IUNI Web of Science Data Enclave 101

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

Val Pentchev and Matt Hutchinson
University Network Science Institute
Indiana University, USA

March 25, 2016

IUNI Web of Science (WoS) Data Enclave Access Instructions

Compiled by the IUNI WoS Data Advisory Board

Introduction

The IUNI WoS Data Enclave is a secure repository containing Thompson Reuters Web of Science XML raw data.

Access will be granted to the user's Karst account. IU students, faculty, staff, and qualifying sponsored affiliates can request accounts on Karst.

More about the IUNI WoS Data Enclave can be found at IUNI WoS web page at http://www.indiana.edu/~iuni/resources/wos.html

Overview

- The Data
- How to Request Access to Enclave
- How to Login to Enclave
- How to use the WoS Database to Run Queries
- How to use Excel and Word-like Programs
- Running Sci2 Tool
- How to Extract Results from Enclave

- Enclave Logging
- Policies

The Data

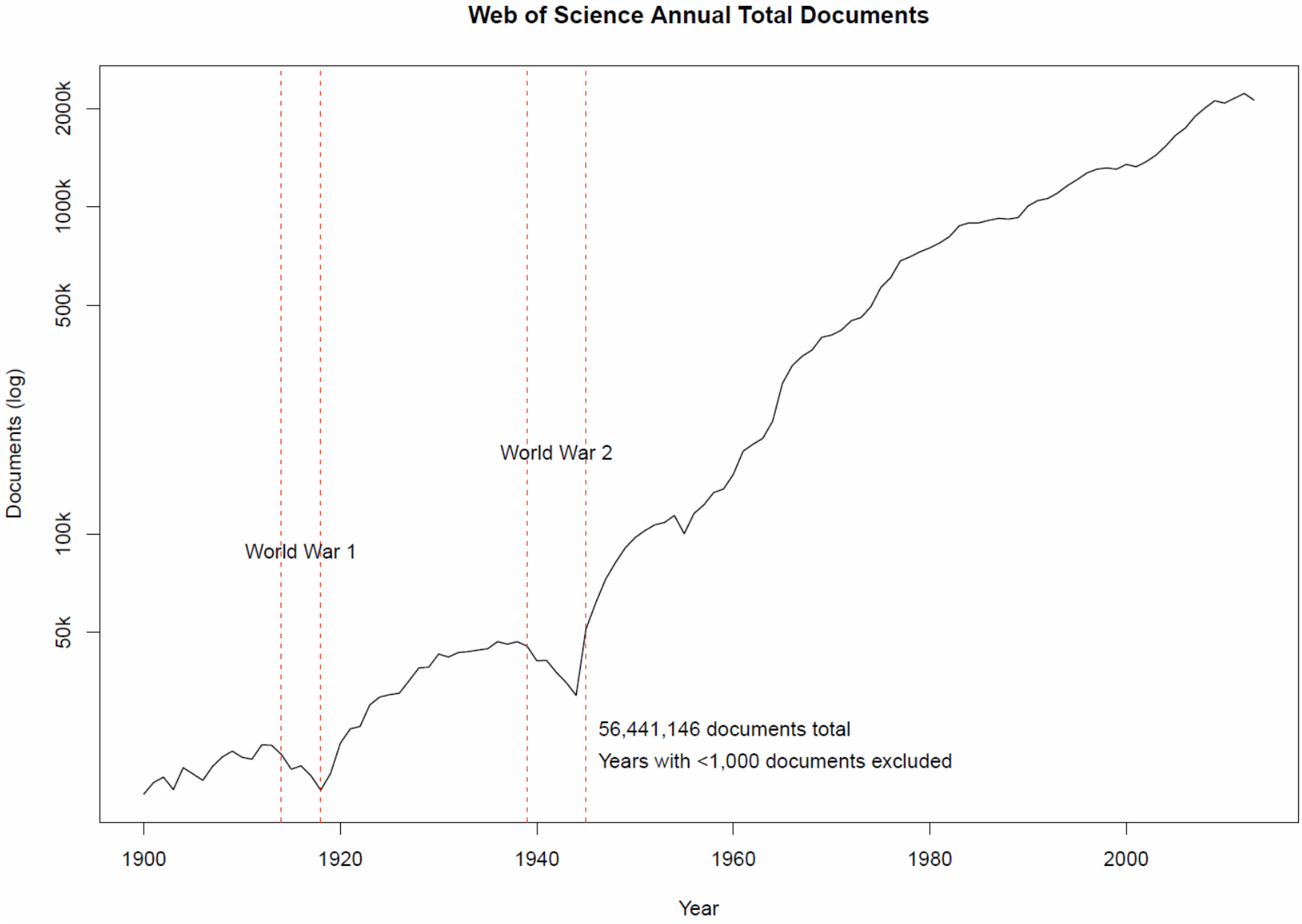
The IUNI Science of Science Hub acquired the complete set of Thomson Reuters' Web of Science XML raw data (Web of Knowledge version 5) comprising

- 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

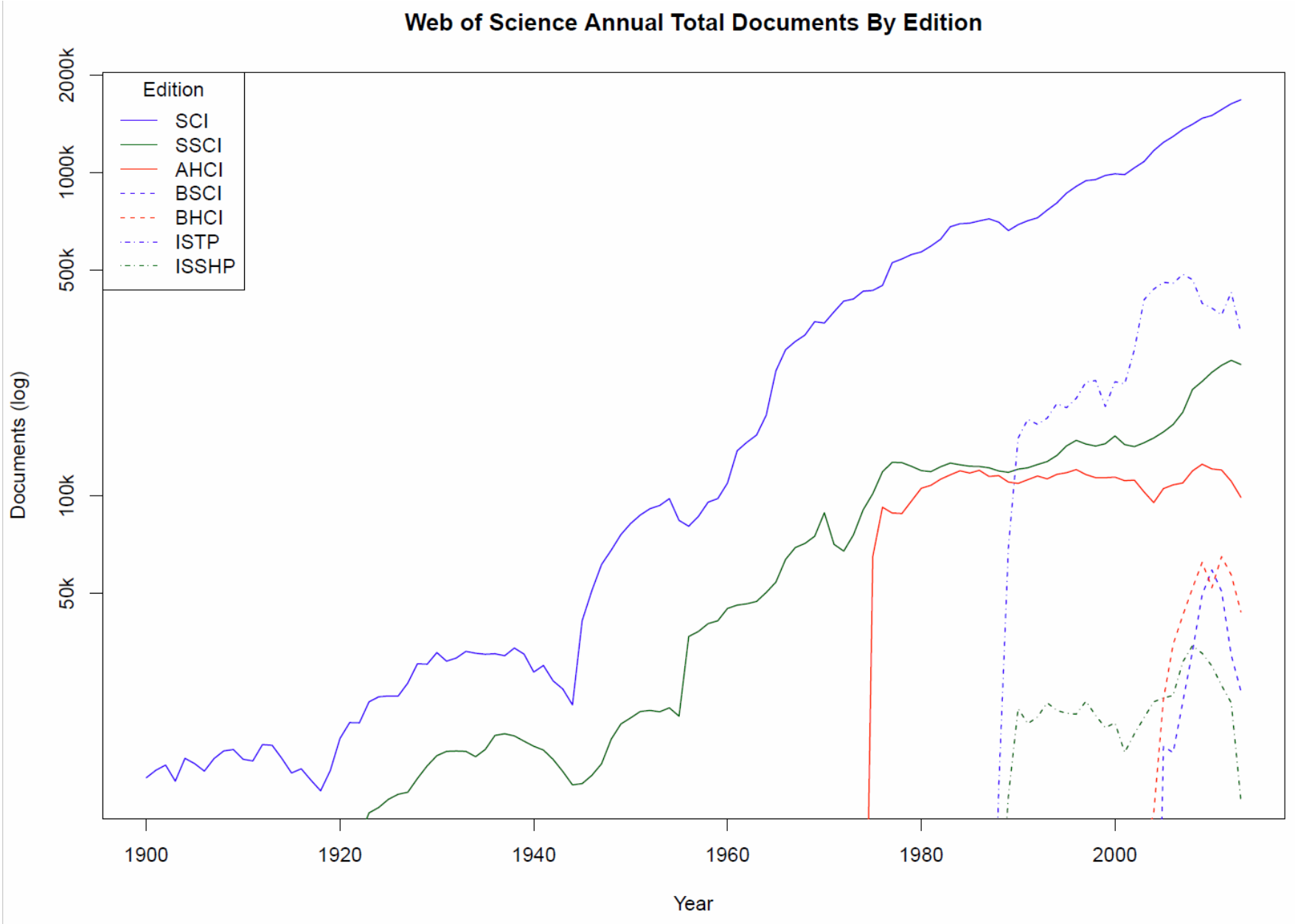
Basic Statistics:

