

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>



CNS

Cyberinfrastructure for
Network Science Center

IUNI Web of Science Data



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

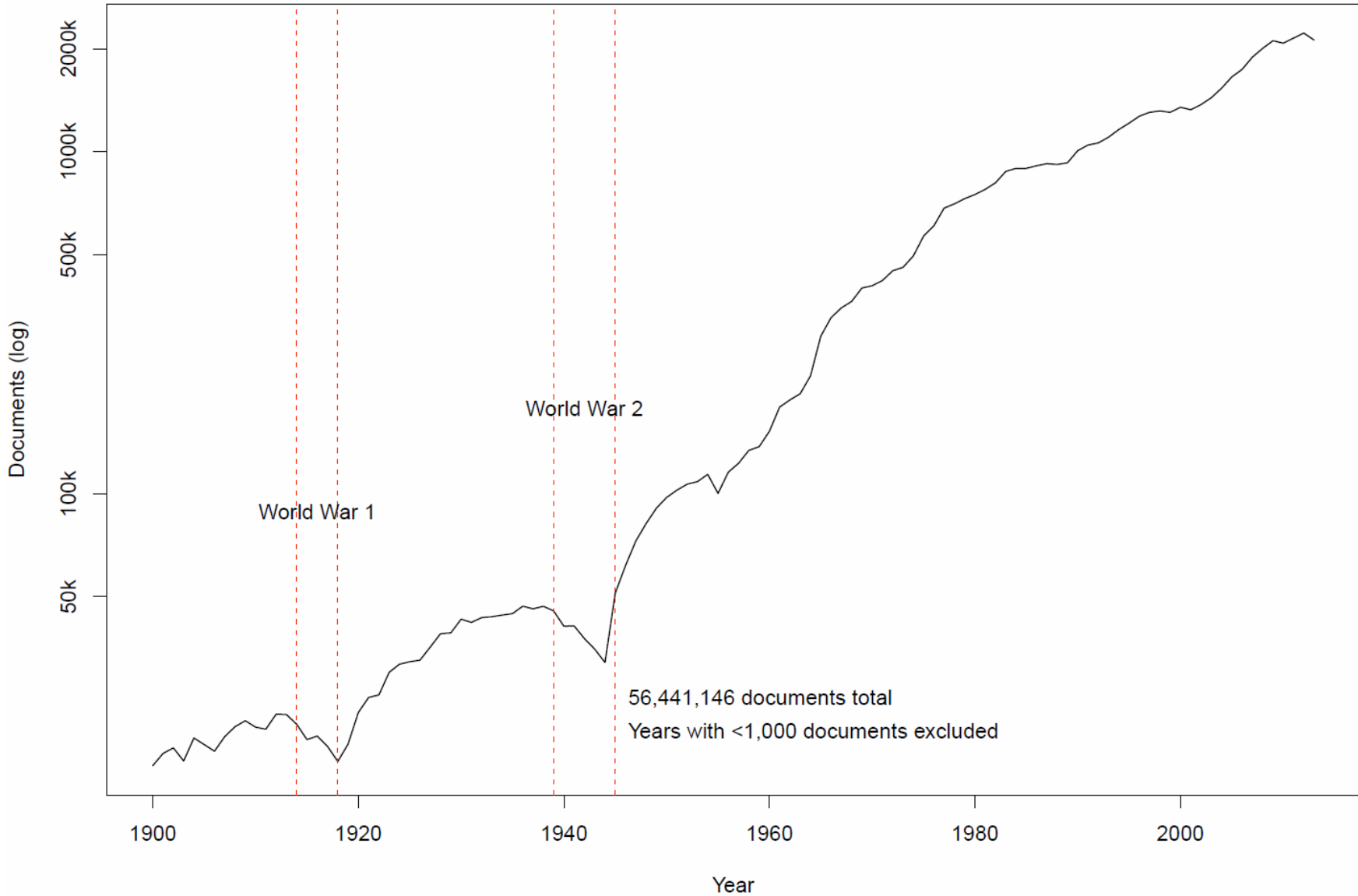
- Science Citation Index Expanded from 1900-2013
- Social Sciences Citation Index from 1900-2013
- Arts & Humanities Citation Index from 1975-2013
- Book Citation Index -- Science from 2005-2013
- Book Citation Index -- Social Sciences & Humanities from 2005-2013
- Conference Proceedings Citation Index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

