eResearch workflows for studying free and open source software development

James Howison, Andrea Wiggins, & Kevin Crowston

Syracuse University School of Information Studies

9 September 2008 ~ IFIP 2.13 - OSS 2008

eResearch

- Scientific practices and technologies permitting distributed research collaborations using:
  - Large data sets
  - Computational resources
  - Analysis tools and workflows
  - Replicable research with provenance metadata
- FLOSS research community is starting to move in this direction

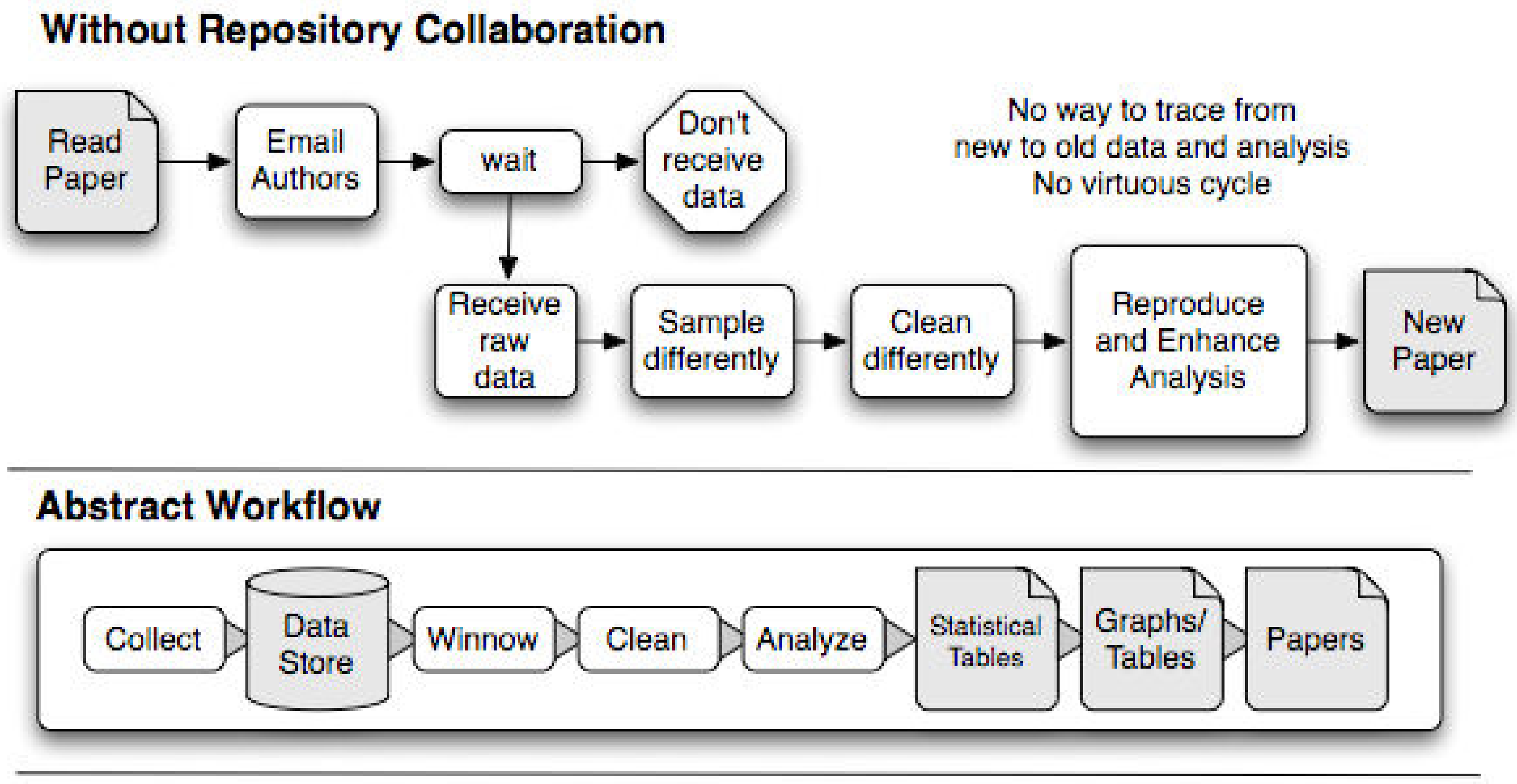
Current FLOSS Research Practices

- Data increasingly available in “repositories of repositories”
  - FLOSSmole
  - FLOSSmetrics
  - SRDA (Notre Dame)
- Not much sharing of analysis or methods for calculating measures; mostly bespoke scripts