- 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]

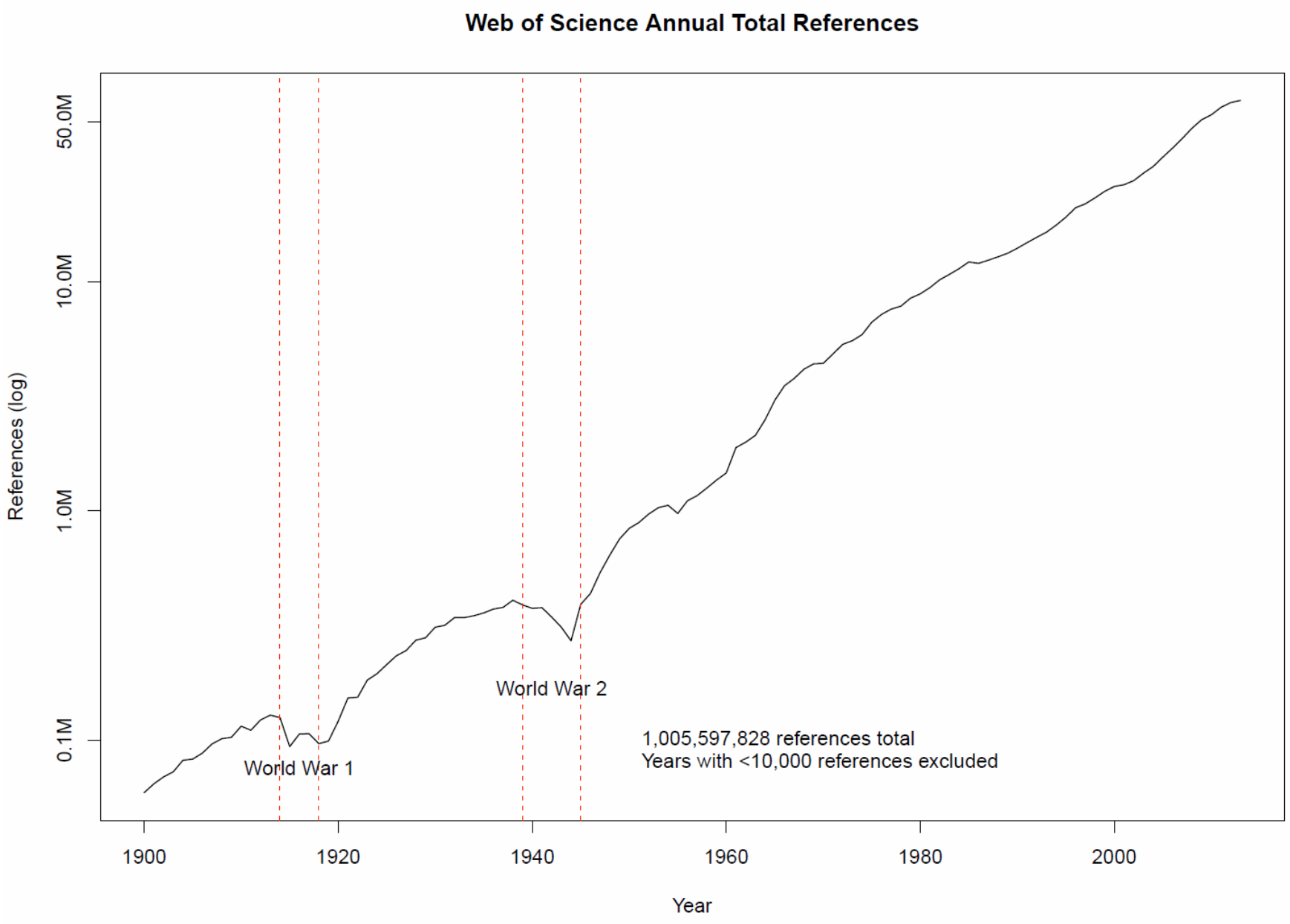
IUNI Web of Science Data



IUNI Web of Science Data

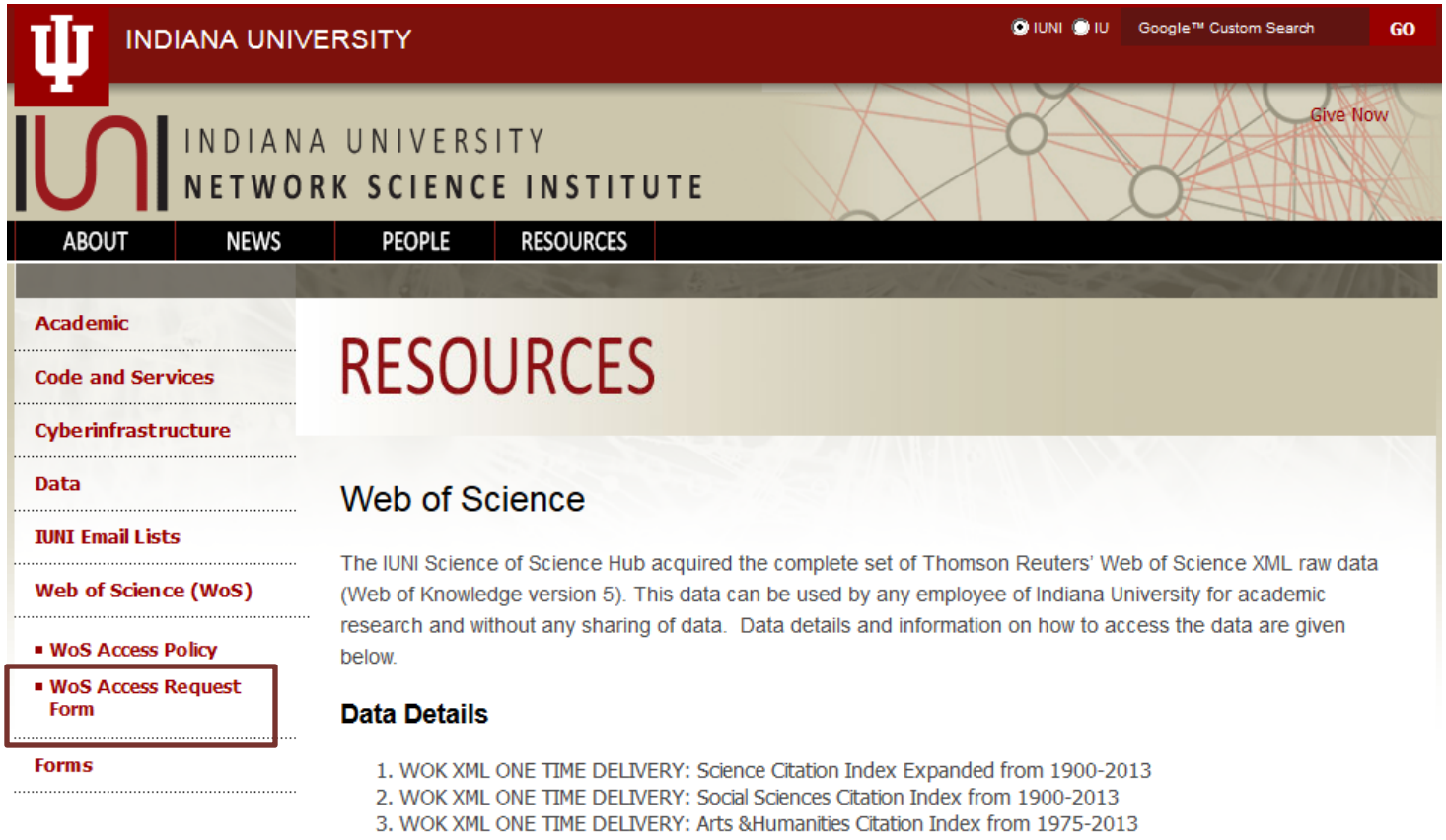


IUNI Web of Science Data



How to Request Access to Enclave

On http://www.indiana.edu/~iuni/resources/wos.html select WoS Access Request Form



