Preparing, Analyzing, and Visualizing Humanities Data

Dr. Katy Börner and Scott Weingart
Cyberinfrastructure for Network Science Center
Information Visualization Laboratory
School of Library and Information Science
Indiana University, Bloomington, IN
http://cns.slis.indiana.edu

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Angela M. Zoss, and Sean Lind

Networks and Network Analysis for the Humanities Workshop
Rice Room, Institute for Pure and Applied Mathematics
UCLA, Los Angeles, CA
9:00-12:00 on August 17, 2010

Workshop Overview

9:00 Marcoscope Design and Usage & CICShell Powered Tools
9:15 Sci2 Tool Basics
- Download and run the tool.
10:00 Sci2 Tool – Using Text and Database Support
- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.
10:15-10:30 Break
10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance
**Workshop Overview**

9:00 Marcoscope Design and Usage & CShell Powered Tools
- 9:15 Sci2 Tool Basics
  - Download and run the tool.
10:00 Sci2 Tool – Using Text and Database Support
  - Studying Four Major NetSci Researchers.
  - Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
  - Find basic statistics and run various algorithms over the network.
  - Visualize the networks as either a graph or a circular hierarchy.
10:15-10:30 Break
10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance

---

**The Changing Scientific Landscape**

*Star Scientist -> Research Teams:* In former times, science was driven by key scientists. Today, science is driven by effectively collaborating co-author teams often comprising expertise from multiple disciplines and several geospatial locations (Börner, Dall’Asta, Ké, & Vespignani, 2005; Shneiderman, 2008).

*Users -> Contributors:* Web 2.0 technologies empower anybody to contribute to Wikipedia or to exchange images and videos via Flickr and YouTube. WikiSpecies, WikiProfessionals, or WikiProteins combine wiki and semantic technology in support of real time community annotation of scientific datasets (Mons et al., 2008).

*Cross-disciplinary:* The best tools frequently borrow and synergistically combine methods and techniques from different disciplines of science and empower interdisciplinary and/or international teams of researchers, practitioners, or educators to fine-tune and interpret results collectively.

*One Specimen -> Data Streams:* Microscopes and telescopes were originally used to study one specimen at a time. Today, many researchers must make sense of massive streams of multiple types of data with different formats, dynamics, and origin.

*Static Instrument -> Evolving Cyberinfrastructure (CI):* The importance of hardware instruments that are rather static and expensive decreases relative to software infrastructures that are highly flexible and continuously evolving according to the needs of different sciences. Some of the most successful services and tools are decentralized increasing scalability and fault tolerance.
Just as the microscope empowered our naked eyes to see cells, microbes, and viruses thereby advancing the progress of biology and medicine or the telescope opened our minds to the immensity of the cosmos and has prepared mankind for the conquest of space, macrosopes promise to help us cope with another infinite: the infinitely complex. Macrosopes give us a ‘vision of the whole’ and help us ‘synthesize’. They let us detect patterns, trends, outliers, and access details in the landscape of science. Instead of making things larger or smaller, macrosopes let us observe what is at once too great, too slow, or too complex for our eyes.

Desirable Features of Macrosopes

Core Architecture & Plugins/Division of Labor: Computer scientists need to design the standardized, modular, easy to maintain and extend “core architecture”. Dataset and algorithm plugins, i.e., the “filling”, are provided by those that care and know most about the data and developed the algorithms: the domain experts.

Ease of Use: As most plugin contributions and usage will come from non-computer scientists it must be possible to contribute, share, and use new plugins without writing one line of code. Users need guidance for constructing effective workflows from 100+ continuously changing plugins.

Modularity: The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

Standardization: Adoption of (industry) standards speeds up development as existing code can be leveraged. It helps pool resources, supports interoperability, but also eases the migration from research code to production code and hence the transfer of research results into industry applications and products.

Open Data and Open Code: Lets anybody check, improve, or repurpose code and eases the replication of scientific studies.

Macroscopes are similar to Flickr and YouTube and but instead of sharing images or videos, you freely share datasets and algorithms with scholars around the globe.

Macroscope Design

Custom Tools for Different Scientific Communities
Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu

Network Workbench Tool + Community Wiki
http://nwb.slis.indiana.edu

Science of Science (Sci2) Tool and Portal
http://sci.slis.indiana.edu
Epidemics Cyberinfrastructure
http://epic.slis.indiana.edu/

180+ Algorithm Plugins and Branded GUIs

Core Architecture
Open Services Gateway Initiative (OSGi) Framework.
http://orgi.org
Cyberinfrastructure Shell (CIShell)
http://cishell.org

CIShell Powered Tools:
Network Workbench (NWB)
The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In February 2009, the tool provides more 169 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 50 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 65,000 times since December 2006.


Project Details

Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert

Software Team: Lead: Micah Linnemeier
Members: Patrick Phillips, Russell Duhon, Tim Kelley & Ann McCranie
Previous Developers: Weixia (Bonnie) Huang, Bruce Herr, Heng Zhang, Duygu Balcan, Bryan Hook, Ben Markines, Santo Fortunato, Felix Terkhor, Ramya Sabineni, Vivek S. Thakre & Cesar Hidalgo

Goal: Develop a large-scale network analysis, modeling and visualization toolkit for physics, biomedical, and social science research.

