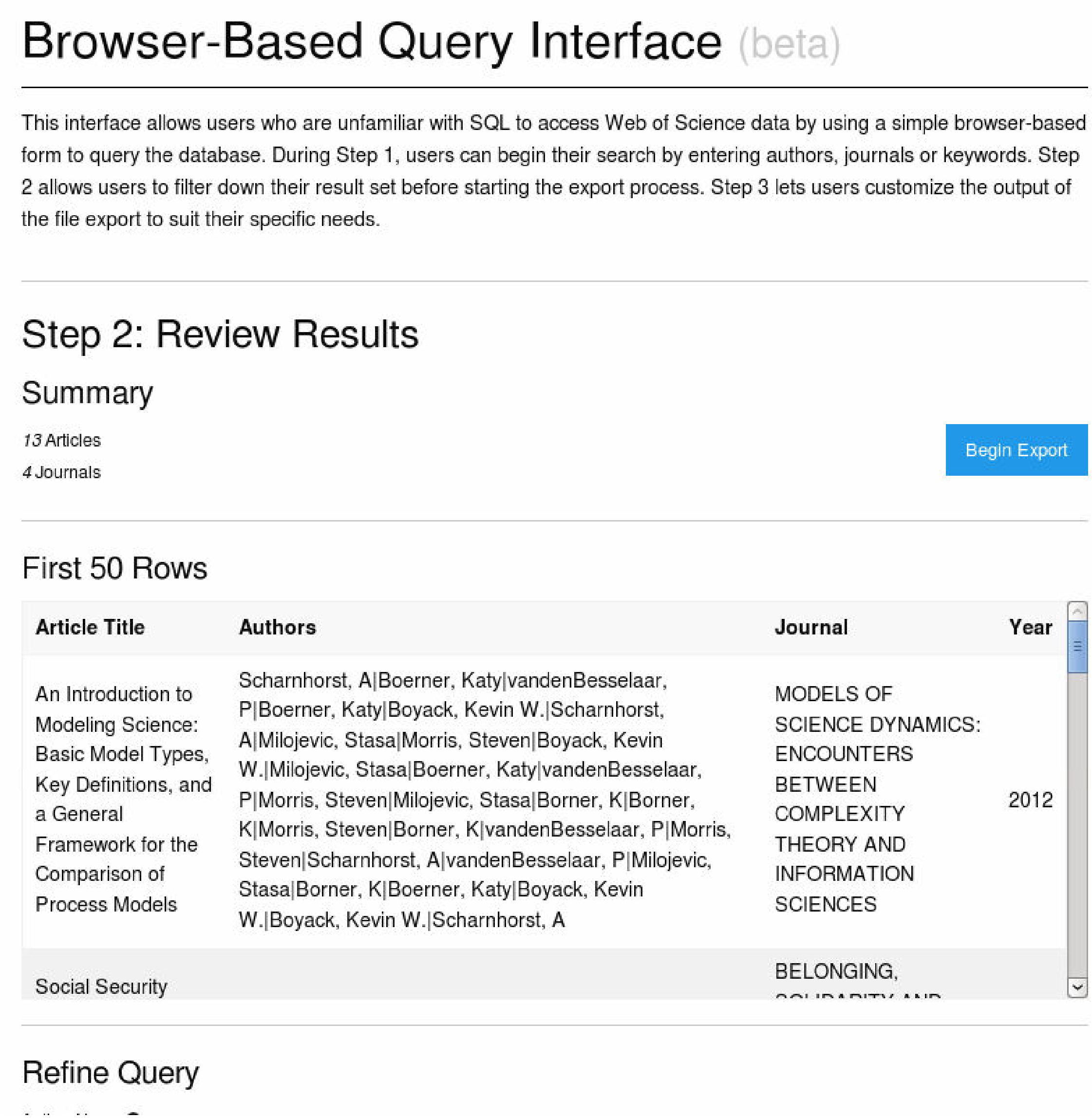
© 2016 Indiana University Network Science Institute