Web of Science - Access Request Form

- updated 11.09.2015 -

Personal Information

Full Name: Last _____ First _____ M.I. _____

Department/ School: _____ - **Full listing of IUB Academic Departments, Centers, and Institutes**

IU Campus: IU Bloomington

Research Area: _____

IU Email Address : _____

Expected research Period: start date _____ end date _____

CNS Cyberinfrastructure for Network Science Center

Project Information

Project Title:

Project Abstract: *(How will the data be used?*)*

Project Weblink:

Project Funding: *(Current or planned)*

What additional data and tools do you plan to use?

CNS Cyberinfrastructure for Network Science Center

Data Request Information

What WoS data do you plan to use:

(Years, fields, etc.)

Data Format ☐ CSV ☐ XML ☐ Other Specify format:

Type of Access:

☐ Raw XML & WoS DB via Data Enclave

☐ Custom Custom Datasets Retrieved from WoS DB

What data do you plan to extract from the enclave?

What other team members will need access?

CNS Cyberinfrastructure for Network Science Center

Data Access & Usage Conditions

Confirm that you understand that:

- ☐ **Data Sharing and Storage:**
  Data has to be stored on secure media. The data should not be shared with anyone. Usernames and passwords should not be shared with anyone. You agree to safeguard your username/password. Any unauthorized use of that account will be your individual responsibility. Any loss of your personal control over the username or password has to be reported to the WoS Data Stewart. For a full list of information security polices please visit http://policies.iu.edu/policies/categories/information-it.

- ☐ **Data Usage:**
  This data may not be used for any other purpose besides that in your approved request for access.

- ☐ **Acknowledgement:**
  All publications that result from using the data need to have the acknowledgement text: *"This work uses Web of Science data by Thomson Reuters provided by the Network Science Institute and the Cyberinfrastructure for Network Science Center at Indiana University."*

- ☐ **Publications:** Scholarly works that result from using the data have to be submitted to the WoS Data Stewart as soon as they are published.

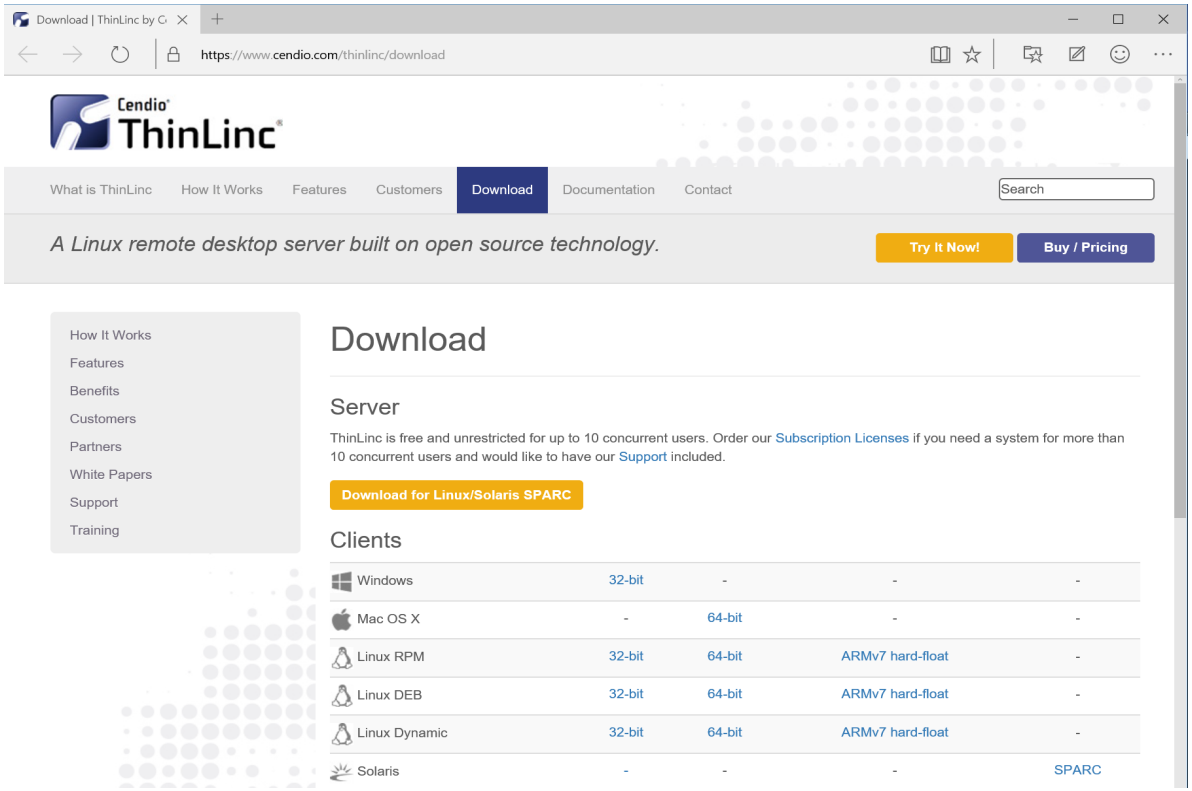
- ☐ **Communication:** You will be subscribed to the IUNIWoS@iu.edu email list that provides information on WoS data updates and user meetings and can be used to ask WoS relevant questions.

- ☐ ***I do affirm that to the best of my knowledge all information provided above is complete and true, and that I will comply with all Indiana University policies with regards to the use of sensitive data.***