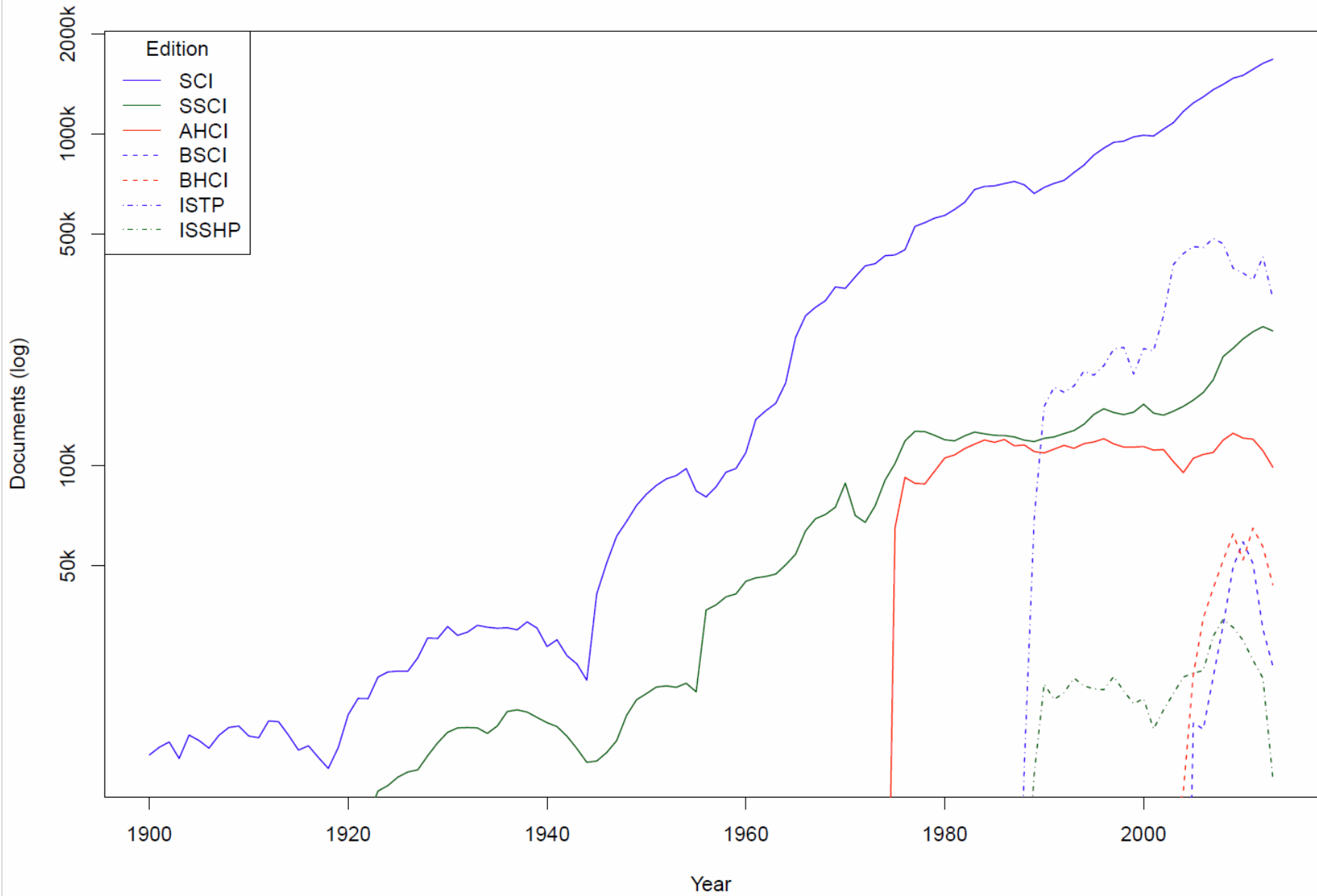


Web of Science Annual Total Documents





Web of Science Annual Total Documents By Edition





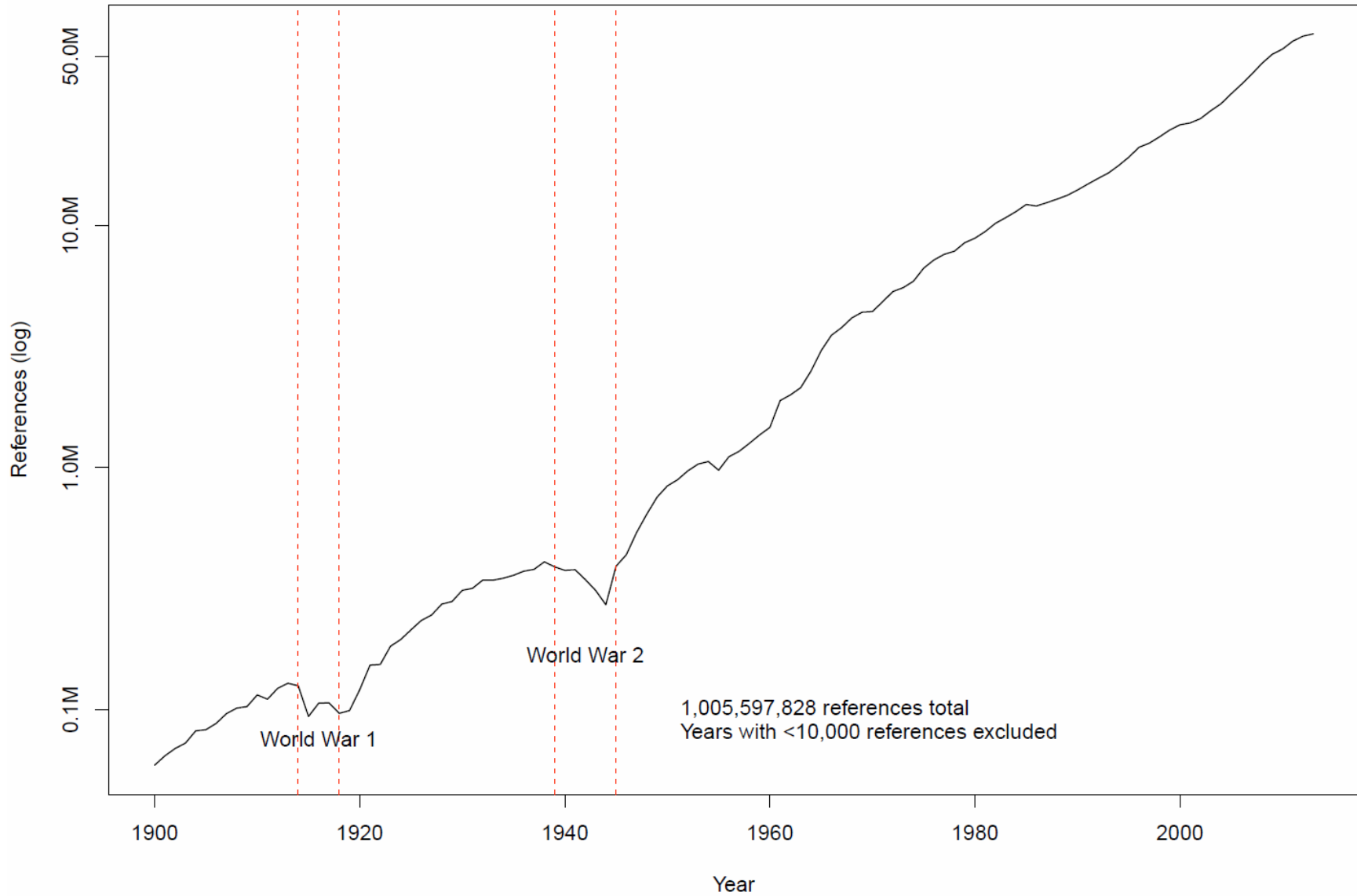
CNS

Cyberinfrastructure for
Network Science Center

IUNI Web of Science Data

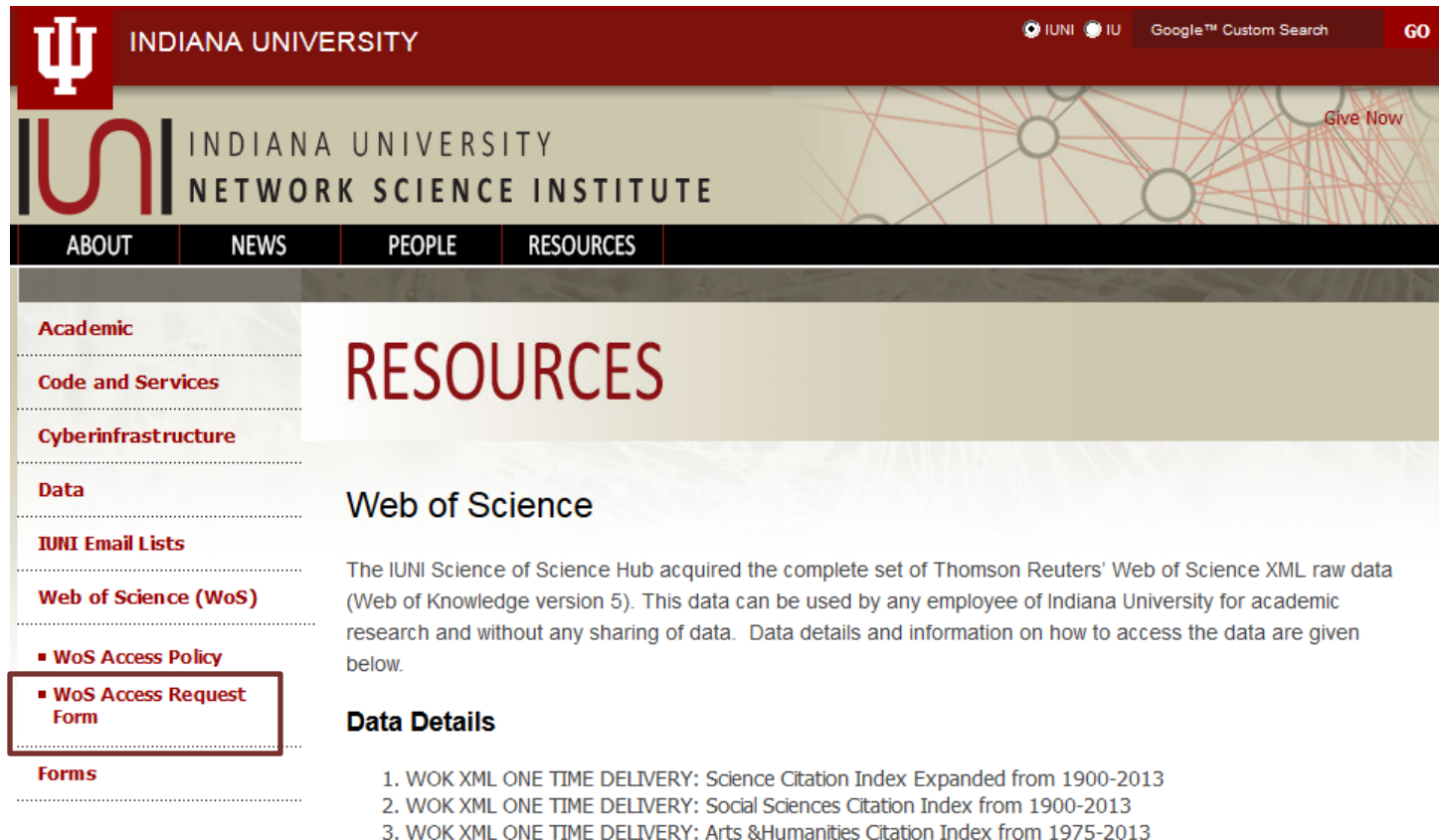


Web of Science Annual Total References



How to Request Access to Enclave

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



The screenshot shows the IUNI website interface. At the top, there is a dark red header with the Indiana University Psi logo and the text 'INDIANA UNIVERSITY'. To the right of the header are links for 'IUNI', 'IU', 'Google™ Custom Search', and a 'GO' button. Below the header is a light-colored banner with the IUNI logo and the text 'INDIANA UNIVERSITY NETWORK SCIENCE INSTITUTE'. A 'Give Now' button is visible on the right side of the banner. Below the banner is a dark navigation bar with links for 'ABOUT', 'NEWS', 'PEOPLE', and 'RESOURCES'. The 'RESOURCES' link is selected, and a sidebar on the left lists various resource categories: 'Academic', 'Code and Services', 'Cyberinfrastructure', 'Data', 'IUNI Email Lists', 'Web of Science (WoS)', and 'Forms'. Under the 'Web of Science (WoS)' category, there are two sub-links: 'WoS Access Policy' and 'WoS Access Request Form'. The 'WoS Access Request Form' link is highlighted with a red rectangular box. The main content area displays the 'RESOURCES' title and the 'Web of Science' section, which includes a paragraph of text and a 'Data Details' section with a list of three items.