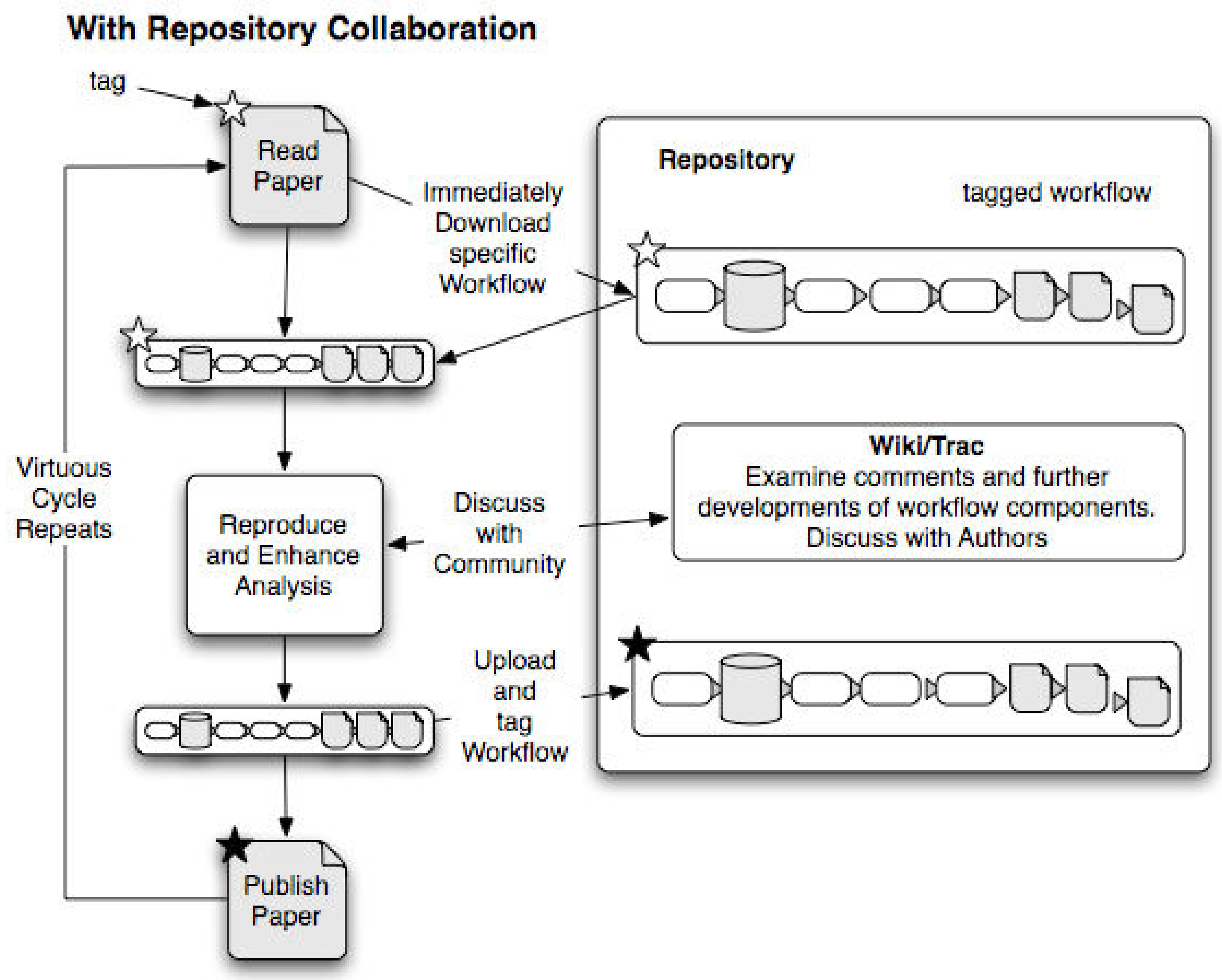
FLOSS Research Repositories

	FLOSSmole	**SRDA (Notre Dame)**	**FLOSSmetrics & CVSAnalY**	**Qualoss & SQO-OSS**	**Source kibitzer**
Project Demographics	Basic data	Basic data	Basic data & confirmed locations		
Developer demographics	Memberships & roles	Memberships & roles			Memberships & roles
Communication venues	Releases; in progress: mail lists, forums, trackers	Forums & trackers	Planned/pilot: mail lists & trackers		
Software venues	SVN/CVS count, packages, releases & dates	SVN/CVS count, packages, releases & dates	SVN/CVS full, size, packages, releases & dates	Planned/pilot: size & complexity metrics	Size & complexity metrics
Use & popularity	Downloads, pageviews, ratings, in Debian, partial: actual use	Downloads & pageviews	In Debian		
Sample collected	Sourceforge, Rubyforge, ObjectWeb, Debian, freshmeat	Sourceforge	Partial: Sourceforge, ObjectWeb, Apache, GNOME	Pilot: KDE	Java projects plus user contributions

FLOSS Research Today



FLOSS Research Tomorrow?



With Repository Collaboration
tag
Read Paper
Immediately Download specific Workflow
Repository
tagged workflow
Virtuous Cycle Repeats
Reproduce and Enhance Analysis
Discuss with Community
Wiki/Trac
Examine comments and further developments of workflow components.
Discuss with Authors
Upload and tag Workflow
Publish Paper

Scientific Workflow Tools

- Tools for scientific analysis (e.g. Kepler, Taverna)
- Self-documenting analysis
  - Analysis conditions recorded at time of execution
- Steps in workflow executed by components with multiple input and output ports
- Components linked by joining input and output ports
- Supports modular analysis development, associated with easier collaboration and higher quality products
- Represented as flow diagram, stored and shared as single XML file