Submit

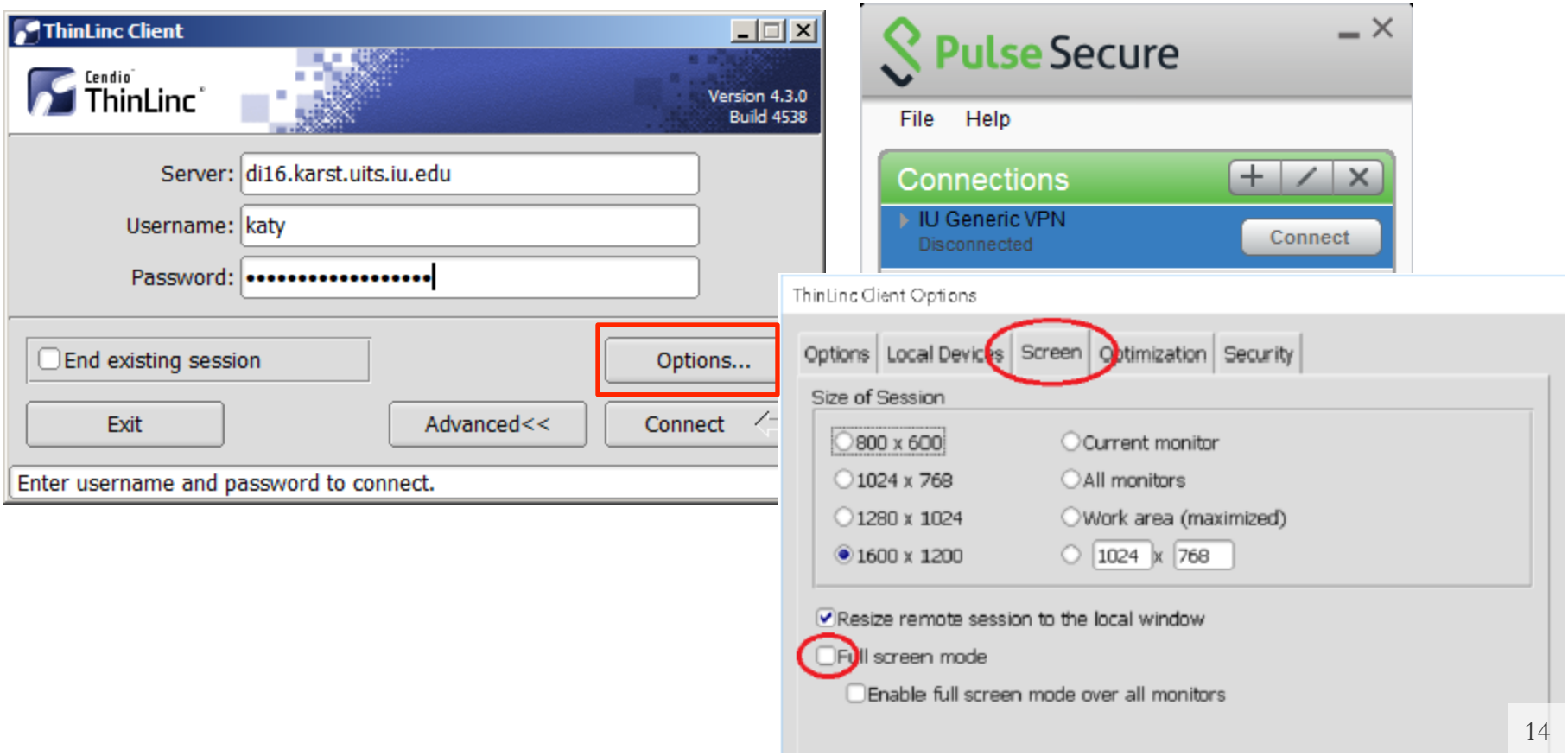
How to Login to the Enclave

Download free ThinLinc VNC client from Cendio.



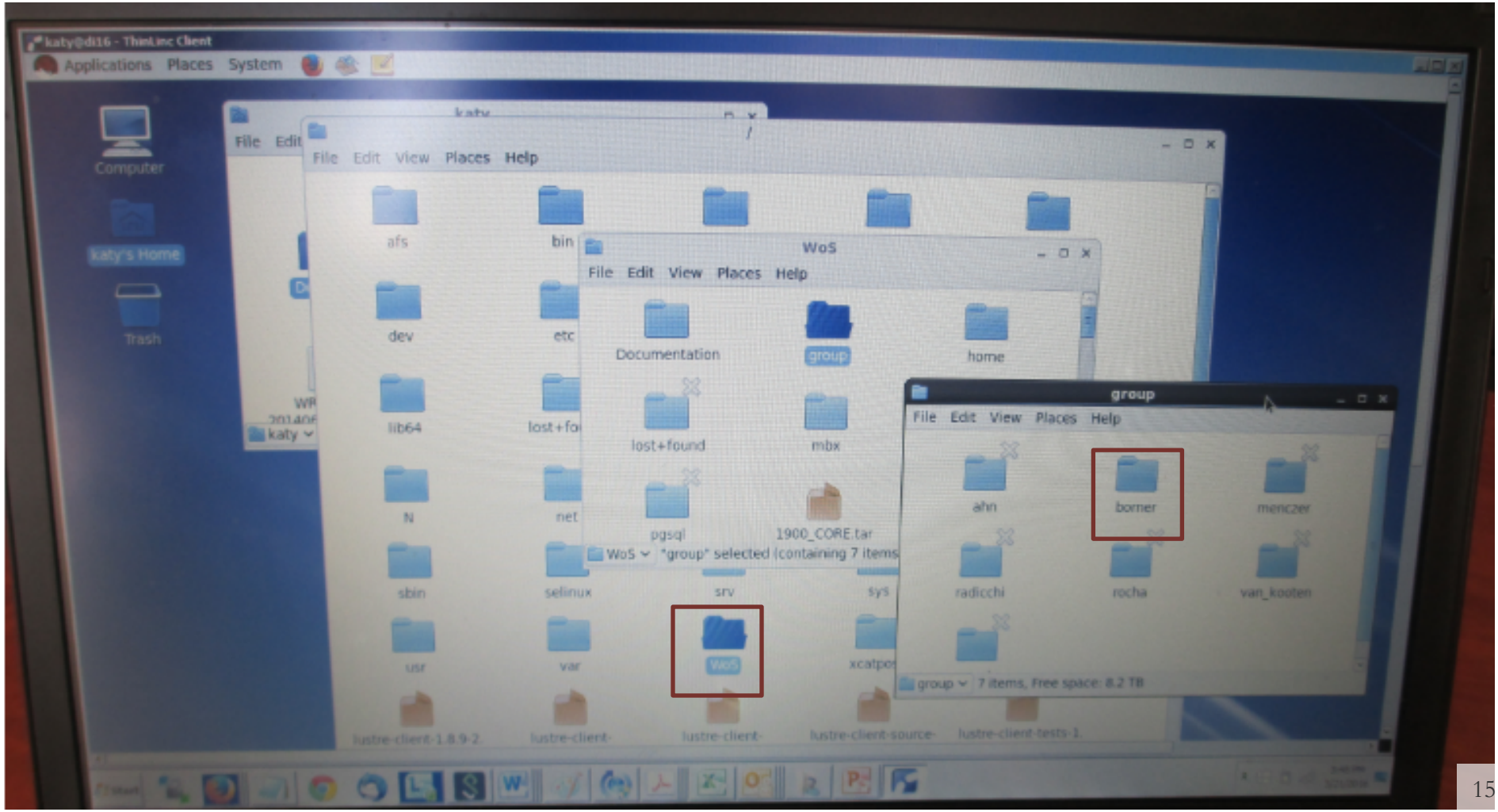
How to Login to the Enclave

Once installed, ThinLinc uses regular IU login name and password that has been granted access to the data enclave.



