Plug-and-Play Macroscopes
That Empower Science

Joseph Biberstine
Senior Software Developer and Research Analyst
Cyberinfrastructure for Network Science Center
School of Library and Information Science
Indiana University, Bloomington
jrbibers@indiana.edu

With special thanks to Katy Börner and all members and collaborators of the Cyberinfrastructure for Network Science Center

Sustaining the Digital Research Enterprise
Hosted by AAMC/GIR, AAMC/GRAND, AAHSL, and NLM
NIH Fishers Lane Conference Center

October 18, 2011
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.
Decision making in science, industry, and politics, as well as in daily life, requires that we make sense of data 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. Macroscopes 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, **macroscopes 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

Explain how cyberinfrastructures like CShell **empower domain scientists** to assemble their own continuously evolving macroscopes.

While microscopes and telescopes are physical instruments, **macroscopes 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 plug-ins, 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.

IS
CS
Bio
SNA

Physics
Changing Scientific Landscape—General Observations

As science becomes more data-driven and computational, more collaborative and interdisciplinary, there is increased need for tools that are easy to extend, share, and customize

- **Star scientist** —> **Research teams**
  Today science is driven by teams that are often larger, more interdisciplinary, and more geographically diverse.

- **Users** —> **Contributors**
  Web 2.0 technologies empower users to contribute to Wikipedia and exchange images, videos, and code via Flickr, YouTube, and SourceForge.net.

- **Disciplinary** —> **Cross-disciplinary**
  The best tools frequently borrow and synergistically combine methods and techniques from different disciplines of science, empowering interdisciplinary and/or international teams to collectively fine-tune and interpret results.

- **Single specimen** —> **Data streams**
  Microscopes and telescopes were originally used to study a single specimen at a time. Today, many researchers must make sense of massive data streams comprising multiple data types and formats from different origins.

- **Static instrument** —> **Evolving cyberinfrastructure**
  The importance of hardware instruments that are static and expensive tends to decrease relative to software tools and services that are highly flexible and evolving to meet the needs of different sciences.
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 for 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.
A framework for easy integration of new and existing algorithms written in any programming language Using CIShell, an algorithm writer can fully concentrate on creating their own algorithm in whatever language they are comfortable with. Simple tools are provided to then take their algorithm and integrate it into CIShell with no additional coding.

A well-defined pool of algorithms and datasets CIShell clearly defines how algorithms and datasets are integrated into the system to create a pool of algorithms and data. An application may then query for algorithms in this pool and execute them. Many applications/tools can be built and customized for different user groups by utilizing the same pool of algorithms.

Leveraging open standards CIShell avoids re-inventing wheels by building on other standards for its specification and reference implementations. It benefits most from the Eclipse family of projects (in particular, the Rich Client Platform and Equinox) and the Open Services Gateway Initiative (OSGi). All CIShell algorithms are integrated as OSGi services and can be used by any OSGi compliant system (including any Eclipse 3.0 or newer based products).

Choose the way you work CIShell offers reference applications that build on the pool of algorithms defined by CIShell. Scripting and a GUI are offered initially with a remoting (peer-to-peer and client-server) architecture, a web front-end, and other interfaces planned. We invite other toolkit developers to build their own applications on top of CIShell's algorithm pool.

Open source, community-driven project CIShell is released under the Apache 2.0 License. Community input is welcome to create a piece of software that advances science and education.
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 (API). 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 0: 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
   5. Mini-Tutorial: Integrating 3rd-party libraries
   6. Where to Learn More
3. Reference
   2. Accessing the OSGi Console in CIShell tools
CIShell Portal (http://cishell.org/home.html)

**Cyberinfrastructure Shell (CIShell)**
CIShell supports the plug-and-play of datasets and algorithms and their bundling into custom tools that serve the specific needs of a user group or research community. It has been applied to develop diverse custom tools; see below. Feel free to take plugins from any of these tools to design your personal dream tool.

Provided by the **Cyberinfrastructure for Network Science Center** at Indiana University.

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 512-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

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 plug-ins that support the preprocessing, analysis, modeling, and visualization of networks.

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

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

Computational Proteomics

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


Figure 2 Drug–target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.
Computational Epidemics
Forecasting (and preventing the effects of) the next pandemic.


- Tutorial #01: Science of Science Research
- Tutorial #02: Network Science / Information Visualization
- Tutorial #03: C1 Shell Powered Tools: Network Workbench and Science of Science Tool
- Tutorial #04: Temporal Analysis—Burst Detection
- Tutorial #05: Geospatial Analysis and Mapping
- Tutorial #06: Topical 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

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

<table>
<thead>
<tr>
<th>Type of Analysis</th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101–10,000 records)</th>
<th>Macro/Global (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 (When)</strong></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 (Where)</strong></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 (What)</strong></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 (With Whom?)</strong></td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network</td>
<td>NSF’s core competency</td>
</tr>
</tbody>
</table>
Sci² Tool – “Open Code for S&T Assessment”

OSGi/CIShell powered tool with NWB plug-ins and many new scientometrics and visualizations plug-ins.