Version: 0.1.14d

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

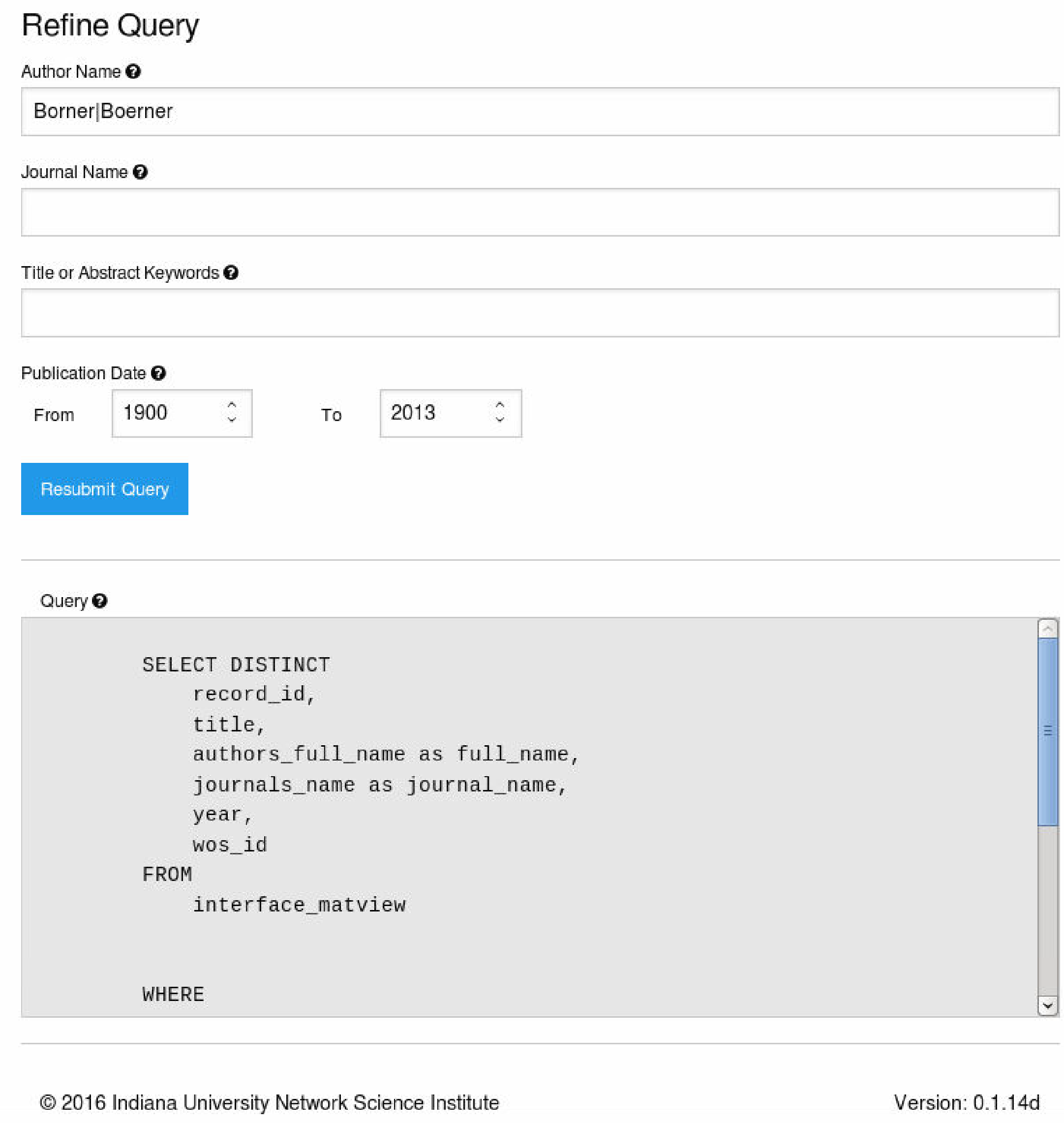
First 50 Rows

Article Title	Authors	Journal	Year
An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General Framework for the Comparison of Process Models	Scharnhorst, A|Boerner, Katy|vandenBesselaar, P|Boerner, Katy|Boyack, Kevin W.|Scharnhorst, A|Milojevic, Stasa|Morris, Steven|Boyack, Kevin W.|Milojevic, Stasa|Boerner, Katy|vandenBesselaar, P|Morris, Steven|Milojevic, Stasa|Borner, K|Borner, K|Morris, Steven|Borner, K|vandenBesselaar, P|Morris, Steven|Scharnhorst, A|vandenBesselaar, P|Milojevic, Stasa|Borner, K|Boerner, Katy|Boyack, Kevin W.|Boyack, Kevin W.|Scharnhorst, A	MODELS OF SCIENCE DYNAMICS: ENCOUNTERS BETWEEN COMPLEXITY THEORY AND INFORMATION SCIENCES	2012
Social Security		BELONGING, SOLIDARITY AND	

Refine Query

Review basic statistics regarding result set. Revise search terms as needed.