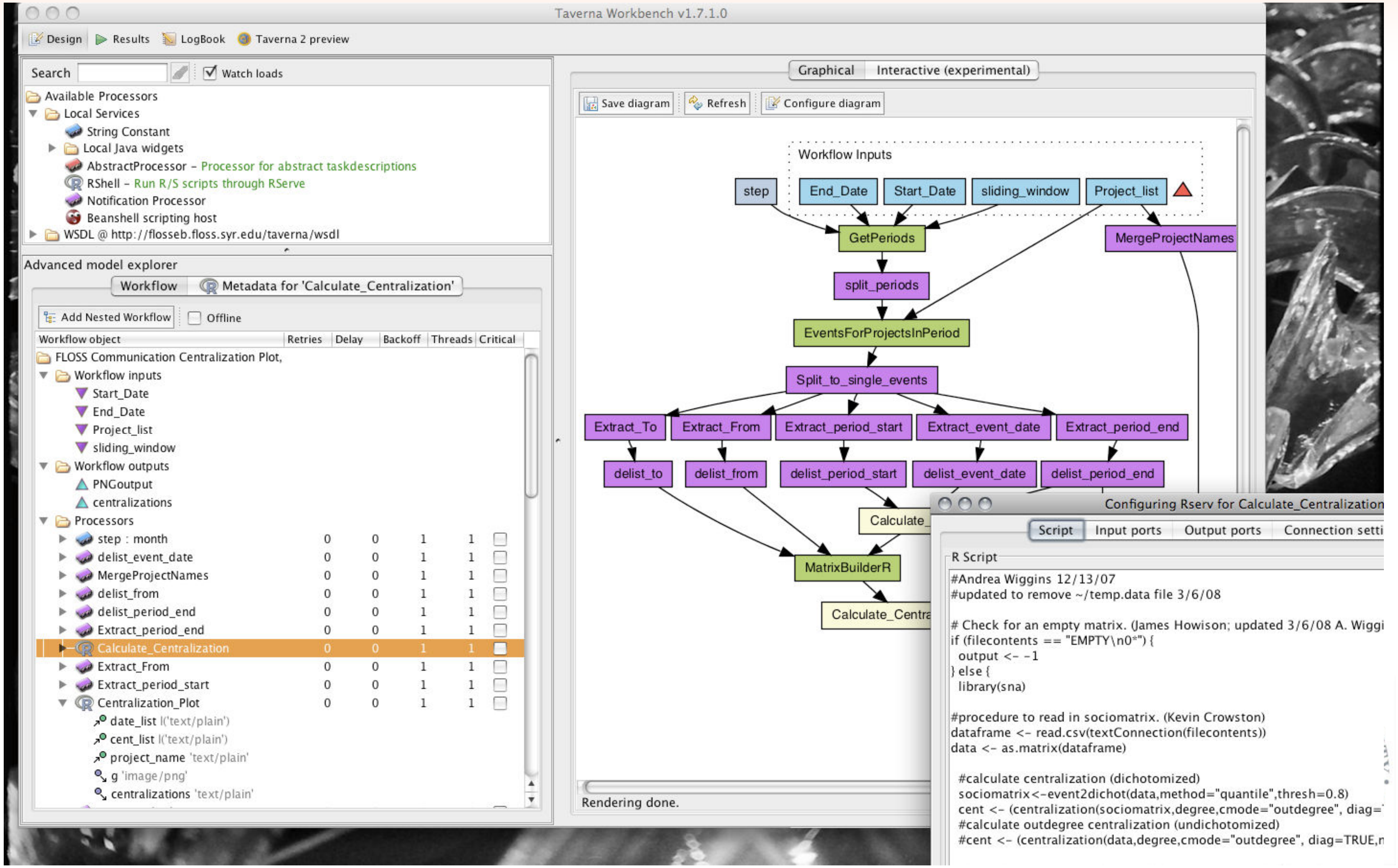
Why Workflows Instead of Scripts?

- Differences and advantages over scripting languages
  - Wider accessibility
    - Programming skills helpful but not necessary to start
  - Compatibility
  - Easier integration of heterogeneous components
    - Mash up and reuse existing scripts
  - Standardized metadata
  - Out-of-the-box operation
    - Requires workflow software to execute, but little other configuration is required
- Example: BioPerl evolution from Perl scripts to libraries to workflow suites

Taverna Workbench

- Open source stand-alone desktop application in Java
- Can also be run via “headless” server application
- Two main interface modes (v 1.71)
  - Design mode: workflow definition, automatically rendered diagram, available component types
  - Execution mode: process monitor, intermediate values, results, XML reports of status and all values
- Local and remote components are connected through input/output ports in MIME typing
  - Can also be grouped into subworkflows

Taverna Design Mode



Workflow object	Retries	Delay	Backoff	Threads	Critical
FLOSS Communication Centralization Plot,					
Workflow inputs					
Start_Date					
End_Date					
Project_list					
sliding_window					
Workflow outputs					
PNGoutput					
centralizations					
Processors					
step : month	0	0	1	1	☐
delist_event_date	0	0	1	1	☐
MergeProjectNames	0	0	1	1	☐
delist_from	0	0	1	1	☐
delist_period_end	0	0	1	1	☐
Extract_period_end	0	0	1	1	☐
Calculate_Centralization	0	0	1	1	☐
Extract_From	0	0	1	1	☐
Extract_period_start	0	0	1	1	☐
Centralization_Plot	0	0	1	1	☐
date_list l('text/plain')					
cent_list l('text/plain')					
project_name 'text/plain'					
g 'image/png'					
centralizations 'text/plain'					