INDIANA UNIVERSITY IUNI IU Google™ Custom Search GO

INDIANA UNIVERSITY NETWORK SCIENCE INSTITUTE Give Now

ABOUT NEWS PEOPLE RESOURCES

Academic
Code and Services
Cyberinfrastructure
Data
IUNI Email Lists
Web of Science (WoS)
▪ WoS Access Policy
▪ **WoS Access Request Form**
Forms

RESOURCES

Web of Science

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

Data Details

1. WOK XML ONE TIME DELIVERY: Science Citation Index Expanded from 1900-2013
2. WOK XML ONE TIME DELIVERY: Social Sciences Citation Index from 1900-2013
3. WOK XML ONE TIME DELIVERY: Arts & Humanities Citation Index from 1975-2013

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:

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:

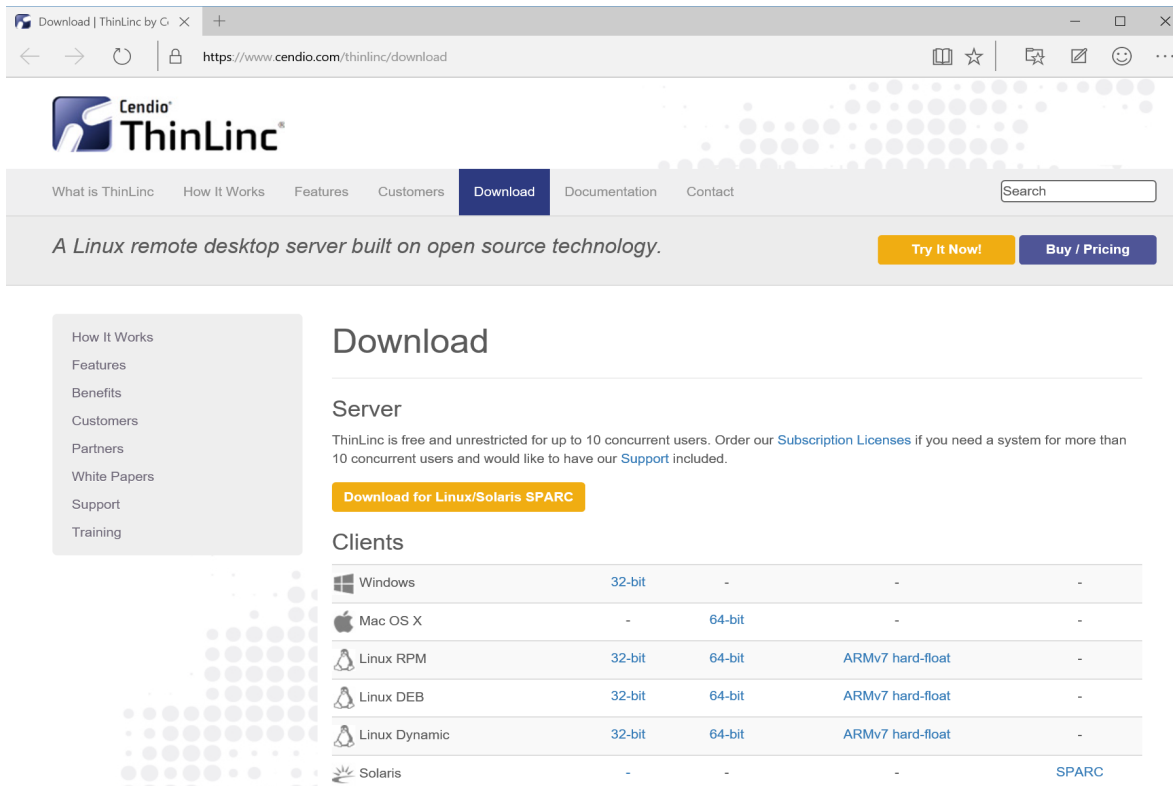
- Data Sharing and Storage:**
Data has to be stored on secure media. The data should not be shared with anyone. Usernames and passwords should not be shared with anyone. You agree to safeguard your username/password. Any unauthorized use of that account will be your individual responsibility. Any loss 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](https://www.cendio.com/thinlinc/download).