Step 2: Review results and refine search



Refine Query

Author Name: Borner|Boerner

Journal Name:

Title or Abstract Keywords:

Publication Date: From 1900 To 2013

Query

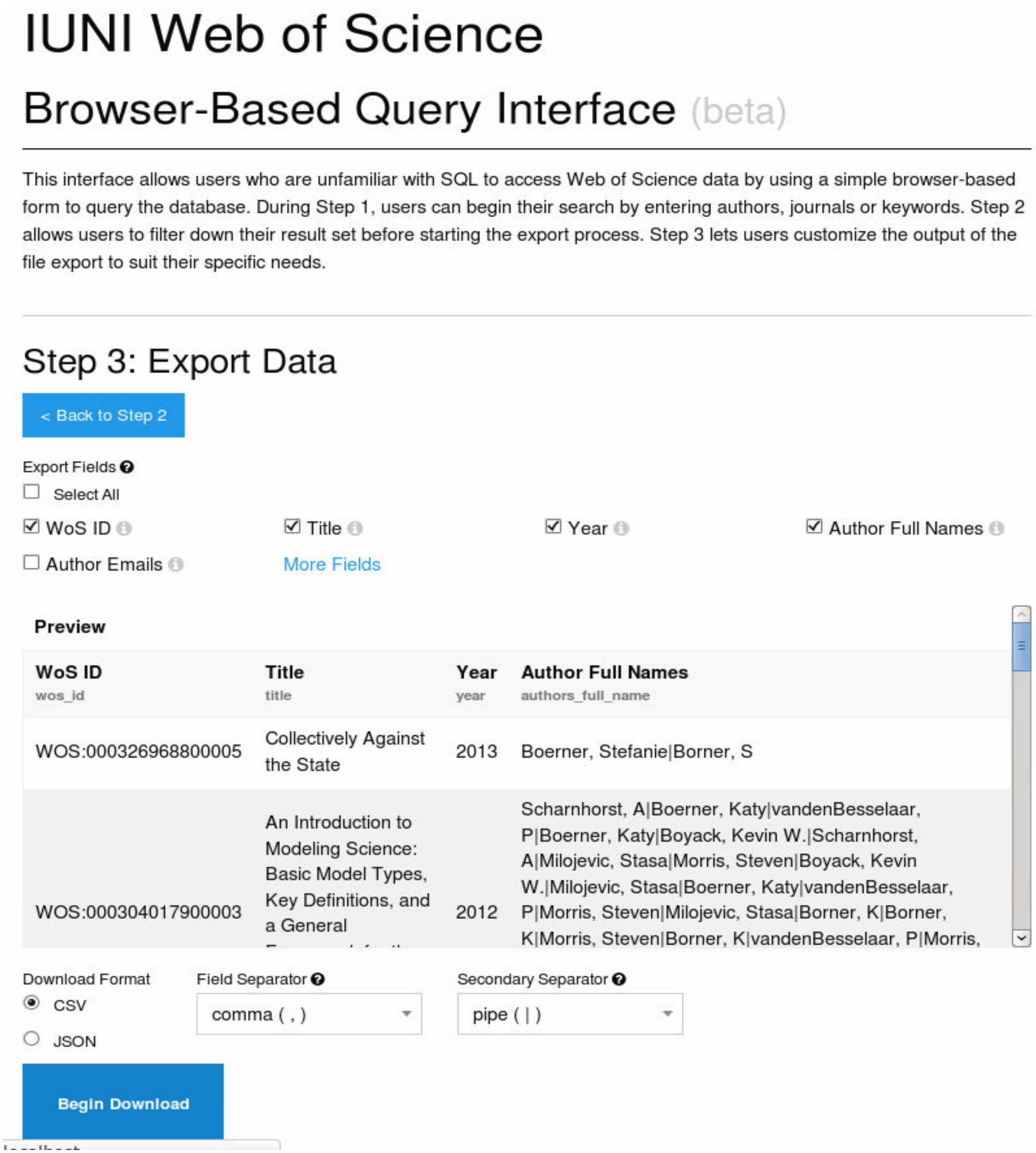
```
        SELECT DISTINCT
            record_id,
            title,
            authors_full_name as full_name,
            journals_name as journal_name,
            year,
            wos_id
        FROM
            interface_matview


        WHERE
```

© 2016 Indiana University Network Science Institute

Version: 0.1.14d

Refine query and see query that was used to generate the preview dataset.