Amount: $1,120,926, NSF IIS-0513650 award
Website: http://nwb.slis.indiana.edu
NWB Advisory Board:

James Hendler (Semantic Web)  http://www.cs.umd.edu/~hendler/
Jason Leigh (CI)  http://www.evl.uic.edu/spiff/
Neo Martinez (Biology)  http://online.sfsu.edu/~webhead/
Michael Macy, Cornell University (Sociology)  http://www.soc.cornell.edu/faculty/macy.shtml
Ulrik Brandes (Graph Theory)  http://www.inf.uni-konstanz.de/˜brandes/
Mark Gerstein, Yale University (Bioinformatics)  http://bioinfo.mbb.yale.edu/
Tom Snijders, University of Groningen  http://stat.gamma.rug.nl/snijders/
Noshir Contractor, Northwestern University  http://www.spcomm.uiuc.edu/nosh/

Computational Proteomics

What relationships exist between protein targets of all drugs and all disease-gene products in the human protein–protein interaction network?

Computational Economics
Does the type of product that a country exports matter for subsequent economic performance?


Computational Social Science
Studying large scale social networks such as Wikipedia

Computational Epidemics
Forecasting (and preventing the effects of) the next pandemic.


NWB Tool Download, Install, and Run

**NWB Tool 1.0.0**
Can be freely downloaded for all major operating systems from http://nwb.slis.indiana.edu
Select your operating system from the pull down menu and download.
Unpack into a /nwb directory.
Run /nwb/nwb.exe

Session log files are stored in ‘*yournwbdirectory*/logs’ directory.

**Cite as**
Console shows references to seminal works. Workflows are recorded into a log file, and soon can be re-run for easy replication. All algorithms are documented online; workflows are given in tutorials.


**Analysis Menu and Submenus**

**Integrated Tools**

**Gnuplot**

**GUESS**
exploratory data analysis and visualization tool for graphs and networks.
[https://nwb.slis.indiana.edu/community/?n=VisualizeData.GUESS](https://nwb.slis.indiana.edu/community/?n=VisualizeData.GUESS).
Supported Data Formats

The NWB tool supports loading the following input file formats:

- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net) & Pajek .Matrix (*.mat)
- NWB (*.nwb)
- TreeML (*.xml)
- Edge list (*.edge)
- CSV (*.csv)
- ISI (*.isi)
- Scopus (*.scopus)
- NSF (*.nsf)
- Bibtex (*.bib)
- Endnote (*.enw)

and the following network file output formats:

- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)

Formats are documented at https://nwb.slis.indiana.edu/community/?n=DataFormats.HomePage.

---

Workshop Overview

9:00 Marcoscope Design and Usage & CIShell Powered Tools

9:15 Sci2 Tool Basics

- Download and run the tool.


10:00 Sci2 Tool – Using Text and Database Support

- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.

11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.

11:30-12:00 Q&A and Technical Assistance
CIShell Powered Tools:
Science of Science (Sci2) Tool

- Explicitly designed for SoS research and practice, well documented, easy to use.
- Empowers many to run common studies while making it easy for exports to perform novel research.
- Advanced algorithms, effective visualizations, and many (standard) workflows.
- Supports micro-level documentation and replication of studies.
- Is open source—anybody can review and extend the code, or use it for commercial purposes.

Science of Science (Sci2) Tool
http://sci.slis.indiana.edu

**SUMMARY**
- Existing metrics have known flaws
- A reliable, open, joined-up data infrastructure is needed
- Data should be collected on the full range of scientists’ work
- Social scientists and economists should be involved

**OPINION**

Let’s make science metrics more scientific

To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says Julia Lane.
Sci² Tool – “Open Code for S&T Assessment”

OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.

Sci\textsuperscript{2} Tool: Download, Install, and Run

Sci\textsuperscript{2} Tool \textbf{Alpha 3 (March 2010)}
Can be freely downloaded for all major operating systems from http://sci.slis.indiana.edu/sci2
Select your operating system from the pull down menu and download.
Unpack into a /sci2 directory.
Run /sci2/sci2.exe

Tutorial is linked from web page.

Cite as

Sci\textsuperscript{2} Tool: Download, Install, and Run

Sci\textsuperscript{2} Tool \textbf{Alpha 4 (Aug 2010)}
Has new features such as
\begin{itemize}
  \item STAR database
  \item Colored Horizontal Bar Graphs
  \item Supports ASCII UTF-8 characters
  \item Bug fixes, streamlined workflows
\end{itemize}

NEH Tutorial beta version was made available on DVD for Windows, Mac, and Linux. Unzip and run /sci2/sci2.exe

Pre-release of tutorial is at http://ella.slis.indiana.edu/~katy/outgoing/Sci2TutorialAlpha4-NEH-Workshop.pdf
Sci2 Tool Interface Components

Use

- **Menu** to read data, run algorithms.
- **Console** to see work log, references to seminal works.
- **Data Manager** to select, view, save loaded, simulated, or derived datasets.
- **Scheduler** to see status of algorithm execution.