R Script

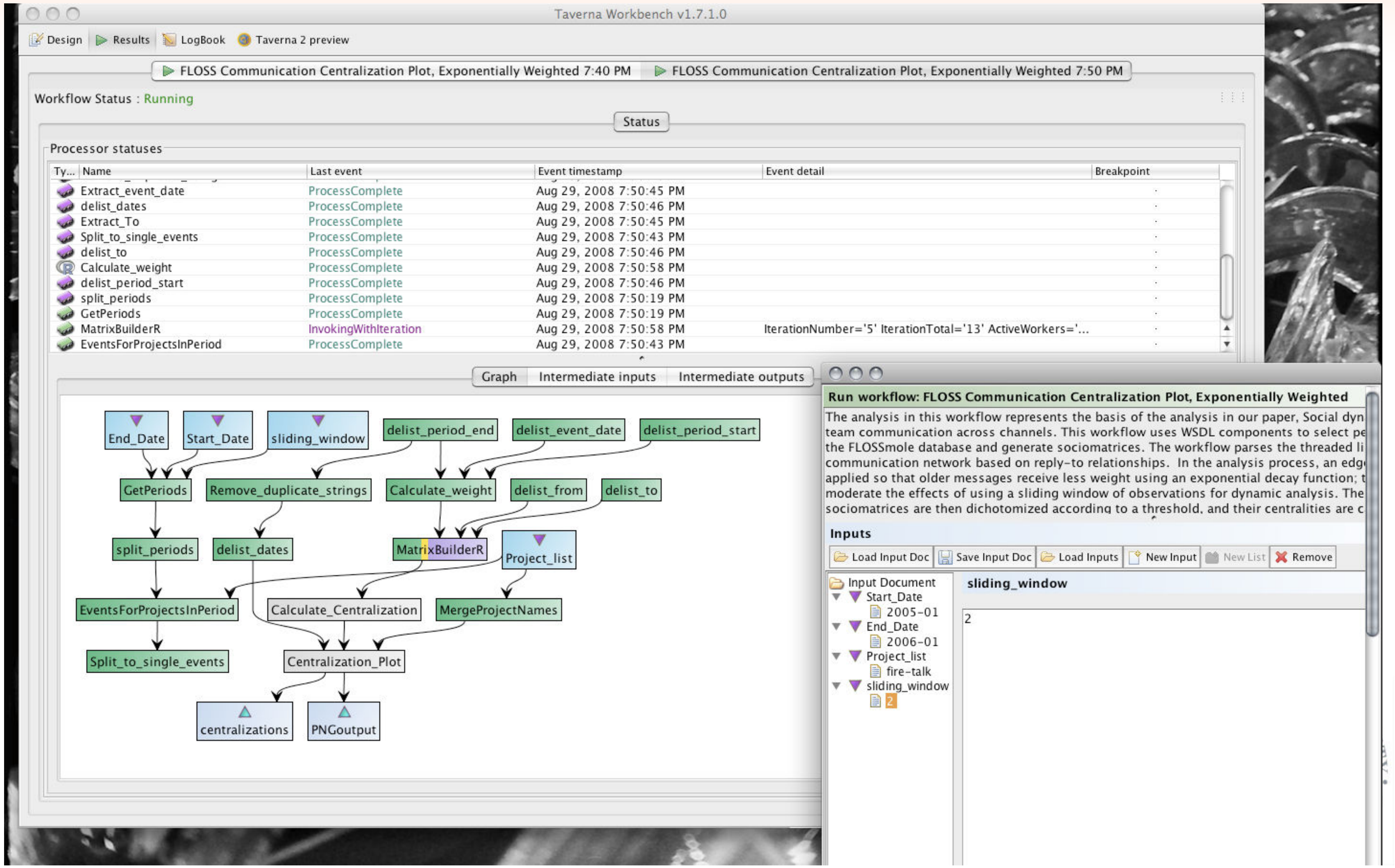
```
#Andrea Wiggins 12/13/07
#updated to remove ~/temp.data file 3/6/08

# Check for an empty matrix. (James Howison; updated 3/6/08 A. Wiggi
if (filecontents == "EMPTY\n0*") {
 output <- -1
} else {
 library(sna)

#procedure to read in sociomatrix. (Kevin Crowston)
dataframe <- read.csv(textConnection(filecontents))
data <- as.matrix(dataframe)

 #calculate centralization (dichotomized)
 sociomatrix<-event2dichot(data,method="quantile",thresh=0.8)
 cent <- (centralization(sociomatrix,degree,cmode="outdegree", diag=
 #calculate outdegree centralization (undichotomized)
 #cent <- (centralization(data,degree,cmode="outdegree", diag=TRUE,r
```