Step 3: Export



IUNI Web of Science

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 3: Export Data

Export Fields

□ Select All

☑ WoS ID ☑ Title ☑ Year ☑ Author Full Names

□ Author Emails

Preview

WoS ID wos_id	Title title	Year year	Author Full Names authors_full_name
WOS:000326968800005	Collectively Against the State	2013	Boerner, Stefanie|Borner, S
WOS:000304017900003	An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General	2012	Scharnhorst, A|Boerner, Katy|vandenBesselaar, P|Boerner, Katy|Boyack, Kevin W.|Scharnhorst, A|Milojevic, Stasa|Morris, Steven|Boyack, Kevin W.|Milojevic, Stasa|Boerner, Katy|vandenBesselaar, P|Morris, Steven|Milojevic, Stasa|Borner, K|Borner, K|Morris, Steven|Borner, K|vandenBesselaar, P|Morris,

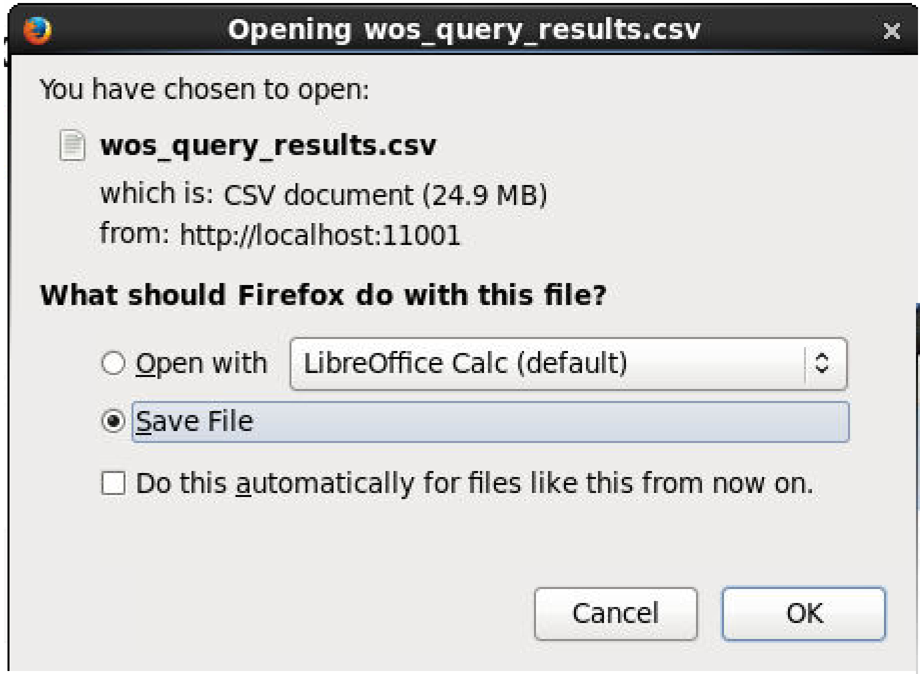
Download Format
● CSV
○ JSON

Field Separator: comma (,)

Secondary Separator: pipe (|)

Select relevant meta-data. Save in dir within data enclave.

Step 4: Save Resulting Data



Opening wos_query_results.csv

You have chosen to open:

wos_query_results.csv

which is: CSV document (24.9 MB)

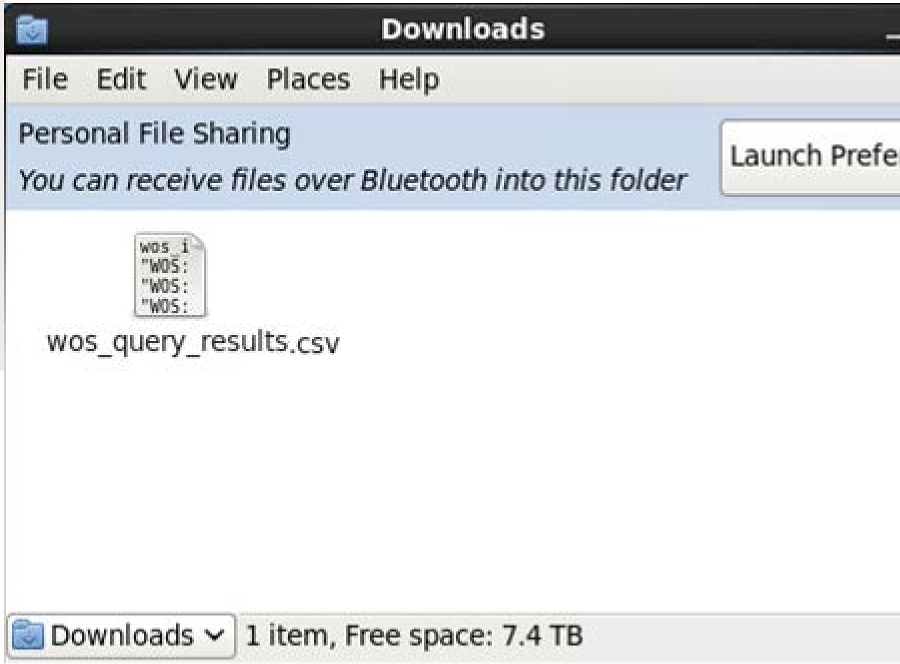
from: http://localhost:11001

What should Firefox do with this file?