All workflows are recorded into a log file (see `/sci2/logs/…`), and soon can be re-run for easy replication. If errors occur, they are saved in a error log to ease bug reporting.

All algorithms are documented online; workflows are given in tutorials, see [http://sci.slis.indiana.edu/sci2](http://sci.slis.indiana.edu/sci2) and [http://nwb.slis.indiana.edu](http://nwb.slis.indiana.edu) > Community
Use ‘File > Read Directory Hierarchy’ with parameters

Visualize resulting ‘Directory Tree - Prefuse (Beta) Graph’ using
- ‘Visualization > Tree View (prefuse beta)’
- ‘Visualization > Tree Map (prefuse beta)’
- ‘Visualization > Balloon Graph (prefuse alpha)’
- ‘Visualization > Radial Tree/Graph (prefuse alpha)’

Different views of the /nwb directory hierarchy.

Note the size of the /plugin directory.
Workshop Overview

9:00 Marcoscope Design and Usage & CIShell Powered Tools
9:15 Sci2 Tool Basics
   ➢ Download and run the tool.


10:00 Sci2 Tool – Using Text and Database Support
   ➢ Studying Four Major NetSci Researchers.
   ➢ Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
   ➢ Find basic statistics and run various algorithms over the network.
   ➢ Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance

Padgett's Florentine Families - Compute Basic Network Properties & View in GUESS

➢ Florentine families related through business ties (specifically, recorded financial ties such as loans, credits and joint partnerships) and marriage alliances.

➢ Node attributes
   - Wealth: Each family's net wealth in 1427 (in thousands of lira)
   - Priorates: The number of priorates (seats on the civic council) held between 1282-1344
   - Totalties: The total number of business or marriage ties in the total dataset of 116 families.

➢ “Substantively, the data include families who were locked in a struggle for political control of the city of Florence around 1430. Two factions were dominant in this struggle: one revolved around the infamous Medicis, the other around the powerful Strozzis.”

Load `yoursci2directory*/sampledata/socialscience/florentine.nwb`

Run ‘Analysis > Network Analysis Toolkit (NAT)’ to get basic properties.

This graph claims to be undirected.
Nodes: 16
Isolated nodes: 1
Node attributes present: label, wealth, totalities, priorates
Edges: 27
No self loops were discovered.
No parallel edges were discovered.
Edge attributes:
Nonnumeric attributes:
Example value
marriage_T
business_F
Average degree: 3.375
There are 2 weakly connected components. (1 isolates)
The largest connected component consists of 15 nodes.
Did not calculate strong connectedness because this graph was not directed.
Density (disregarding weights): 0.225

Optional: Run ‘Analysis > Unweighted & Undirected > Node Betweenness Centrality’ with default parameters.
Select network and run ‘Visualization > GUESS’ to open GUESS with file loaded.
Apply ‘Layout > GEM’.
Pan: “grab” the background by holding left-click and moving your mouse.

Zoom:
Using scroll wheel, press the “+” and “-” buttons in the upper-left hand corner, or right-click and move the mouse left or right. Center graph by selecting ‘View -> Center’.

Select to select/move single nodes. Hold down ‘Shift’ to select multiple.

Right click to modify Color, etc.

**Graph Modifier:**
Select “all nodes” in the Object drop-down menu and click ‘Show Label’ button.

Select ‘Resize Linear > Nodes > totalities’ drop-down menu, then type “5” and “20” into the From” and To” Value box separately. Then select ‘Do Resize Linear’.

Select ‘Colorize > Nodes > totalities’, then select white and enter (204,0,51) in the pop-up color boxes on in the “From” and “To” buttons.

Select “Format Node Labels”, replace default text {originallabel} with your own label in the pop-up box ‘Enter a formatting string for node labels.’
Workshop Overview

9:00 Marcoscope Design and Usage & CIShell Powered Tools
9:15 Sci2 Tool Basics
  ➢ Download and run the tool.

10:00 Sci2 Tool – Using Text and Database Support
  ➢ Studying Four Major NetSci Researchers.
  ➢ Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
  ➢ Find basic statistics and run various algorithms over the network.
  ➢ Visualize the networks as either a graph or a circular hierarchy.

10:15-10:30 Break

10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance

Interpreter:
Uses Jython a combination of Java and Python.
Try
colorize(wealth, white, red)
resizeLinear(sitebetweenness, 5, 25)
Thomson Reuter’s Web of Knowledge (WoS) is a leading citation database cataloging over 10,000 journals and over 120,000 conferences. Access it via the “Web of Science” tab at http://www.isiknowledge.com (note: access to this database requires a paid subscription). Along with Scopus, WoS provides some of the most comprehensive datasets for scientometric analysis.

To find all publications by an author, search for the last name and the first initial followed by an asterisk in the author field.
Comparison of Counts
No books and other non-WoS publications are covered.

