Plug-and-Play Macroscopes That Support Replicable Science of Science Studies

Joseph Biberstine & Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
katy@indiana.edu

With special thanks to the members at the Cyberinfrastructure for Network Science Center, the NWB team, the Sci2 team, the EpiC team, and all other teams that use OSGi/CIShell.

JSMF Workshop on Standards for Science Metrics, Classifications, and Mapping
Indiana University, Bloomington, IN
August 11, 2011
Designing “Dream Tools”

Many of the best micro-, tele-, and macrosopes are designed by scientists keen to observe and comprehend what no one has seen or understood before. Galileo Galilei (1564–1642) recognized the potential of a spyglass for the study of the heavens, ground and polished his own lenses, and used the improved optical instruments to make discoveries like the moons of Jupiter, providing quantitative evidence for the Copernican theory.

Today, scientists repurpose, extend, and invent new hardware and software to create “macrosopes” that may solve both local and global challenges.

The tools I will show you today empower me, my students, colleagues, and 100,000 others that downloaded them.
Decision making in science, industry, and politics, as well as in daily life, requires that we make sense of data sets representing the structure and dynamics of complex systems. Analysis, navigation, and management of these continuously evolving data sets require a new kind of data-analysis and visualization tool we call a macroscope (from the Greek macros, or “great,” and skopein, or “to observe”) inspired by de Rosnay’s futurist science writings. Macrosopes provide a “vision of the whole,” helping us “synthesize” the related elements and enabling us to detect patterns, trends, and outliers while granting access to myriad details. Rather than make things larger or smaller, macrosopes let us observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend.
Goal of This Talk

**Inspire computer scientists** to implement software frameworks that **empower domain scientists** to assemble their own continuously evolving macrosopes, adding and upgrading existing (and removing obsolete) plug-ins to arrive at a set that is truly relevant for their work—with little or no help from computer scientists.

While microscopes and telescopes are physical instruments, **macrosopes resemble continuously changing bundles of software plug-ins**. Macrosopes make it easy to select and combine algorithm and tool plug-ins but also interface plug-ins, workflow support, logging, scheduling, and other plug-ins needed for scientifically rigorous yet effective work.

They make it easy to share plug-ins via email, flash drives, or online. To use new plugins, simply copy the files into the plug-in directory, and they appear in the tool menu ready for use. No restart of the tool is necessary. **Sharing algorithm components, tools, or novel interfaces becomes as easy as sharing images on Flickr or videos on YouTube. Assembling custom tools is as quick as compiling your custom music collection.**
Changing Scientific Landscape—Personal Observations

Different datasets/formats.
Diverse algorithms/tools written in many programming languages.

Physics
IS
CS
Bio
SNA
Related Work

Google Code and SourceForge.net provide special means for developing and distributing software

➢ In August 2009, SourceForge.net hosted more than 230,000 software projects by two million registered users (285,957 in January 2011);
➢ In August 2009 ProgrammableWeb.com hosted 1,366 application programming interfaces (APIs) and 4,092 mashups (2,699 APIs and 5,493 mashups in January 2011)

Cyberinfrastructures serving large biomedical communities

➢ Cancer Biomedical Informatics Grid (caBIG) (http://cabig.nci.nih.gov)
➢ Biomedical Informatics Research Network (BIRN) (http://nbirn.net)
➢ Informatics for Integrating Biology and the Bedside (i2b2) (https://www.i2b2.org)
➢ HUBzero (http://hubzero.org) platform for scientific collaboration uses
➢ myExperiment (http://myexperiment.org) supports the sharing of scientific workflows and other research objects.

Missing so far is a common standard for

➢ the design of modular, compatible algorithm and tool plug-ins (also called “modules” or “components”)
➢ that can be easily combined into scientific workflows (“pipeline” or “composition”),
➢ and packaged as custom tools.
CIShell (http://cishell.org) 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://osgi.org), a standardized, component oriented, computing environment for networked services widely used in industry since more than 10 years.

Specifically, CIShell provides “sockets” into which existing and new datasets, algorithms, and tools can be plugged using a wizard-driven process.
Learn more about existing CIShell-powered tools below.