○ Open with LibreOffice Calc (default)

● Save File

□ Do this automatically for files like this from now on.



Downloads

Personal File Sharing

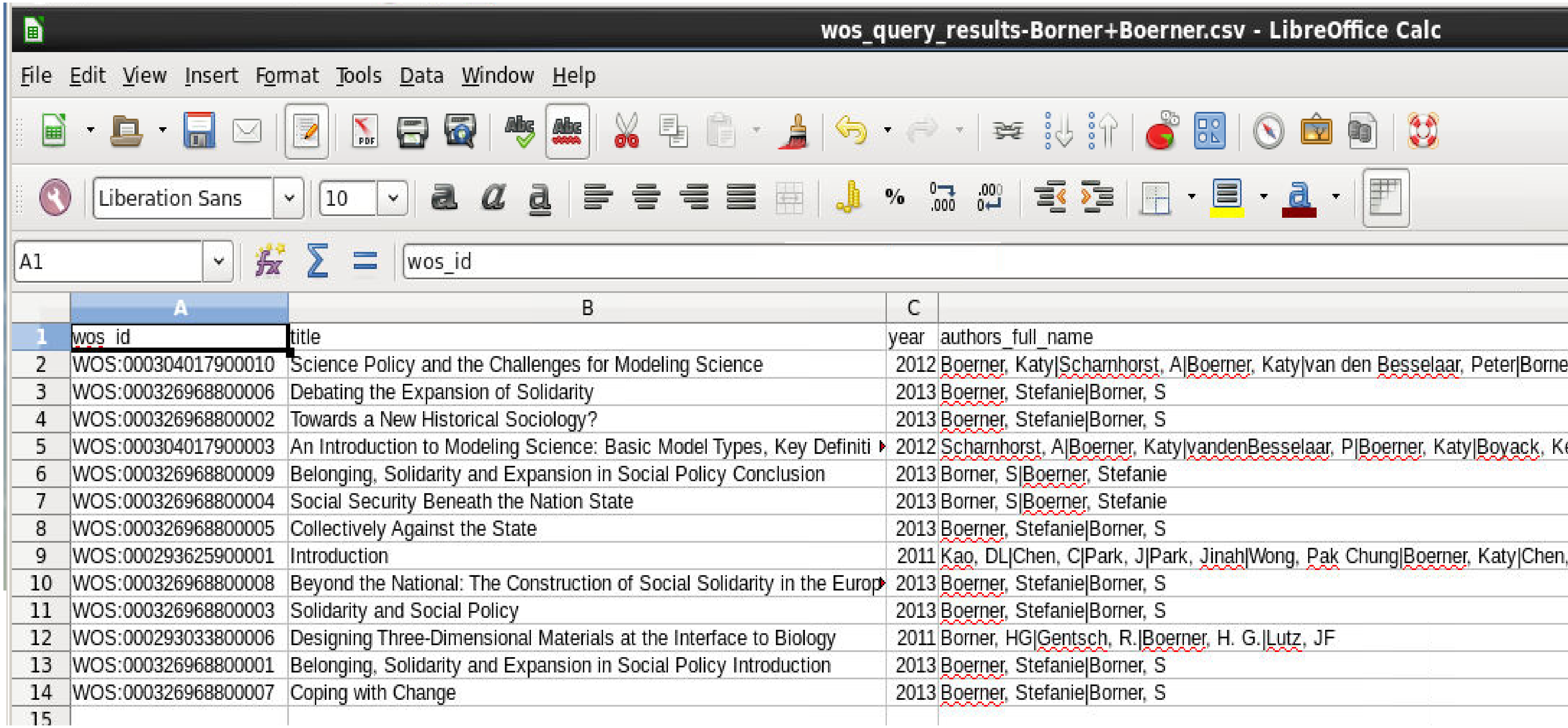
You can receive files over Bluetooth into this folder

wos_query_results.csv

Downloads 1 item, Free space: 7.4 TB

Select relevant meta-data. Save in dir within data enclave.

