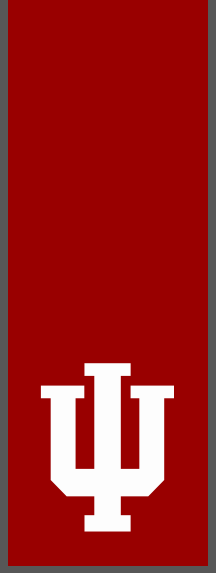


Indiana University
Network Science Institute

Web of Science as a Research Data Set

Valentin Pentchev, Director of IT
Matthew Hutchinson, Data Manager

Indiana University Network Science Institute

- IUNI is cross-campus entity that fosters interdisciplinary, collaborative research in network science.
- Over 150 IU faculty are affiliated with IUNI representing library science, sociology, informatics, neuroscience, public health, and more.
- IUNI Affiliates work with IT and research staff and have access to IUNI IT infrastructure and data.

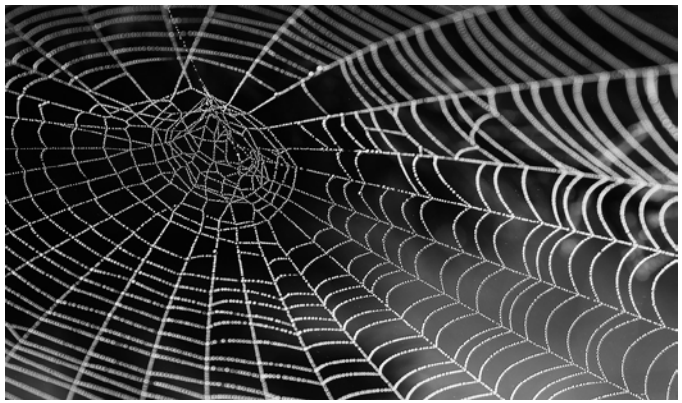
Indiana University
Network Science Institute

Web of Science Enclave

A Secure Research Environment

The Data Set

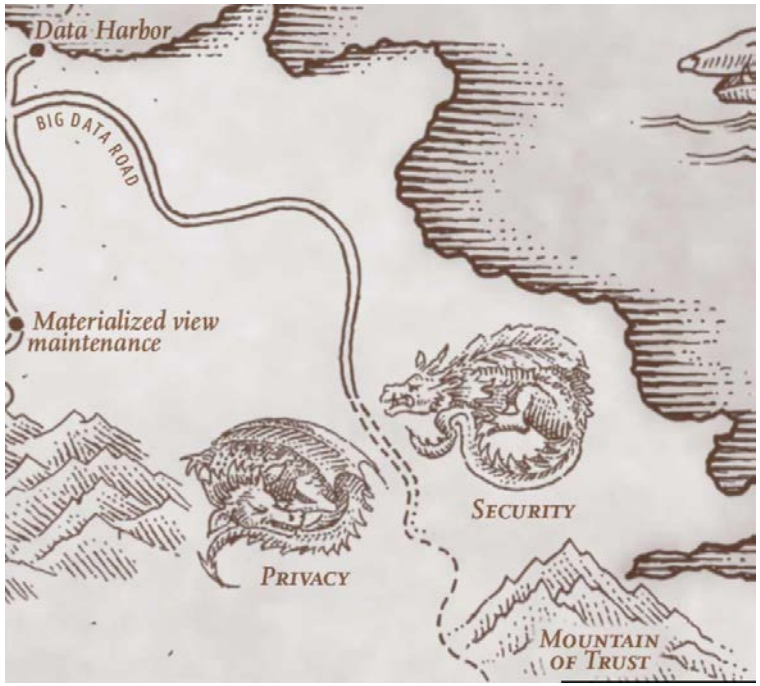
- Over 56,000,000 unique records
- More than 1,000,000,000 references
- Includes records from 1898 -2013 with 2014 and 2015 coming soon.