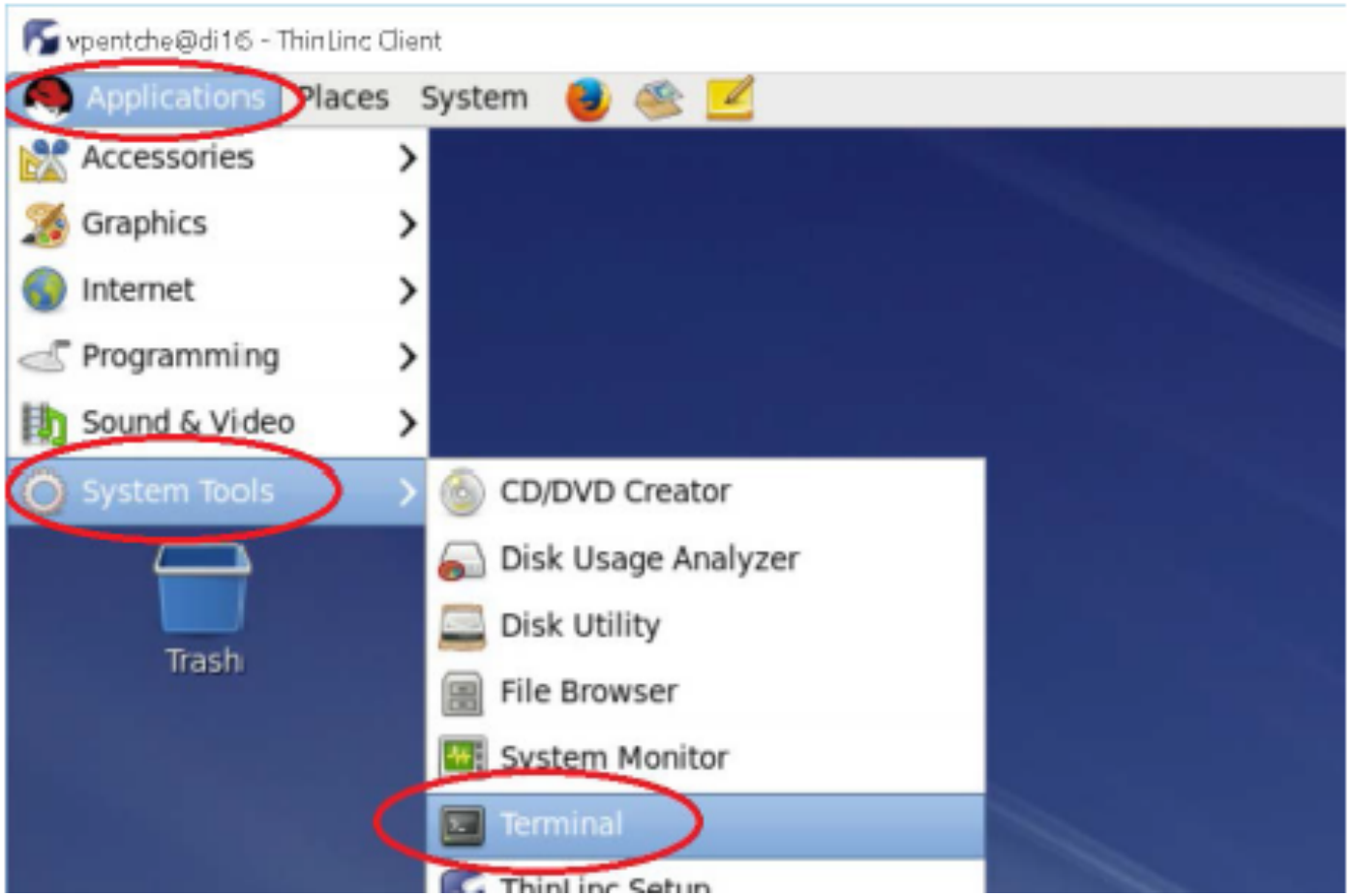
Enclave Virtual Desktop

GNOM graphical interface of Red Hat Enterprise Server 6.6.



Enclave Virtual Desktop

To access the WoS folder using the command line – please open Terminal by selecting Applications > System Tools > Terminal and issue the command *cd /WoS* at the prompt.