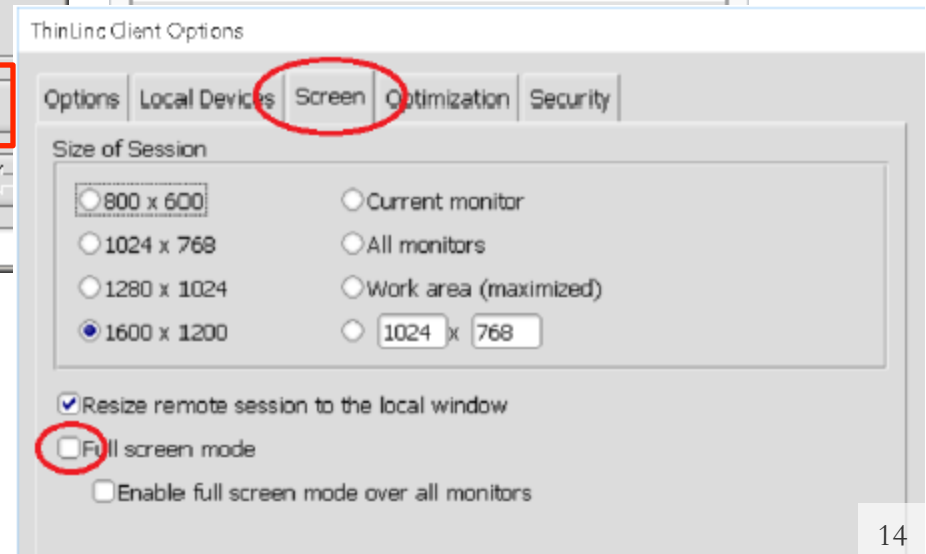
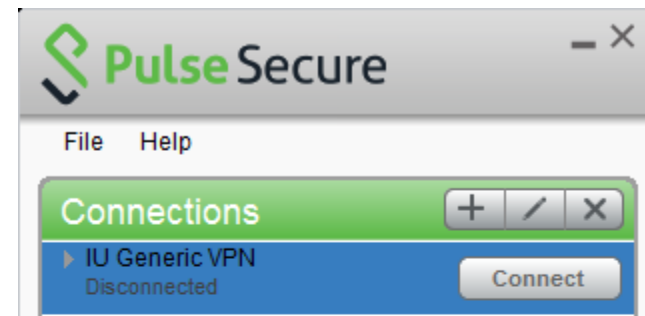
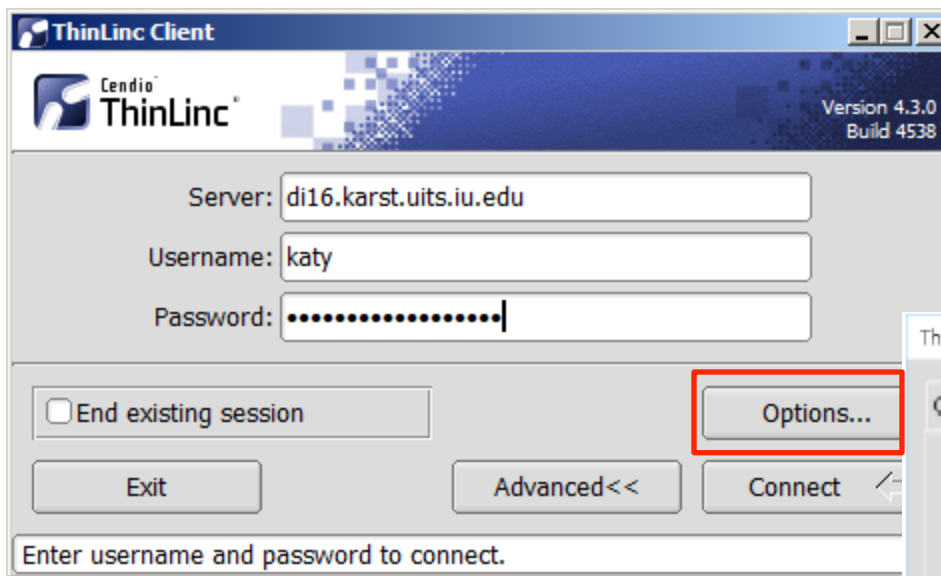


The screenshot shows the Cendio ThinLinc website's download page. The browser address bar shows the URL <https://www.cendio.com/thinlinc/download>. The page features a navigation menu with 'Download' selected, a search bar, and a tagline: 'A Linux remote desktop server built on open source technology.' Below this, there are 'Try It Now!' and 'Buy / Pricing' buttons. A sidebar on the left lists navigation options like 'How It Works', 'Features', 'Benefits', etc. The main content area is titled 'Download' and includes a 'Server' section with a 'Download for Linux/Solaris SPARC' button. A 'Clients' section contains a table with the following data:

Client	32-bit	64-bit	ARMv7 hard-float	SPARC
Windows	32-bit	-	-	-
Mac OS X	-	64-bit	-	-
Linux RPM	32-bit	64-bit	ARMv7 hard-float	-
Linux DEB	32-bit	64-bit	ARMv7 hard-float	-
Linux Dynamic	32-bit	64-bit	ARMv7 hard-float	-
Solaris	-	-	-	SPARC