- 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

Balancing Access and Security

- How to allow researchers access to the data without risking the data being released **onto the web?**
- The University classified the data as ‘Restricted’ adding additional security requirements
- Initial plan had been to store the data on a single machine in a locked office



Reproduced with permission from *Designing Data-Intensive Applications* by Martin Kleppman

The WoS Enclave

1. **WoS Enclave is managed by a dedicated Data Manager, with support from a team of data specialists**

2. **A hardened software environment that does not allow the export of any data by individual users. All exports of data must be performed by the Data Manager**

Hardware:

- A dedicated node on KARST - IU's newest high-throughput computing cluster

- **CPU:** 2x Intel® Xeon ® E5-2650 v4 24 Cores 30MB L3 Cache
  Memory: 512GB
  HDD: 30TB + 1.2 TD SSD

Software:

- **OS:** Redhat Enterprise Linux 6.8

- **Software:** All tools that are available on the University's Karst cluster are available on the enclave and includes standard statistical and development tools including R, SAS, SPSS, MatLab, Python, Stata, LibreOffice and many more

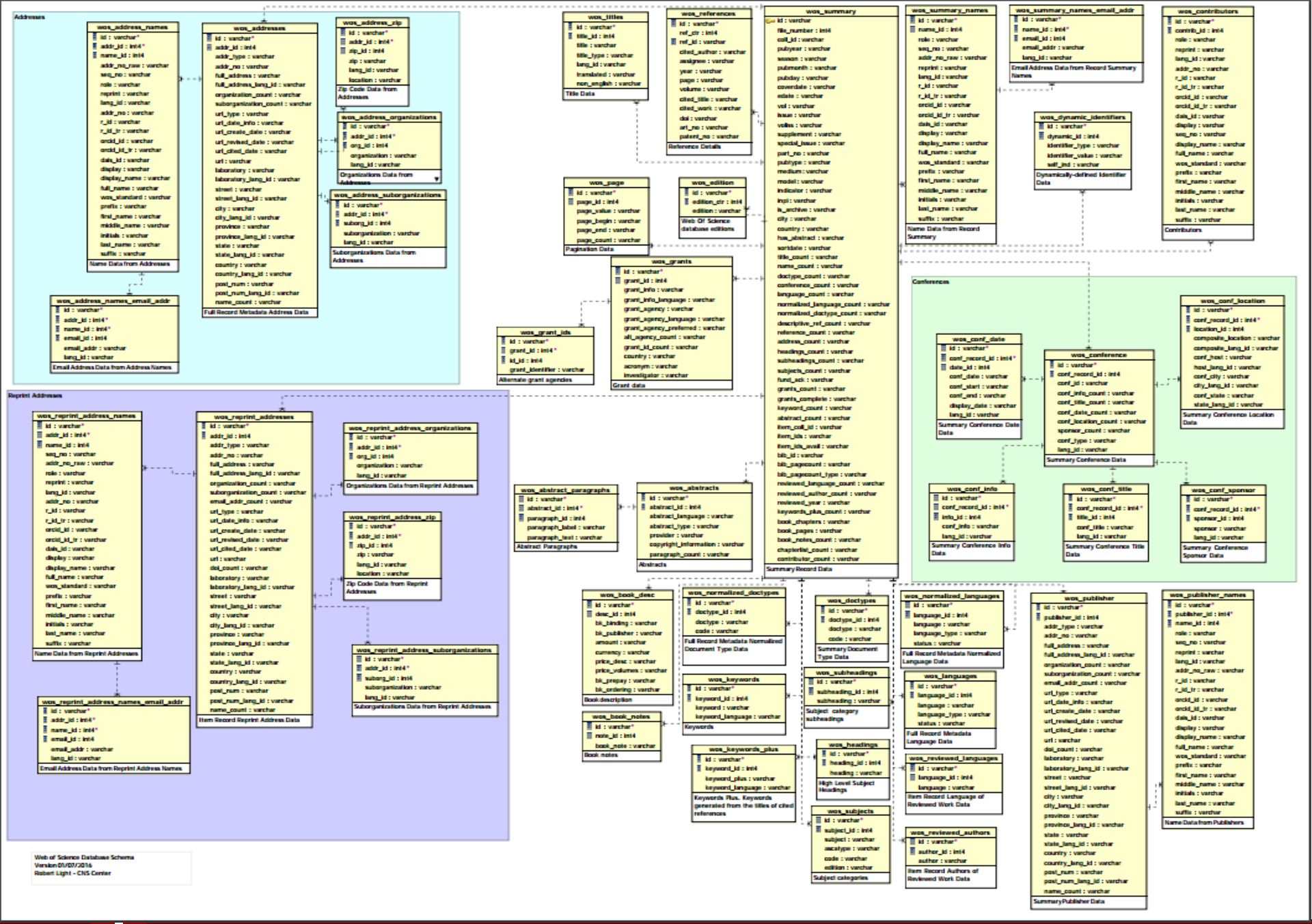
THE DATA - Raw XML

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Created from PDF via Acrobat SaveAsXML -->