Sci² Tool

Welcome to the Science of Science Tool (Sci²). The development of this tool is supported in part by the Indiana University Research Network Science Center and the School of Library and Information Science, Indiana University, the National Science Foundation under Grant No. IIS-0715303, and the James S. McDonnell Foundation through the Cyberinfrastructure portal (http://sci.slis.indiana.edu). The primary investigators are Katy Börner, Indiana University and SciTech Strategies Inc. The Sci² tool was developed by John Duhan, Patrick A. Phillips, Chintan Tank, and the Cyberinfrastructure Shell (http://cshell.org) for Network Science Center (http://cns.iu.edu). Many algorithm plugins were derived from the DrL (http://drl.cs.nyu.edu) and Specified (http://nwb.slis.indiana.edu).

Please cite as follows:

Geo Maps

Circular Hierarchy
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

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)
---------------------------------------------
Horizontal Bar Graph
Circular Hierarchy
Geo Map (Circle Annotation Style)
Geo Map (Colored-Region Annotation Style)
Science Map (Circle Annotation)

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

Soon:
Database support for ISI and NSF data.
<table>
<thead>
<tr>
<th>Type of Analysis</th>
<th>Micro/Individual (1-100 records)</th>
<th>Meso/Local (101–10,000 records)</th>
<th>Macro/Global (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 (When)</strong></td>
<td>Funding portfolio of one individual</td>
<td>Mapping topic bursts in 20-years of PNAS</td>
<td>113 Years of PNAS Research</td>
</tr>
<tr>
<td><strong>Geospatial Analysis (Where)</strong></td>
<td>Career trajectory of one individual</td>
<td>Mapping a state’s intellectual landscape</td>
<td>PNAS publications</td>
</tr>
<tr>
<td><strong>Topical Analysis (What)</strong></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 (With Whom?)</strong></td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network NIH’s core competency</td>
<td>NIH’s Co-authorship Map</td>
</tr>
</tbody>
</table>
Individual Co-PI Network

Ke & Börner, (2006)

Temporal/Network Analysis
2001-2006, US, InfoVis Scope
Evolving project-PI networks
Mapping Indiana’s Intellectual Space

Geospatial/Network Analysis
2001-2006, BioMed, IN Scope

Academic-Industry collaborations and knowledge diffusion

Identify
- Pockets of innovation
- Pathways from ideas to products
- Interplay of industry and academia
Mapping Topic Bursts

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions


Research questions:

1. Does space still matter in the Internet age?
2. Does one still have to study and work at major research institutions in order to have access to high quality data and expertise and to produce high quality research?
3. Does the Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research institutions?

Contributions:

- Answer to Qs 1 + 2 is YES.
- Answer to Qs 3 is NO.
- Novel approach to analyzing the dual role of institutions as information producers and consumers and to study and visualize the diffusion of information among them.
Topical Composition and Knowledge Flow Patterns in Chemistry Research for 1974 and 2004

Kevin W. Boyack, Katy Börner, & Richard Klavans (2007)

Chemistry - Biology Interface

Temporal/Network Analysis
1974-2004, US, NSF Chemistry Funding Scope
Mapping the main structure, topic evolution, and knowledge diffusion

Number of papers by cluster
40,000
20,000
10,000
5,000

Fraction of papers by cluster
Biochemistry
Chemistry
Biology
Other

Knowledge flows cluster to cluster
cited
citing
Mapping Transdisciplinary Tobacco Use Research Centers Publications

Compare R01 investigator based funding with TTURC Center awards in terms of number of publications and evolving co-author networks.

Zoss & Börner, forthcoming.

Temporal/Network Analysis

1998-2009, US, Tobacco research scope

Evolving co-author networks
MEDLINE Publication Output by The National Institutes of Health (NIH) Using Nine Years of ExPORTER Data

Katy Börner, Nianli Ma, Joseph R. Biberstine, Cyberinfrastructure for Network Science Center, SLIS, Indiana University, Robin M. Wagner, Rediet Berhane, Hong Jiang, Susan E. Ivey, Katrina Pearson and Carl McCabe, Reporting Branch, Division of Information Services, Office of Research Information Systems, Office of Extramural Research, Office of the Director, National Institutes of Health (NIH), Bethesda, MD.

Topical Analysis
2001-2009, US, All NIH funding and associated papers
Changes in publication patterns

Copyright (c) 2008 The Regents of the University of California
Interactive Science Map of NIH Funding

Topic/Network Analysis
2007, US, NIH Funding Scope
Map main structure of NIH funding and institutes’ funding profiles

Grants Awarded

Discover Grants by clicking on map or using the options below.

- United States
- NIH/NSF
- Year(s): 2000-2009
- Amount: 0.00-130.00m

Refine results by selecting institutions or topics:

Top Research Institutions

http://rd-dashboard.nitrd.gov
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 plug-ins 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 plug-ins increases, the capabilities of custom tools will expand.
Acknowledgements

Software development benefits greatly from the open-source community. Full software credits are distributed with the source, but we would especially like to acknowledge OSGi, Sun/Oracle, Jython, Prefuse, GUESS, Apache Derby, JUNG, and GnuPlot.

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


VIVO 1.0 source code was publicly released on April 14, 2010
87 downloads by June 11, 2010.
The more institutions adopt VIVO, the more high quality data will be available to understand, navigate, manage, utilize, and communicate progress in science and technology.
All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.iu.edu