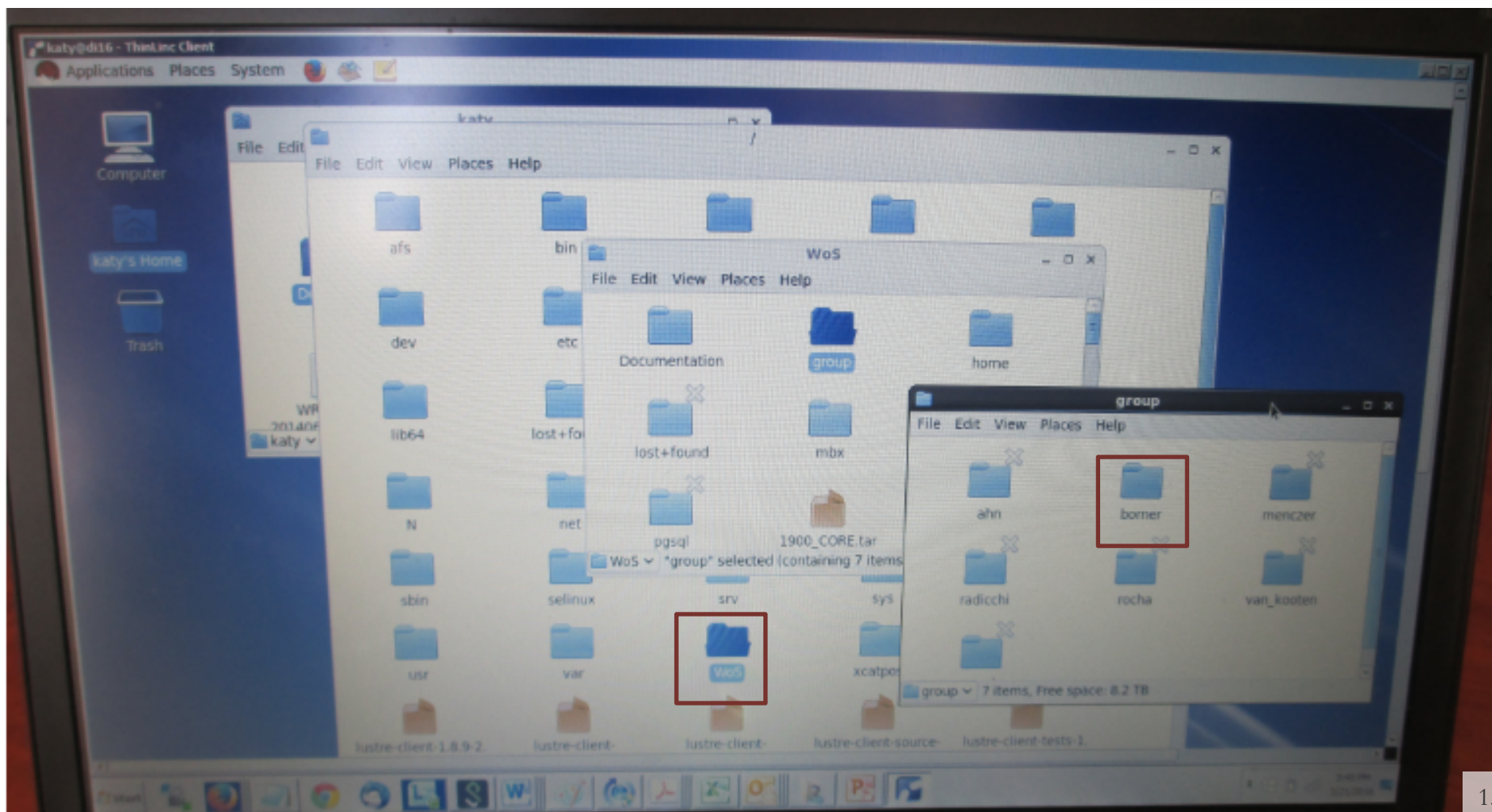
How to Login to the Enclave

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



Enclave Virtual Desktop

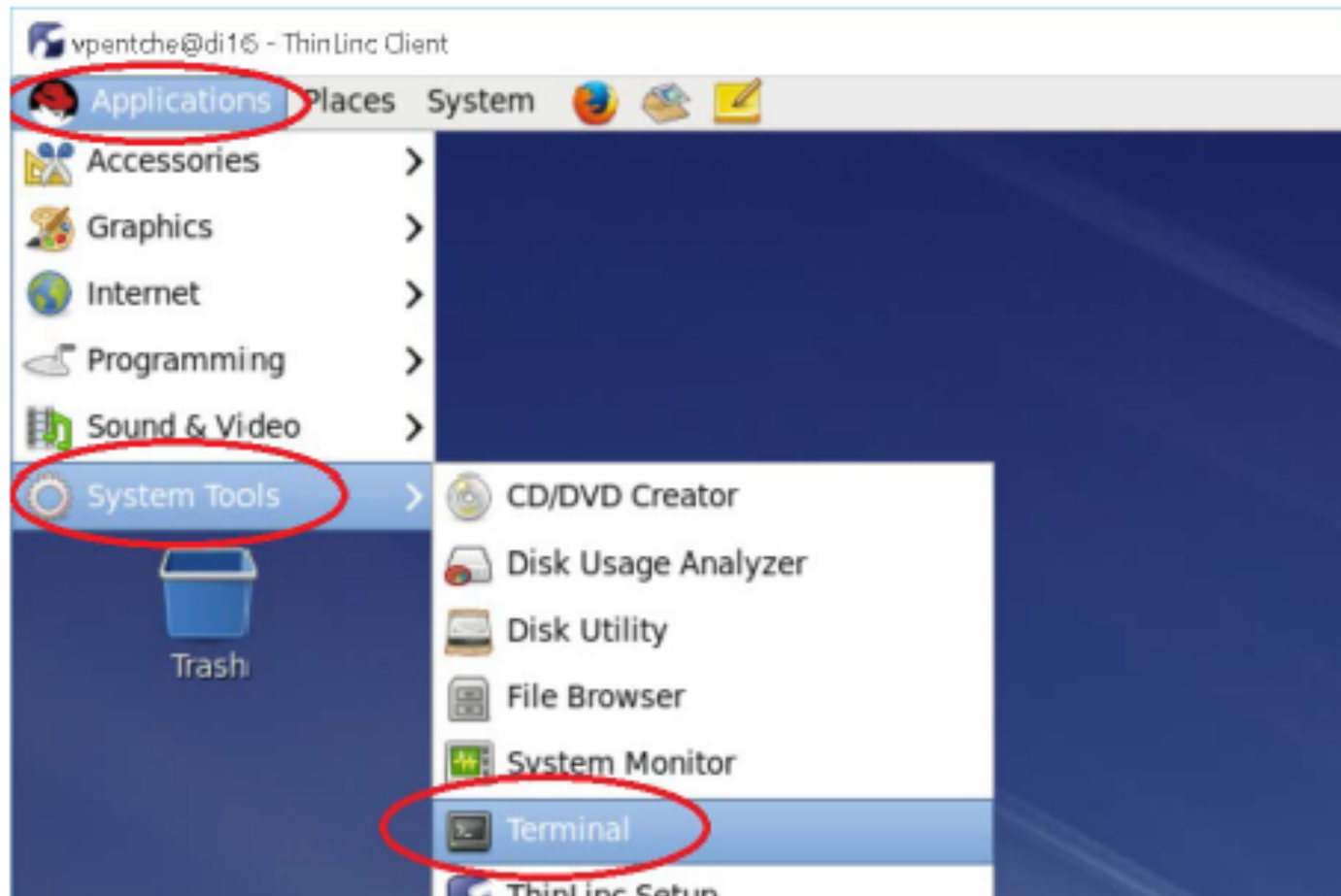
GNOM graphical interface of Red Hat Enterprise Server 6.6.





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.