Taverna Execution Mode



Taverna Workbench v1.7.1.0

Design | Results | LogBook | Taverna 2 preview

FLOSS Communication Centralization Plot, Exponentially Weighted 7:40 PM | FLOSS Communication Centralization Plot, Exponentially Weighted 7:50 PM

Workflow Status : Running

Status

Processor statuses

Ty...	Name	Last event	Event timestamp	Event detail	Breakpoint
	Extract_event_date	ProcessComplete	Aug 29, 2008 7:50:45 PM		·
	delist_dates	ProcessComplete	Aug 29, 2008 7:50:46 PM		·
	Extract_To	ProcessComplete	Aug 29, 2008 7:50:45 PM		·
	Split_to_single_events	ProcessComplete	Aug 29, 2008 7:50:43 PM		·
	delist_to	ProcessComplete	Aug 29, 2008 7:50:46 PM		·
	Calculate_weight	ProcessComplete	Aug 29, 2008 7:50:58 PM		·
	delist_period_start	ProcessComplete	Aug 29, 2008 7:50:46 PM		·
	split_periods	ProcessComplete	Aug 29, 2008 7:50:19 PM		·
	GetPeriods	ProcessComplete	Aug 29, 2008 7:50:19 PM		·
	MatrixBuilderR	InvokingWithIteration	Aug 29, 2008 7:50:58 PM	IterationNumber='5' IterationTotal='13' ActiveWorkers='...	·
	EventsForProjectsInPeriod	ProcessComplete	Aug 29, 2008 7:50:43 PM		·

Graph | Intermediate inputs | Intermediate outputs

Run workflow: FLOSS Communication Centralization Plot, Exponentially Weighted

The analysis in this workflow represents the basis of the analysis in our paper, Social dyn
team communication across channels. This workflow uses WSDL components to select pe
the FLOSSmole database and generate sociomatrices. The workflow parses the threaded li
communication network based on reply-to relationships. In the analysis process, an edg
applied so that older messages receive less weight using an exponential decay function; t
moderate the effects of using a sliding window of observations for dynamic analysis. The
sociomatrices are then dichotomized according to a threshold, and their centralities are c

Inputs

Load Input Doc | Save Input Doc | Load Inputs | New Input | New List | Remove

Input Document
- Start_Date
  - 2005-01
- End_Date
  - 2006-01
- Project_list
  - fire-talk
- sliding_window
  - 2