<table>
<thead>
<tr>
<th>Age</th>
<th>Total # Cites</th>
<th>Total # Papers</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>1,525</td>
<td>672</td>
<td>31</td>
</tr>
<tr>
<td>122</td>
<td>122</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>42</td>
<td>451</td>
<td>101</td>
<td>33</td>
</tr>
<tr>
<td>40</td>
<td>2,218</td>
<td>126</td>
<td>47</td>
</tr>
<tr>
<td>41</td>
<td>16,920</td>
<td>159</td>
<td>52</td>
</tr>
</tbody>
</table>

(Dec 2007)

Extract Co-Author Network

Load `*yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi` using 'File > Load and Clean ISI File' and parameters.

And file with 361 records appears in Data Manager.
To extract the co-author network, select the ‘361 Unique ISI Records’ table and run ‘Data Preparation > Text Files > Extract Co-Author Network’ using isi file format:

The result is an undirected but weighted network of co-authors in the Data Manager. Run ‘Analysis > Network > Network Analysis Toolkit (NAT)’ to calculate basic properties: the network has 247 nodes and 891 edges.

Use ‘Analysis > Network > Unweighted and Undirected > Node Degree’ to calculate the number of neighbors for each node.

To view the complete network, select the ‘Extracted Co-Authorship Network’ and run ‘Visualization > Networks > GUESS’.

Network is loaded with random layout. In GUESS, run ‘Layout > GEM’ and ‘Layout > Bin Pack’ to improve layout.
Co-Author Network of all Four NetsSci Researchers

Use the GUESS Graph Modifier to change color and size coding.

Calculate node degrees in Sci2 Tool.

Use a graphic program to add legend.

Individual Co-Author Networks
(Read/map 4 files separately)
Load and Clean ISI File was selected.
Loaded 361 records.
Removed 0 duplicate records.
Author names have been normalized.
361 records with unique ISI IDs are available via Data Manager.

Extract Co-Author Network was selected.
Input Parameters:
File Format: isi

Network Analysis Toolkit (NAT) was selected.
Nodes: 247
Edges: 891

GUESS was selected.
Weak Component Clustering was selected.
Implementer(s): Russell Duhon
Integrator(s): Russell Duhon

Input Parameters:
Number of top clusters: 10
3 clusters found, generating graphs for the top 3 clusters.

Node Degree was selected.
Documentation:
https://nwb.slis.indiana.edu/community/?n=AnalyzeData.NodeDegree
Network Visualization:
Color/Size Coding by Betweeness Centrality

Node Betweenness Centrality was selected.
Author(s): L. C. Freeman
Implementer(s): Santo Fortunato
Integrator(s): Santo Fortunato, Weixia Huang

Input Parameters:
Number of bins: 10

Network Visualization:
Reduced Network After Pathfinder Network Scaling

MST-Pathfinder Network Scaling was selected.
Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight
Network Visualization: Circular Hierarchy Visualization

Select Co-Author Network and run Blondel Community detection:

With parameter values

Network Visualization: Circular Hierarchy Visualization

Visualize resulting file using ‘Visualization > Networks > Circular Hierarchy’ with parameter values
Network Visualization: Circular Hierarchy Visualization

Nodes that are interlinked/clustered are spatially close to minimize the number of edge crossings.

Node labels, e.g., author names.

Network structure using edge bundling.

Color coded cluster hierarchy according to Blondel community detection algorithm.

Note:
Header/footer info, legend, and more meaningful color coding are under development.

Paper-Citation Network Layout

To extract the paper-citation network, select the ‘361 Unique ISI Records’ table and run ‘Data Preparation > Text Files > Extract Paper Citation Network.’

The result is a unweighted, directed network of papers linked by citations, named Extracted paper-citation network in the Data Manager.

Run NAT to calculate that the network has 5,342 nodes and 9,612 edges. There are 15 weakly connected components. (0 isolates)

Run ‘Analysis > Networks > Unweighted and Directed > Weak Component Clustering’ with parameters

to identify top-10 largest components. The largest (giant) component has 5,151 nodes.

To view the complete network, select the network and run ‘Visualization > GUESS’.
Data:
WoS and Scopus for 2001–2005, 7.2 million papers, more than 16,000 separate journals, proceedings, and series

Similarity Metric:
Combination of bibliographic coupling and keyword vectors

Number of Disciplines:
554 journal clusters further aggregated into 13 main scientific disciplines that are labeled and color coded in a metaphorical way, e.g., Medicine is blood red and Earth Sciences are brown as soil.
Using Database Support – Load a File

Load `*yoursci2directory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi`, using ‘File > Load’ and select

It might take several minutes to load the data into the database.

Select the ISI Database in the Data Manager and run ‘Data Preparation > Database > ISI > *’ to calculate statistics, e.g., annual counts, extract networks, prepare data for burst detection, etc.