Step 5: Explore Data



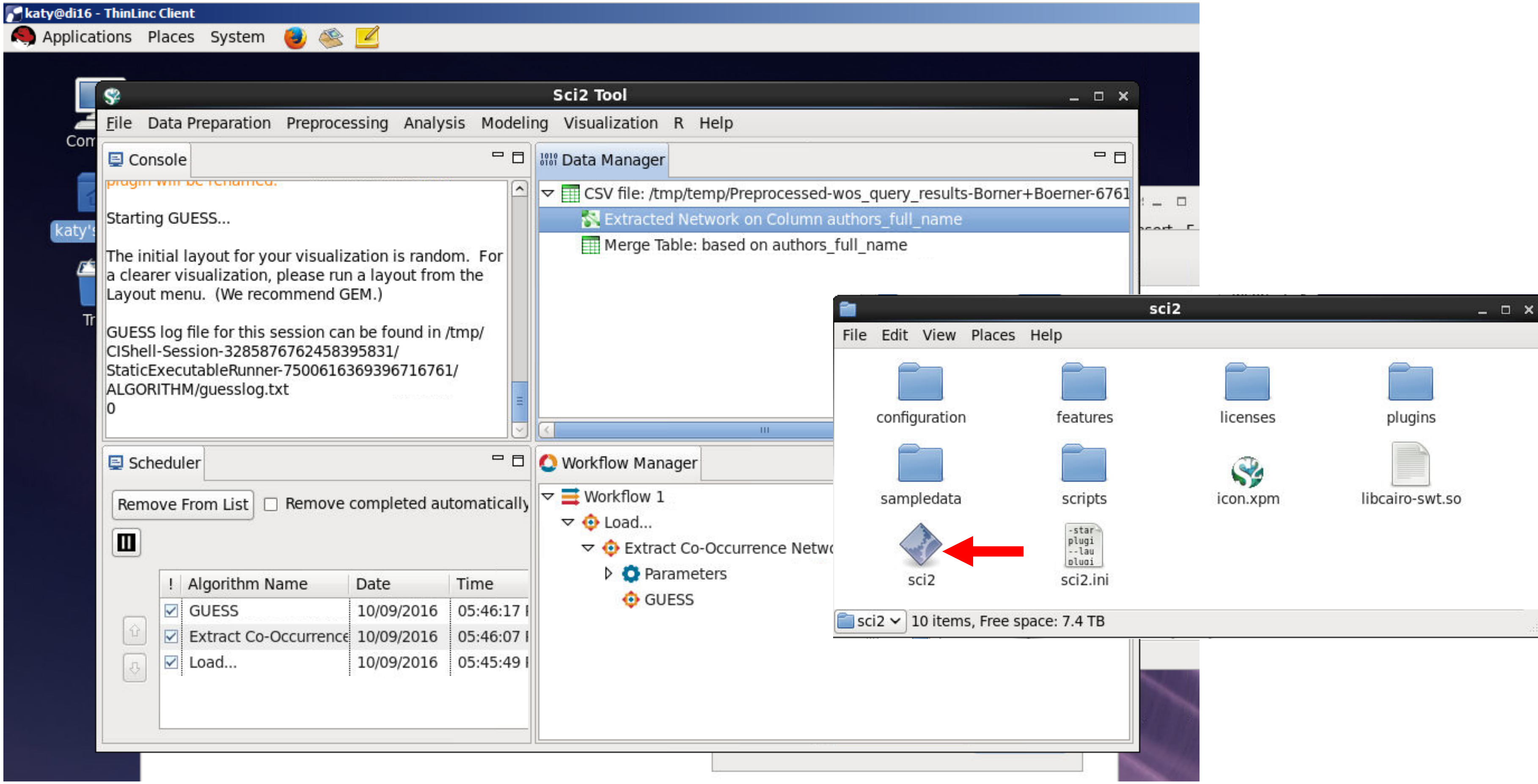
wos_query_results-Borner+Boerner.csv - LibreOffice Calc
File Edit View Insert Format Tools Data Window Help
Liberation Sans
10
A1
wos_id