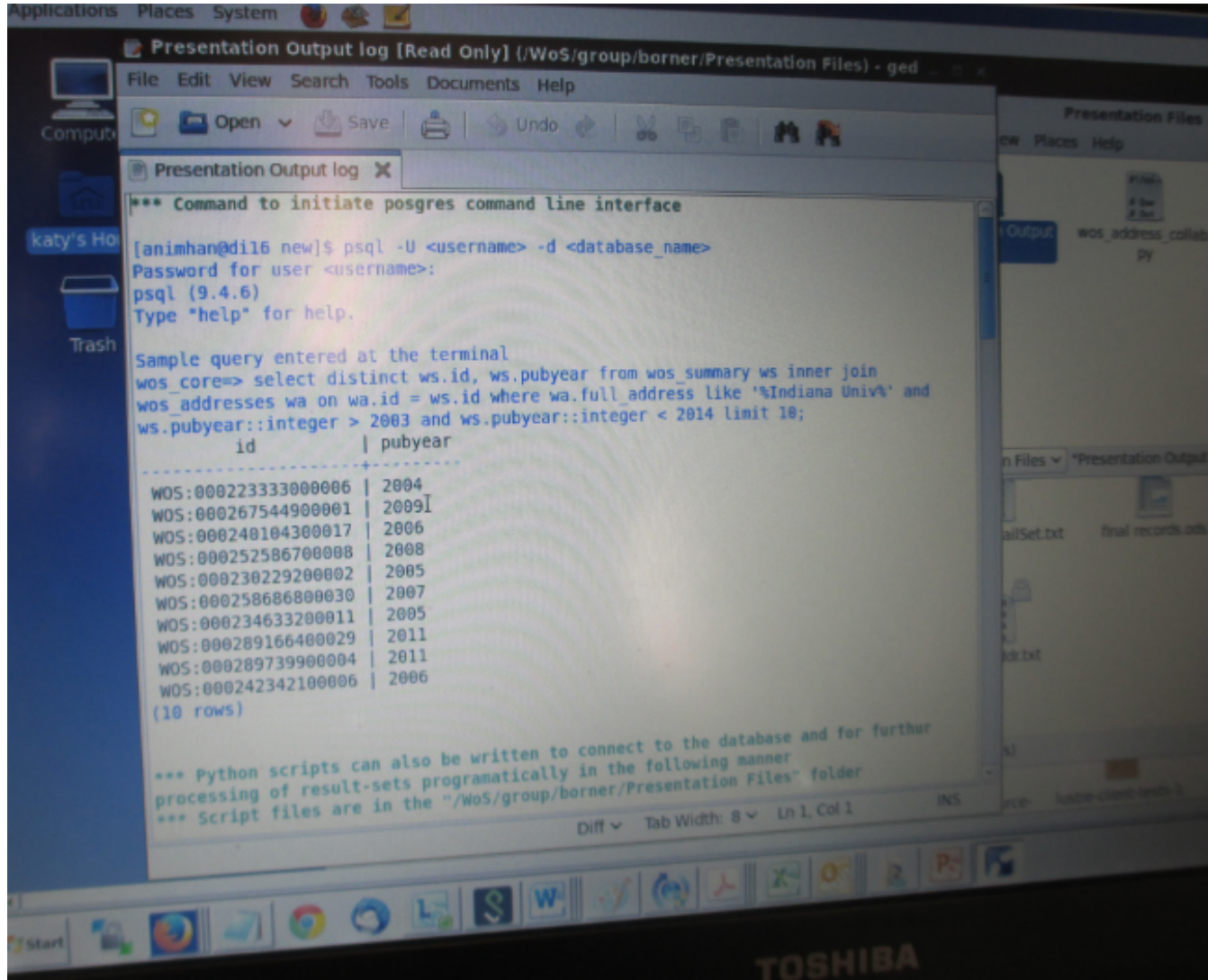


Manipulating the Web of Science Data

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



*** Command to initiate postgres command line interface

```
[animhan@dil6 new]$ psql -U <username> -d <database_name>
Password for user <username>:
psql (9.4.6)
Type "help" for help.
```

Sample query entered at the terminal

```
wos_core=> select distinct ws.id, ws.pubyear from wos_summary ws inner join
wos_addresses wa on wa.id = ws.id where wa.full_address like '%Indiana Univ%' and
ws.pubyear::integer > 2003 and ws.pubyear::integer < 2014 limit 10;
```

id	pubyear
WOS:000223333000006	2004
WOS:000267544900001	2009
WOS:000240104300017	2006
WOS:000252586700008	2008
WOS:000230229200002	2005
WOS:000258686800030	2007
WOS:000234633200011	2005
WOS:000289166400029	2011
WOS:000289739900004	2011
WOS:000242342100006	2006

(10 rows)

*** Python scripts can also be written to connect to the database and for further processing of result-sets programmatically in the following manner
*** Script files are in the "/WoS/group/borner/Presentation Files" folder