sliding_window

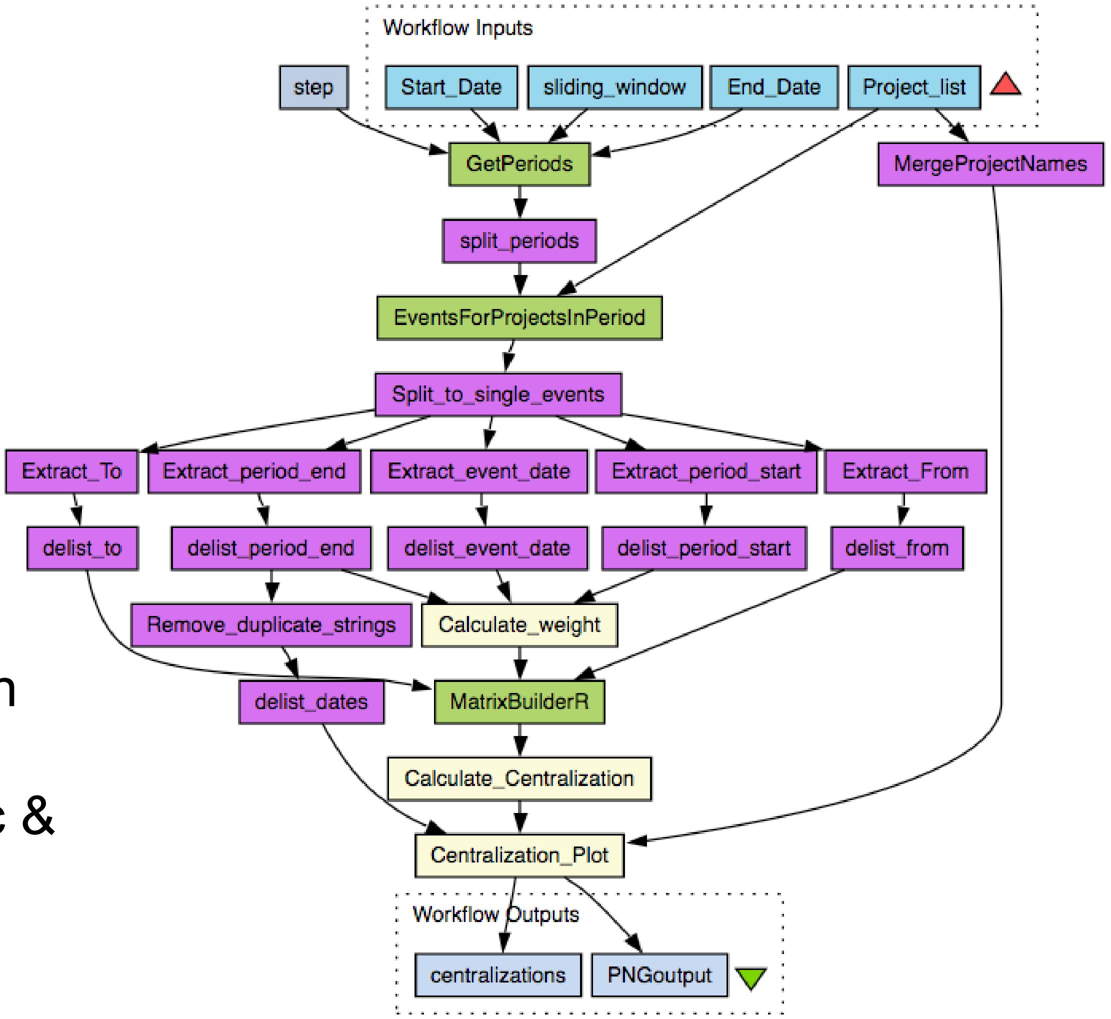
2

Workflow Component Types

- Abstract & notification processors
- String constants: file locations, threshold values, etc.
- Beanshell: simplified Java
- Rshell: R statistical program running through Rserve
- Java widgets/shims for common operations
  - i/o, lists, text manipulation, JDBC, etc.
  - Command-line
- Web services
  - WSDL with SOAP
  - Can be “scavenged” from URLs or other workflows