<!-- Mapping Table version: 28-February-2003 -->
- <TaggedPDF-doc>
      <?xpacket begin='' id='W5M0MpCehiHzreSzNTczkc9
      <?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc
  - <x:xmpmeta x:xmptk="Adobe XMP Core 5.2-c001 63
    - <rdf:RDF xmlns:rdf="http://www.w3.org/1999/
      - <rdf:Description xmlns:xmp="http://ns.adobe
          <xmp:CreateDate>2011-04-06T17:22:05
          <xmp:CreatorTool>ESRI ArcSOC 9.2.0.13
          <xmp:ModifyDate>2011-04-07T08:17:15
          <xmp:MetadataDate>2011-04-07T08:17:
        </rdf:Description>
      - <rdf:Description rdf:about="" xmlns:xmpMM="h
          <xmpMM:DocumentID>uuid:323e04f1-10
          <xmpMM:InstanceID>uuid:6329c275-adc
        </rdf:Description>
      - <rdf:Description rdf:about="" xmlns:dc="http:/
          <dc:format>xml</dc:format>
        </rdf:Description>
      </rdf:RDF>
    </x:xmpmeta>
      <?xpacket end="w"?>
      <?xpacket end='r'?>
  - <Figure>
      <ImageData src="images/WA_Dayton_2011040
    </Figure>
</TaggedPDF-doc>
```

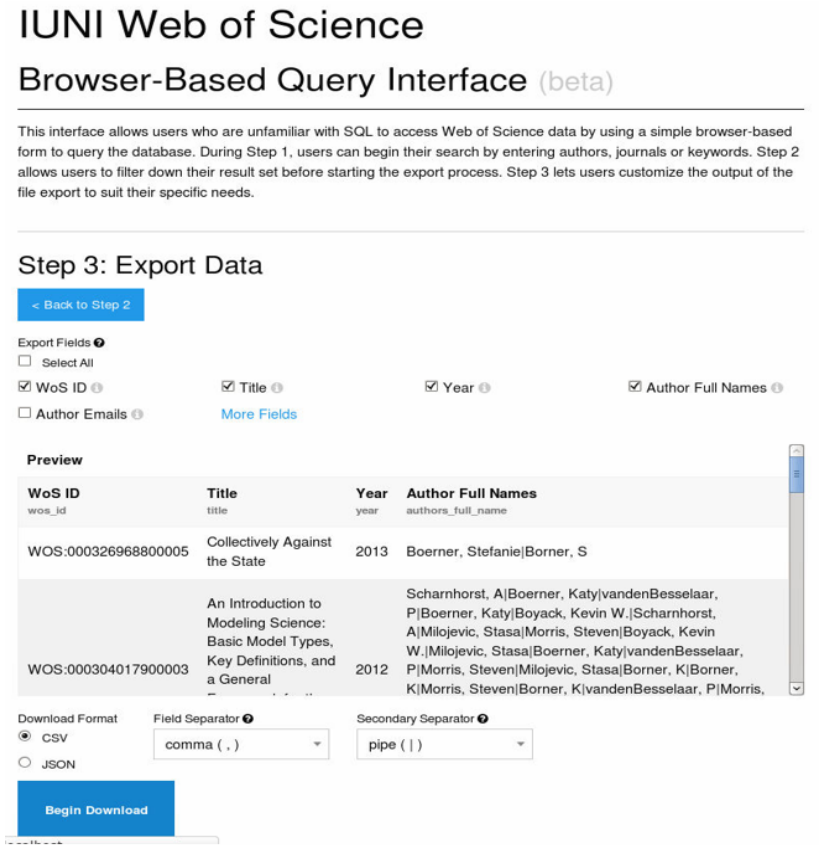
1. **Raw XML data from Thomson Reuters/Clarivate Analytics was received and placed in the enclave**
2. **Unprocessed data available for researchers**
3. **Parsing software developed by Robert Light and Daniel Halsey at IU's Cyberinfrastructure for Network Science (CNS)**
   - **Developed in Python and available for use under the Apache License 2.0 from: https://github.com/cns-iu/generic_parser**