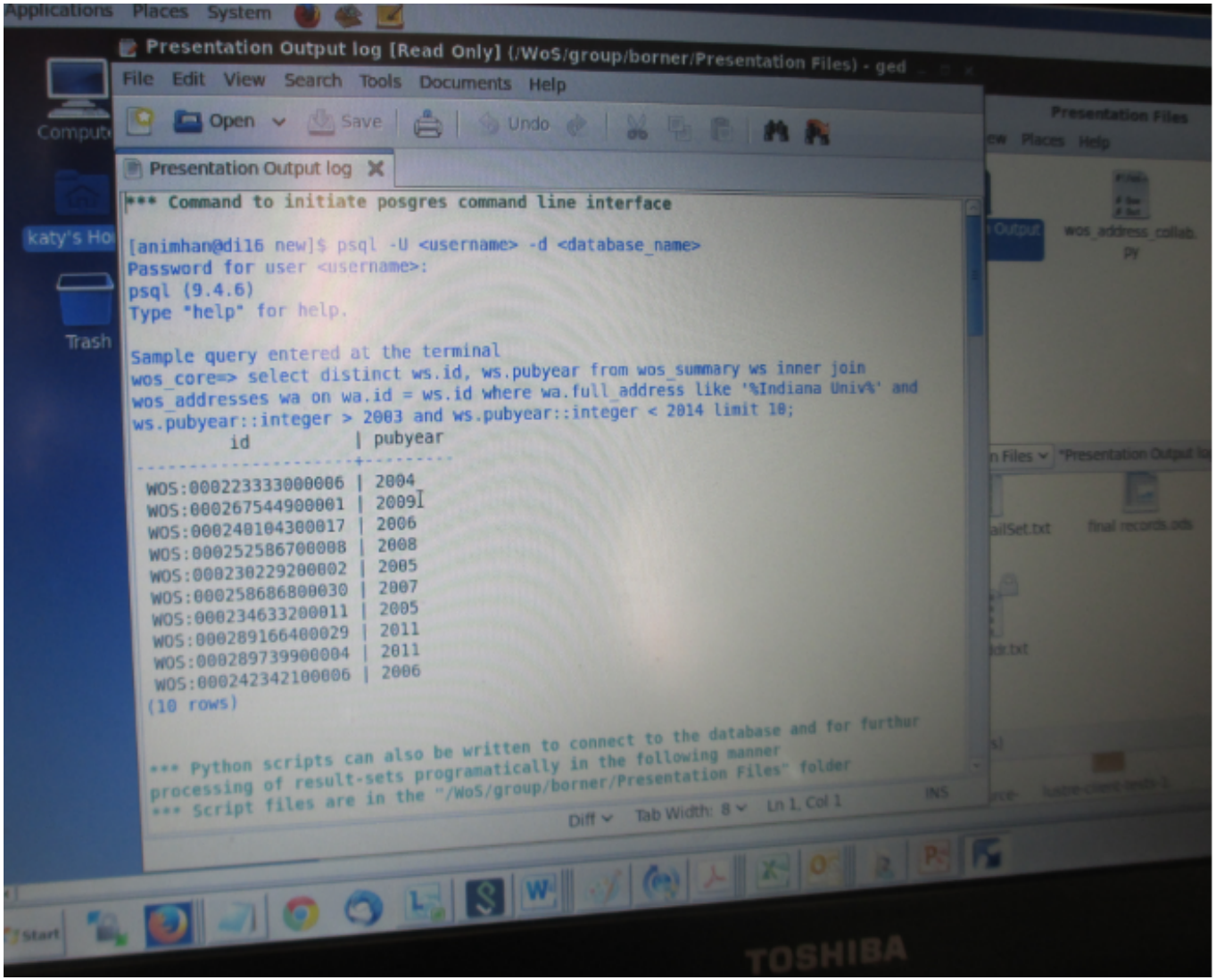


Manipulating the Web of Science Data

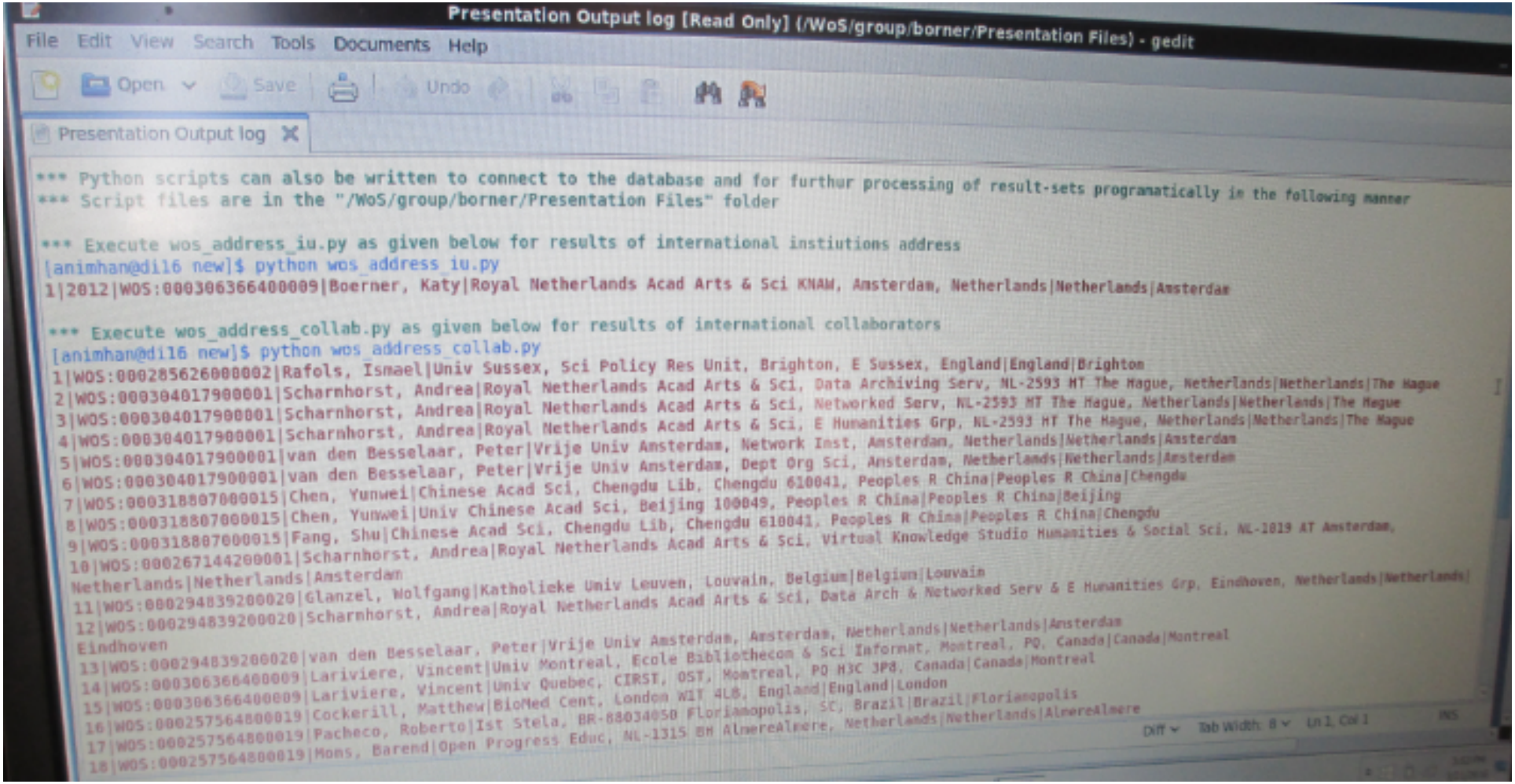
The IUNI WoS Data Enclave is a part of the Karst high-throughput computing cluster and offers access to all software currently in the Modules environment management system. Below is a list of the available packages as of August 2015.

am3	chezscheme	fsl	hmmer	matlab	netcdf-c	pvm	stattransfer
amber	clustalw2	ga	hpss	meep	netcdf-fortran	python	swig
Ansys	cmake	gamess	imod	mentalray	netlib	Qt	Tau
Arpack	collectl	gap	impute2	migrate	ngsutils	R	Tiff
Atlas	cpmd	gauss	imsl	mimar	nltk-data	rats	timbl
auto07p	cryoem	gaussian	intel	minimac	openfoam	raxml	totalview
autoconf	cufflinks	gcc	intel-mpi	miso	openmpi	rcs	Tpp
autodocksuite	curl	geant	itk	molcas	oracleds	rosetta	Tre
automake	dl_poly	git	java	mono	otf	rsem	valgrind
bedtools	dose2geno	gmp	kpp	mothur	p7zip	rstudio	vampirtrace
Bfast	doxygen	gnuplot	lammps	motif	papi	ruby	vcftools
Bigjob	dyninst	grads	lastools	mpc	pari	samtools	vmd
bioconductor	emboss	grass	ldhat	mpfr	paup	sas	Vxl
biopython	expat	gromacs	leptonica	mpich	pdt	score-p	weka
Blat	farsight	gsl	lincrna	mrbayes	perl	siesta	wgrib
Boost	fastqc	gulp	local-utils	mummer	petitechezscheme	splus	Wrf
bowtie	fftw	gurobi	mach	muscle	pgi	spm	wxpython
breakdancer	flexbar	harminv	macs	nag	phast	spss	xpdf
breseq	freesurfer	hdf4	maple	namd	phylip	sra-toolkit	xplor-nih
cableswig-itk	freetype2	hdf5	mathematica	ncbiblast+	plink	stata	

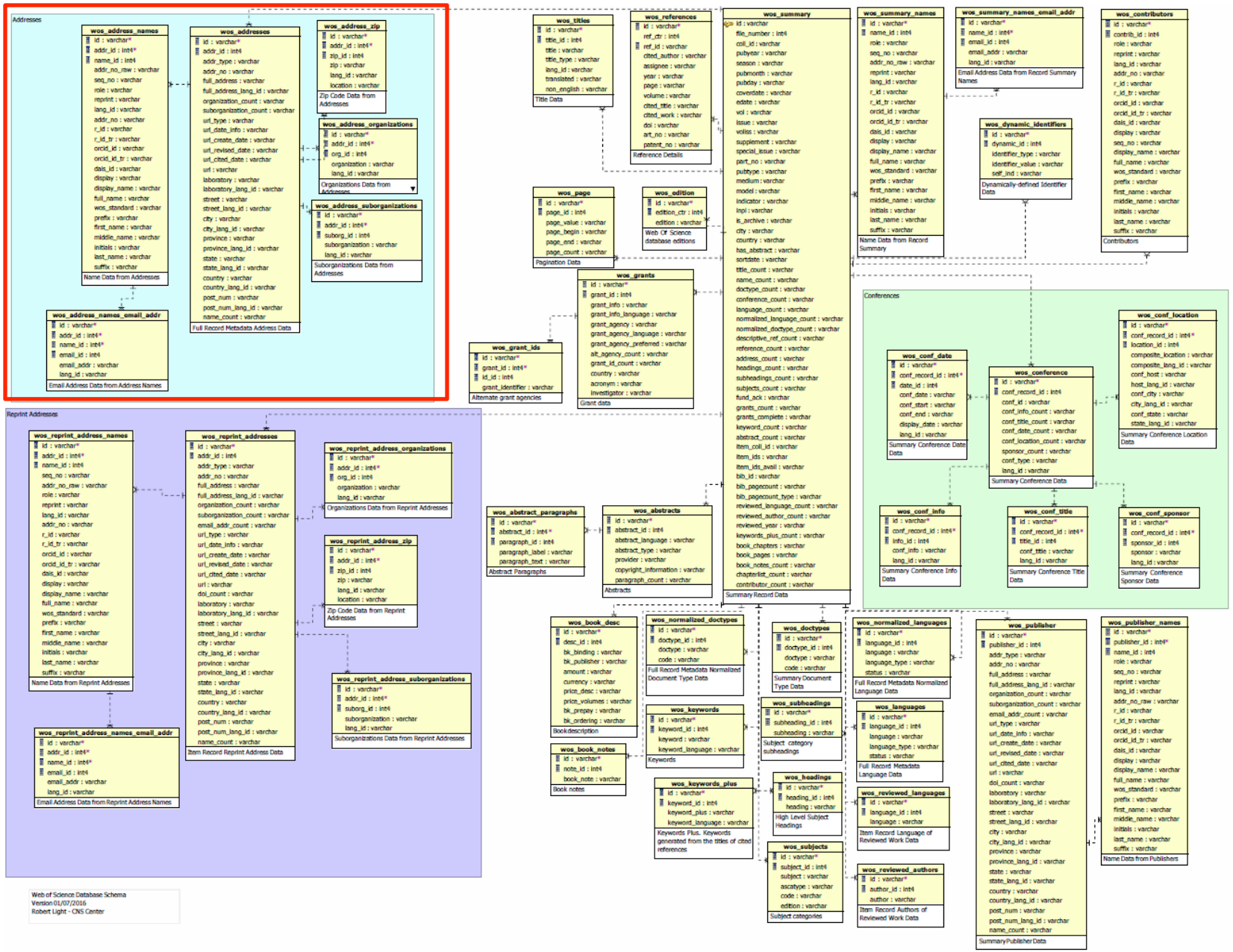
How to use the WoS Database to run queries