Network Workbench Tool (NWB)
The NWB Tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks. It comes with a 77-page user manual.

Science of Science Tool (Sci²)
The Sci² Tool was specifically developed for science policy makers and researchers that study science by scientific means. It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels. There exists a 168-page user manual and 24 hours of NIH tutorials in this tool.
Science of Science Cyberinfrastructures

Scholarly Database: 25 million scholarly records
http://sdb.slis.indiana.edu

James S. McDonnell Foundation

VIVO Research Networking
http://vivoweb.org

Information Visualization Cyberinfrastructure
http://iv.slis.indiana.edu

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

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

Epidemics Cyberinfrastructure
http://epic.slis.indiana.edu/
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.
Sci² Tool
A tool for science of science research & practice

Email Address

Password

Login

Forgot your password?
To recover your account password, please visit our password recovery page.

Not registered yet?

Register now

Tutorials

Scott Weingart, BBVA Compass Science 2010, Science, Indiana

- Tutorial #01: Science of Science Research
- Tutorial #02: Network Science / Information Visualization
- Tutorial #03: C|Shelf Powered Tools: Network Workbench and Science of Science Tool
- Tutorial #04: Temporal Analysis—Burst Detection
- Tutorial #05: Geospatial Analysis and Mapping
- Tutorial #06: Temporal Analysis & Mapping
- Tutorial #07: Tree Analysis and Visualization
- Tutorial #08: Network Analysis and Visualization
- Tutorial #09: Large Network Analysis and Visualization
- Tutorial #10: Using the Scholarly Database at IU
- Tutorial #11: VIVO National Researcher Networking
- Tutorial #12: Future Developments


Gootha Senthil (2010) Multidisciplinary Nature of Work With Reference to PIs and ICs Within a Portfolio. PA Group at NIH.


http://sci2.cns.iu.edu
http://sci2.wiki.cns.iu.edu
## Type of Analysis vs. Level of Analysis

<table>
<thead>
<tr>
<th></th>
<th><strong>Micro/Individual</strong> (1-100 records)</th>
<th><strong>Meso/Local</strong> (101–10,000 records)</th>
<th><strong>Macro/Global</strong> (10,000 &lt; records)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical Analysis/Profiling</strong></td>
<td>Individual person and their expertise profiles</td>
<td>Larger labs, centers, universities, research domains, or states</td>
<td>All of NSF, all of USA, all of science.</td>
</tr>
<tr>
<td><strong>Temporal Analysis</strong> <em>(When)</em></td>
<td>Funding portfolio of one individual</td>
<td>Mapping topic bursts in 20-years of PNAS</td>
<td>113 Years of Physics Research</td>
</tr>
<tr>
<td><strong>Geospatial Analysis</strong> <em>(Where)</em></td>
<td>Career trajectory of one individual</td>
<td>Mapping a states intellectual landscape</td>
<td>PNAS Publications</td>
</tr>
<tr>
<td><strong>Topical Analysis</strong> <em>(What)</em></td>
<td>Base knowledge from which one grant draws.</td>
<td>Knowledge flows in Chemistry research</td>
<td>VxOrd/Topic maps of NIH funding</td>
</tr>
<tr>
<td><strong>Network Analysis</strong> <em>(With Whom?)</em></td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network</td>
<td>NSF’s core competency</td>
</tr>
</tbody>
</table>
OSGi/CIShell powered tool with NWB plugins and many new scientometrics and visualizations plugins.

Sci Maps

GUESS Network Vis

Horizontal Time Graphs

Preprocessing
Extract Top N% Records
Extract Top N Records
Normalize Text
Slice Table by Line
Extract Top Nodes
Extract Nodes Above or Below Value
Delete Isolates
Extract top Edges
Extract Edges Above or Below Value
Remove Self Loops
Trim by Degree
MST-Pathfinder Network Scaling
Fast Pathfinder Network Scaling
Snowball Sampling (in nodes)
Node Sampling
Edge Sampling
Symmetrize
Dichotomize
Multipartite Joining
Geocoder
Extract ZIP Code