THE DATA - Relational Databases

1. **PostgreSQL database with a fully parsed XML data**
   - **wos_core:** 46 tables, joined on ‘wos_id’
   - **wos_simple:** 13 tables with serial
     Public and Foreign key relations
2. **Database schema and Data Dictionary documentation**
   - **wos_core:** fully documented and available to users
   - **wos_simple:** under development, available for upon request

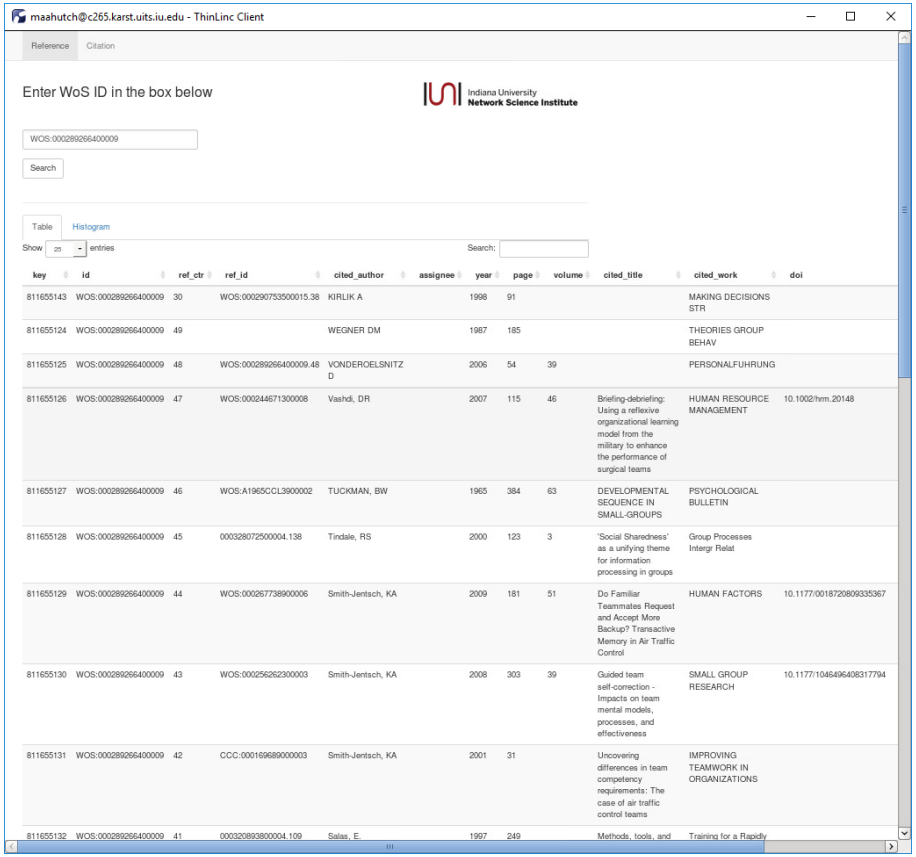

Addresses
wos_address_names
wos_addresses
wos_address_zip
wos_address_organizations
wos_address_suborganizations
wos_address_names_email_addr
Name Data from Addresses
Full Record Metadata Address Data
Zip Code Data from Addresses
Organizations Data from Addresses
Suborganizations Data from Addresses
Email Address Data from Address Names
wos_titles
wos_references
wos_summary
wos_summary_names
wos_summary_names_email_addr
wos_contributors
wos_dynamic_identifiers
Title Data
Reference Details
Summary Record Data
Name Data from Record Summary
Email Address Data from Record Summary Names
Contributors
Dynamically-defined Identifier Data
wos_page
wos_edition
wos_grants
wos_grant_ids
Pagination Data
Web Of Science database editions
Grant data
Alternate grant agencies
Conferences
wos_conf_date
wos_conference
wos_conf_location
wos_conf_info
wos_conf_title
wos_conf_sponsor
Summary Conference Date Data