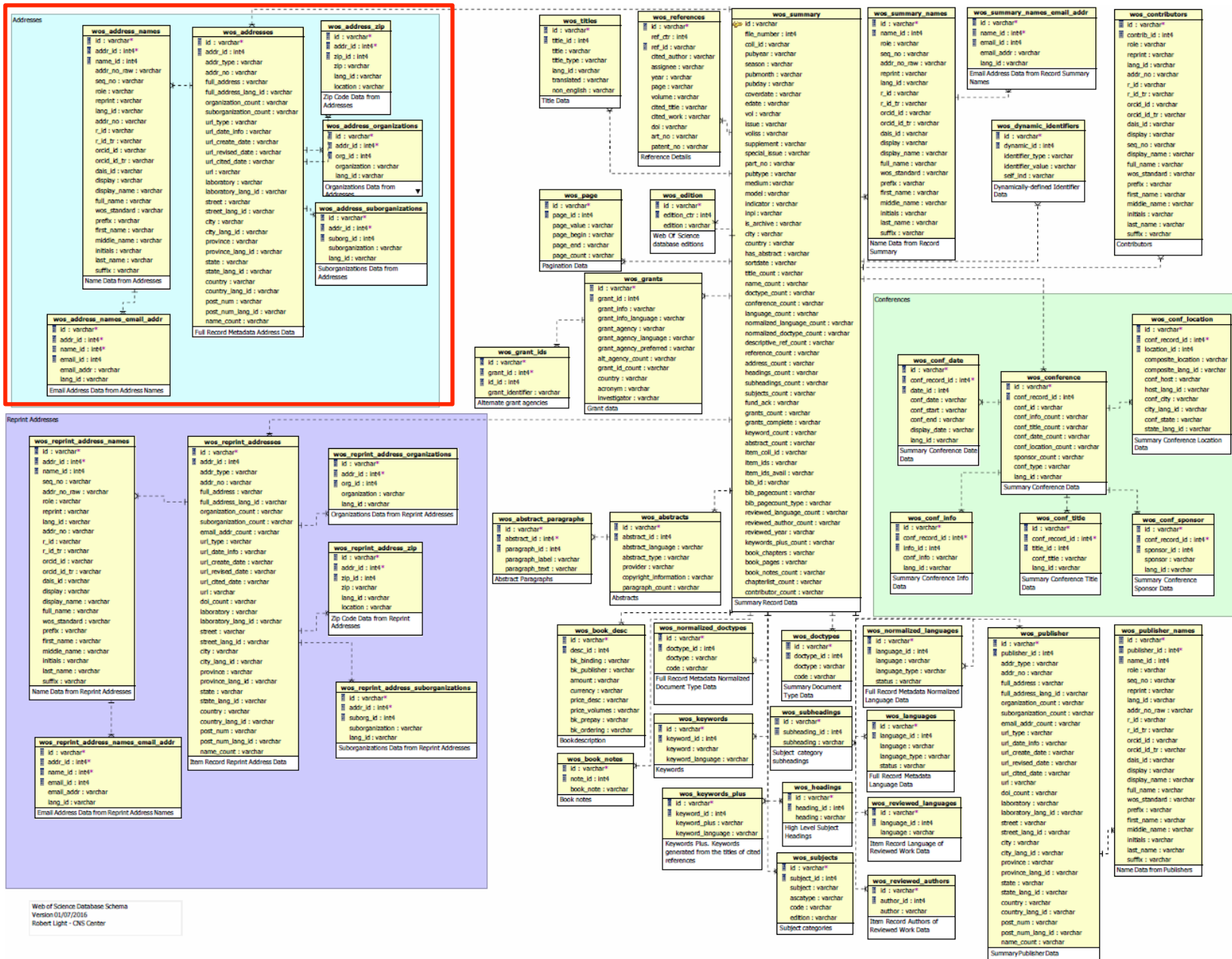


```

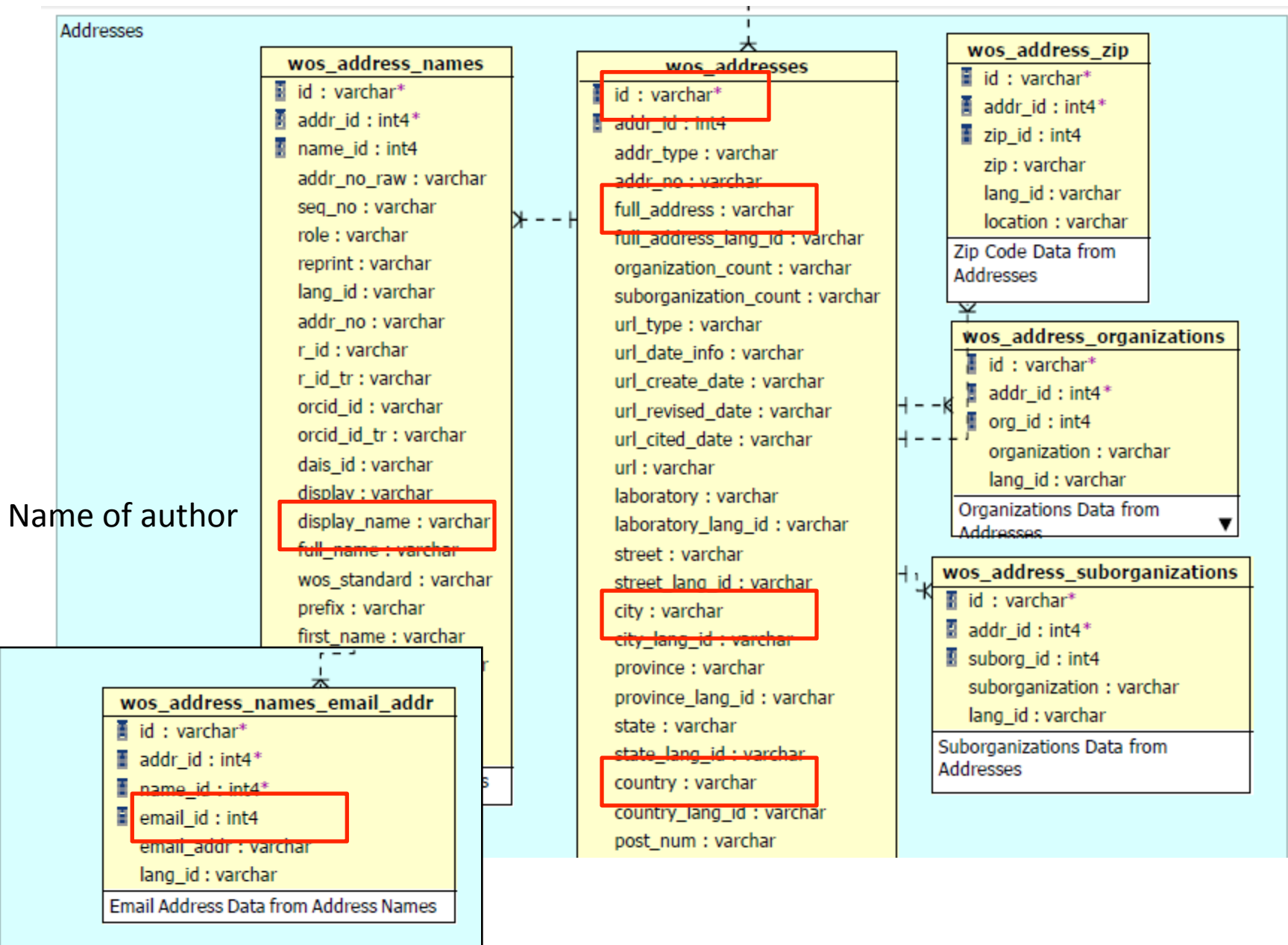
Presentation Output log [Read Only] (/WoS/group/borner/Presentation Files) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Presentation Output log X
*** Python scripts can also be written to connect to the database and for further processing of result-sets programatically in the following manner
*** Script files are in the "/WoS/group/borner/Presentation Files" folder

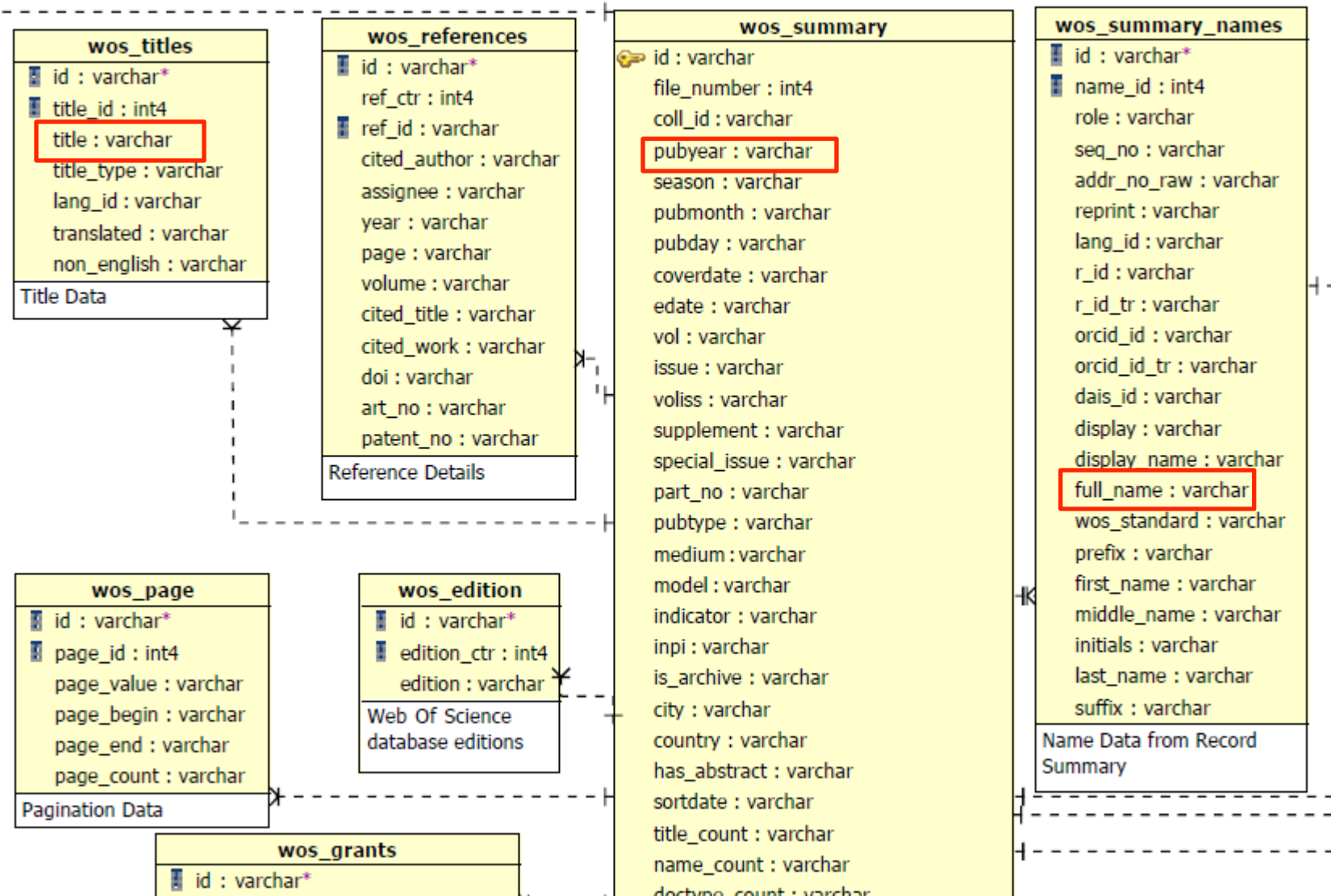
*** Execute was_address_iu.py as given below for results of international instiutions address
[animhan@d116 new]$ python was_address_iu.py
1|2012|WOS:000306366400009|Boerner, Katy|Royal Netherlands Acad Arts & Sci KNAM, Amsterdam, Netherlands|Netherlands|Amsterdam

*** Execute was_address_collab.py as given below for results of international collaborators
[animhan@d116 new]$ python was_address_collab.py
1|WOS:000285626000002|Rafols, Ismael|Univ Sussex, Sci Policy Res Unit, Brighton, E Sussex, England|England|Brighton
2|WOS:000304017900001|Scharnhorst, Andrea|Royal Netherlands Acad Arts & Sci, Data Archiving Serv, NL-2593 HT The Hague, Netherlands|Netherlands|The Hague
3|WOS:000304017900001|Scharnhorst, Andrea|Royal Netherlands Acad Arts & Sci, Networked Serv, NL-2593 HT The Hague, Netherlands|Netherlands|The Hague
4|WOS:000304017900001|Scharnhorst, Andrea|Royal Netherlands Acad Arts & Sci, E Humanities Grp, NL-2593 HT The Hague, Netherlands|Netherlands|The Hague
5|WOS:000304017900001|van den Besselaar, Peter|Vrije Univ Amsterdam, Network Inst, Amsterdam, Netherlands|Netherlands|Amsterdam
6|WOS:000304017900001|van den Besselaar, Peter|Vrije Univ Amsterdam, Dept Org Sci, Amsterdam, Netherlands|Netherlands|Amsterdam
7|WOS:000318807000015|Chen, Yunwei|Chinese Acad Sci, Chengdu Lib, Chengdu 610041, Peoples R China|Peoples R China|Chengdu
8|WOS:000318807000015|Chen, Yunwei|Univ Chinese Acad Sci, Beijing 100049, Peoples R China|Peoples R China|Beijing
9|WOS:000318807000015|Fang, Shu|Chinese Acad Sci, Chengdu Lib, Chengdu 610041, Peoples R China|Peoples R China|Chengdu
10|WOS:000267144200001|Scharnhorst, Andrea|Royal Netherlands Acad Arts & Sci, Virtual Knowledge Studio Humanities & Social Sci, NL-1019 AT Amsterdam, Netherlands|Netherlands|Amsterdam
11|WOS:000294839200020|Glanzel, Wolfgang|Katholieke Univ Leuven, Louvain, Belgium|Belgium|Louvain
12|WOS:000294839200020|Scharnhorst, Andrea|Royal Netherlands Acad Arts & Sci, Data Arch & Networked Serv & E Humanities Grp, Eindhoven, Netherlands|Netherlands|Eindhoven
13|WOS:000294839200020|van den Besselaar, Peter|Vrije Univ Amsterdam, Amsterdam, Netherlands|Netherlands|Amsterdam
14|WOS:000306366400009|Lariviere, Vincent|Univ Montreal, Ecole Bibliothecon & Sci Informat, Montreal, PQ, Canada|Canada|Montreal
15|WOS:000306366400009|Lariviere, Vincent|Univ Quebec, CIRSI, OST, Montreal, PQ H3C 3P8, Canada|Canada|Montreal
16|WOS:000257564800019|Cockerill, Matthew|BioMed Cent, London W1T 4LS, England|England|London
17|WOS:000257564800019|Pacheco, Roberto|Ist Stela, BR-88034050 Florianopolis, SC, Brazil|Brazil|Florianopolis
18|WOS:000257564800019|Mons, Barend|Open Progress Educ, NL-1315 BH AlmereAlmere, Netherlands|Netherlands|Almere
  