Modeling
Random Graph
Watts-Strogatz
Small World
Barabási-Albert Scale-Free
TARL

Analysis
Network Analysis Toolkit (NAT)
Unweighted & Undirected
Node Degree
Degree Distribution
K-Nearest Neighbor (Java)
Watts-Strogatz Clustering Coefficient
Watts Strogatz Clustering Coefficient over K
Diameter
Average Shortest Path
Shortest Path Distribution
Node Betweenness Centrality
Weak Component Clustering
Global Connected Components
Extract K-Core
Annotate K-Coreness
HITS

Weighted & Undirected
Clustering Coefficient
Nearest Neighbor Degree
Strength vs Degree
Degree & Strength
Average Weight vs End-point Degree
Strength Distribution
Weight Distribution
Randomize Weights
Blondel Community Detection
HITS
Unweighted & Directed
Node Indegree
Node Outdegree
Indegree Distribution
Outdegree Distribution
K-Nearest Neighbor
Single Node in-Out Degree Correlations
Dyad Reciprocity
Arc Reciprocity
Adjacency Transitivity
Weak Component Clustering
Strong Component Clustering

See https://nwb.slis.indiana.edu/community

**Visualization**
- GnuPlot
- GUESS
- Image Viewer
- Radial Tree/Graph (prefuse alpha)
- Radial Tree/Graph with Annotation (prefuse beta)
- Tree View (prefuse beta)
- Tree Map (prefuse beta)
- Force Directed with Annotation (prefuse beta)
- Fruchterman-Reingold with Annotation (prefuse beta)
- DrL (VxOrd)
- Specified (prefuse beta)

**Scientometrics**
- Remove ISI Duplicate Records
- Remove Rows with Multitudinous Fields
- Detect Duplicate Nodes
- Update Network by Merging Nodes
- Extract Directed Network
- Extract Paper Citation Network
- Extract Author Paper Network
- Extract Co-Occurrence Network
- Extract Word Co-Occurrence Network
- Extract Co-Author Network
- Extract Reference Co-Occurrence (Bibliographic Coupling) Network
- Extract Document Co-Citation Network

**Textual**
- Burst Detection

**Soon:**
Database support for ISI and NSF data.
Contents

1 Introduction ................................................................................................................................. 3
   1.1 Background .......................................................................................................................... 3
   1.2 Objectives ........................................................................................................................... 4
   1.3 Status .................................................................................................................................. 4
   1.4 Data ..................................................................................................................................... 5
   1.5 Assumptions ......................................................................................................................... 5

2 Getting Started ............................................................................................................................. 6
   2.1 Download, Install, Uninstall ............................................................................................... 6
   2.2 User Interface ...................................................................................................................... 7
      2.2.1 Menus .......................................................................................................................... 7
      2.2.2 Console ......................................................................................................................... 8
      2.2.3 Data Manager ............................................................................................................... 8
      2.2.4 Scheduler ...................................................................................................................... 9
   2.3 Data Formats ....................................................................................................................... 10
   2.4 Saving Visualizations for Publication .................................................................................. 10
   2.5 Sample Datasets .................................................................................................................. 11

3 Algorithm and Tool Plugins ....................................................................................................... 12
   3.1 Sci2 Tool Plugins ............................................................................................................... 12
   3.2 Load, View, and Save Data ................................................................................................. 13
   3.3 Memory Allocation ............................................................................................................. 14
      3.3.1 Windows and Linux ...................................................................................................... 14
      3.3.2 Mac .............................................................................................................................. 14
   3.4 Memory Limits .................................................................................................................... 15

4 Workflow Design ....................................................................................................................... 16
   4.1 Overview .............................................................................................................................. 16
   4.2 Data Acquisition and Preparation ....................................................................................... 17
      4.2.1 Datasets: Publications ................................................................................................. 17
      4.2.2 Datasets: Funding .......................................................................................................... 18
      4.2.3 Datasets: Scholarly Database ....................................................................................... 18
   4.3 Database Loading and Manipulation ..................................................................................... 19
   4.4 Summaries and Table Extractions ....................................................................................... 20
   4.5 Statistical Analysis/Profiling ............................................................................................... 20
   4.6 Temporal Analysis (When) ................................................................................................... 21
      4.6.1 Burst Detection .............................................................................................................. 21
      4.6.2 Slice Table by Time ....................................................................................................... 22
   4.7 Geospatial Analysis (Where) ............................................................................................... 22
   4.8 Topical Analysis (What) ....................................................................................................... 23
      4.8.1 Word Co-Occurrence Network ..................................................................................... 23