Summary Conference Data
Summary Conference Location Data
Summary Conference Info Data
Summary Conference Title Data
Summary Conference Sponsor Data
Reprint Addresses
wos_reprint_address_names
wos_reprint_addresses
wos_reprint_address_organizations
wos_reprint_address_zip
wos_reprint_address_suborganizations
wos_reprint_address_names_email_addr
Name Data from Reprint Addresses
Item Record Reprint Address Data
Organizations Data from Reprint Addresses
Zip Code Data from Reprint Addresses
Suborganizations Data from Reprint Addresses
Email Address Data from Reprint Address Names
wos_abstract_paragraphs
wos_abstracts
Abstract Paragraphs
Abstracts
wos_book_desc
wos_normalized_doctypes
wos_doctypes
wos_normalized_languages
wos_publisher
wos_publisher_names
wos_keywords
wos_subheadings
wos_languages
wos_book_notes
wos_keywords_plus
wos_headings
wos_reviewed_languages
wos_subjects
wos_reviewed_authors
Book description
Full Record Metadata Normalized Document Type Data
Summary Document Type Data
Full Record Metadata Normalized Language Data
Summary Publisher Data
Name Data from Publishers
Keywords
Subject category subheadings
Full Record Metadata Language Data
Book notes
Keywords Plus. Keywords generated from the titles of cited references
High Level Subject Headings
Item Record Language of Reviewed Work Data
Subject categories
Item Record Authors of Reviewed Work Data