Using Database Support – View Database Structure

View the database schema by right-clicking on the loaded database in the Data Manager and clicking “View”
Using Database Support – Data Unification

Run ‘Data Preparation > Database > ISI > Merge Identical ISI People’, followed by ‘Data Preparation > Database > ISI > Merge Document Sources’ and ‘Data Preparation > Database > ISI > Match References to Papers’. Make sure to wait until each cleaning step is complete before beginning the next one. Read red warnings.

Successfully merged 64 entries into other entries, leaving 2003 entries in the database.

Using Database Support – Extract Basic Properties

Run ‘Data Preparation > Database > ISI > Extract Authors’ and right-click on the resulting table to view all the authors from FourNetSciResearchers.isi. The table also has columns with information on how many papers each person in the dataset authored, their Global Citation Count (how many times they have been cited according to ISI), and their Local Citation Count (how many times they were cited in the current dataset).
Using the ISI Database (section 3.1)

**Database > ISI**

- Extract Authors – Outputs a table containing one row per author in the database, which includes Papers Authored in Dataset, Global Citation Count, and Local Citation Count.
- Extract Documents – Outputs a table containing one row per document in the database, with all bibliographic information related to the document (including abstract text, if available), as well as Times Cited, Cited Reference Count, Digital Object ID (if available), and Funding Information.
- Extract Keywords – Outputs a table containing one row per keyword in the database which includes occurrences of each keyword in the dataset.
- Extract Document Sources – Outputs a table containing one row per document source in the database, which includes the Full Title, Number of Papers Contained from the Dataset, ISSN, and other bibliographic information (if available).

Using Database Support – Records over time

Aggregate data by year by running ‘Data Preparation > Database > ISI > Extract Longitudinal Summary.’ Result is a table which lists metrics for every year mentioned in the dataset. The longitudinal study table contains the volume of documents and references published per year, as well as the total amount of references made, the amount of distinct references, distinct authors, distinct sources, and distinct keywords per year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Documents References</th>
<th>Total References</th>
<th>Distinct References</th>
<th>Distinct Authors</th>
<th>Distinct Sources</th>
<th>Distinct Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>90</td>
<td>19,610</td>
<td>1,453</td>
<td>361</td>
<td>30</td>
<td>119</td>
</tr>
<tr>
<td>2004</td>
<td>91</td>
<td>20,200</td>
<td>1,587</td>
<td>388</td>
<td>31</td>
<td>123</td>
</tr>
<tr>
<td>2005</td>
<td>92</td>
<td>20,800</td>
<td>1,723</td>
<td>415</td>
<td>32</td>
<td>131</td>
</tr>
<tr>
<td>2006</td>
<td>93</td>
<td>21,400</td>
<td>1,865</td>
<td>440</td>
<td>33</td>
<td>137</td>
</tr>
<tr>
<td>2007</td>
<td>94</td>
<td>22,000</td>
<td>1,999</td>
<td>461</td>
<td>34</td>
<td>143</td>
</tr>
<tr>
<td>2008</td>
<td>95</td>
<td>22,600</td>
<td>2,131</td>
<td>485</td>
<td>35</td>
<td>150</td>
</tr>
<tr>
<td>2009</td>
<td>96</td>
<td>23,200</td>
<td>2,267</td>
<td>507</td>
<td>36</td>
<td>157</td>
</tr>
<tr>
<td>2010</td>
<td>97</td>
<td>23,800</td>
<td>2,403</td>
<td>531</td>
<td>37</td>
<td>164</td>
</tr>
<tr>
<td>2011</td>
<td>98</td>
<td>24,400</td>
<td>2,543</td>
<td>554</td>
<td>38</td>
<td>171</td>
</tr>
<tr>
<td>2012</td>
<td>99</td>
<td>25,000</td>
<td>2,685</td>
<td>578</td>
<td>39</td>
<td>178</td>
</tr>
<tr>
<td>2013</td>
<td>100</td>
<td>25,600</td>
<td>2,829</td>
<td>603</td>
<td>40</td>
<td>185</td>
</tr>
</tbody>
</table>
Using Database Support – Records over time
Aggregate data by year by running ‘Data Preparation > Database > ISI > Extract Longitudinal Study.’ Result is a table which lists metrics for every year mentioned in the dataset. The longitudinal study table contains the volume of documents and references published per year, as well as the total amount of references made, the amount of distinct references, distinct authors, distinct sources, and distinct keywords per year.

Using Database Support – Burst Analysis for References
The queries can also output data specifically tailored for the burst detection algorithm (see Section 4.6.1 Burst Detection). Run ‘Data Preparation > Database > ISI > Extract Authors > Extract References by Year for Burst Detection’ on the cleaned database followed by ‘Analysis > Topical > Burst Detection’ with parameters on left and then run ‘Visualize > Temporal > Horizontal Bar Graph’ with parameters on right.