```



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

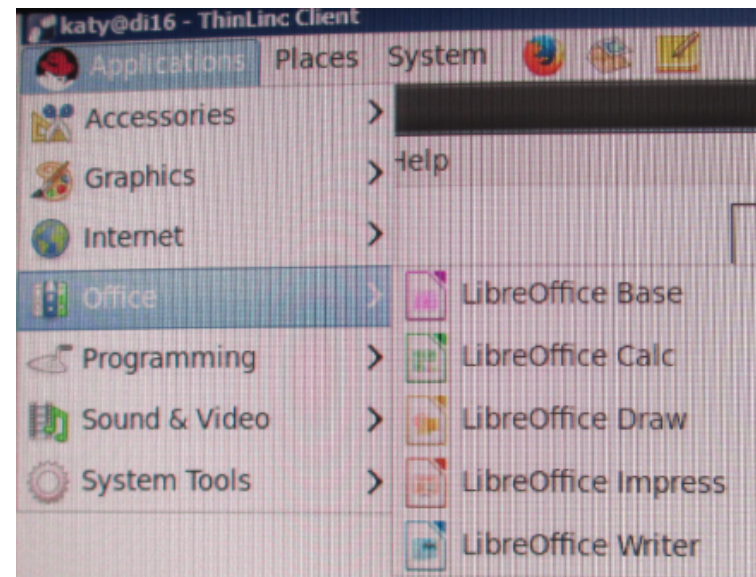




How to use Excel and Word-like Programs

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.



CNS

Cyberinfrastructure for
Network Science Center

IUNI Web of Science Data



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