Web of Science Database Schema
Version 01/07/2016
Robert Light - CNS Center

THE DATA – Browser-based Interface



1. **Created to allow users unfamiliar with SQL to access data**
2. **Queries the wos_simple database**
3. **Allows for text searching on Author Names, Journal Names, Article Names, Abstracts and Years**
4. **Python backend based on the Flask micro server package**

THE DATA – Citation Interface



1. **Created to allow users unfamiliar with SQL to explore the reference and citations networks for individual records**
2. **Also queries the wos_simple database**
3. **Allows searching both citations and references**
4. **Created using R Shiny**

Super Computing and the Web of Science

1. Scripts can be run against the raw WoS XML data on both Karst and Big Red II
2. Simply submit your script along with any necessary instructions to the IUNI Data Manager

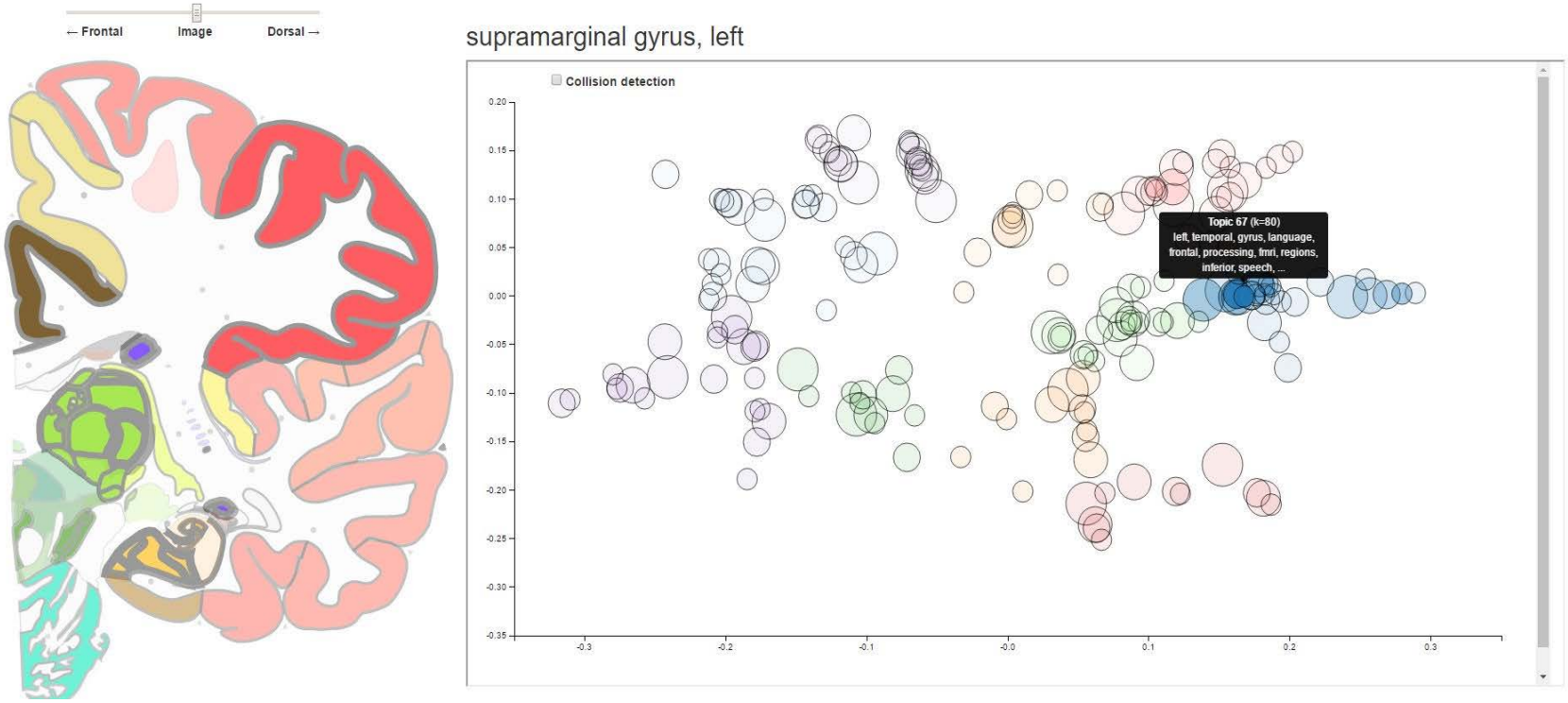
Research:

1. **Not just Bibliometrics and Scientometrics!**

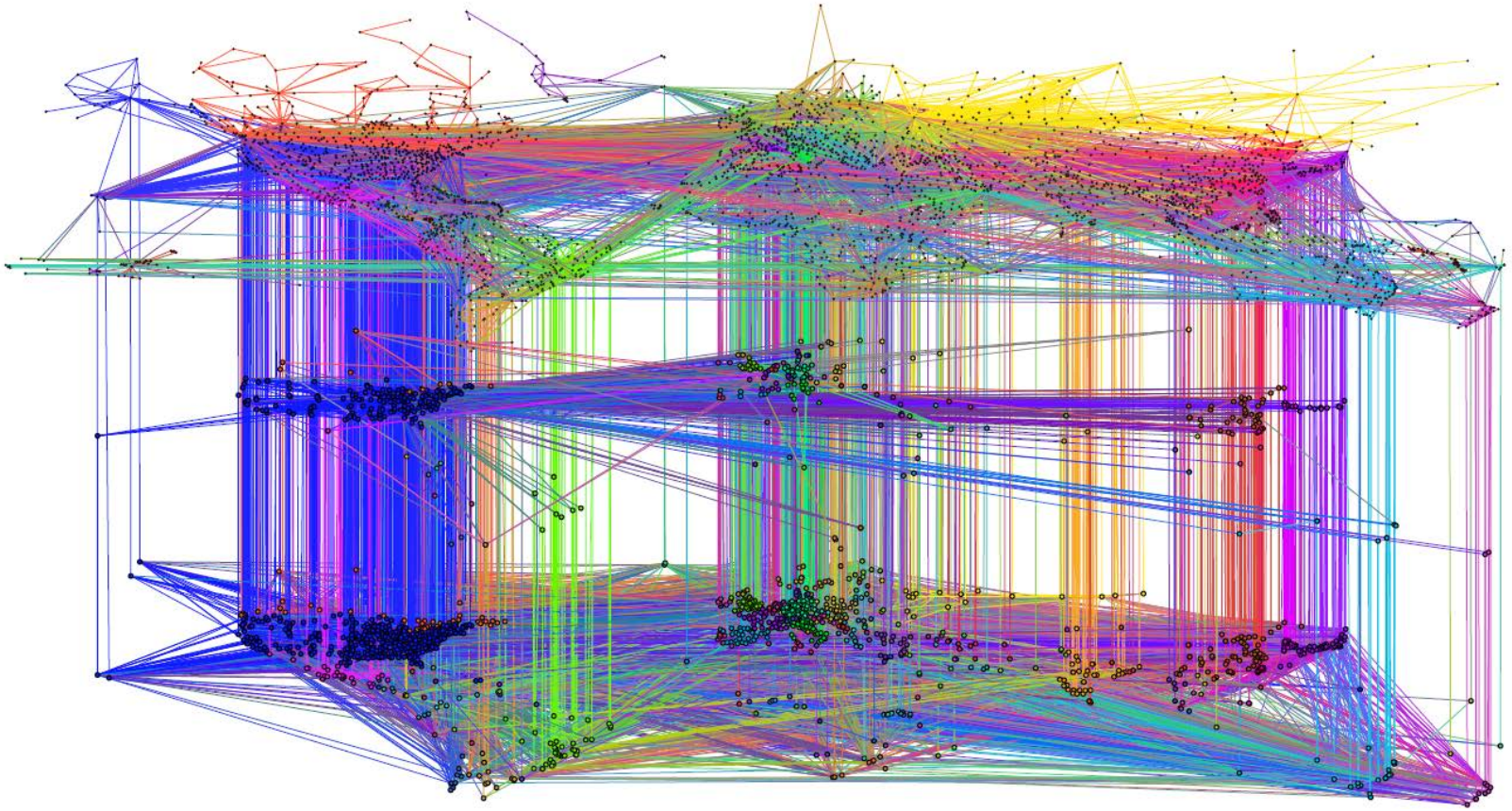
2. **Cognitive Science – Hyperbrain project**

3. **Network Science – Correlating Transportation Networks with Academic Collaborations**

4. **And many more…**



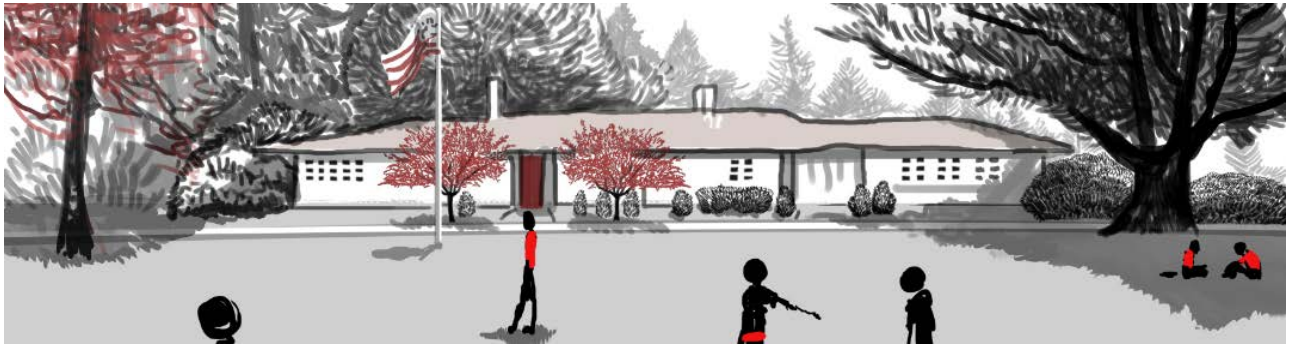
Hyperbrain.org, reproduced with permission by Jaimie Murdock & Franco Pestilli



“Mapping Six Years of Research by Three Universities” Reproduced with permission by Xiaoran Yan, Indiana University