Watch those red warnings!
Studying Four Major NetSci Researchers (ISI Data) using Database (section 5.1.5)

Using Database Support – Burst Analysis Result

| Bishop YMM, 1975, DISCRETE MULTIVARIAT | Fienberg SE, 1980, ANAL CROSS CLASSIFIE |
| White HC, 1976, AM J SOCIOL, V81, P730 | Wasserman S, 1986, BRIT J MATH STAT PSY, V39, F |
| Garfield E, 1977, CURR CONTENTS, P5 | Wasserman S, 1984, SOC NETWORKS, V6, P177 |

early bursts
Workshop Overview

9:00 Marcoscope Design and Usage & C1Shell Powered Tools
9:15 Sci2 Tool Basics
  ➢ Download and run the tool.
10:00 Sci2 Tool – Using Text and Database Support
  ➢ Studying Four Major NetSci Researchers.
  ➢ Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
  ➢ Find basic statistics and run various algorithms over the network.
  ➢ Visualize the networks as either a graph or a circular hierarchy.
10:15-10:30 Break
10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance

Indiana Philosophy Ontology Project – Mapping Concepts in Philosophy

The Indiana Philosophy Ontology (InPhO) project offers a variety of tools for students, researchers, programmers and scholars. The ontology currently covers philosophical ideas, thinkers, and journals, and is powered by expert feedback and cutting-edge machine learning techniques over the 13 million word Stanford Encyclopedia of Philosophy (SEP).
SEPrevRelatedness.nwb – Using the philosophy ontology to find related articles in the Stanford Encyclopedia of Philosophy

Open SEPrevRelatedness.nwb in a text editor.

```
#Nodesid*int label*string xpos*real ypos*real
1  "law-language" 0.648 0.2027
2  "constructive-empiricism" 0.7694 0.3133
3  "patrizi" 0.45 0.6043
4  "habermas" 0.1964 0.3122
5  "marcel" 0.5689 0.31996
...

#UndirectedEdgessource*int target*int weight*float
154 433 0.45
417 161 0.62
428 85 0.53
709 349 0.45
647 652 0.68
```

NWB Network file includes nodes, edges, and attributes for each

Load SEPrevRelatedness.nwb

Run ‘Analysis > Network Analysis Toolkit (NAT)’ to get basic properties.

Nodes: 868
Isolated nodes: 4
Node attributes present: label, xpos, ypos
Edges: 2539
No self loops were discovered.
No parallel edges were discovered.
Edge attributes:
Did not detect any nonnumeric attributes
Numeric attributes:
  min max mean
  weight 0.43 0.85 0.53 677
This network seems to be valued.
Average degree: 5.8502

Select network and run ‘Preprocessing > Networks > Delete Isolates’

Run ‘Analysis > Networks > Weighted & Undirected > Degree & Strength’ selecting ‘weight’ as the Weight Attribute, followed by ‘Analysis > Networks > Weighted & Undirected > Node Betweenness Centrality’ with default values.

Run ‘Visualization > GUESS’ to open GUESS with file loaded.

Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’
Resize Linear > Nodes > Site Betweenness > From 1 to 50 > Do Resize Linear.

Colorize > Nodes > strength > > Do Colorize.

Resize Linear > Edges > Weight > 1 to 10 > Do Resize Linear.

Colorize > Edges > weight > > Do Colorize.

Zoom in to view cliques & right-click to toggle labels.
Workshop Overview

9:00 Marcoscope Design and Usage & CIShell Powered Tools
9:15 Sci2 Tool Basics
➢ Download and run the tool.
10:00 Sci2 Tool – Using Text and Database Support
➢ Studying Four Major NetSci Researchers.
➢ Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
➢ Find basic statistics and run various algorithms over the network.
➢ Visualize the networks as either a graph or a circular hierarchy.
10:15-10:30 Break
10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance

The Letters of Athanasius Kircher – Glen Worthey

Kircher.csv – 362 letters and documents from Stanford’s Kircher collection
➢ Open Kircher.csv

ID,TITLE,NAME,OCUPATION,DAY,MONTH,DATE,PLACE,LANGUAGE,Subs
2487,Brahe @ Copernicus @ Appenzeller @ Bidermann @ Adriansens @ Clavius,astronomer,27,June,1600,Ingolstadt,Latin,calendar @ sundials @ astronomy
2541,Gellibrand,mathematician,09,March,1635,no place given,Latin,longitude @ magnetic declination @ magnetism
…
Load Kircher.csv in ‘Standard csv format’.

Run ‘File > Load Generic CSV-File Into Database’ with the following parameters:

- Core entity = LETTERS
- IDTITLE=Integer
- NAMES = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type '@' as the separator
- OCCUPATIONS = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type '@' as the separator
- DAY = Integer
- MOUNTH = String
- DATE = Integer
- CITY = String, Create a separate leaf table, merge if identical
- LANGUAGE = String, Create a separate leaf table, merge if identical
- Subs = String, Create a separate leaf table, merge if identical, contains multi-valued fields, type '@' as the separator

These attributes can also be loaded from columnAttributes.gcl

‘I’m Finished!’

Run ‘Data Preparation > Database > General > Extract Raw Tables From Database’ to view various database tables.

Run ‘Data Preparation > Database > Generic CSV > Extract Co-Occurrence Network’ with the following parameters:

Run ‘Visualization > GUESS’ to open GUESS with file loaded.

Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’

Results not surprising for an egocentric network.
Run ‘Data Preparation > Database > Generic CSV > Extract Co-Occurrence Network’ with the following parameters:

- Run ‘Visualization > GUESS’ to open GUESS with file loaded.
- Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’

Resize Linear > Edges > edge_count_idtitle > From 1 to 10 > Do Resize Linear.

Colorize > Edges > edge_count_idtitle > From: To: > Do Colorize.

Resize Linear > Nodes > node_count_idtitle > 0 to 25 > Do Resize Linear.

Object: ‘Nodes Based On’
Property: ‘node_count_idtitle’
ID: ‘>=’
Value: ‘15’
Show Label
Run ‘Data Preparation > Database > Generic CSV > Extract Bi-Partite Network’ with the following parameters:

- Choose the Source for your bipartite network extraction: LANGUAGE
- Choose the Target for your bipartite network extraction: OCCUPATIONS

Node Attributes
- Add Another Node Attribute Field
- Remove All Node Fields

Edge Attributes
- Add Another Edge Attribute Field
- Remove All Edge Fields

Run ‘Visualization > GUESS’ to open GUESS with file loaded.
Apply ‘Layout > GEM’ followed by ‘Layout > Bin Pack’
Show labels and color nodes based on entity_type Language

Workshop Overview

9:00 Marcoscope Design and Usage & CIShell Powered Tools
9:15 Sci2 Tool Basics
- Download and run the tool.
10:00 Sci2 Tool – Using Text and Database Support
- Studying Four Major NetSci Researchers.
- Load and clean a dataset as text file or using the Sci2 Database; process raw data into networks.
- Find basic statistics and run various algorithms over the network.
- Visualize the networks as either a graph or a circular hierarchy.
10:15-10:30 Break
10:30-11:00 Sci2 Research Demonstration I: Indiana Philosophy Ontology Project - Map concepts and influence in the field of philosophy.
11:00-11:30 Sci2 Research Demonstration II: The Letters of Athanasius Kircher.
11:30-12:00 Q&A and Technical Assistance
Extraneous Slides

- Adding Plugins to CIShell Powered Tools
- OSGi/CIShell Adoption

Cyberinfrastructure Shell (CIShell)
http://cishell.org

- CIShell is an open source software specification for the integration and utilization of datasets, algorithms, and tools.
- It extends the Open Services Gateway Initiative (OSGi) (http://www.osgi.org), a standardized, component oriented, computing environment for networked services widely used in industry since 10 years.
- Specifically, CIShell provides “sockets” into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.
CIShell is built upon the Open Services Gateway Initiative (OSGi) Framework.

**OSGi** ([http://www.osgi.org](http://www.osgi.org)) is
- A standardized, component-oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 8 years.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Widely adopted in open source realm, especially since Eclipse 3.0 that uses OSGi R4 for its plugin model.

**Advantages of Using OSGi**
- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can connected via RPC/RMI supporting peer-to-peer sharing of data, algorithms, and computing power.

Ideally, CIShell becomes a standard for creating OSGi Services for algorithms.

---

No central data format.
- Sci² Tool has 26 external and internal data formats and 35 converters.
- Their relationships can be derived by running ‘File > Converter Graph’ and plotted as shown here. Note that some conversions are symmetrical (double arrow) while others are one-directional (arrow).
CIShell – Add new Plugins, e.g., UCSD Science Map

- Not all code can be shared freely (yet).
- To make the UCSD Science Map and new geomaps available via the Sci² menu, simply add

  
<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Type</th>
<th>Date Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>edu.cncl.visualization.geomap_v0.0.1.jar</td>
<td>4.6 Kb</td>
<td>Executable Jar File</td>
<td>6/18/2010 5:41 PM</td>
</tr>
<tr>
<td>edu.cncl.visualization.geomap_v0.0.1.jar</td>
<td>4.6 Kb</td>
<td>Executable Jar File</td>
<td>6/18/2010 5:41 PM</td>
</tr>
<tr>
<td>eng.cncl.ontology.visualization.geomaps_v0.0.0.jar</td>
<td>61 Kb</td>
<td>Executable Jar File</td>
<td>6/18/2010 5:41 PM</td>
</tr>
<tr>
<td>eng.cncl.utilitites_1.0.6.jar</td>
<td>75 Kb</td>
<td>Executable Jar File</td>
<td>6/18/2010 5:41 PM</td>
</tr>
</tbody>
</table>

  to the ‘yourdirectory/plugin’ directory and restart the tool.

The rights to the UCSD map are owned by the Regents of UCSD. Usage does not require a separate, signed agreement or an additional request to our office if consistent with the permission. As a courtesy, please send information on how the map is being used to

William J. Decker, Ph.D., Associate Director, Technology Transfer Office
University of California, San Diego, 9500 Gilman Drive Dept. 0910, La Jolla, CA 92093
phone: 858-822-5128, fax: 858-534-7345, e-mail: wjdecker@ucsd.edu

- To delete algorithms that you do not use, simply delete the corresponding *.jar files in the plugin directory.
- Customize your menu structure accordingly—see next slide.