	A	B	C	
1	wos_id	title	year	authors_full_name
2	WOS:000304017900010	Science Policy and the Challenges for Modeling Science	2012	Boerner, Katy|Scharnhorst, A|Boerner, Katy|van den Besselaar, Peter|Borne
3	WOS:000326968800006	Debating the Expansion of Solidarity	2013	Boerner, Stefanie|Borner, S
4	WOS:000326968800002	Towards a New Historical Sociology?	2013	Boerner, Stefanie|Borner, S
5	WOS:000304017900003	An Introduction to Modeling Science: Basic Model Types, Key Definiti	2012	Scharnhorst, A|Boerner, Katy|vandenBesselaar, P|Boerner, Katy|Boyack, Ke
6	WOS:000326968800009	Belonging, Solidarity and Expansion in Social Policy Conclusion	2013	Borner, S|Boerner, Stefanie
7	WOS:000326968800004	Social Security Beneath the Nation State	2013	Borner, S|Boerner, Stefanie
8	WOS:000326968800005	Collectively Against the State	2013	Boerner, Stefanie|Borner, S
9	WOS:000293625900001	Introduction	2011	Kao, DL|Chen, C|Park, J|Park, Jinah|Wong, Pak Chung|Boerner, Katy|Chen,
10	WOS:000326968800008	Beyond the National: The Construction of Social Solidarity in the Europ	2013	Boerner, Stefanie|Borner, S
11	WOS:000326968800003	Solidarity and Social Policy	2013	Boerner, Stefanie|Borner, S
12	WOS:000293033800006	Designing Three-Dimensional Materials at the Interface to Biology	2011	Borner, HG|Gentsch, R.|Boerner, H. G.|Lutz, JF
13	WOS:000326968800001	Belonging, Solidarity and Expansion in Social Policy Introduction	2013	Boerner, Stefanie|Borner, S
14	WOS:000326968800007	Coping with Change	2013	Boerner, Stefanie|Borner, S
15				

Open in LibreOffice Calc

Save results within data enclave or ask Data Steward to extract results.

Step 6: Data Analysis and Visualization



```
Starting GUESS...

The initial layout for your visualization is random.  For
a clearer visualization, please run a layout from the
Layout menu.  (We recommend GEM.)

GUESS log file for this session can be found in /tmp/
CIShell-Session-3285876762458395831/
StaticExecutableRunner-7500616369396716761/
ALGORITHM/guesslog.txt
0
```

- CSV file: /tmp/temp/Preprocessed-wos_query_results-Borner+Boerner-6761
  - Extracted Network on Column authors_full_name
  - Merge Table: based on authors_full_name

□ Remove completed automatically

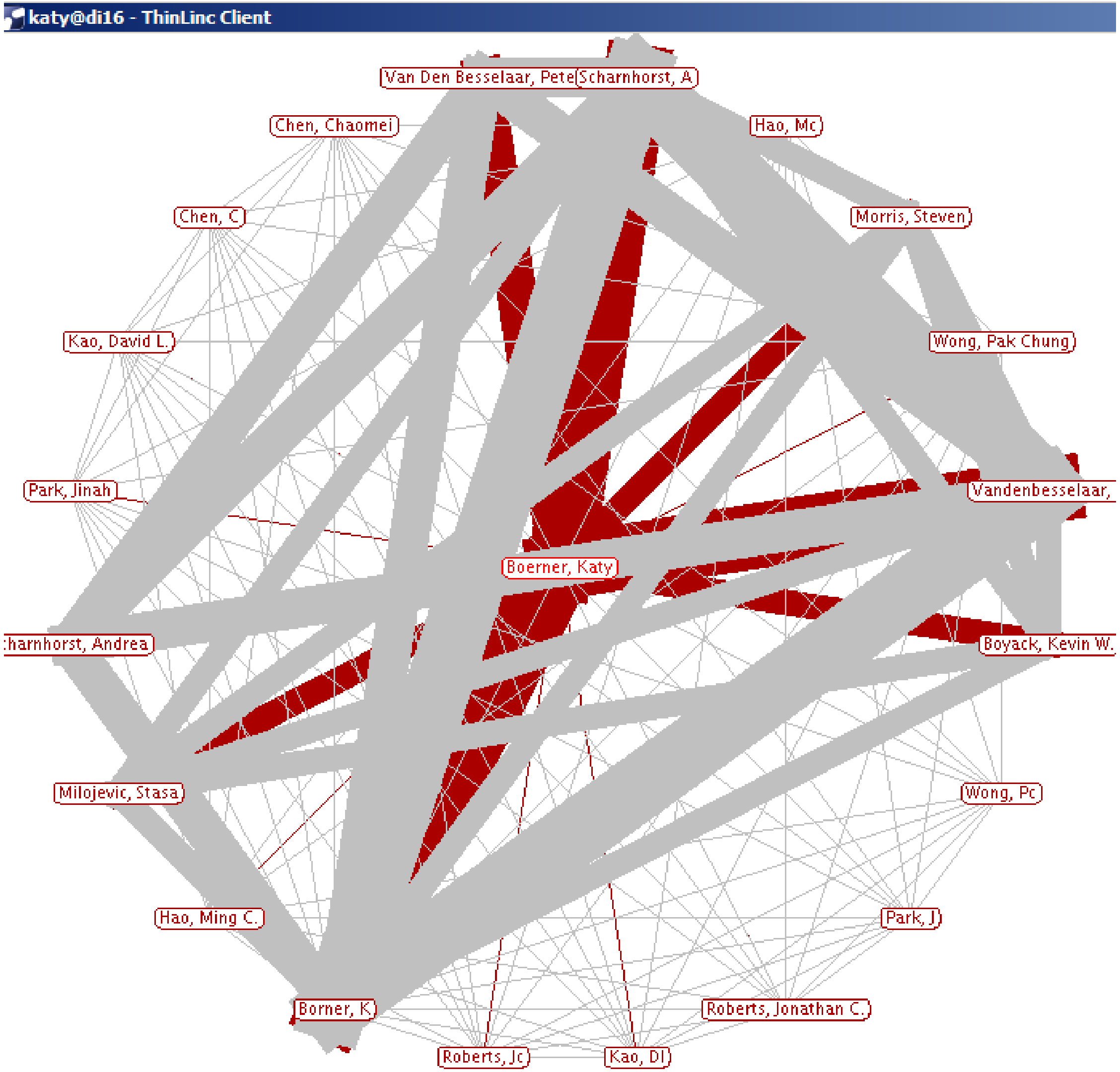
!	Algorithm Name	Date	Time
☑	GUESS	10/09/2016	05:46:17
☑	Extract Co-Occurrence	10/09/2016	05:46:07
☑	Load...	10/09/2016	05:45:49