5 Sample Workflows ..................................................................................................................... 24
   5.1 Sample Workflows: Micro ................................................................................................. 24
      5.1.1 Mapping Collaboration, Publication and Funding Profies of One Researcher ......... 24
      5.1.2 Time Slicing of Co-Authorship Networks (SI Data) .................................................... 25
   5.2 Sample Workflows: Macro ................................................................................................. 26
      5.2.1 Funding Profiles of Three Universities (NSF Data) ...................................................... 26
      5.2.2 Mapping the Field of RNAi Research (SDB Data) ..................................................... 27
      5.2.3 Geo USPTO (SDB Data) .............................................................................................. 28

6 Sample Science Studies & Online Services ................................................................................ 29
   6.1 Science Dynamics ............................................................................................................... 29
   6.2 Local Impact-Output / ROI Studies ..................................................................................... 30
      6.2.1 Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers (2003) .... 30
      6.2.2 Mapping Transdisciplinary Tobacco Use Research Centers Publications (forthcoming) .......................................................... 31
   6.3 Local and Global Science Studies ....................................................................................... 31
      6.3.1 Mapping the Evolution of Co-Authorship Networks (2004) ..................................... 31

7 Extending the Sci2 Tool .............................................................................................................. 33
   7.1 CShell Basics ....................................................................................................................... 33
   7.2 Read New Data ................................................................................................................... 34
   7.3 Creating and Sharing New Algorithm Plugins ................................................................. 34
   7.4 Tools That Use GSGI and/or CShell .................................................................................. 35

8 Relevant Datasets and Tools ..................................................................................................... 36
   8.1 Datasets ............................................................................................................................... 36
   8.2 Network Analysis Tools ...................................................................................................... 36

9 References .................................................................................................................................. 37

See Sci2 Tool Wiki
http://sci2.wiki.cns.iu.edu
Changing Scientific Landscape—Personal Observations Cont.

Common algorithm/tool pool
Easy way to share new algorithms
Workflow design logs
Custom tools

EpiC
Converters
Sci2
NWB

TexTrend

- IS
- CS
- Bio
- SNA
- Phys
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),
- Network Science (NWB Tool) (http://nwb.slis.indiana.edu),
- Scientometrics and Science Policy (Sci² Tool) (http://sci.slis.indiana.edu), and
- Epidemics (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) 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) 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) 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) 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.
Acknowledgements

- Micah Linnemeier and Russell J. Duhon Bruce W. Herr II, George Kamps, Gregory J. E. Rawlins, Geoffrey Fox, Shawn Hampton, Carol Goble, Mike Smoot, Yanbo Han for stimulating discussions and comments.

- The Cyberinfrastructure for Network Science Center (http://cns.iu.edu), the Network Workbench team (http://nwb.cns.iu.edu), and Science of Science project team (http://sci2.cns.iu.edu) for their contributions toward the work presented here.

- Software development benefits greatly from the open-source community. Full software credits are distributed with the source, but I would especially like to acknowledge Jython, JUNG, Prefuse, GUESS, GnuPlot, and OSGi, as well as Apache Derby, used in the Sci2 tool.

This research and development is based on work supported by National Science Foundation grants SBE-0738111, IIS-0513650, IIS-0534909 and National Institutes of Health grants R21DA024259 and 5R01MH079068.
References


[http://www.pnas.org/content/vol101/suppl_1/](http://www.pnas.org/content/vol101/suppl_1/)


[http://scimaps.org/atlas](http://scimaps.org/atlas)

Scharnhorst, Andrea, Börner, Katy, van den Besselaar, Peter (2011) **Models of Science Dynamics.** Springer Verlag.
All papers, maps, tools, talks, press are linked from http://cns.iu.edu

CNS Facebook: http://www.facebook.com/cnscenter
Mapping Science Exhibit Facebook: http://www.facebook.com/mappingscience