After you added the new plugins, load an ISI file using ‘File > Load and Clean ISI File > EugeneGarfield.isi.’

The file can be found in the /sampledata/scientometrics/isi directory.

Select ‘99 Unique ISI Records’ file in Data Manager and run ‘Visualization > Topical > Science Map via Journals’ with parameters:

The result is a science map overlay of Garfield’s papers and a listing of journals in 13 fields of science below.
Adding more alayout algorithms and network visualization interactivity via Cytoscape [http://www.cytoscape.org].

Simply add `org.textrend.visualization.cytoscape_0.0.3.jar` into your `/plugin` directory.

Restart Sci2 Tool.

Cytoscape now shows in the Visualization Menu.

Select a network in Data Manager, run Cytoscape and the tool will start with this network loaded.
CIShell – Integrate New Algorithms

Algorithm Developer’s Guide

Overview
The CyberInfrastructure Shell (CIShell) is an open-source, community-driven platform for the integration and utilization of datasets, algorithms, tools, and computing resources. Algorithm integration support is built in for Java and most other programming languages. Being Java based, it will run on almost all platforms. The software and specification is released under an Apache 2.0 License.

This guide attempts to aid algorithm developers in creating algorithms for CIShell (and applications built on CIShell).

This guide tries to contain all the information a new developer needs, but where necessary, it may cite the CIShell 1.0 Specification (API) or the OSGi Service Platform Specification, Release 4 (ADP). While the guide tries to make beginning algorithm development easier, the CIShell Specification has the last word on how the CIShell Platform works.

Table of Contents
1. CIShell Basics
2. Getting Started
   1. Tutorial 1: Setting Up the Development Environment
   2. Tutorial 1: Creating a Hello World Java Algorithm
   3. Tutorial 2: Practical Java Algorithm Development
   4. Tutorial 3: Integrating a Non-Java Program As An Algorithm
   6. Where to Learn More
3. Reference
   2. Accessing the OSGi Console in CIShell tools


94
The file ‘yourtooldirectory/configuration/default_menu.xml’ encodes the structure of the menu system.

In NWB Tool, the Modeling menu (left) is encoded by the following piece of XML code:

```xml
<top_menu name="Modeling">
  <menu pids="edu.ui.web.modeling,endoskeletongraph"/>
  <menu pids="edu.ui.web.modeling,grababastalbert"/>
  <menu type="break"/>
  <menu pids="edu.ui.web.modeling,ppp,can,CanAlgorithm"/>
  <menu pids="edu.ui.web.modeling,ppp,quod,ChordAlgorithm"/>
  <menu pids="edu.ui.web.modeling,ppp,quad,QuadAlgorithm"/>
  <menu type="break"/>
  <menu pids="edu.ui.web.modeling,quad1,QuadAlgorithm"/>
  <menu type="break"/>
</top_menu>
```

### OSGi/CIShell Adoption

CIShell/OSGi is at the core of different CIs and a total of 169 unique plugins are used in the
- Information Visualization ([http://iv.slis.indiana.edu](http://iv.slis.indiana.edu)),
- Network Science (NWB Tool) ([http://nwb.slis.indiana.edu](http://nwb.slis.indiana.edu)),
- Scientometrics and Science Policy (Sci² Tool) ([http://sci.slis.indiana.edu](http://sci.slis.indiana.edu)), and
- Epidemics ([http://epic.slis.indiana.edu](http://epic.slis.indiana.edu)) research communities.

Most interestingly, a number of other projects recently adopted OSGi and one adopted CIShell:

**Cytoscape** ([http://www.cytoscape.org](http://www.cytoscape.org)) lead by Trey Ideker, UCSD is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data (Shannon et al., 2002).

**Taverna Workbench** ([http://taverna.sourceforge.net](http://taverna.sourceforge.net)) lead by Carol Goble, University of Manchester, UK is a free software tool for designing and executing workflows (Hull et al., 2006). Taverna allows users to integrate many different software tools, including over 30,000 web services.

**MAEviz** ([https://wiki.ncsa.uiuc.edu/display/MAE/Home](https://wiki.ncsa.uiuc.edu/display/MAE/Home)) managed by Shawn Hampton, NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.

**TEXTrend** ([http://www.textrend.org](http://www.textrend.org)) lead by George Kampis, Eötvös University, Hungary develops a framework for the easy and flexible integration, configuration, and extension of plugin-based components in support of natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component.

As the functionality of OSGi-based software frameworks improves and the number and diversity of dataset and algorithm plugins increases, the capabilities of custom tools will expand.
TEXTrend adds R bridge, WEKA, Wordij, CFinder, and more. See the latest versions of TEXTrend Toolkit modules at http://textrend.org/index.php?option=com_content&view=article&id=47&Itemid=53

Epidemics Marketplace and EpiC Tool

http://dev.epic.slis.indiana.edu
All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.slis.indiana.edu