Interested in Working with IUNI's copy of the WoS data?

1. Simply go to http://iuni.iu.edu/resources/web-of-science
2. Read the details of the data available and the WoS Access Policy
3. Complete the application for 'General Data User'
4. Request an account on Karst if you do not already have one
5. Wait to be contacted by IUNI staff (which should occur within 24 hrs)



The Team

Valentin Pentchev
Director of Information Technology
IUNI

Matt Hutchinson
Data Manager
IUNI

Ben Serrette
Web Developer
IUNI

Patricia Mabry
Executive Director
IUNI

Katy Börner
Faculty, Center Director
CNS

Robert Light
Senior Systems Analyst, Database Administrator
CNS

Daniel Halsey
Senior System Architect, Project Manager
CNS

Joe Rinkovsky
Lead Systems Administrator
High Performance Systems, UITS

Acknowledgements: Jennifer Adams, IU General Counsel, Sara Chambers and the IU Committee of Data Stewards and Timothy Otto, Clarivate Analytics

Indiana University
Network Science Institute

Thank You

Valentin Pentchev
Director of IT

1001 E SR 45/46 Bypass **|** Bloomington,
IN 47408-1415
Email: vpentche@iu.edu **|**
Phone: (812) 856-7087 **|** Fax: (812) 856-1192

Matthew Hutchinson
Data Manager

1001 E SR 45/46 Bypass **|** Bloomington,
IN 47408-1415
Email: maahutch@iu.edu **|**
Phone: (812) 855-1404 | Fax: (812) 856-1192

INDIANA UNIVERSITY BLOOMINGTON
FULFILLING *the* PROMISE