- Workflow 1
  - Load...
    - Extract Co-Occurrence Netw
      - Parameters
      - GUESS

sci2 10 items, Free space: 7.4 TB

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

IUNI Web of Science Data

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.

IUNI Web of Science Data

Planned WoS Data Usage

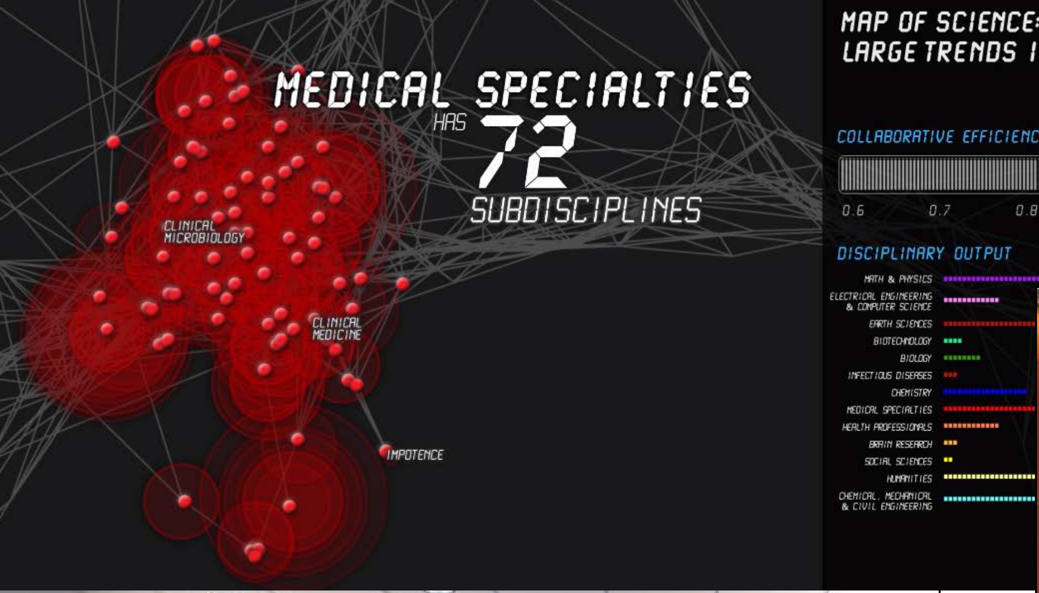
Special agreements exist for using WoS publications by IU faculty. Data can be used to

- 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 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.

Science Forecast S1:E1, 2015



SCIENTIFIC COLLABORATIONS BETWEEN WORLD CITIES
BY OLIVIER H. BEAUCHESNE
SCIENTIFIC COLLABORATIONS (MILLIONS)

Web of Science as a Research Dataset

Date:

November 14-15, 2016

Meeting Place:

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

Web Indiana University Campus Map »

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.
- Researcher 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.

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