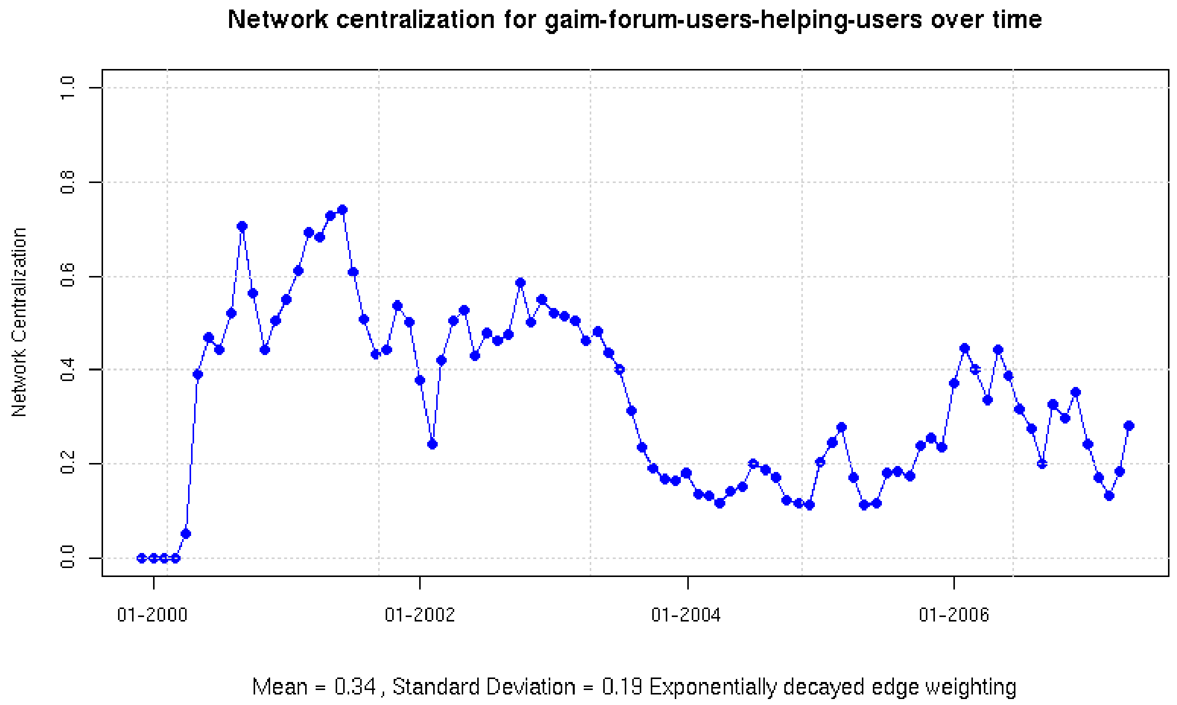
Example Workflow

- Teal: inputs
- Light blue: outputs
- Other light blue: string constant
- Green: web services
- Purple: Java shims
- Yellow: RShell
- Calculates weighted network centralization in dynamic networks, generates both numeric & graphical output



Example Output

- Inputs
  - Sliding window size
  - Project venue
  - Date range
- Outputs
  - Time series graph of centralizations with summary stats
  - CSV of dates and centralization values for additional analysis



Benefits of Using Workflows

- Modular design yields benefits in flexibility, transparency, and ease of reuse
  - Easier to co-develop and debug components, and to integrate independent efforts
  - Can quickly change strategies with minimal adjustment to existing workflow structure
  - Can reuse existing scripts and workflow components
- Can also conduct exhaustive sensitivity testing
- Multiple ways to achieve analysis tasks

Conclusion

- Combination of growing large-scale data sets and workflow tools present great opportunity for eResearch on FLOSS
- Work needed for eResearch infrastructure:
  - Access to data
  - Ontologies for naming data and defining relationships
  - Incorporating metadata & social science data, e.g. content analysis schemes
- Trade-offs involved in standardizing on tools to benefit collaboration, but much potential gain

More

- Taverna demo screencast
  - Long version (24 minutes): floss.syr.edu/Presentations/tavernaDemoScreencast.mov
  - Short version (14 minutes): floss.syr.edu/Presentations/TavernaDemoRedux.m4v
- MyExperiment FLOSS group
  - www.myexperiment.org/groups/64
- 16:30 - 17:30 presentation today: Social Dynamics of FLOSS Team Communication across Channels
- Tomorrow: Workshop on Public Data about Software Development (WoPDaSD 2008)