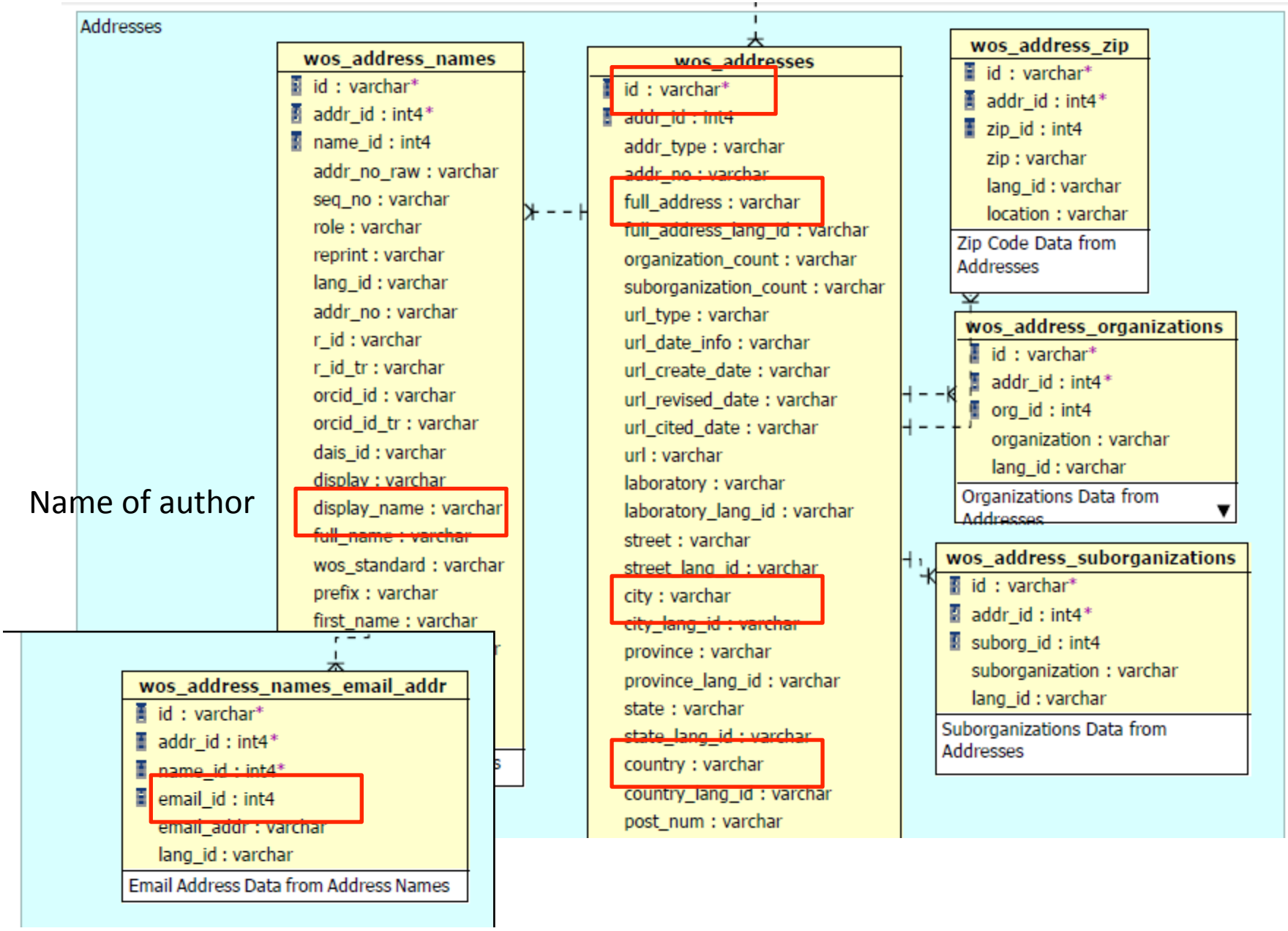


IUNI Web of Science Data



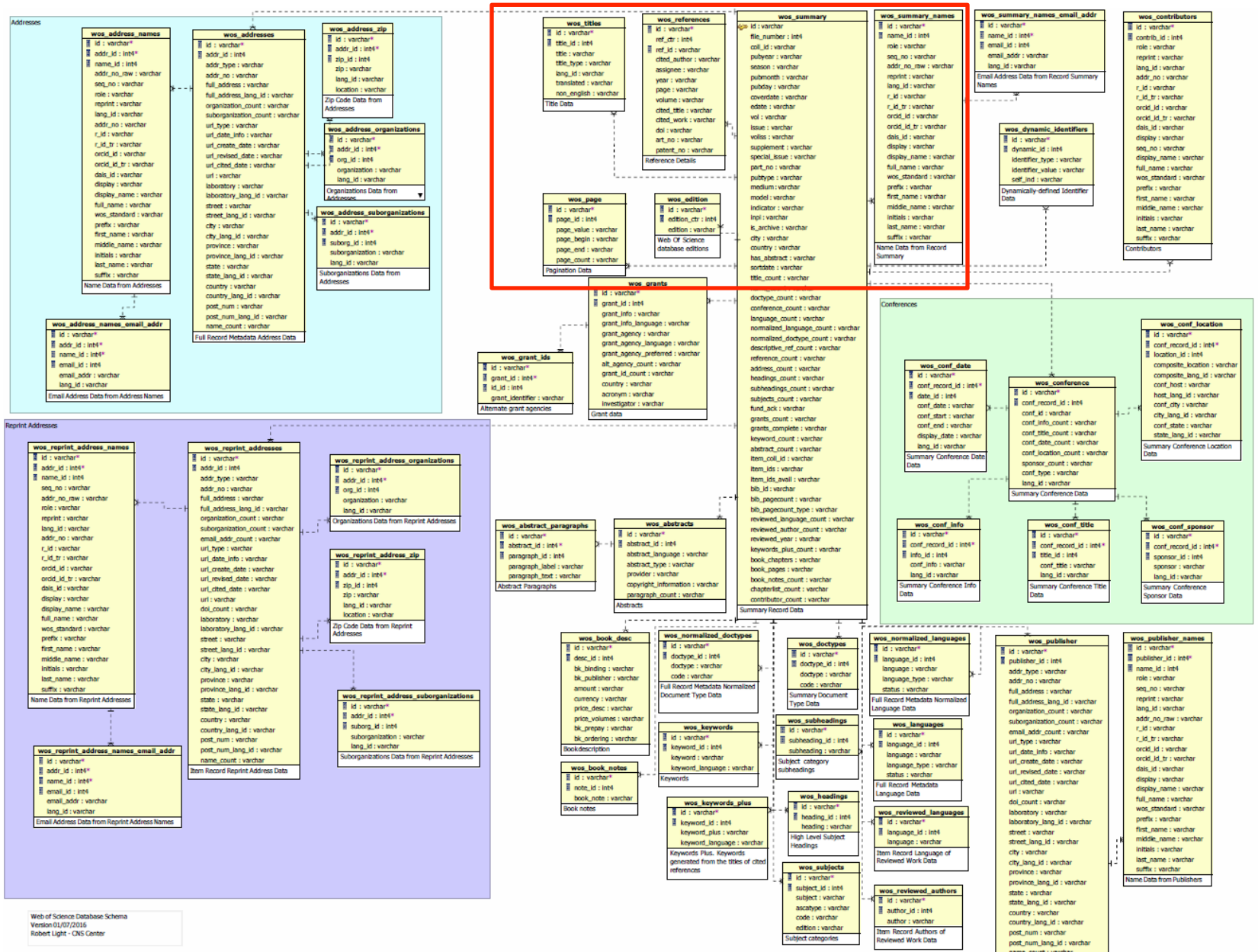
IUNI Web of Science Data

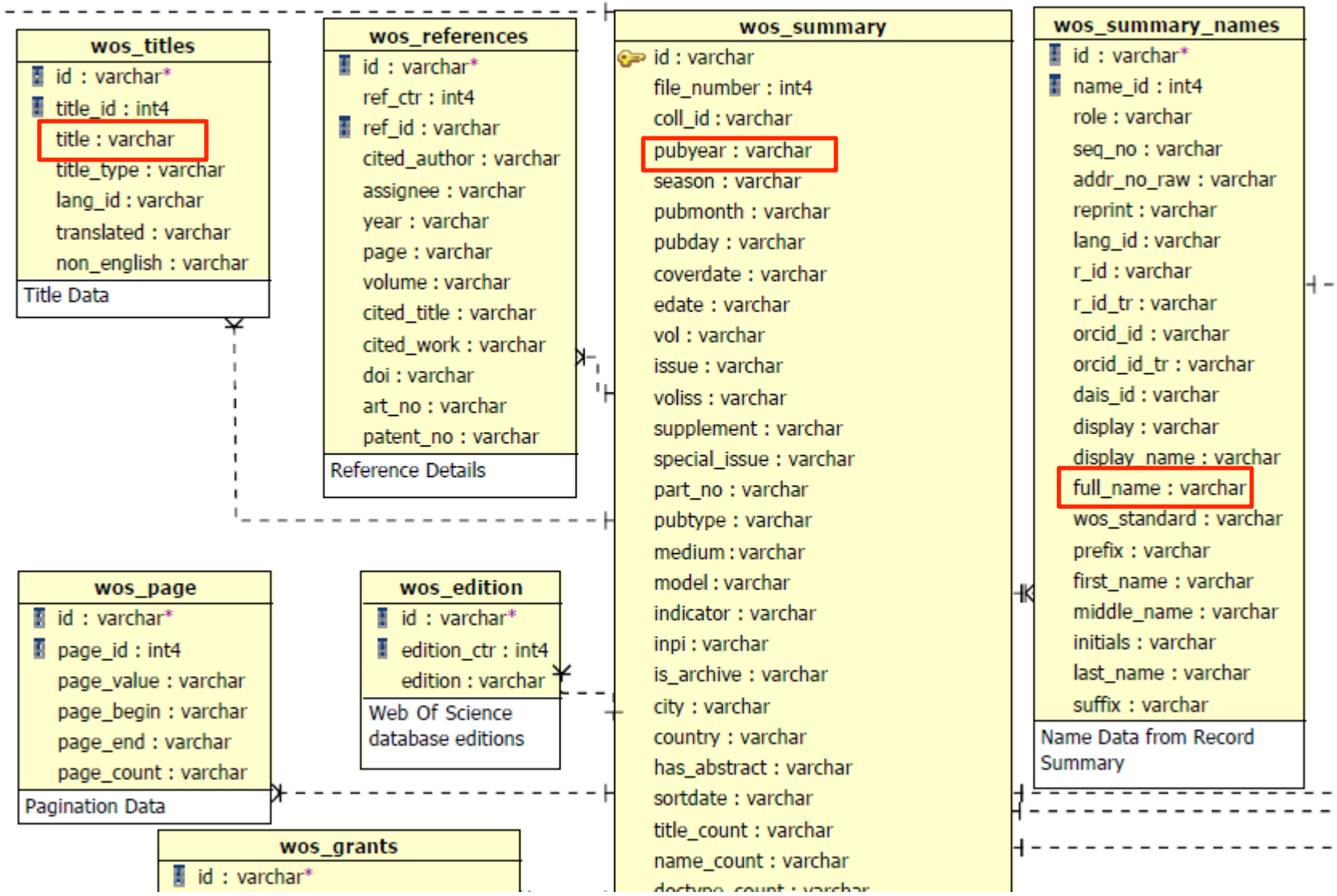




Name of author

IUNI Web of Science Data

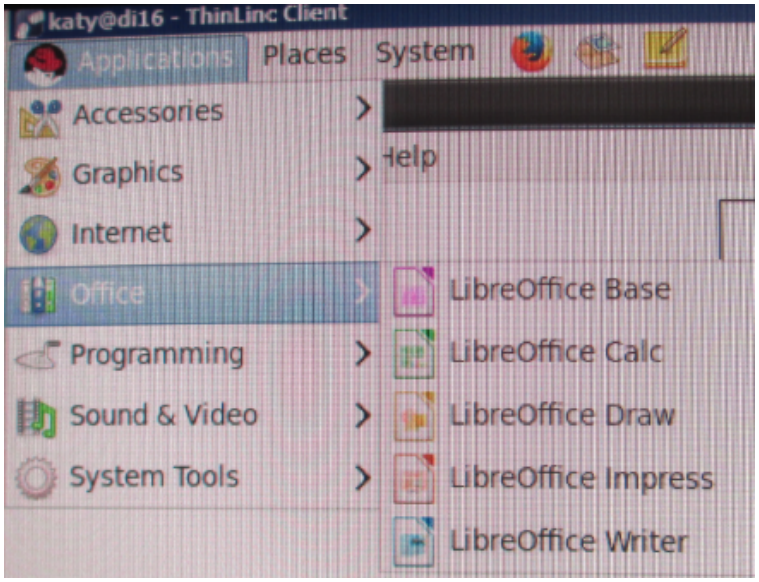



wos_titles
id : varchar*
title_id : int4
title : varchar
title_type : varchar
lang_id : varchar
translated : varchar
non_english : varchar
Title Data
wos_references
id : varchar*
ref_ctr : int4
ref_id : varchar
cited_author : varchar
assignee : varchar
year : varchar
page : varchar
volume : varchar
cited_title : varchar
cited_work : varchar
doi : varchar
art_no : varchar
patent_no : varchar
Reference Details
wos_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
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_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*

How to use Excel and Word-like Programs

Open data files in Excel like program.

Write paper inside the Enclave.



Additional software can be requested by contacting the IUNI WoS Data Steward. Upon review, new software may be installed on the Enclave.

Scripts and temporary files should be stored in either users' home folders or project team folders on the enclave.

How to Extract Results from Enclave

Each research team will have a mailbox directory established to facilitate data communication in and out of the system. This space is readable only by the team and the Data Steward and can be found at **/WoS/mbx/PI-last-name**. When a team has research artifacts (tables, graphs, aggregated data) that they wish to remove, these files should be copied into that team's mailbox.

Then, email the Data Steward (iuniwosd@indiana.edu) with the description of what files you intend to remove. Please provide complete file name and brief description of file content (e.g., author-names.txt; list of top-100 high reputation authors; graph.gif, degree distribution of co-author network). All requests will be reviewed by the Data Steward to ensure compliance with

Thompson Reuters' data agreement terms and conditions. Upon approval, data will be securely delivered to the requestor. Further instructions will be provided from that point, based on the size and nature of the data to be removed as to the best way to deliver it.

Enclave Logging

All user activity in the IUNI WoS Data Enclave is monitored by UITS in real time to identify possible security violations and abnormal usage.

Aggregate statistics of user activity (e.g., time logged into the system, number of programs used, total amount of space consumed by Enclave users) will be compiled at the aggregate level to communicate Enclave usage.

Policies

- Only current versions of tools are provided.
- Max processor usage per account is limited to 60% of total.
- Students should list a faculty or staff mentor.

See also IUNI WoS Data Policy

Slides from the Jan 11, 2016 presentation about this unique dataset and the data enclave functionality are at http://cns.iu.edu/docs/presentations/2016-borner-wos-enclave.pdf

Questions and Answers